PPA: Principal parcellation analysis for brain connectomes and multiple traits

Rongjie Liu\textsuperscript{a}, Meng Li\textsuperscript{b,}\textsuperscript{\ast}, David B. Dunson\textsuperscript{c}

\textsuperscript{a} Department of Statistics, Florida State University, Tallahassee, FL, USA
\textsuperscript{b} Department of Statistics, Rice University, Houston, TX, USA
\textsuperscript{c} Department of Statistical Science, Duke University, Durham, NC, USA

\begin{abstract}
Our understanding of the structure of the brain and its relationships with human traits is largely determined by how we represent the structural connectome. Standard practice divides the brain into regions of interest (ROIs) and represents the connectome as an adjacency matrix having cells measuring connectivity between pairs of ROIs. Statistical analyses are then heavily driven by the (largely arbitrary) choice of ROIs. In this article, we propose a human trait prediction framework utilizing a tractography-based representation of the brain connectome, which clusters fiber endpoints to define a data-driven white matter parcellation targeted to explain variation among individuals and predict human traits. This leads to Principal Parcellation Analysis (PPA), representing individual brain connectomes by compositional vectors building on a basis system of fiber bundles that captures the connectivity at the population level. PPA eliminates the need to choose atlases and ROIs a priori, and provides a simpler, vector-valued representation that facilitates easier statistical analysis compared to the complex graph structures encountered in classical connectome analyses. We illustrate the proposed approach through applications to data from the Human Connectome Project (HCP) and show that PPA connectomes improve power in predicting human traits over state-of-the-art methods based on classical connectomes, while dramatically improving parsimony and maintaining interpretability. Our PPA package is publicly available on GitHub, and can be implemented routinely for diffusion image data.
\end{abstract}

1. Introduction

Image-based brain parcellation is a fundamental tool for understanding brain organization and function, in which the brain is divided into multiple interacting regions (Eickhoff et al., 2018). Modern neuroimaging techniques enable the collection of whole-brain magnetic resonance (MR) scans in large samples of individuals. Several large studies, such as the Human Connectome Project (Van Essen et al., 2013), have collected such data along with human behavioral and cognitive information, leading to a surge of interest in relating structural brain networks with various human traits (e.g., cognitive abilities, emotional feelings, and motor performance) (Akparian et al., 2020; Glasser et al., 2016; Lin et al., 2020; Park and Friston, 2013; Roine et al., 2019; Yeung et al., 2022; Zhang et al., 2019b). To predict individual traits using the structural brain connectome, the typical approach in most existing literature involves two steps: (i) extracting efficient representations of brain structural connectivity from preprocessed neuroimaging data, and (ii) training supervised learning models to identify the link between the extracted structural connectivity representations and individual traits, which can be adopted to predict the traits in a testing dataset (see, e.g., Girault et al., 2019; Greening and Mitchell, 2015; Powell et al., 2018; Yeung et al., 2022, and references therein).

There are distinct representations of the brain connectome obtained from different brain imaging modalities (Bijsterbosch et al., 2021). For example, functional connectivity is obtained from functional magnetic resonance imaging and determines a network using pairwise correlation measures in the BOLD signal. Structural connectivity aims to measure the strength of connectivity between regions based on the white matter fiber tracts, which can be estimated and extracted through tractography analysis on diffusion-weighted images (see, e.g., Bijsterbosch et al., 2021; Sporns et al., 2005; Toga et al., 2012; Yao et al., 2015, and references therein). In this paper, we focus on the latter. Then, a key issue in the above individual traits prediction task, as noted by Park and Friston (2013) and Yeung et al. (2022), is the way structural connectivity is measured, represented, and modeled from the white matter fiber tracts. Classically, the intricate spatial locations of all the fibers in the brain are summarized via an adjacency matrix, with the cells containing summaries of connections between pairs of regions of interest (ROIs)

\textsuperscript{\ast} Corresponding author.

E-mail address: meng@rice.edu (M. Li).

https://doi.org/10.1016/j.neuroimage.2023.120214.
Received 4 May 2023; Accepted 31 May 2023
Available online 5 June 2023.
1053-8119/© 2023 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
(O’Donnell et al., 2013). The associated anatomical parcellation analysis (APA)—one of the most popular approaches in connectivity-based parcellation analysis (Yao et al., 2015)—obtains the connectivity map by calculating connectivities between all pairs of regions in a predetermined anatomical parcellation scheme. Based on the selected atlas, we may build the connectivity matrix by counting the number of fibers passing between each pair of ROIs after fiber tracking. This connectivity matrix can then be used as a matrix-valued predictor in statistical analyses studying relationships with human traits (Lin et al., 2020; de Reus and Van den Heuvel, 2013; Wang et al., 2019; Zhang et al., 2018c).

However, such APA analyses require a particular anatomical ROI definition (Eickhoff et al., 2015), and many different schemes are available involving different numbers and locations of ROIs, such as FreeSurfer-DKT, automated anatomical labeling (AAL), automatic nonlinear imaging matching and anatomical labeling (ANIMAL) atlases, and many others (He et al., 2007; 2008; Klein and Tourville, 2012; Tzourio-Mazoyer et al., 2002; Wang et al., 2019; Yao et al., 2015). Choosing which scheme to use in practice is challenging. Several studies have reported impacts of different atlases on brain networks (Messé, 2020; Zalesky et al., 2010), and evidence suggests that not only the connectivity maps but also the inferences relating connectomes to human traits are strongly sensitive to the parcellation strategy.

An additional major issue is that APA leads to connectome representations corresponding to high-dimensional adjacency matrices. While there is a growing literature focused on statistical analysis of such replicated graph or network data (Bansal et al., 2018; Schiffer et al., 2017), such methods are under-developed and poorly understood relative to the rich literature on methods for high-dimensional vector-valued predictors, and computationally efficient methods that scale well in practice are lacking. For these reasons, it is common to simply vectorize the upper-triangular part of the adjacency matrix prior to statistical analyses. However, this fails to exploit the network structure in performing dimension reduction (Hochberg et al., 2007; O’Malley and Marsden, 2008; Wang et al., 2017), and can suffer from substantial loss of efficiency and accuracy (Wang et al., 2019).

