Spatially-enhanced clusterwise inference for testing and localizing intermodal correspondence

Sarah M. Weinstein[a], Simon N. Vandekar[b], Erica B. Baller[c], Danni Tu[a], Azeez Adebimpe[c], Tinashe M. Tapera[c], Ruben C. Gur[c], Raquel E. Gur[c], John A. Detre[d], Armin Raznahan[e], Aaron F. Alexander-Bloch[c,f], Theodore D. Satterthwaite[c], Russell T. Shinohara[a], Jun Young Park[g,*]

[a]*Penn Statistics in Imaging and Visualization Center, Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA 19104, USA*
[b]*Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, 37232, USA*
[c]*Department of Psychiatry, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA 19104, USA*
[d]*Department of Neurology, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA 19104, USA*
[e]*Section on Developmental Neurogenomics, National Institute of Mental Health Intramural Research Program, Bethesda, MD 20892, USA*
[f]*Department of Child and Adolescent Psychiatry and Behavioral Science, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA*
[g]*Department of Statistical Sciences and Department of Psychology, University of Toronto, Toronto, ON M5S, Canada.*

Abstract

With the increasing availability of neuroimaging data from multiple modalities—each providing a different lens through which to study brain structure or function—new techniques for comparing, integrating, and interpreting information within and across modalities have emerged. Recent developments include hypothesis tests of associations between neuroimaging modalities, which can be used to determine the statistical significance of intermodal associations either throughout the entire brain or within anatomical subregions or functional networks. While these methods provide a crucial foundation for inference on intermodal relationships, they cannot be used to answer questions about where in the brain these associations are most pronounced. In this paper, we introduce a new method, called CLEAN-R, that can be used both to test intermodal correspondence throughout the brain and also to localize this correspondence. Our method involves first adjusting for the underlying spatial autocorrelation structure within each modality before aggregating information within small clusters to construct a map of spatially enhanced test statistics. Using structural and functional magnetic resonance imaging data from a subsample of children and adolescents from the Philadelphia Neurodevelopmental Cohort, we conduct simulations and data analyses where we illustrate the high statistical power and nominal type I error levels of our method. By constructing an interpretable map of group-level correspondence using spatially-enhanced test statistics, our method offers insights beyond those provided by earlier methods.

*Corresponding author
Email address: junjy.park@utoronto.ca (Jun Young Park)

Keywords: Clusterwise inference, Data integration, Intermodal correspondence, Permutation

1. Introduction

A growing emphasis on extracting, synthesizing, and interpreting patterns from multiple imaging modalities simultaneously has prompted innovation in methods development. One area of expanding interest is intermodal coupling, where researchers have sought to study and quantify correspondence between two modalities (e.g. between structural and functional brain measurements). Methodological developments in this area have included statistical estimation of individual-level maps measuring intermodal coupling [1, 2, 3, 4, 5] and hypothesis tests for establishing statistical significance of intermodal associations [6, 7, 8, 9].

Previous methods for statistical testing of intermodal correspondence have been primarily focused on addressing the question of whether or not coupling exists throughout the entire brain or within pre-defined subregions of the brain. Developed by Alexander-Bloch et al. [6] the spin test formalized the notion of using spatially-constrained randomization for testing intermodal correspondence using group-level imaging measurements. A few years later, Burt et al. [7] developed Brain Surrogate Maps with Autocorrelated Spatial Heterogeneity (BrainSMASH), using spatially constrained null modeling based on reconstructing variograms in spatially permuted data. More recently, we described the simple permuation-based intermodal correspondence test (SPICE), the first method to leverage subject-level as opposed to group-averaged brain maps and a traditional permutation framework—that is, permuting individual subjects' maps rather than using spatially-derived null models from group-level data [9].

Each of these recent methodological advancements for testing intermodal correspondence involves computing global test statistics of correspondence across the brain (or within subregions of the brain), without differentiating the degree of correspondence across the brain. Localizing where coupling is most prominent could provide important insight for what it means for two modalities to be strongly related—for instance, whether structure-function coupling across the brain is driven primarily by a subset of brain regions. The primary existing approach that has been used to localize correspondence involves a mass-univariate analysis, where measurements at each location in the brain are modeled separately. Without accounting for spatial autocorrelation, hypothesis tests may suffer inflated type I error rates or reductions in statistical power [8].

In our current work, we propose a novel hypothesis testing method that involves a combination of approaches from Weinstein et al. [9] (SPICE) and recent work by Park and Fiecas [10] that explicitly models spatial autocorrelation in cluster-level inference. Using Park and Fiecas [10]'s spatial modeling framework, our method leverages spatial information through cluster-level inference after adjusting for differences in the spatial autocorrelation structure of each modality. Using a permutation framework similar to our previous work [9], our new method controls type I error rates

at the nominal level and exhibits high statistical power for both global testing and localization of intermodal correspondence, as assessed in simulations and analyses of structure-function coupling in data from the Philadelphia Neurodevelopmental Cohort.

2. Methods

2.1. Notation and null hypothesis

We consider a data set consisting of subject-level measurements from two different neuroimaging modalities X and Y; for example, X could be a measurement derived from structural magnetic resonance imaging (MRI) data and Y a measure of brain function from functional MRI (fMRI) data. We define $X(v) = (X_1(v), \ldots, X_N(v))$ and $Y(v) = (Y_1(v), \ldots, Y_N(v))$ to be vectors of data from each modality at a single image location v (e.g. vertex or voxel), and let $X_i(\mathbf{v})$ and $Y_i(\mathbf{v})$ be the ith participant's two images including all location indices $\mathbf{v} = (1, \ldots, V)$. From here, we express our global null hypothesis that there is no correspondence between X and Y as follows:

$$H_0 : r_{X(v),Y(v)} = 0 \quad \text{for all } v = 1, \ldots, V, \tag{1}$$

where $r_{U,V}$ is the Pearson correlation between vectors U and V. In the presence of subject-level covariates $\mathbf{z}$ (e.g. age and sex), then the global null hypothesis is expressed in terms of partial correlation:

$$H_0 : r_{X(v),Y(v)|\mathbf{z}} = 0 \quad \text{for all } v = 1, \ldots, V,$$

where $r_{U,V|W}$ is the partial correlation between vectors U and V given nuisance covariates W. For notational simplicity, we use (1) throughout this paper unless mentioned explicitly. If H_0 is true, then there is no location v in the entire brain where there is an association between $X(v)$ and $Y(v)$. We will use permutation to test H_0.

2.2. CLEAN-R: Clusterwise inference leveraging spatial autocorrelations in neuro-imaging for testing intermodal correlations r

2.2.1. Leveraging modality-specific spatial autocorrelation

To our knowledge, among the existing methods for testing intermodal correspondence, Burt et al. [7]'s BrainSMASH is the only widely used method that includes a model of the spatial autocorrelation structure (although Alexander-Bloch et al. [6], Weinstein et al. [9] handle this in other ways that do not require explicit spatial modeling). Burt et al.'s null hypothesis for testing similarity between group-level (i.e. average) brain maps, denoted by $\overline{X}(\mathbf{v})$ and $\overline{Y}(\mathbf{v})$ states that any observed correspondence between $\overline{X}(\mathbf{v})$ and $\overline{Y}(\mathbf{v})$ is due to spatial autocorrelation and not genuine