Different from the predetermined atlas in the APA method, there are some approaches that generate fiber cluster based brain parcellation (He et al., 2022; O’Donnell and Westin, 2007; Wang et al., 2020; Zhang et al., 2018a; 2018b). First, a whole brain tractography registration tool (e.g., O’Donnell et al., 2012) is adopted to align all subjects tractography into a common space; next, a data-driven white matter parcellation atlas is created via applying some unsupervised learning approaches (e.g., spectral clustering and K-Means) on the aligned tractography across subjects; and then the generated parcellation atlas is applied to each individual subject to obtain a subject-specific white matter parcellation. It has been revealed that the fiber clustering method does not need to focus on an atlas-dependent adjacency matrix representation (Zhang et al., 2018a). In addition, the fiber clustering strategy also outperforms the APA method in terms of both segmentation consistency and test-retest reproducibility (Zhang et al., 2017; 2019a).

Due to the distinct advantages of fiber clustering approaches, in this paper, we propose a human trait prediction framework utilizing a tractography-based representation of the brain connectome, which clusters fibers to define a data-driven white matter parcellation targeted to explain variation among individuals and predict human traits. This leads to Principal Parcellation Analysis (PPA), representing individual brain connectomes by compositional vectors building on a basis system of fiber bundles that captures the connectivity at the population level. To the best of our knowledge, our PPA is the first to compare fiber clustering based structural connectivity representations to atlas-dependent adjacency matrix representations in the prediction of human traits. Unlike APA connectomes, PPA connectomes do not rely on any anatomical atlas or choosing ROIs a priori and lead to a substantially different representation of the connectome. This representation of structural brain connectomes facilitates analyses using well-established statistical methods designed for vector data. The PPA representation provides an alternative to the current standard ROI-based adjacency matrix representation in analyses studying how structural connectomes vary across individuals, both randomly and in relation to individual traits, and can accomplish these same inference goals at a fraction of the cost for implementing APA-based counterparts. We illustrate the proposed approach through applications to data from the Human Connectome Project (HCP) and show that PPA connectomes, when combined with standard high-dimensional regression methods, improve power in predicting human traits over state-of-the-art methods based on classical connectomes, while dramatically improving parsimony and maintaining interpretability. Our proposed method is general for any data that consist of a collection of fibers. Our PPA package is publicly available on GitHub, and can be implemented routinely for diffusion image based parcellation analysis.

The rest of the paper is structured as follows. Section 2 introduces the proposed parcellation approach and PPA formulation. In Section 3, we compare PPA to state-of-the-art APA-based methods using HCP data, and focus on prediction, visualization, and interpretability. Section 4 contains a discussion.

2. Methods

2.1. The PPA framework

Suppose that we observe structural MRI, diffusion MRI, and human traits from n individuals. The PPA pipeline, as illustrated in Fig. 1, consists of three modules: (i) reconstruction of fibers; (ii) representation of fibers and unsupervised clustering; and (iii) high-dimensional supervised learning adaptive to human traits. Each module of PPA encompasses a variety of choices, equipping PPA with easy extensibility. We first describe these modules using initial default settings, followed by a discussion on extensions.

Let \( F = \{ f_{ik}, k = 1, ..., m_i; i = 1, ..., n \} \), where \( f_{ik} \) is the \( i \)-th fiber in the \( k \)-th individual’s brain, and \( m_i \) is the total number of fibers in the \( i \)-th subject. In addition, let \( y_i(s) \) denote the \( s \)-th ‘trait’ of the \( i \)-th individual with \( y_i(s) \) = \( (y_1(s), ..., y_{S}(s)) \) for \( s = 1, ..., S \); traits can range from demographic characteristics, alcohol and drug exposures, to scores on cognitive, psychological and behavioral assessments.

In Module (i), we reconstruct fibers using the recently proposed TractoFlow method (Theaud et al., 2020); alternative fiber tracking algorithms can be used instead without changing the subsequent steps in the PPA pipeline. TractoFlow takes raw diffusion weighted images as the input, consists of 14 steps for the diffusion weighted image (DWI) processing and 8 steps for the T1 weighted image processing, and outputs classical diffusion imaging measures. The outlier fiber tracts are detected and removed using the method proposed in Garyfallidis et al. (2012). Module (i) is also a key step in estimating APA connectomes.

In Module (ii), we formulate connectomes at the population level through basis networks in the form of fiber bundles that represent clusters of streamlines, and represent individual connectomes via compositional vectors. The clustering is performed on pooled fibers from all subjects. In particular, let \( \{a_{ik} \}_{k=1}^{m_i} \) and \( \{b_{ik}^{(h)} \}_{h=1}^{H} \) be the 3D coordinates of two endpoints for each fiber from the \( i \)-th subject. Let \( Z \) be a \( 6 \times m \) matrix stacking all \( \{a_{ik}^T, b_{ik}^{(h)}^T \} \) as columns for \( k = 1, ..., m_i \) and \( i = 1, ..., n \), where \( m = \sum m_i \) is the total number of fibers from all subjects. We perform a cluster analysis at the fiber level using the matrix \( Z \), outputting a collection of partitions of \( F \), denoted by \( A_K = \{ A_K^{(1)}, ..., A_K^{(K)} \} \), where \( K \) is the number of clusters and each set \( A_K^{(k)} \) can be interpreted as a fiber bundle. The enormous number of fibers (e.g., around 2.8 billion in our data application) presents substantial computational challenges in clustering; for example, traditional K-Means does not scale well and requires large memory, leading to prohibitive computational cost. As such, we adopt mini-batch K-Means (Sculley, 2010), which reduces the memory use and converges to the optimal value orders of magnitude faster than the full-batch K-Means. For a fiber and cluster center, we define their distance as the minimum of their Euclidean distances considering...
two orderings of fiber endpoints, accounting for fiber bi-directionality similarly to fiber flipping (Garyfallidis et al., 2012). The number of clusters \( K \) is an important hyperparameter. Our experiments have provided guidance on good values of \( K \) in practice. To automate the choice of \( K \), we can rely on cross validation as illustrated in Fig. 2. Alternatively, we can decrease sensitivity to \( K \) by choosing multiple values in a multi-scale representation of the brain connectome.

For a given \( K \), an individual’s connectome can be represented by the proportions of the individual’s fibers belonging to each of the inferred population-level fiber bundles. In particular, the \( i \)th individual’s connectome is represented via the \( K \)-dimensional compositional vector \( \omega_i = (\omega^1_i, \ldots, \omega^K_i) \), with \( \omega^k_i \) the proportion of fibers belonging to the \( k \)th fiber bundle \( A_k^{(K)} \), for \( k = 1, \ldots, K \). The connectome data for all \( n \) subjects is then contained in the matrix \( \omega = (\omega^1_1, \ldots, \omega^n_K)^T \). This provides a much simpler representation than the adjacency matrix-based APA approach. Note that the framework does not require all clusters \( K \) to be present in all subjects since for certain individuals the number of fibers belonging to a particular cluster can be zero.

In Module (iii), we relate the connectome \( \omega_i \) to traits \( y_i(s) \). For simplicity in interpretation, we initially focus on trait-specific linear regression models:

\[
y_i(s) = \beta_0(s) + \sum_{k=1}^{K-1} \omega^k_i \beta^k(s) + e_i(s),
\]

where \( \beta_0(s) \) is a trait specific intercept, which can be expanded to include non-connectome covariates, and \( \beta^k(s) (k = 1, \ldots, K) \) is a regression coefficient characterizing the relationship between the density of connections in the \( k \)th fiber bundle and the \( s \)th trait. For a sufficiently flexible specification, one may choose \( K \) to be large in which case many of the \( \beta^k(s) \) coefficients are expected to be zero or close to zero. Standard sparse learning methods can be used to estimate the coefficients while learning the sparsity pattern. This yields a set of estimated non-zero coefficients \( \left\{ \beta^k : k \in \mathcal{K}(s) \right\} \), where \( \mathcal{K}(s) = \{k_1, \ldots, k_{m(s)}\} \subset \{1, \ldots, K-1\} \) collects the indices of the \( m(s) \) fiber bundles having non-zero coefficients. We refer to these fiber bundles as “active” for the \( s \)th trait. Active bundles impact the response \( y_i(s) \) via equation (1), while inactive bundles have no impact on the response. For each \( K \) and the \( s \)th trait, the total number of fibers in active bundles is \( \sum_{k \in \mathcal{K}(s)} |A^k_s| \), where \( |A^k_s| \) is the number of fibers in the bundle \( A^k_s \).

In our numerical experiments, we use LASSO (Tibshirani, 1996), one of the most popular high-dimensional regression methods, as a representative example. LASSO has been very widely studied and relatively efficient algorithms are readily available. However, applying LASSO to the vectorized upper triangle portion of APA connectome adjacency matrices produces less reliable estimation and has worse predictive performance than an identical analysis using PPA instead of APA connectomes. This is consistent with previous results motivating complex statistical methods that take into account the graph structure of the APA connectomes (Wang et al., 2019).

2.2. Extensions of PPA

In Module (i), other fiber tracking algorithms alternatives to TractoFlow can be considered, such as Euler Delta Crossings (EuDX) in Garyfallidis et al. (2012) and Sparse Fascicle Model (SFM) in Rokem et al. (2015). We will compare various fiber tracking algorithms in our analyses of HCP data. In Module (ii), other clustering or factorization methods, including spectral clustering and non-negative matrix factorization (NMF), can also be adopted. In addition to the endpoints, the length and shape of the fibers may contain useful information (Zhang et al., 2018c), which can be incorporated in clustering analyses. Module (iii) can be modified building on the rich literature on high-dimensional supervised learning methods. Instead of LASSO, other sparse shrinkage methods, such as elastic net (Zou and Hastie, 2005) and smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001), can be used without complication.

3. Human connectome project data analyses

In this section we use the HCP dataset to compare PPA-based methods with state-of-the-art APA-based approaches using various human traits, demonstrate how to choose \( K \) in a data-driven manner, assess the robustness of PPA with respect to the fiber tracking algorithms and regularization strategies, and illustrate the interpretability of PPA.

3.1. HCP data description

Data collection and sharing for this project was provided by the MGH-USC Human Connectome Project (HCP; Principal Investigators: Bruce Rosen, M.D., Ph.D., Arthur W. Toga, Ph.D., Van J. Weeden, MD). HCP funding was provided by the National Institute of Dental and Craniofacial Research (NIDCR), the National Institute of Mental Health (NIMH), and the National Institute of Neurological Disorders and Stroke (NINDS). HCP data are disseminated by the Laboratory of Neuro Imaging at the University of Southern California.

We use the same set of 1065 HCP subjects as in Wang et al. (2019), including dMRI data along with human traits, downloaded from HCP 1200 Subjects Data Release. Details about the dMRI data acquisition and preprocessing can be found in Sotiropoulos et al. (2013); Van Essen et al. (2012). For the human traits data, seven different scores were selected: receptive vocabulary, oral reading, list sorting, flanker, picture sequence memory, card sort, and processing speed. These scores

---

1. https://www.humanconnectome.org/study/hcp-young-adult/document/1200-subjects-data-release.
Fig. 2. Comparison of 5-fold cross validation MSE (first row) and cross-validated $R^2$ (second row) of trait prediction based on PPA and three APA-based methods (LASSO, SBL, and MultiGraphPCA) for three traits: (a) PicVocab, (b) ReadEng, and (c) ListSort. The last two rows are the bar-plots of MSE and $R^2$ for LASSO, SBL, MultiGraphPCA with $K = 400$, and PPA with $K = 400$. APA methods are in blue and PPA method in cyan. The red horizontal line in the third row indicates the MSE of the null model, which is 230.92 in PicVocab; 219.02 in ReadEng; 174.66 in ListSort. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
can be used to study human cognition. Note that although we use task-based scores in this section, the proposed methods are broadly applicable for any measurement reflecting human traits. All the scores are age-adjusted, and their details can be found on the HCP website. A brief description of each trait is also included in the Appendix for easy reference.