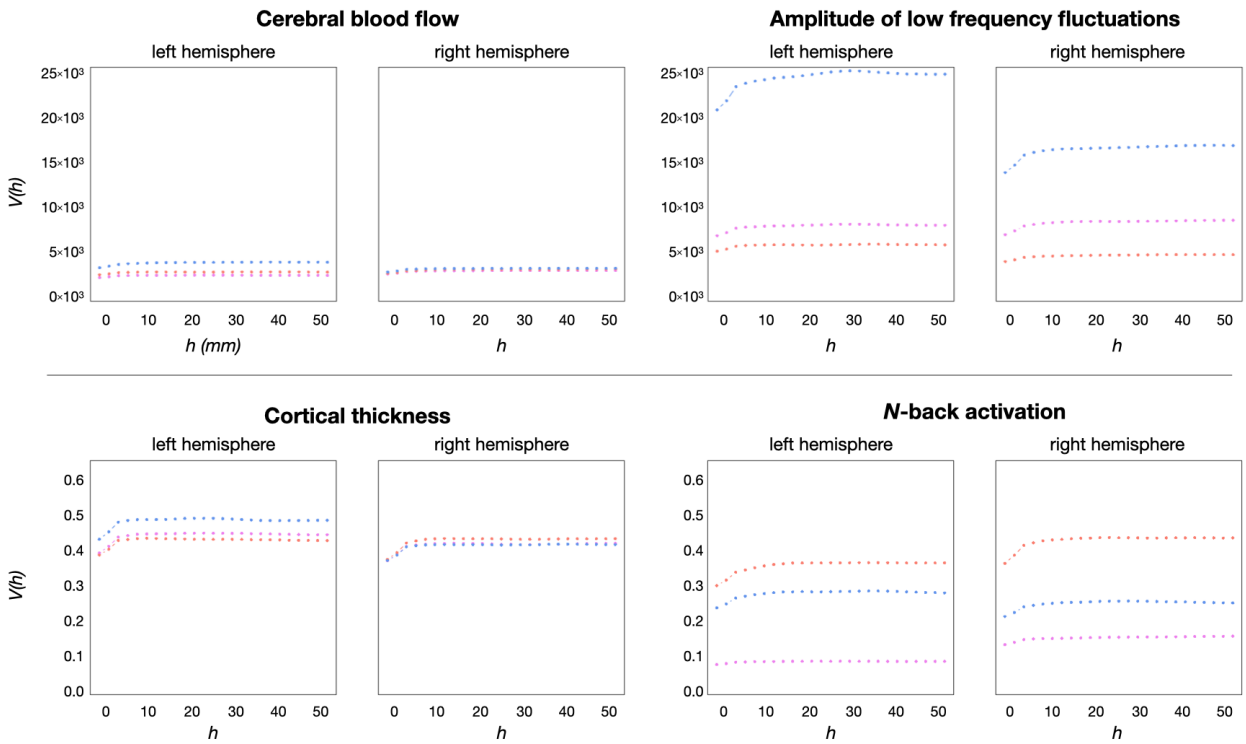

Figure 1: Empirical variograms for three different subjects (each represented by different color lines/dots) in two different pairs of modalities from the Philadelphia Neurodevelopmental Cohort, described in Section 3. Empirical variograms are computed separately for each subject for image measurements within at most h mm from one another.

relatedness of the two modalities. Their observed test statistic $\hat{r}_{\overline{X}(\mathbf{v})),\overline{Y}(\mathbf{v}))}$, the estimated correlation between $\overline{X}(\mathbf{v})$ and $\overline{Y}(\mathbf{v})$, is compared against a distribution of null test statistics, $\hat{r}_{\overline{X}(\mathbf{v}^{(k)}),\overline{Y}(\mathbf{v})}$ (for $k = 1, \ldots, K$ permutations of spatial locations), where each $\overline{X}(\mathbf{v}^{(k)})$ is a spatial autocorrelation-preserving surrogate brain map for modality X. Each surrogate map is obtained by first randomly shuffling vertices $\mathbf{v}$ into $\mathbf{v}^{(k)}$ in $\overline{X}(\mathbf{v})$ (thus destroying both the spatial autocorrelation structure and any genuine associations with $\overline{Y}(\mathbf{v})$), and then applying Gaussian smoothing over $\overline{X}(\mathbf{v}^{(k)})$ so that its spatial autocorrelation structure nearly matches that of $\overline{X}(\mathbf{v})$. While Burt et al.'s method is an important development in hypothesis testing involving neuroimaging data, several limitations warrant further consideration, including the implicit assumption that the spatial autocorrelation structure of both modalities is similar, and that stationarity can be assumed throughout the brain, as well as within anatomical subregions and functional networks. Additionally, Burt et al.'s BrainSMASH and other methods involving spatial null models (summarized by Markello and Misic [8]), are much more computationally intensive than "refitting-free" approaches (e.g. Weinstein et al. [9]'s SPICE test).

As illustrated in Figure 1, spatial autocorrelation structures can vary across subjects and modalities. Therefore, in this paper we advocate accounting for this variability in the context of inter-

modal correspondence testing by employing a framework recently introduced by Park and Fiecas [10]. Their method, called CLEAN (**c**lusterwise inference **le**veraging spatial **a**utocorrelations in **n**euroimaging) involves testing associations between a single neuroimaging modality $X_i(\mathbf{v})$ ($i = 1, \ldots, N$) (e.g. fMRI task activation) and a phenotype of interest (e.g. behavior measurement), while adjusting for nuisance variables $\mathbf{z}_i$ (e.g. age and sex). Park and Fiecas [10] first explicitly model the spatial autocorrelation structure of the entire image using subject-level data. Next, they adjust for this source of variation by scaling the multivariate score test statistic with an approximation to the inverse of the spatial covariance matrix. Finally, they aggregate test statistics across locations within clusters (such as those illustrated in Figure 2), thus mitigating the redundancy of spatially correlated observations. After combining cluster-level information, it is possible to construct an interpretable map of spatially enhanced test statistics for quantifying, testing, and localizing image-phenotype associations.

In our current work, we extend CLEAN to a multimodal setting and propose CLEAN for testing the correlation r between two modalities, or CLEAN-R. We consider one model for each modality:

$$X_i(\mathbf{v}) = \mathbf{Z}_i \boldsymbol{\alpha}_X + \epsilon_{X_i}(\mathbf{v}) \quad \text{and} \quad Y_i(\mathbf{v}) = \mathbf{Z}_i \boldsymbol{\alpha}_Y + \epsilon_{Y_i}(\mathbf{v}),$$

where $\mathbf{Z}_i = \mathbf{z}_i \otimes \mathbf{I}_V$ with the Kronecker product $\otimes$, $\epsilon_{X_i}(\mathbf{v}) \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_X)$, $\epsilon_{Y_i}(\mathbf{v}) \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_Y)$. From here, we further decompose $\boldsymbol{\Sigma}_X$ and $\boldsymbol{\Sigma}_Y$ into the sum of the their respective spatial and non-spatial variations:

$$\boldsymbol{\Sigma}_X = \sigma_X^2 \Phi_X(\phi_X) + \tau_X^2 \mathbf{I}_V \quad \text{and} \quad \boldsymbol{\Sigma}_Y = \sigma_Y^2 \Phi_Y(\phi_Y) + \tau_Y^2 \mathbf{I}_V. \tag{2}$$

Here, $\Phi_X(\phi_X)$ and $\Phi_Y(\phi_Y)$ are the assumed spatial autocorrelation functions and τ_X^2 and τ_Y^2 are the nugget effects measuring residual non-spatial variability in $X_i(\mathbf{v})$ and $Y_i(\mathbf{v})$. For each spatial autocorrelation function $\Phi_X(\phi_X)$ and $\Phi_Y(\phi_Y)$, we specify an assumed structure (e.g. exponential or squared exponential), which can be determined via exploratory analysis. Based on Figure 1, we assume the exponential function for each of the modalities considered in this study (described in Section 3); that is, we assume that the (v, v^*)th element of $\Phi_X(\phi_X)$ and $\Phi_Y(\phi_Y)$ are $\exp(-\phi_X \times d_{v,v^*})$ and $\exp(-\phi_Y \times d_{v,v^*})$ respectively, where d_{v,v^*} is the geodesic distance between locations v and v^*. Park and Fiecas [10] provide a detailed description of the estimation process for Equation (2) via a method-of-moments approach.

Equation (2) provides an explicit random effects formulation of the error process, which is equivalent to setting $\epsilon_{X_i}(\mathbf{v}) = b_{X_i}(\mathbf{v}) + \delta_{X_i}(\mathbf{v})$ where $b_{X_i}(\mathbf{v}) \sim MVN(0, \sigma_X^2 \Phi(\phi_X))$, $\delta_{X_i}(\mathbf{v}) \sim MVN(0, \tau_X^2 I)$ and $b_{X_i}(\mathbf{v})$ and $\delta_{X_i}(\mathbf{v})$ are independent. As $b_{X_i}(\mathbf{v})$ and $b_{Y_i}(\mathbf{v})$ are components that are explained solely by each modality's spatial autocorrelation and not by any other factors such

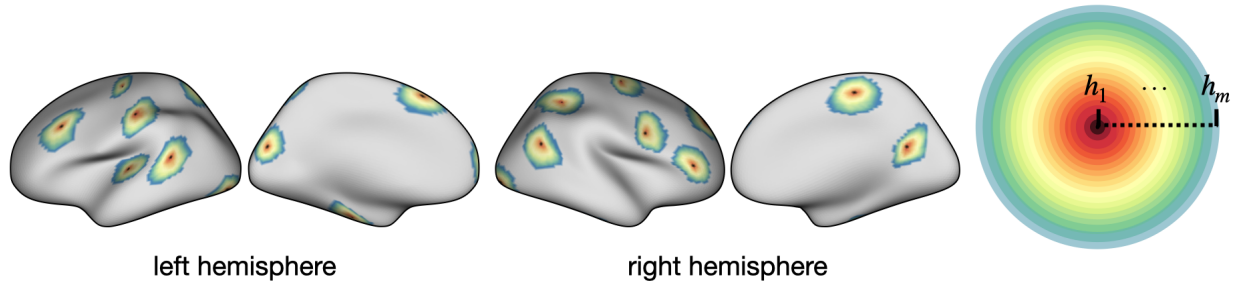

Figure 2: Example clusters with maximum radius of $h_m = 20$mm in the left and right hemispheres. Our proposed method for clusterwise inference compares test statistics aggregated across $v^* \in N_{h_j}(v)$, for all locations $v*$ in the neighborhood of a specific location v, such that $d_{v,v^*} < h_j), j = 1, \ldots, m$. Note that for images represented as cortical surfaces, we use geodesic distance.

as intermodal correspondence, we view these as unwanted variations and remove them by instead using $\hat{\delta}_{X_i}(\mathbf{v})$ and $\hat{\delta}_{Y_i}(\mathbf{v})$, defined by:

$$\hat{\delta}_{X_i}(\mathbf{v}) = \hat{\tau}_X^2 \widehat{\mathbf{\Sigma}}_X^{-1} \left(X_i(\mathbf{v}) - \mathbf{Z}_i \widehat{\boldsymbol{\alpha}}_X\right) \quad \text{and} \quad \hat{\delta}_{Y_i}(\mathbf{v}) = \hat{\tau}_Y^2 \widehat{\mathbf{\Sigma}}_Y^{-1} \left(Y_i(\mathbf{v}) - \mathbf{Z}_i \widehat{\boldsymbol{\alpha}}_Y\right) \tag{3}$$

for each subject ($i = 1, \ldots, N$). Thus, we characterize $\hat{\delta}_{X_i(\mathbf{v})}$ and $\hat{\delta}_{Y_i(\mathbf{v})}$ to be transformations of $X_i(\mathbf{v})$ and $Y_i(\mathbf{v})$, respectively, which have been adjusted for their unique underlying spatial autocorrelation structures. Note that $\hat{\delta}_{X_i}(\mathbf{v})$ is equivalent to $\hat{\epsilon}_{X_i}(\mathbf{v}) - \hat{b}_{X_i}^{\text{BLUP}}(\mathbf{v})$ if the variance components are estimated by maximizing the restricted maximum likelihood (REML) [11], which in practice is very close to the method-of-moment approach.

While inverting $\widehat{\mathbf{\Sigma}}_X$ and $\widehat{\mathbf{\Sigma}}_Y$ is still computationally intensive with high dimensional imaging data, we approximate the inverse of $\widehat{\mathbf{\Sigma}}_X$ and $\widehat{\mathbf{\Sigma}}_Y$ with $\widehat{\mathbf{\Sigma}}_{X,\text{NNGP}}^{-1}$ and $\widehat{\mathbf{\Sigma}}_{Y,\text{NNGP}}^{-1}$, respectively, which are estimated via a nearest-neighbor Gaussian process (NNGP) [12, 13]. NNGP is a computationally efficient way to obtain a close approximation of the inverse of the parametric spatial covariance matrix using information from a subset of neighbors, which has been adopted in CLEAN [10]. For our extension of CLEAN to test intermodal correspondence described in the next section (CLEAN-R), estimation of and inference on correspondence between the two modalities proceeds using $\hat{\delta}_{X(v)} = (\hat{\delta}_{X_1(v)}, \hat{\delta}_{X_2(v)}, \ldots, \hat{\delta}_{X_N(v)})$ and $\hat{\delta}_{Y(v)} = (\hat{\delta}_{Y_1(v)}, \hat{\delta}_{Y_2(v)}, \ldots, \hat{\delta}_{Y_N(v)})$ for each vertex.

2.2.2. Proposed spatially-enhanced clusterwise inference for testing intermodal correspondence

In this section we describe the next steps for implementing spatially-enhanced clusterwise inference for testing intermodal correspondence, using the spatially-adjusted imaging data ($\hat{\delta}_X(v)$ and $\hat{\delta}_Y(v)$) as input to compute location-specific and cluster-level test statistics.

First, we compute location-specific association measures using the Fisher Z-transformed Pearson correlation between the two modalities at each location:

$$\hat{\gamma}(v) = \operatorname{atanh}\left(\hat{r}_{\hat{\delta}_X(v),\hat{\delta}_Y(v)}\right) = \frac{1}{2}\log\left(\frac{1+\hat{r}_{\hat{\delta}_X(v),\hat{\delta}_Y(v)}}{1-\hat{r}_{\hat{\delta}_X(v),\hat{\delta}_Y(v)}}\right), \tag{4}$$

where each $\hat{\gamma}(v)$ is normally distributed with mean 0 under the null hypothesis (Equation (1)).

Second, we compute cluster-level test statistics using $\hat{\gamma}(v)$ in Equation (4) as a building block. Let $N_h(v)$ denote the set of vertices that fall within radius h from a central vertex v. $N_h(v)$ is defined for each location and different possible radii, allowing us to select an optimal cluster size, as described below. Suppose for each vertex, we consider clusters defined by m different radii, $h_1, \ldots, h_m$. Assuming $h_1 = 0$, $N_{h_1}(v)$ is the cluster consisting of only the central vertex v. Thus, the correspondence measure for this cluster is equal to $\hat{\gamma}(v)$. Test statistics for clusters with radius up to h_j ($j = 2, \ldots, m$) are computed as the sums of univariate correspondence measures $\hat{\gamma}(v^*)$ for all $v^* \in N_{h_j}(v)$ (e.g. $\hat{\gamma}(v^*)$, $\sum_{v^* \in N_{h_2(v)}} \hat{\gamma}(v^*)$, $\ldots$, $\sum_{v^* \in N_{h_m(v)}} \hat{\gamma}(v^*)$).

Third, we construct an empirical null distribution for our cluster-based test statistics using permutations. Because our null hypothesis is expressed in terms of within-person intermodal correlations, randomly re-ordering subjects within one modality destroys any associations measured for two maps derived from the same individual; thus, an empirical null distribution made up of inter-individual correlations is appropriate for testing the null hypothesis (Equation (1)). In the presence of covariate information $\mathbf{z}$, we preserve the order of the covariates and the shuffled modality in each permutation because our hypothesis addresses residual (partial) correlation of two modalities after regressing out non-imaging covariates.

As a further illustration, let $\hat{\gamma}^{(k)}(v)$ denote the estimated Fisher Z-transformed Pearson correlation after the kth permutation of $\hat{\delta}_X(v)$. This is analogous to randomly shuffling participant information in $X(v)$, given that $\hat{\delta}_X(v)$ and $\hat{\delta}_Y(v)$ are functions of $X(v)$ and $Y(v)$, and permuting the entries of $\hat{\delta}_X(v)$ to destroy any associations with $\hat{\delta}_Y(v)$ achieves the same goal as permuting the entries of $X(v)$ to destroy associations with $Y(v)$. Cluster-level test statistics forming the empirical null distribution used to test the null hypothesis are, like in the observed data, the sums of univariate test statistics $\hat{\gamma}^{(k)}(v^*)$ for all $v^* \in N_{h_j}(v)$ for each radius h_j ($j = 1, \ldots, m$) (e.g. $\hat{\gamma}^{(k)}(v)$, $\sum_{v^* \in N_{h_2(v)}} \hat{\gamma}^{(k)}(v^*)$, $\ldots$, $\sum_{v^* \in N_{h_m(v)}} \hat{\gamma}^{(k)}(v^*)$).