### 3.2. Analysis using PPA and APA

PPA and APA provide distinct representations of human brain connectomes. Performance in studying relationships between connectomes and traits depends on the downstream analysis methods after the connectomes are obtained. As such, we chose state-of-the-art methods developed under APA connectomes, and adopted one of the most standard analysis methods, LASSO, under PPA connectomes. Such comparisons give APA an advantage. We also implemented LASSO for the vectorized APA connectomes.

We implemented PPA using the default choices in Section 2. In particular, we used TractoFlow for fiber tracking, which depends on two main technologies: Nextflow and Singularity (Avants et al., 2009; Di Tommaso et al., 2017; Garyfallidis et al., 2014; Jenkinson et al., 2012; Kurtzer et al., 2017; Tournier et al., 2019). We obtained around 2.8 billion fibers for the HCP subjects. The fiber tracking results for two randomly selected subjects are displayed in Figure S1 in the Appendix. We clustered the fibers using mini-batch K-Means (Sculley, 2010). We set the batch size to 1000 and varied the number of clusters K from 10 to 500 across 10, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500. For each choice of K, we conducted analyses relating the PPA connectome to the trait of interest. We can then assess, based on cross validation, which choice of K is best for that particular trait. For each K, we obtained K fiber bundles A_k for k = 1, …, K, leading to PPA connectome \( \omega_k \) for each individual \( i = 1, \ldots, n \).

Figures S2 and S3 in the Appendix show examples of the inferred fiber clusters (\( K = 10 \)) with each color denoting one cluster. The same clusters, and corresponding colors, are used for the different subjects, and some heterogeneity is apparent across subjects. The numbers of fibers in each cluster are shown in Figure S4 in the Appendix for these two subjects. The profile of these counts across clusters is similar for the two subjects, but subject 2 has a considerably greater proportion of fibers in cluster 10. A close inspection of the inferred clusters indicates that the anatomical locations of fiber bundles produced by PPA often correspond to known fascicles in the literature (Friederici and Gierhan, 2013; Shin et al., 2019; Gupta et al., 2017). Taking the 10 clusters in Figure S3 as an example, cluster 2 and cluster 10 cover the inferior longitudinal fasciculus; cluster 5 overlaps with corticospinal tract; cluster 3 and cluster 5 contain corpus callosum and uncinate fasciculus; cluster 8 and 9 contain superior longitudinal fasciculus, arcuate fasciculus and inferior fronto-occipital fasciculus (Friederici and Gierhan, 2013), which are language function related ROIs.

For APA connectomes, we provide the same set of fibers generated by the TractoFlow fiber tracking method as used in PPA, ensuring a fair comparison between the two approaches. APA connectomes further require one to choose brain ROIs according to an atlas template, and then select a summary of connectivity between each pair of ROIs, such as the number of connections. APA-based methods represent the ith individual’s brain connectome as a \( p \times p \) matrix \( W_i \). Each cell of this matrix contains a summary of the strength of connection between a pair of brain ROIs; here, we use the number of fibers connecting the regions. Different atlas templates lead to different connectivity matrices having different dimensionality \( p \). We chose the FreeSurferDKT cerebral cortex atlas. FreeSurferDKT is an atlas manually labeled in the macroscopic anatomy in magnetic resonance images by the Desikan-Killiany-Tourville (DKT) protocol (Klein and Tourville, 2012), which includes 31 cortical regions per hemisphere; see Tables S1 and S2 in the Appendix for descriptions of these 62 cortical regions. Note that one can also obtain different count values using different tractography algorithms (Knösche et al., 2015) as well as many filtering approaches, e.g., scale-invariant feature transform (SIFT) (Burger and Burge, 2016) to make streamline counting more quantitative.

For PPA connectomes, we used LASSO to fit Model (1); the hyperparameter in LASSO was selected using 5-fold cross validation. For APA connectomes, we implemented two recently proposed methods: symmetric bilinear regression (SBL) (Wang et al., 2019) and multi-graph principal component analysis (MultiGraphPCA) (Winter et al., 2020). SBL investigates the relationship between human traits and connectivity matrices through a symmetric bilinear regression model, and MultiGraphPCA proposes a tensor network factorization method that links the scale-specific brain structural connectivity matrices through a common set of individual-specific scores, which are further used for human trait prediction. Tuning in SBL and MultiGraphPCA followed the recommendations by the authors. In particular, for SBL we use K = 14; gamma = 6.9; fullit = 50; maxit = 10000; tol = 1e-6; Replicates = 5. There is a single tuning parameter K in MultiGraphPCA that corresponds to the number of latent factors; we compare results for K = {2, 10, 20, 50, 70, 200, 400, 500}. Note that in our visualizations for MultiGraphPCA with varying K, we use the same “number of cluster” label on the x-axis as PPA for simplicity. We also implemented LASSO on the vectorized (only keeping upper-triangular elements) connectivity matrix \( W_i \).

In order to assess the robustness of the proposed method, we tested different versions of PPA with respect to the fiber tracking algorithms and regularization strategies. In particular, we adopted another two fiber tracking algorithms in Module (I), Euler Delta Crossings or EuDX (Garyfallidis et al., 2014), and local tracking with Sparse Fascicle Model or SFM (Rokem et al., 2015). For the regularization strategy, we also considered elastic net (Zou and Hastie, 2005), which combines the \( L_2 \) penalty of LASSO with the \( L_2 \) penalty of ridge regression (Hoerl and Kennard, 1970), with the goal of simultaneous selection of correlated predictors.

Figures S5 and S6 in the Appendix show the comparison of various versions of PPA using different fiber tracking algorithms and regularization strategies, respectively. The results are similar to those shown above, and demonstrate the robustness of the PPA method to the fiber tracking algorithm and regularization approach.

### 3.3. Predictive performance & parsimony

We implemented three methods under APA connectomes, which are coded as LASSO, SBL, and MultiGraphPCA, and the version of PPA using Tractoflow fiber tracking algorithm and LASSO regularization to analyze the 1065 HCP subjects. We calculated the 5-fold cross validation mean squared error (MSE) to compare their predictive performance for seven human traits. We also included the performance of a null model, which only contains an intercept term and uses the sample average of responses in the training set as the prediction; this model assumes no significance of the connectome in explaining selected traits and serves as a reference model.

Fig. 2 plots the prediction performance for three traits. The upper row of Fig. 2 shows that the proposed PPA, using the simple LASSO method, clearly outperforms SBL, LASSO, and MultiGraphPCA in most scenarios in terms of cross validation MSE, especially as the number of clusters K increases. The performance of SBL and LASSO does not depend on K; although K has different meanings in PPA and MultiGraphPCA, our figures use the same x-axis label for both methods for simplicity. The third row of Fig. 2 compares the MSEs of our PPA-based methods at K = 400 to the three APA methods, while the red horizontal line represents the performance of the null model. Since all APA-based methods use Tractoflow, we first focus on the PPA method using the same fiber tracking algorithm. In this case, MSEs of PPA are smaller.

---

3 https://wiki.humanconnectome.org/display/PublicData/.
than the three APA-based methods, uniformly across the three traits. In sharp contrast to the excellent performance of LASSO for the PPA connectomes, LASSO predictions based on vectorized APA connectomes did no better than the null model. SBL and MultiGraphPCA improve the MSEs over LASSO, as a result of a better utilization of the network structure of APA connectomes. The comparison of MSEs suggests that fundamentally changing the connectome representation based on defining population fiber bundles can perhaps lead to even bigger gains.

We additionally evaluate various methods using cross-validated $R^2$ (Gao et al., 2019) as the metric, which is a variation of the conventional $R^2$ but under cross-validation settings. In particular, for responses $\{y_i\}_{i=1}^n$ with sample mean $\bar{y}$, the cross-validated $R^2$ for a given model is defined as $R^2_{cv} = 1 - \sum_{i=1}^n (y_i - \hat{y}_i)^2 / \sum_{i=1}^n (y_i - \bar{y})^2$, where $\hat{y}_i$ is the predicted value of $y_i$ in a 5-fold cross-validation procedure, with the model fitted on the four-fold training data and $y_i$ being in the one-fold test data. Since $R^2_{cv}$ can be negative, we set negative values to 0 in practice. The second and fourth rows of Fig. 2 replicate the comparison in the first and third rows, but with the cross-validated $R^2$ metric. Our analysis suggests that the comparison between methods remains consistent, irrespective of whether we use the MSE or $R^2$ metrics. This is expected as MSE and $R^2$ are closely related measures of model performance, with an approximate relationship between them. Fig. 2 shows that defining large numbers of fiber bundles may lead to predictive gains.

For the other four traits (Flanker, PicSeq, CardSort, ProcSpeed), all methods tend to give a MSE close to the null model (Fig. 3), indicating limited predictive power of structural connectivity for these traits. It is reassuring that the proposed method is consistent with APA-based methods in these cases. We remark that the lack of predictive power might be caused in part by a weak relationship between these measured traits and actual innate abilities in the test subjects.

While Fig. 2 shows how MSE varies with $K$ for the three different fiber tracking algorithms, TractoFlow, EuDX and SFM, Fig. 4 plots the MSEs against the total number of fibers in active bundles to provide additional insight into the impact of $K$. Taking picture vocabulary test (PicVocab) as an example, the best MSE is achieved when the total fiber counts in active bundles is around $0.8 \times 10^7$. This optimal number varies from trait to trait, but a U-shaped curve typically emerges with $K$ varied up to 500. For the other traits in Fig. 3, MSEs do not change much as we vary the total fiber counts in active bundles, which is expected as the MSE curve is flat when plotted against $K$; we omit these curves here as they are redundant.

Table 1 reports the number of selected parameters in PPA and APA-based methods, which shows parsimony and effectiveness of PPA-based methods compared to LASSO applied to APA connectomes and SBL. In particular, LASSO for APA connectomes selects nearly zero active connections, which explains its poor predictive performance in Fig. 2. PPA selects substantially fewer non-zero parameters than SBL; this com-
Fig. 4. Comparison of 5-fold cross validation MSE of trait prediction based on PPA for three traits: (a) PicVocab, (b) ReadEng, and (c) ListSort.

Table 1
Number of selected parameters in different methods. For MultiGraphPCA, the number of parameters is set to be $K$ used in PPA.

| Method                  | PicVocab | ReadEng | ListSort | Flanker | PicSeq | CardSort | ProcSpeed |
|-------------------------|----------|---------|----------|---------|--------|----------|-----------|
| PPA ($K = 50$)          | 23       | 20      | 12       | 11      | 15     | 0        | 1         |
| PPA ($K = 100$)         | 38       | 33      | 19       | 3       | 29     | 3        | 2         |
| PPA ($K = 200$)         | 64       | 47      | 12       | 7       | 6      | 0        | 0         |
| PPA ($K = 300$)         | 53       | 31      | 2        | 3       | 1      | 0        | 0         |
| PPA ($K = 400$)         | 19       | 20      | 18       | 18      | 15     | 2        | 1         |
| PPA ($K = 500$)         | 56       | 39      | 22       | 1       | 3      | 0        | 1         |
| LASSO + FreeSurferDKT   | 1        | 1       | 0        | 0       | 5      | 1        | 1         |
| SBL + FreeSurferDKT     | 576      | 542     | 338      | 405     | 601    | 371      | 670       |

Fig. 5. Visualization for trait PicVocab: each column represents the visualization for a different number of clusters ($K = 50$; $K = 100$; $K = 200$; $K = 400$) in PPA; First row represents the connectivity matrix between any two ROI regions in the FreeSurferDKT cortical atlas. Second row shows anatomy of connections in an axial view. Third row shows anatomy of connections in a sagittal view.
combined with the better MSEs in Fig. 2 shows the effectiveness of PPA connectomes in representing key features of brain networks predictive of traits. The last four traits show little to no signal for any of the methods and selecting few if any features for these traits seems appropriate.

To evaluate generalizability of our comparisons between PPA and APA and assess robustness of APA methods, we tested multiple different versions of APA with respect to the atlases and summary of connectivity used in defining the connectivity matrices. In our analyses as we vary these two aspects when implementing APA methods, the improved predictive performance of PPA over APA that was shown earlier is consistently observed, indicating the way connectomes are represented (PPA vs APA) appears to be a more important factor in explaining the performance gain.