Location-specific test statistics are defined as the maximum of the square of observed cluster-level test statistics (sums of $\hat{\gamma}(v)$ across vertices in each cluster) are scaled elementwise by their empirical standard deviations (from the permutation null distribution), giving our proposed "spatially-enhanced" test statistic at vertex v:

$$T(v) = \max\left\{ \frac{\left(\sum_{v^* \in N_h(v)} \hat{\gamma}(v)\right)^2}{\widehat{\mathrm{Var}}_{H_0}\left(\sum_{v^* \in N_h(v)} \hat{\gamma}(v)\right)} \,\middle|\, h = h_1, \ldots, h_m \right\}, \tag{5}$$

and a global statistic for testing H_0 (Equation (1)) across all locations is

$$T = \max_{v=1,\ldots,V} T(v). \tag{6}$$

The empirical null distribution of T is comprised of $T^{(1)}, \ldots, T^{(K)}$ which, similar to $T(v)$ (Equation (5)), use the sums of $\hat{\gamma}^{(k)}(v)$'s as building blocks:

$$T^{(k)}(v) = \max \left\{ \frac{\left(\sum_{v^* \in N_h(v)} \hat{\gamma}^{(k)}(v)\right)^2}{\widehat{\mathrm{Var}}_{H_0}(\sum_{v^* \in N_h(v)} \hat{\gamma}(v))} \middle| h = h_1, \ldots, h_m \right\}, \tag{7}$$

and similar to T (Equation (6)) uses the the maximum

$$T^{(k)} = \max_{v=1,\ldots,V} T^{(k)}(v) \tag{8}$$

to form a global null distribution for testing at any individual location v (using any $T(v)$ (Equation (5)) as a test statistic), or across all locations in the brain using T (Equation (6) as the test statistic. For a pre-specified family-wise error rate (FWER) α, let t_α denote the $(1-\alpha) \times 100$th percentile of the empirical null distribution $(T^{(1)}, T^{(2)}, \ldots, T^{(K)})$, which serves as a critical value for both global and vertex-level hypothesis tests of H_0, which maintains FWER smaller than $\alpha \times 100\%$ by using a global null distribution for each tests. A global test rejects H_0 if $T > t_\alpha$, and to test H_0 at a single vertex v, we reject the null hypothesis if $T(v) > t_\alpha$. Using $\alpha = 0.05$, we expect to reject H_0 approximately 5% of the time when H_0 is true.

We note that the proposed permutation strategy is computationally efficient because shuffling subject information (and corresponding covariate information) does not affect the variance component estimates of the original model. Therefore, permutation is not limited to shuffling $X_i(\mathbf{v})$ but applicable to $\hat{\delta}_i(\mathbf{v})$ from the original model, which is very practical in implementation. In practice, applying CLEAN-R to a sample of N=50—with nearly 10,000 vertices in each hemisphere for each person and using $K = 10,000$ permutations—takes under 14 minutes. This includes the time to fit four covariance regression models (one for each modality in each hemisphere), and it can be even more efficient with parallelization.

2.2.3. Testing and localizing group differences in intermodal correspondence

CLEAN-R can be easily extended to test and localize group differences in intermodal correspondence. In this setting, we restate the null hypothesis (Equation (1)) in terms of the difference in intermodal correlations between groups A and B at vertex v (e.g., $H_0 : r_{X_A(v),Y_A(v)} - r_{X_B(v),Y_B(v)} = 0$ for all v). Let $\hat{\gamma}_A(v)$ and $\hat{\gamma}_B(v)$ denote separately estimated location-specific association measures (as in Equation (4)) for subgroups A and B, respectively. To test and localize differences in intermodal correspondence these two groups, we replace $\hat{\gamma}(v)$ with $(\gamma_A(v) - \gamma_B(v))$ in Equations (5) and (7). Specifically, the spatially enhanced statistic used for testing differences between groups at

vertex v is given by

$$T_{A-B}(v) = \max \left\{ \frac{\left(\sum_{v^* \in N_h(v)} \left(\hat{\gamma}_A(v) - \hat{\gamma}_B(v) \right) \right)^2}{\widehat{\mathrm{Var}}_{H_0} \left(\sum_{v^* \in N_h(v)} \left(\hat{\gamma}_A(v) - \hat{\gamma}_B(v) \right) \right)} \middle| h = h_1, \ldots, h_m \right\} \tag{9}$$

and a global test for group differences can be implemented using $T_{A-B} = \max_{v=1,\ldots,V} T_{A-B}(v)$ (similar to Equation (6)). After permuting participants separately within each group, an empirical null distribution for both global and local tests of group differences in correspondence is obtained analogously to Equations (7) and (8); that is,

$$T^{(k)}_{A-B}(v) = \max \left\{ \frac{\left(\sum_{v^* \in N_h(v)} \left(\hat{\gamma}^{(k)}_A(v) - \hat{\gamma}^{(k)}_B(v) \right) \right)^2}{\widehat{\mathrm{Var}}_{H_0} \left(\sum_{v^* \in N_h(v)} \left(\hat{\gamma}_A(v) - \hat{\gamma}_B(v) \right) \right)} \middle| h = h_1, \ldots, h_m \right\}, \tag{10}$$

and similar to $T^{(k)}$ (Equation (8)), $T^{(k)}_{A-B} = \max_{v=1,\ldots,V} T^{(k)}_{A-B}(v)$.

3. Simulations and data analysis

3.1. Multimodal neuroimaging data from the Philadelphia Neurodevelopmental Cohort

The Philadelphia Neurodevelopmental Cohort (PNC) is a large-scale research initiative led by the Brain Behavior Laboratory at the University of Pennsylvania and Children's Hospital of Philadelphia. PNC researchers recruited N=9,498 children and adolescents ages 8-21 years old who consented or received parental consent to participate. A subset of 1,601 participants from the PNC underwent multimodal neuroimaging scans at the Hospital of University of Pennsylvania. All participants in this subset were scanned in the same MRI scanner, following protocols described in Satterthwaite et al. [14].

Cortical reconstruction of structural and functional images was implemented using Freesurfer version 5.3 [15]. Voxelwise measurements from each imaging modality were resampled to the fsaverage5 atlas, consisting of 10,242 vertices per hemisphere per subject. For all analyses, the number of vertices in the left and right hemispheres reduces to 9,354 and 9,361, respectively after we exclude portions of each hemisphere corresponding to the medial wall, a byproduct of working with a surface-based representation of brain images [16]. Vertices in the medial wall are added back for visualization purposes, which is done using the `fsbrain` package in `R` [17].

3.1.1. H_0 in context

In this study, we consider tests of correspondence in two different pairs of modalities from the PNC. In each context, we test the global null hypothesis from Equation (1), which says that

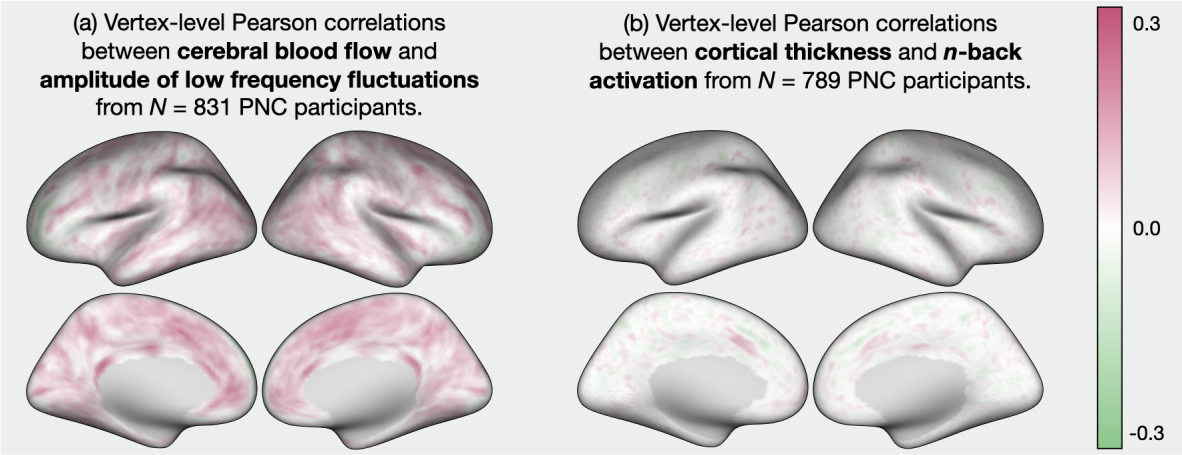

Figure 3: Maps of vertex-level partial correlations (adjusting for age and sex) between (a) CBF and ALFF in N=831 PNC participants and (b) cortical thickness and n-back task activation in N=789 PNC participants.

among all locations in the brain, there is no location where there is an association between the two modalities.