In particular, for all APA related methods (SBL, LASSO, MultiGraphPCA), we checked two different atlases: FreeSurferDKT and AAL2. FreeSurferDKT is an atlas manually labeled in the macroscopic anatomy in magnetic resonance images by the DKT protocol (Klein and Tourville, 2012), and we chose the settings used previously in obtaining the results in Figs. 2 and 3. AAL2 stands for automated anatomical labeling atlas 2, providing an alternative parcellation of the orbitofrontal cortex (Rolls et al., 2015). From the results in Figure S7 in the Appendix, we can see that changing the atlas has little impact on the predictive performance of APA methods.

For analyses assessing sensitivity of the APA results to the summary of connectivity between regions, we focused on the FreeSurferDKT atlas and tested three different ways of calculating the connectivity matrix: “count”, “ncount”, and “ncount2”. For each entry in the connectivity matrix, these three summaries correspond to counting the number of tracts that pass two ROIs (“count”), which is used in the previous comparison in Figs. 2 and 3, normalizing the count by the median length of the fibers (“ncount”), and multiplying the count by the sum of the inverse of the length (“ncount2”). Figure S8 the Appendix shows that the performance of APA methods is robust to the summary used in calculating connectivity matrices.

3.4. Integrating PPA and atlas

The proposed PPA connectome does not rely on any tractography atlas in defining the connectome or building a regression model for traits. This provides vector-valued connectomes that can be analyzed by many standard statistical methods with improved prediction and parsimony, as shown in the preceding sections. In this section, we demonstrate another feature of PPA in terms of its compatibility with traditional ROIs—as anab initio, tractography-based representation of connectomes, PPA can be integrated with any existing atlas templates to borrow the ROI information encoded therein in a straightforward manner. Such integration allows us to relate the interpretation of PPA results to traditional ROIs. Through visualization, we find the proposed PPA leads to interesting and interpretable findings.

We align active fiber bundles produced by PPA to an atlas. In particular, based on a selected atlas, we build the connectivity matrix at the population level with each matrix entry generated by counting the number of fibers in active bundles passing between each pair of ROIs.
Note that similarly to deriving the connectivity matrix in APA, one can obtain many different summaries of connectivity between two regions by using various filtering approaches such as SIFT and different ways of normalization, with the count just one simple choice. We use the FreeSurferDKT cortical atlas (Klein and Tourville, 2012) as in our implementation of APA-based methods, which segments the brain into 62 regions. For each human trait, we visualize the PPA-induced connectivity matrix and anatomy of connections through DSI Studio (http://dsi-studio.labsolver.org), a tractography software tool that maps brain connections and correlates findings with traits. We adopt the default setting in DSI Studio by thresholding matrix entries with a small number of connecting tracks relative to the maximum connecting tracks in the connectivity matrix. We use 0.5 as the threshold for this ratio.

According to the visualization plots in Figs. 5–7, PPA discovers some insightful connections of various anatomical regions that are related to the human traits. Some interesting findings are listed as follows. For most human traits, the primary pattern in the connectivity matrix does not vary much as the number of clusters (fiber bundles) increases. For example, when analyzing the trait PicVocab, two subgraphs are consistently identified in the left and right brain respectively across all four cluster settings, i.e., $K = 50, 100, 200, 400$ (see Fig. 5). Specifically, the subgraph in the left brain includes connections among ROI 5 (left_fusiform), ROI 7 (left_inferior_temporal), ROI 9 (left_lateral_occipital), ROI 10 (left_lateral_orbitofrontal), ROI 11 (left_lingual), ROI 28 (left_superior_temporal), and ROI 31 (left_insula), while the subgraph in the right brain includes connections among ROI 37 (right_fusiform), ROI 39 (right_inferior_temporal), and ROI 43 (right_lingual). The trait ListSort that is related to human working memory (Fig. 6) also shows a common pattern of two subgraphs in each hemisphere according to the connectivity matrix across cluster settings. Compared to the network detected using the trait PicVocab, the network related to the trait ListSort contains some additional regions, such as ROI 14 (left_parahippocampal) and ROI 19 (left_pericalcarine), which have been found to be highly correlated with working memory performance (Owens et al., 2018; Yonelinas, 2013).

For language associated human traits (e.g., the trait PicVocab is related to language and vocabulary comprehension while the trait ReadEng is related to language and reading decoding), the regions with strong connections are mainly located in the left hemisphere (see Figs. 5 and 7). This finding indicates that the left hemisphere is particularly important for language, which has been consistently verified in clinical and experimental settings (Riès et al., 2016).

We find that a particularly important subgraph including ROI 7 (left_inferior_temporal) and ROI 28 (left_superior_temporal) is detected for most human traits, as these two ROIs are consistently prominent nodes (ROI 7: light green; ROI 28: pink) in the identified subgroup for most values of $K$ (see the second row in Figs. 5–7). These results obtained from PPA are in agreement with previous findings in the field. The temporal lobes are located beneath the lateral fissure on both cerebral hemispheres of the brain. They are most commonly involved in processing sensory input into derived meanings for the appropriate retention of visual memory, language comprehension, and emotion association (Eichenbaum et al., 2007; Squire et al., 2004). In particular, for most individuals, the left temporal lobe is involved in understanding...
language as well as learning and remembering verbal information (Flick et al., 2018; Trimml et al., 2018).

4. Discussion

In this article, we propose a new human trait prediction framework using a fiber clustering based representation of structural brain connectomes as an alternative to the widely used atlas-dependent structural connectivity representations. This framework leads to Principal Parcelation Analysis (PPA), where we represent individual brain connectomes by compositional vectors building on a basis system of fiber bundles that captures the connectivity at the population level.

PPA eliminates the need to choose atlases and ROIs a priori when predicting human traits. Unlike the traditional connectomes where data objects are of complex graph-structure and ultra-dimension, our PPA connectomes can be analyzed using existing statistical tools for high-dimensional vector-valued features. Our application to HCP data indicates that PPA is robust to the specific choice of fiber tracking algorithm. We propose an approach to integrate the parcellation produced by PPA with an atlas, so that the results can be visualized and interpreted using traditional ROIs.

There are several interesting next directions building on our initial PPA approach. Firstly, the methods used in each of the three modules can be refined. For example, we can consider different fiber clustering algorithms that take into account more than just the endpoint locations. Also, instead of just applying LASSO within a linear regression model for the trait responses, we can use more elaborate predictive algorithms and inferential approaches. Particularly for large datasets, improved predictive accuracy may be possible with flexible non-linear regression methods. We have focused on using the proposed connectomes to analyze human traits via a regression model, and it is interesting to consider other data problems beyond explaining and predicting task-based scores. For example, one may consider both structural and functional connectivity (Tian and Zalesky, 2018) and utilize other dynamical features (Kobeleva et al., 2021), with the proposed PPA connectomes serving as a building block to represent structural connectivity.

Data and code availability statement

The dMRI and human traits data from the original study are available from HCP: https://www.humanconnectome.org/. Code for implementing PPA is freely available online at https://github.com/ylximeng/PPA.

Declaration of Competing Interest

Authors declare that they have no conflict of interest.

Credit authorship contribution statement

Rongjie Liu: Conceptualization, Methodology, Software, Writing – original draft. Meng Li: Conceptualization, Methodology, Validation, Supervision, Writing – review & editing. David B. Dunson: Conceptualization, Methodology, Validation, Writing – review & editing.

Data availability

I have shared the link to data/code in the paper.

Acknowledgments

We thank Lu Wang and Zhengwu Zhang for helpful discussion, and Yerong Li for helping with numerical examples in an earlier version of PPA. We thank the editors and referees whose constructive comments led to substantial improvements in this article. This work was partially funded by grant DMS-2015569 from the National Science Foundation and grant R01MH118927 of the National Institute of Mental Health of the United States National Institutes of Health.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2023.120214.

References

Apkarian, V., Bonin Pinto, C., Bielefeld, J., Jabbarian, R., Reckziedl, D., Griffith, J., 2020. Neural and genetic bases for human ability traits. Front. Hum. Neurosci. 14, 565.
Avants, B., Tustison, N., Song, G., 2009. Advanced normalization tools (ANTS). Insight J. 2, 1–35.
Bansal, K., Nakuci, J., Muldoon, S., 2018. Personalized brain network models for assessing structure-function relationships. Curr. Opin. Neurobiol. 52, 42–47.
Bijlroosv, J., Vait, S., Wang, D., Glasser, M., 2021. Recent developments in representations of the connectome. NeuroImage 243, 118533.
Burger, W., Burge, M., 2016. Scale-invariant feature transform (SIFT). In: Digital Image Processing. Springer, pp. 609–664.
Di Tommaso, P., Chatrou, M., Floden, E., Barja, P., Palumbo, E., Notredame, C., 2017. Nextflow enables reproducible computational workflows. Nat. Biotechnol. 35 (4), 316.
Eichenbaum, H., Yonelinas, A., Ranganath, C., 2007. The medial temporal lobe and recognition memory. Annu. Rev. Neurosci. 30, 123.
Eickhoff, S., Thirion, B., Varoquaux, G., Bzdok, D., 2015. Connectivity-based parcellation: critique and implications. Hum. Brain Mapp. 36 (12), 4771–4792.
Eickhoff, S., Ye, Y., Genon, S., 2018. Imaging-based parcellations of the human brain. Nat. Rev. Neurosci. 19 (11), 672–686.
Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. 96 (456), 1348–1360.
Flick, G., Oseki, Y., Kaczmarek, A., Al Kaabi, M., Maranta, A., Pykälniemi, L., 2016. Building words and phrases in the left temporal lobe. Cortex 106, 213–236.
Friederici, A., Gierhan, S., 2013. The language network. Curr. Opin. Neurobiol. 23 (2), 250–254.
Gao, S., Greene, A., Constable, T., Schenod, C., 2019. Combining multiple connectomes improves predictive modeling of phenotypic measures. Neuroimage 201, 116038.
Garyfallidis, E., Brett, M., Amirbekian, B., Rokem, A., Van Der Walt, S., Descomtes, M., Nimmo-Smith, I., 2014. Dipy, a library for the analysis of diffusion MRI data. Front. Neuroinform. 8, 8.
Garyfallidis, E., Brett, M., Correia, M., Williams, G., Nimmo-Smith, I., 2012. Quickbundles, a method for tractography simplification. Front. Neurosci. 6, 175.
Girault, J., Munsell, B., Puechmaill, D., Goldman, B., Prieto, J., Styn, M., Gilmore, J., 2019. White matter connectomes at birth accurately predict cognitive abilities at age 2. NeuroImage 192, 145–155.
Glasser, M., Smith, S., Marcus, D., Andersson, J., Auerbach, E., Behrens, T., Coalson, T., Harms, M., Jenkinson, M., Moeller, S., Robinson, E., Sotiropoulos, S., Xu, J., Yacoub, E., Ugurbil, K., Van Essen, D., 2016. The human connectome project’s neuroimaging approach. Nat. Neurosci. 19 (9), 1175–1187.
Greenhill, S., Mitchell, D., 2015. A network of amygdala connectivities predict individual differences in trait anxiety. Hum. Brain Mapp. 36 (12), 4819–4830.
Gupta, V., Thomopoulos, S.I., Ruthr, F.M., Thompson, P.M., 2017. FiberNET: An Ensemble Deep Learning Framework for Clustering White Matter Fibers. In: Descomtes, M., Maier-Hein, L., Franz, A., Janin, P., Collins, D., Duchesne, S. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2017, vol 10433 Lecture Notes in Computer Science. Springer, Cham doi:10.1007/978-3-319-66182-7_63.
He, H., Zhang, F., Piefer, M., Krubs, N., Rathi, Y., Wells, W., O'Donnell, L., 2022. Model and predict age and sex in healthy subjects using brain white matter features: a deep learning approach. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). IEEE, pp. 1–5.
He, Y., Chen, Z., Evans, A., 2007. Small-world anatomical networks in the human brain revealed by cortical thickness from MRI. Cereb. Cortex 17 (10), 2407–2419.
He, Y., Chen, Z., Evans, A., 2008. Structural insights into aberrant topological patterns of large-scale cortical networks in Alzheimer's disease. J. Neurosci. 28 (18), 4756–4766.
Hochberg, Y., Ljungqvist, A., Ly, U., 2007. Whom you know matters: venture capital networks and investment performance. J. Finance 62 (1), 251–301.
Hoei, A.E., Kennard, R.W., 1970. Ridge regression: applications to nonorthogonal problems. Technometrics 12 (1), 69–82.
Jenkinson, M., Beckmann, C., Behrens, T., Woolrich, M., Smith, S., 2012. FSL Neuroimage 62 (2), 792–790.
Klein, A., Tourville, J., 2012. 101 labeled brain images and a consistent human cortical labeling protocol. Front. Neurosci. 6, 171.
Knothe, A., Anwander, A., Liptrot, M., Dyrby, T., 2015. Validation of tractography: comparison with manganese tracing. Hum. Brain Mapp. 36 (10), 4116–4134.
Kobeleva, A., López-González, D., Gao, W., 2021. Revisiting the relevant spatiotemporal scale underlying Whole-Brain dynamics. Front. Neurosci. 15, 715861. doi:10.3389/fnins.2021.715861.
Kurtzer, G., Sochat, V., Bauer, M., 2017. Singularity: scientific containers for mobility of computational workflows. PloS One 12 (5), e0177459.
Lin, Y., Baste, S., Wang, X., Bouda, F., 2020. Mapping brain-behavior networks using functional and structural connectome fingerprinting in the HCP dataset. Brain Behav. 10 (6), e01647.
Menze, B., 2020. Parcellation influence on the connectivity-based structure-function relationship in the human brain. Hum. Brain Mapp. 41 (5), 1167–1180.
O'Donnell, L., Golby, A., Westin, C., 2013. Fiber clustering versus the parcellation-based connectome. NeuroImage 80, 283–289.
O’Donnell, L., Westin, C., 2007. Automatic tractography segmentation using a high-dimensional white matter atlas. IEEE Trans. Med. Imaging 26 (11), 1562–1575.