Previous work has highlighted the relationship between cerebral blood flow (CBF) and measures of brain function. For instance, Baller et al. [18] documented the evolution of coupling between CBF and resting-state fMRI measurements of amplitude of low frequency fluctuations (ALFF) over the course of childhood and adolescence among N=831 participants from the PNC (a subset meeting quality control criteria for both modalities and other health-related inclusion criteria). Baller et al. found significant coupling between CBF and ALFF throughout the cortex, noting differences in the degree of coupling among different age groups and by sex. In our current study, we apply our proposed method to test our null hypothesis CBF-ALFF coupling. Given the findings from Baller et al., we expect to reject H_0 throughout the cortex, and our method will also allow us to localize regions where coupling is strongest.

It is also informative to consider a setting in which we do not expect to see strong intermodal associations. Thus, we consider correspondence tests between cortical thickness and n-back task activation, as previous work has not suggested we would expect either global or local correspondence between these modalities [9]. Cortical thickness is quantified as the minimum distance between pial and white matter surfaces. In the n-back task, participants are presented with a sequence of visual stimuli and asked to indicate (with a press of a button) if the stimulus before them matches the nth previous stimulus. In this study, we quantify task activation as the percent change in functional brain activation comparing the n=2 and n=0-back sequences. We include a subset of N=789 participants from the PNC meeting quality-control criteria for both cortical thickness and n-back activation.

For exploratory purposes, vertex-level Pearson correlations from the full subsamples of both

imaging modality pairs are mapped in Figure 3. As these correlations are estimated in relatively large samples of around 800 participants each, the maps in Figure 3 may provide insight into which regions we might expect to find evidence for localized intermodal correspondence, and whether our method makes it possible to replicate such associations in smaller sample sizes. In small-sample settings, a mass-univariate approach might not reveal patterns of correspondence like those observed in larger samples. However, we expect our proposed spatially-enhanced testing procedure to be powerful even in small samples.

3.2. Simulation study in the Philadelphia Neurodevelopmental Cohort

We conduct a plasmode simulation study by taking 1,000 repeated random subsamples of sizes N=50 and 100 from the PNC to test correspondence first between CBF and ALFF and then between cortical thickness and n-back. For each pair of modalities, we compare global type I error rates and statistical power from the proposed CLEAN-R approach to three other methods for testing intermodal correspondence using participant-level data: a mass-univariate approach, a clusterwise approach that does not adjust for spatial autocorrelation, and the SPICE permutation test described in our previous work [9].

For CLEAN-R, candidate clusters are constructed using radii ranging from $h_1 = 0$ to $h_m =$ 20mm in increments of 1mm (as in Figure 2), so that each $T(v)$ (Equation (5)) is the maximum test statistic centered at a given vertex v, which aggregates information from other vertices at most 20mm from v. For computational efficiency, a sparse distance matrix restricted to the top 50% of distances is used to estimate modality-specific spatial parameters while using information from nearby vertices. To assess the added benefit of adjusting for spatial autocorrelation, a "spatially un-adjusted" clusterwise approach is also considered. This involves a similar set of test statistics and null distribution as in CLEAN-R, but we replace the spatially adjusted data ($\hat{\delta}_{X_i}(\mathbf{v})$ and $\hat{\delta}_{Y_i}(\mathbf{v})$) with spatially un-adjusted data ($\hat{\epsilon}_{X_i}(\mathbf{v})$ and $\hat{\epsilon}_{Y_i}(\mathbf{v})$). A mass-univariate approach involves computing location-specific associations based on spatially un-adjusted residuals (i.e. $T(v) = \text{atanh}(\hat{r}_{\hat{\epsilon}_X(v),\hat{\epsilon}_Y(v)})$). In the SPICE permutation test [9], the test statistic of interest is the mean within-subject correlation between two modalities (i.e. $T_{\text{SPICE}} = \frac{1}{N}\sum_{i=1}^{N} \hat{r}_{X_i(\mathbf{v}),Y_i(\mathbf{v})}$), and an empirical null distribution is computed based on permuting one of the modalities across subjects and re-computing T_{SPICE}.

For each method, we use 999 permutations in each test and compute global statistical power (the rate of rejecting H_0 in the observed data) as the proportion of simulations where the observed p-value is smaller than $\alpha = 0.05$. The type I error rate for each method is based on a setting where each test is conducted using a draw from the null distribution (i.e. a test statistic computed in permuted data).

To identify specific vertices and broader regions of the brain where statistical power for rejecting H_0 is high, we compare the spatially enhanced vertex-level test statistics $T(v)$ (from Equation (5))

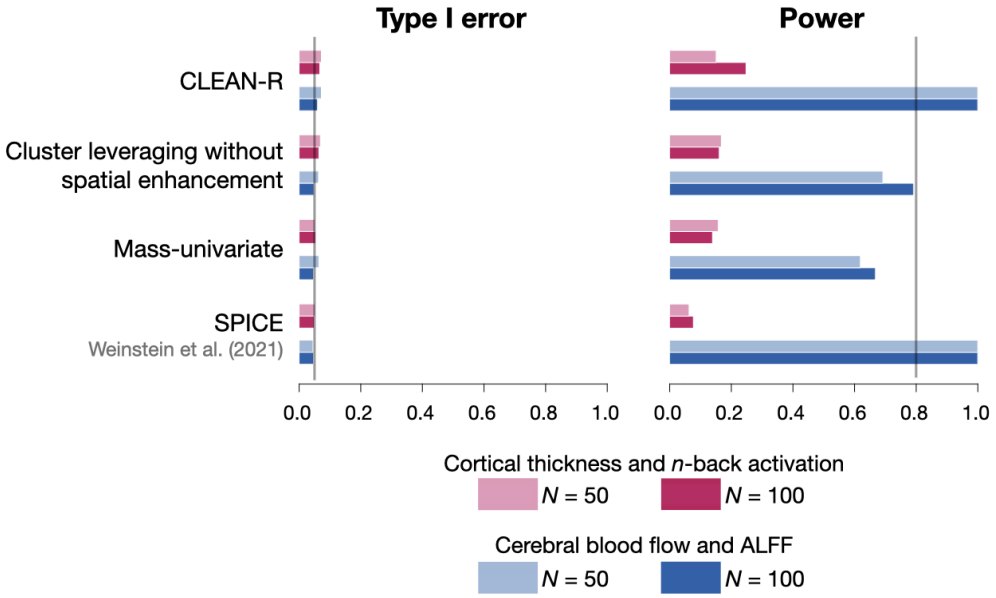

Figure 4: Rate of rejecting global null hypothesis in plasmode simulation settings for testing correspondence between cortical thickness and n-back activation (red bars) and cerebral blood flow and ALFF (blue bars) in 1,000 random subsamples of N=50 and 100 from the Philadelphia Neurodevelopmental Cohort (PNC). Type I error rates (left) are based on a simulation setting in which the "observed" data consist of bi-modal data from two different subjects. Observed rejection rates reflect the power of the test if H_0 is false.

to the global threshold t_α. For each $v = 1, \ldots, V$, location-specific power is the proportion of simulations where $T(v) > t_\alpha$. The SPICE test does not estimate location-specific test statistics, so it is not possible to map this method's statistical power throughout the cortex.

3.2.1. Plasmode simulation results: type I error and power

Simulation results for our global tests of intermodal correspondence are included in Figure 4. All four methods show type I error rates close to the nominal 5% level for testing CBF/ALFF and cortical thickness/n-back correspondence in simulation settings where the null hypothesis is true. Statistical power for tests of correspondence between CBF and ALFF suggest that CLEAN-R performs similarly to the SPICE test and outperforms the spatially un-adjusted clusterwise approach and the mass-univariate approach. Because the SPICE test does not provide localization of intermodal coupling, CLEAN-R is especially appealing as it achieves high statistical power and also has the added benefit of spatial localization, which we illustrate and discuss below.

For tests of cortical thickness and n-back, no method exhibited high statistical power. However, CLEAN-R rejected the null hypothesis in more simulations than any other method. Given the small regions exhibiting signal in the full sample correlation maps (Figure 3b), correspondence between cortical thickness and n-back may be localized to extremely small regions, making statistical power lower for methods that do not leverage enhanced cluster-level statistics as in CLEAN-R.

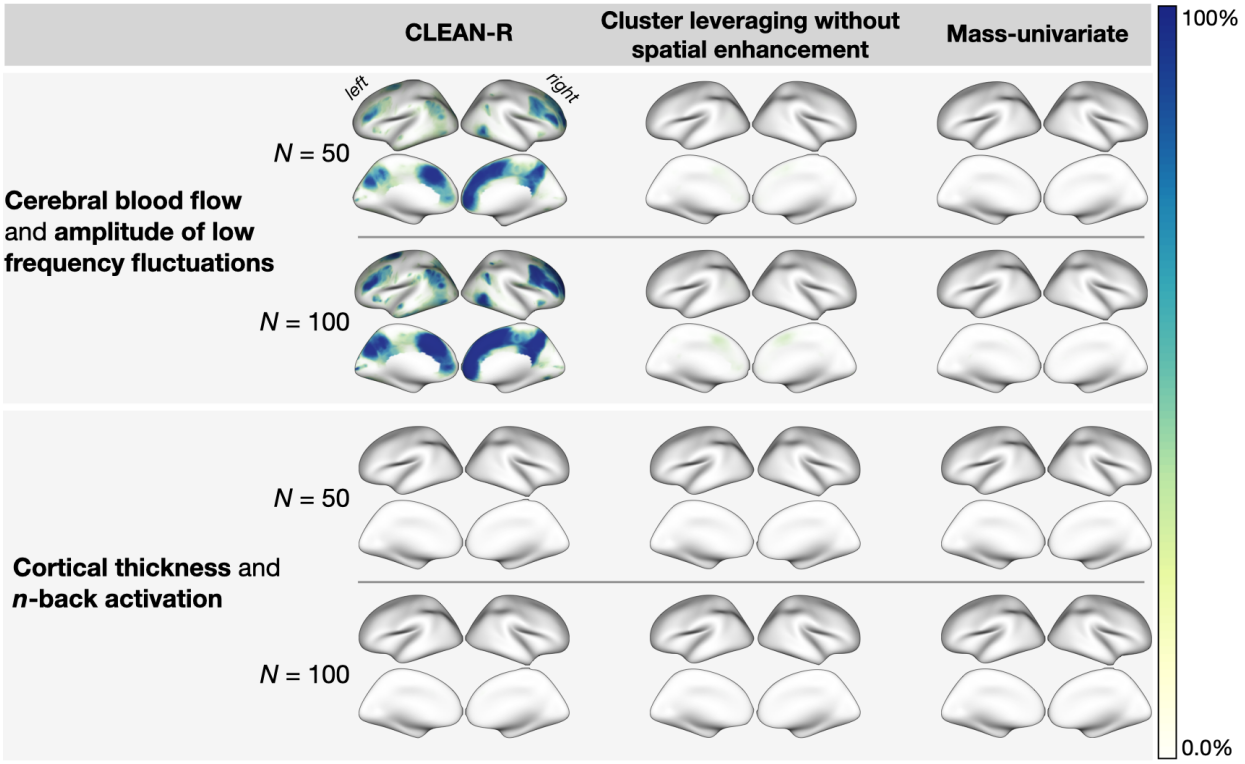

Figure 5: Statistical power from three different methods mapped throughout cortex. Map values between 0 and 100% indicate the proportion of simulations where H_0 was rejected at a particular location. Power for global tests of coupling between each pair of modalities according to each method are provided in Figure 4. At each vertex v, statistical power is the proportion of the 1000 simulations where the vertex-level test statistic (e.g. $T(v)$ as in Equation (5)) exceeds the FWER-controlling threshold t_α.

3.2.2. Plasmode simulation results: localization of intermodal correspondence

Figure 5 maps statistical power from each of the three methods that involve estimating vertex-level test statistics (i.e. $T(v)$ from Equation (5)). For tests of correspondence between CBF and ALFF, CLEAN-R is the only method to outline clear regions of strong statistical power for tests of correspondence, adding to our understanding of anatomical regions and functional networks that may be driving significant results in global correspondence tests. For tests of correspondence between cortical thickness and n-back, no method delineates specific areas of the brain with high statistical power.

3.2.3. Comparison to spatial null modeling approaches

The resampling-based design of the simulations above is a suitable procedure for computing the power and type I error of tests that involve participant-level data; however, this approach does not translate well for computing power of the spin test and BrainSMASH, which use spatial null modeling approaches. In particular, group-averaged maps from tend to be fairly similar across different subsets; thus, we would get very similar results across different simulations based on the

procedure described above, making estimates of statistical power invalid.

Despite inherent differences between CLEAN-R (and other methods that use participant-level data) and spatial null modeling methods, a comparison to these methods is warranted as these approaches have been widely adopted in studies of intermodal correspondence over the last few years [8]. Therefore, we apply the spin test and BrainSMASH to group-averaged data from each subsample of N=50 and 100 considered in our simulations above for comparison. For the spatial null modeling methods, we use only 199 permutations (spherical rotations in the spin test and permutation of vertices in BrainSMASH in this context) due to the computational burden of running many simulations and permutations for each method. To make the spatial model refitting involved in each BrainSMASH permutation computationally feasible to repeat in 1000 simulations, we use parcellated brain maps with $V = 501$ regions per hemisphere described by Schaefer et al. [19], instead of the Freesurfer vertex-level data with nearly 10,000 vertices per hemisphere.

In tests of correspondence between group-level maps of CBF and ALFF, we find $p < 0.05$ from the spin test in 97.7% ($N = 50$) and 100% of ($N = 100$) simulations and in 100% of simulations for both $N = 50, 100$ from BrainSMASH. However, for cortical thickness and n-back, there was no case that resulted in a significant p-value from either method in either sample size.

Although we cannot quantify statistical power of the methods using group-level data using a resampling-based procedure, it is possible to evaluate these methods' type I error rates given the variability across different draws from each method's null distribution, which are used as test statistics for computing a p-value under the null. Under the "rotation null" hypothesis where the rotated mean maps from a modality are compared to the observed mean maps of another, the spin test's type I error rates are 3.7% ($N = 50$) and 3.6% ($N = 100$) between CBF and ALFF, and 5.1% ($N = 50$) and 5% ($N = 100$) between cortical thickness and n-back. When using spatial autocorrelation-preserving surrogate brain maps for testing BrainSMASH's null hypothesis, we find type I error rates of 7.9% ($N = 50$) and 7.2% ($N = 100$) between CBF and ALFF, and 9.1% ($N = 50$) and 8.4% ($N = 100$) between cortical thickness and n-back. As Burt et al. [7] discuss in detail, a higher degree of spatial autocorrelation corresponds to increased false positives. Our finding of type I error rate inflation may be suggestive that, in our current implementation of BrainSMASH, the generated surrogate maps may not properly mimic the spatial structure of the observed data. Furthermore, given the variability in the spatial autocorrelation structures of different modalities (Figure 1), it is conceivable that BrainSMASH may be sensitive to the choice of which modality is used for generating surrogate maps. To our knowledge, the impact of the choice of which modality from which to generate surrogate brain maps on statistical inference and false positive rates has not yet been investigated. However, we note that all methods that involve permutation across participants (CLEAN-R, SPICE, and mass-univariate) are inherently insensitive to the choice of modality involved in permutations.

3.3. Subsample analysis: global testing and localization of sex differences

Finally, we apply our proposed method to test associations between CBF and ALFF first within stratified subgroups of males ($N = 353$) and females ($N = 478$) from the PNC and then to test and localize group differences using the methodology described in Section 2.2.3. Here, we focus on CBF/ALFF only here as cortical thickness/n-back did not show noticeable power or localization. We compare the results of CLEAN-R to the spatially-unadjusted version of our method and to the mass-univariate approach. Implementation of each method is done using the same parameters specified in Section 3.2 (e.g. candidate cluster radii ranging from $h_1 = 0$ to $h_m = 20$mm).