Owens, M., Duda, B., Sweet, L., Mackillop, J., 2018. Distinct functional and structural neural underpinnings of working memory. NeuroImage 174, 463–471.

O’Donnell, L., Wells, W., Golby, A., Westin, C., 2012. Unbiased groupwise registration of white matter tractography. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 123–130.

O’Malley, J., Marsden, P., 2008. The analysis of social networks. Health Serv. Outcomes Res. Methodol. 8 (4), 222–269.

Park, H., Friston, K., 2013. Structural and functional brain networks: from connections to cognition. Science 342 (6158), 1238411.

Powell, M., Garcia, J., Yeh, F., Vettel, J., Verstynen, T., 2018. Local connectome phenotypes predict social, health, and cognitive factors. Natw. Neurosci. 21 (2), 86–105.

de Reus, M., Van den Heuvel, M., 2013. The parcellation-based connectome: limitations and extensions. NeuroImage 80, 397–404.

Ribe, S., Drunker, K., Knight, R., 2016. Choosing words: left hemisphere, right hemisphere, or both? Perspective on the lateralization of word retrieval. Ann. N. Y. Acad. Sci. 1369 (1), 111.

Roine, T., Jeurissen, B., Perrone, D., Acherman, J., Philips, W., Sijbers, J., Lemann, A., 2019. Reproducibility and intercorrelation of graph theoretical measures in structural brain connectivity networks. Med. Image Anal. 52, 56–67.

Rokem, A., Yeatman, J., Pestilli, F., Kay, K., Mezer, A., Van Der Walt, S., Wandell, B., 2015. Evaluating the accuracy of diffusion MRI models in white matter. PLoS One 10 (4), e0125272.

Rolls, E., Joliot, M., Tzourio-Mazoyer, N., 2015. Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas. NeuroImage 122, 1–5.

Schiffer, P., Tenberge, J., Wiendl, H., Meuth, S., 2017. Cortex parcellation associated whole white matter parcellation in individual subjects. Front. Hum. Neurosci. 11, 352.

Scally, D., 2010. Web-scale k-means clustering. In: Proceedings of the 19th International Conference on World Wide Web. ACM, pp. 1177–1178.

Shin, J., Rowley, J., Chowdhury, R., Joliceur, P., Klein, D., Grova, C., Rosa-Neto, P., Kobayashi, E., 2019. Inferior longitudinal fasciculus’ role in visual processing and language comprehension: a combined MEG-DTI study. Front. Neurosci. 13. doi:10.3389/fnins.2019.00875.

Sotiropoulos, S., Ibbadi, S., Xu, J., Andersson, J., Moeller, S., Auerbach, E., Glasser, M., Hernandez, M., Sapiro, G., Jenkinson, M., Feinberg, D., Yacoub, E., Lenglet, C., Van Essen, D., Ugurbil, K., Behrens, T., Consortium, W.-M.H., 2013. Advances in diffusion MRI acquisition and processing in the human connectome project. NeuroImage 80, 125–143.

Sporns, O., Tononi, G., Kötter, R., 2005. The human connectome: a structural description of the human brain. PLoS Comput. Biol. 1 (4), e22.

Squire, L., Stark, C., Clark, R., 2004. The medial temporal lobe. Annu. Rev. Neurosci. 27, 279–306.

Theaud, G., Houade, J., Boré, A., Rhenult, F., Morency, F., Descoteaux, M., 2020. Tractflow: A robust, efficient and reproducible diffusion MRI pipeline leveraging nextflow & singularity. NeuroImage 218, 116889.

Tian, Y., Zalesky, A., 2018. Characterizing the functional connectivity density of the insula cortex: subregions, diversity curves and behavior. NeuroImage 183, 716–733.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. 58 (1), 267–288.

Toga, A., Clark, K., Thompson, P., Shattuck, D., Van Horn, J., 2012. Mapping the human connectome. Neurosurgery 71 (1), 1–5.

Tournier, J., Smith, R., Raffelt, D., Tabbara, R., Dhollander, T., Pietsch, M., Christiaens, D., Jeurissen, B., Yeh, C., Connelly, A., 2019. Mitrix: a fast, flexible and open software framework for medical image processing and visualisation. NeuroImage 202, 116137.