Results from sex-stratified global tests of correspondence using each method as well as maps of test statistics for localization of intermodal correspondence among males and females are shown in Figure 6(a). Un-adjusted global p-values for each method (shown above each set of maps) suggest that all three methods may detect significant overall correspondence between CBF and

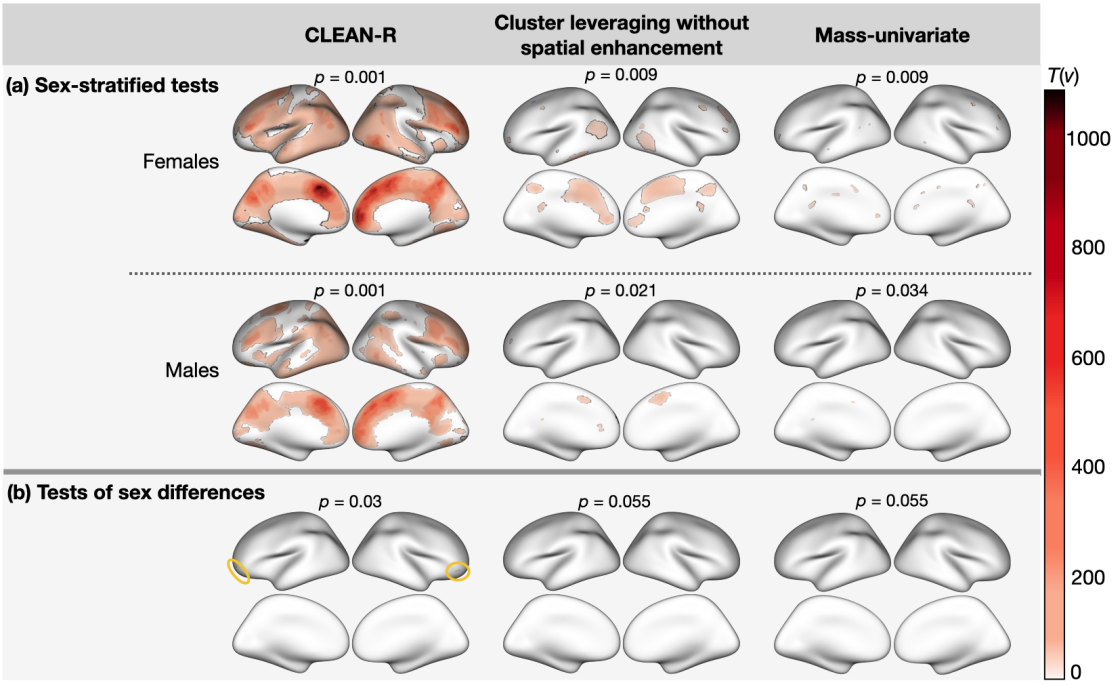

Figure 6: Testing and localizing intermodal correspondence between cerebral blood flow (CBF) and amplitude of low frequency fluctuations (ALFF) among $N = 478$ females and $N = 353$ males in the Philadelphia Neurodevelopmental Cohort. The p-values from global significance tests are provided above each map. Panel (a) includes test statistic maps from CLEAN-R, clusterwise inference without spatial adjustment, and mass-univariate tests of correspondence between CBF and ALFF applied separately in males and females. We used permutation to methods control FWER at the rate of α using permutation, and thresholded vertices if their statistics is less than the FWER-controlling threshold. Panel (b) includes thresholded test statistic maps from testing sex differences in correspondence between CBF and ALFF, using an extension of the original CLEAN-R methodology discussed in Section 2.2.3. CLEAN-R localizes significant sex differences in intermodal correspondence to four vertices (two in each hemisphere), which are encircled in yellow. Neither the clusterwise method without spatial adjustment nor the mass-univariate approach localize intermodal correspondence to any specific vertices.

ALFF in both males and females. However, CLEAN-R provides the most interpretable maps for localizing CBF/ALFF associations, by amplifying the signal of intermodal correspondence and detecting signal even in very small regions that are otherwise undetected by the mass-univariate and spatially un-adjusted implementation of CLEAN-R.

In Figure 6(b), we compare sex differences in correspondence between CBF and ALFF. In global tests for overall sex differences in correspondence throughout the entire brain, CLEAN-R is the only to suggest localized correspondence between CBF and ALFF at any vertices. In a study of the same population, Baller et al. [18] evaluated sex differences in intermodal coupling [3], and using an adaptation of Alexander-Bloch et al. [6]'s spin test for testing network enrichment, found significant sex differences in CBF-ALFF coupling that were especially pronounced within Yeo et al. [20]'s frontoparietal control network. We note that two of the four significant vertices identified by CLEAN-R lie in the frontoparietal control network; however, given the small number of significant vertices, a follow-up study and power analysis would be necessary to confirm such findings.

4. Discussions

In this paper, we proposed CLEAN-R, a new method for statistical inference on and localization of correspondence between two neuroimaging modalities. Compared to existing methods for testing the *existence* of intermodal correspondence (global testing), CLEAN-R offers additional insight on spatial *localization* of correspondence in an interpretable way. While the idea of localization itself is a classical topic in neuroimaging—exemplified by mass-univariate analyses—the proposed method dramatically improves reproducibility (both in global testing and localization) by accounting for modality-specific spatial autocorrelation. Our method expands the scope of research questions including, but not limited to, (i) testing intermodal correspondence after adjusting for covariates and (ii) testing differences in intermodal correspondence between two groups.

Through plasmode simulations and subgroup analyses from the Philadelphia Neurodevelopmental Cohort (PNC), we illustrated that CLEAN-R exhibits high statistical power for detecting global correspondence between cerebral blood flow (CBF) and amplitude of low frequency fluctuations (ALFF). While the finding that coupling between CBF and ALFF exists is not new, our methodology for approaching this question offers insight beyond what has been provided by earlier methods. Unlike previous methods, CLEAN-R provides interpretable maps that may inform our understanding of which parts of the brain are driving statistically significant global tests. This aspect of intermodal correspondence testing was not part of spatial null modeling approaches [6, 7] or the simple permutation method by Weinstein et al. [9], since these methods involved measuring associations across the brain that could not be mapped spatially.

Although we did not find vertices with high reproducibility in the correspondence between cortical thickness and n-back, mostly due to lack of power in testing global hypothesis, CLEAN-R's rate of rejecting the null hypothesis of no intermodal correspondence between cortical thickness

and n-back in PNC subsamples was notably higher than other methods (Figure 4). In our earlier work [9], we also found some evidence for significant associations between n-back activation and cortical thickness for certain age groups when testing correspondence within functional brain networks. While other global tests require pre-defining local brain regions to pick up on highly localized associations, CLEAN-R's spatially-enhanced global test statistic amplifies signal of localized associations, while controlling for false positives.

Outside the realm of intermodal coupling analysis and testing, this paper contributes to ongoing discussions in the neuroimaging research community on clusterwise inference. Expanding on ideas from Park and Fiecas [10], in our current work we argue that adjusting for spatial autocorrelation before aggregating information within clusters can mitigate the problems brought to light by Eklund et al. [21], Noble et al. [22], and others. For instance, the difference in statistical power between clusterwise inference with and without spatial enhancement (Figures 4 and 5) illustrates what we gain in terms of statistical power by adjusting for modality-specific spatial autocorrelation before conducting clusterwise inference. Compared to existing approaches that mostly focus on the spatial autocorrelation of the test statistics map, which could make the parametrization of spatial covariance structure intractable, our proposed approach offers a straightforward and interpretable way to account for parametric spatial Gaussian process via variogram analysis.

One limitation of the spatial modeling involved in CLEAN-R is its assumption of spatial stationarity, which has been subject to criticism in previous work on both cluster-level inference [21] and intermodal correspondence testing [9]. Despite the support for the brain-wise stationary process via variogram analylsis, it might be limited to capture heterogeneous covariance structure for specific areas of the brain. While the stationary assumption warrants further consideration in the context of CLEAN-R, it is conceivable that even with imperfect assumptions about the spatial structure of the data, using the spatially adjusted data (from Equation (3)) allows aggregation of information within clusters to be more tenable than using non-adjusted spatial data. We also look forward to investigating this in future analyses. In addition, it would be interesting to extend the current methodology to localize time-varying correspondence between two imaging modalities—for example, using the Adolescent Brain Cognitive Development (ABCD) study that contains rich collection of longitudinal data from multiple modalities [23, 24]. We leave this for future work.

Finally, a tremendous advantage of CLEAN-R is its computational efficiency. Spatial modeling of each modality is required only once per analysis, and a test involving 1,000 permutations can be completed in a matter of a few minutes for several hundred subjects. In contrast, methods involving spatial null models (e.g. Alexander-Bloch et al. [6], Burt et al. [7]) can be much more computationally intensive by generating an empirical null distribution that involves altering high-dimensional images repeatedly. An R package and detailed documentation are publicly available on GitHub, making CLEAN-R an efficient, powerful, and user-friendly approach to test and localize intermodal correspondence.

Acknowledgments

This research was supported by the National Science Foundation Graduate Research Fellowship Program, the National Institute of Health (R01MH123563, 2T32MH019112, R01MH107235, R01MH119219, R01NS111115, R01MH113550, R01EB022573, R01MH120482, RF1MH116920, K08MH120564, R01MH112847, R01NS060910), the National Institute of Mental Health (1ZIAMH002949), and the Natural Sciences and Engineering Research Council of Canada (RGPIN-2022-04831)

Declaration of interest

Russell T. Shinohara receives consulting income from Octave Bioscience and compensation for reviewership duties from the American Medical Association.

Data and code availability statement

Raw neuroimaging data from the PNC are publicly available at the dbGaP (phs000607.v3.p2). Codes for conducting statistical analyses using CLEAN-R are currently available as an `R` package at `https://github.com/junjypark/CLEAN`.

CRediT authorship contribution statement

Sarah M. Weinstein: Conceptualization, Methodology, Software, Formal analysis, Validation, Visualization, Writing - original draft. **Simon N. Vandekar**: Supervision, Writing - review & editing. **Erica B. Baller**: Supervision, Writing - review & editing. **Danni Tu**: Methodology, Writing - review & editing. **Azeez Adebimpe**: Software, Formal analysis, Supervision, Writing - review & editing. **Tinashe M. Tapera**: Software, Formal analysis, Supervision, Writing - review & editing. **Ruben C. Gur**: Supervision, Writing - review & editing. **Raquel E. Gur**: Supervision, Writing - review & editing. **John A. Detre**: Supervision, Writing - review & editing. **Armin Raznahan**: Supervision, Writing - review & editing. **Aaron F. Alexander-Bloch**: Supervision, Writing - review & editing. **Theodore D. Satterthwaite**: Supervision, Writing - review & editing, Funding acquisition. **Russell T. Shinohara**: Methodology, Supervision, Writing - review & editing, Funding acquisition. **Jun Young Park**: Conceptualization, Methodology, Software, Supervision, Writing - review & editing, Funding acquisition.

References

[1] R. Casanova, R. Srikanth, A. Baer, P. J. Laurienti, J. H. Burdette, S. Hayasaka, L. Flowers, F. Wood, J. A. Maldjian, Biological parametric mapping: a statistical toolbox for multimodality brain image analysis, Neuroimage 34 (2007) 137–143.

[2] X. Yang, L. Beason-Held, S. M. Resnick, B. A. Landman, Biological parametric mapping with robust and non-parametric statistics, Neuroimage 57 (2011) 423–430.

[3] S. N. Vandekar, R. T. Shinohara, A. Raznahan, R. D. Hopson, D. R. Roalf, K. Ruparel, R. C. Gur, R. E. Gur, T. D. Satterthwaite, Subject-level measurement of local cortical coupling, NeuroImage 133 (2016) 88–97.

[4] A. M. Valcarcel, K. A. Linn, S. N. Vandekar, T. D. Satterthwaite, J. Muschelli, P. A. Calabresi, D. L. Pham, M. L. Martin, R. T. Shinohara, MIMoSA: an automated method for intermodal segmentation analysis of multiple sclerosis brain lesions, Journal of Neuroimaging 28 (2018) 389–398.

[5] F. Hu, S. M. Weinstein, E. B. Baller, A. M. Valcarcel, A. Adebimpe, A. Raznahan, D. R. Roalf, T. E. Robert-Fitzgerald, V. Gonzenbach, R. C. Gur, et al., Voxel-wise intermodal coupling analysis of two or more modalities using local covariance decomposition, bioRxiv (2022).

[6] A. F. Alexander-Bloch, H. Shou, S. Liu, T. D. Satterthwaite, D. C. Glahn, R. T. Shinohara, S. N. Vandekar, A. Raznahan, On testing for spatial correspondence between maps of human brain structure and function, Neuroimage 178 (2018) 540–551.

[7] J. B. Burt, M. Helmer, M. Shinn, A. Anticevic, J. D. Murray, Generative modeling of brain maps with spatial autocorrelation, NeuroImage 220 (2020) 117038.

[8] R. D. Markello, B. Misic, Comparing spatial null models for brain maps, NeuroImage 236 (2021) 118052.

[9] S. M. Weinstein, S. N. Vandekar, A. Adebimpe, T. M. Tapera, T. Robert-Fitzgerald, R. C. Gur, R. E. Gur, A. Raznahan, T. D. Satterthwaite, A. F. Alexander-Bloch, et al., A simple permutation-based test of intermodal correspondence, Human brain mapping 42 (2021) 5175–5187.

[10] J. Y. Park, M. Fiecas, CLEAN: Leveraging spatial autocorrelation in neuroimaging data in clusterwise inference, Neuroimage (2022).

[11] C. E. McCulloch, S. R. Searle, Generalized, linear, and mixed models, John Wiley & Sons, 2004.

[12] A. Datta, S. Banerjee, A. O. Finley, A. E. Gelfand, Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets, Journal of the American Statistical Association 111 (2016) 800–812.

[13] A. O. Finley, A. Datta, B. D. Cook, D. C. Morton, H. E. Andersen, S. Banerjee, Efficient algorithms for Bayesian nearest neighbor Gaussian processes, Journal of Computational and Graphical Statistics 28 (2019) 401–414.

[14] T. D. Satterthwaite, M. A. Elliott, K. Ruparel, J. Loughead, K. Prabhakaran, M. E. Calkins, R. Hopson, C. Jackson, J. Keefe, M. Riley, et al., Neuroimaging of the Philadelphia neurodevelopmental cohort, Neuroimage 86 (2014) 544–553.

[15] B. Fischl, Freesurfer, Neuroimage 62 (2012) 774–781.

[16] A. M. Dale, B. Fischl, M. I. Sereno, Cortical surface-based analysis: I. Segmentation and surface reconstruction, Neuroimage 9 (1999) 179–194.

[17] T. Schaefer, C. Ecker, fsbrain: an R package for the visualization of structural neuroimaging data, 2020. URL: `https://www.biorxiv.org/content/10.1101/2020.09.18.302935v1`. doi:`10.1101/2020.09.18.302935`.

[18] E. B. Baller, A. M. Valcarcel, A. Adebimpe, A. Alexander-Bloch, Z. Cui, R. C. Gur, R. E. Gur, B. Larsen, K. A. Linn, C. M. O'Donnell, et al., Developmental coupling of cerebral blood flow and fmri fluctuations in youth (2021).

[19] A. Schaefer, R. Kong, E. M. Gordon, T. O. Laumann, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff, B. T. Yeo, Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI, Cerebral cortex 28 (2018) 3095–3114.

[20] B. T. Yeo, F. M. Krienen, J. Sepulcre, M. R. Sabuncu, D. Lashkari, M. Hollinshead, J. L. Roffman, J. W. Smoller, L. Zöllei, J. R. Polimeni, et al., The organization of the human cerebral cortex estimated by intrinsic functional connectivity, Journal of neurophysiology (2011).

[21] A. Eklund, T. E. Nichols, H. Knutsson, Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates, Proceedings of the national academy of sciences 113 (2016) 7900–7905.

[22] S. Noble, D. Scheinost, R. T. Constable, Cluster failure or power failure? evaluating sensitivity in cluster-level inference, Neuroimage 209 (2020) 116468.

[23] N. D. Volkow, G. F. Koob, R. T. Croyle, D. W. Bianchi, J. A. Gordon, W. J. Koroshetz, E. J. Pérez-Stable, W. T. Riley, M. H. Bloch, K. Conway, et al., The conception of the ABCD study: From substance use to a broad NIH collaboration, Developmental cognitive neuroscience 32 (2018) 4–7.

[24] J. Y. Park, M. Fiecas, A. D. N. Initiative, et al., Permutation-based inference for spatially localized signals in longitudinal MRI data, NeuroImage 239 (2021) 118312.