A simple permutation-based test of intermodal correspondence

Sarah M. Weinstein*[1], Simon N. Vandekar[2], Azeez Adebimpe[3], Tinashe M. Tapera[3], Timothy Robert-Fitzgerald[1], Ruben C. Gur[4], Raquel E. Gur[4], Armin Raznahan[5], Theodore D. Satterthwaite[1,3,6], Aaron F. Alexander-Bloch†[7,8], and Russell T. Shinohara †[1,6]

[1]Penn Statistics in Imaging and Visualization Center, Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Perelman School of Medicine
[2]Department of Biostatistics, Vanderbilt University
[3]Lifespan Informatics and Neuroimaging Center, Department of Psychiatry, University of Pennsylvania, Perelman School of Medicine
[4]Brain and Behavior Laboratory and Penn-CHOP Lifespan Brain Institute, Department of Psychiatry, University of Pennsylvania, Perelman School of Medicine
[5]Developmental Neurogenomics Unit, National Institute of Mental Health
[6]Center for Biomedical Image Computing and Analytics, University of Pennsylvania, Perelman School of Medicine
[7]Department of Child and Adolescent Psychiatry and Behavioral Science, Children's Hospital of Philadelphia
[8]Neurodevelopment and Psychosis Section, Department of Psychiatry, University of Pennsylvania, Perelman School of Medicine

Data availability statement: Neuroimaging data were acquired as part of the Philadelphia Neurodevelopmental Cohort (PNC), a research initiative at the University of Pennsylvania Brain and Behavior Laboratory and Children's Hospital of Philadelphia. PNC data are publicly available in raw format at https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000607.v3.p2. Additionally, R code for reproducing our simulations can be found at https://github.com/smweinst/spice_test.

Funding statement: This work was supported by the following NIH grants: R01MH123563 (S.N.V.), R01MH107235 (R.C.G.), R01MH119219 (R.C.G. and R.E.G.), R01MH113550 (T.D.S.), R01EB022573 (T.D.S.), R01MH120482 (T.D.S.), RF1MH116920 (T.D.S.), K08MH120564 (A.F.A.B.), R01MH112847 (R.T.S. and T.D.S.), and R01NS060910 (R.T.S.). This study was supported by the intramural research program of the National Institute of Mental Health (NIMH) (project funding: 1ZIAMH002949) (A.R.). This work was also supported by the National Science Foundation Graduate Research Fellowship Program (S.M.W.).

Conflict of interest disclosure: The authors declare no competing interests.

Ethics approval statement: The institutional review boards of the Hospital of the University of Pennsylvania and Children's Hospital of Philadelphia approved protocols for the PNC.

Patient consent statement: Participants and/or their parents provided informed consent to be included in the PNC.

Permission to reproduce material from other sources: N/A

Clinical trial registration: N/A

*Corresponding author: Sarah M. Weinstein (smweinst@pennmedicine.upenn.edu; Penn Statistics in Imaging and Visualization Center, University of Pennsylvania, Perelman School of Medicine, 423 Guardian Drive, Philadelphia, PA 19104)

†Contributed equally

Abstract

Many key findings in neuroimaging studies involve similarities between brain maps, but statistical methods used to measure these findings have varied. Current state-of-the-art methods involve comparing observed group-level brain maps (after averaging intensities at each image location across multiple subjects) against spatial null models of these group-level maps. However, these methods typically make strong and potentially unrealistic statistical assumptions, such as covariance stationarity. To address these issues, in this paper we propose using subject-level data and a classical permutation testing framework to test and assess similarities between brain maps. Our method is comparable to traditional permutation tests in that it involves randomly permuting subjects to generate a null distribution of intermodal correspondence statistics, which we compare to an observed statistic to estimate a p-value. We apply and compare our method in simulated and real neuroimaging data from the Philadelphia Neurodevelopmental Cohort. We show that our method performs well for detecting relationships between modalities known to be strongly related (cortical thickness and sulcal depth), and it is conservative when an association would not be expected (cortical thickness and activation on the n-back working memory task). Notably, our method is the most flexible and reliable for localizing intermodal relationships within subregions of the brain and allows for generalizable statistical inference.

1 Introduction

Neuroimaging studies often seek to understand similarities across modalities. However, methods underlying these studies' findings continue to vary. Naive approaches to the so-called "correspondence problem" have included simply visualizing two maps next to each other and observing whether the two maps look similar. Another approach involves the use of parametric p-values; for example, fitting a linear regression or estimating the correlation across locations of the cortex and using a parametric null distribution (e.g., Student's t-distribution), ignoring inherent spatial autocorrelation. A third approach has been to use a partial correlation coefficient controlling for anatomical distance (Honey et al. 2009; Horn et al. 2014).

Methods that ignore spatial autocorrelation have been shown to inflate type I error rates (Markello and Misic 2021); thus, recent methodological advancements that do account for spatial correlation have been widely adopted by the neuroimaging community. A spatial permutation framework was first introduced by Alexander-Bloch et al. (2013), using spatially-constrained randomization to generate null models for testing intermodal developmental synchrony (Alexander-Bloch et al. 2013). Vandekar et al. (2015) then used spherical rotations to obtain null models of spatial alignment in testing correspondence between various cortical measurements in the human brain. The spin test was formally introduced by Alexander-Bloch et al. (2018), who used random spherical rotations of group-level or population-averaged surfaces to generate null models of spatial alignment. A surface-based permutation null model was also developed independently in the context of validating functional magnetic resonance imaging parcellations (Gordon et al. 2016; Gordon et al. 2017a; Gordon et al. 2017b).

Since the advent of the spin test, numerous neuroimaging experiments have adopted it to corroborate claims about intermodal similarities (Lefèvre et al. 2018; Paquola et al. 2019; Schmitt et al. 2019; Shafiei et al. 2020; Cui et al. 2020; Stoecklein et al. 2020) and have extended its implementation to regionally parcellated maps (Váša et al. 2018; Vázquez-Rodríguez et al. 2019; Baum et al. 2020; Cornblath et al. 2020). The spin test was certainly an improvement compared with prior approaches; however, deriving frequentist properties of inference across datasets using this method depends on strong and often unrealistic statistical assumptions, such as stationarity of the covariance structure. Other critiques of the spin test include its limited use in the context of cortical surface maps without a straightforward extension to volumetric data, its lack of flexibility for incorporating subject-level data by accommodating only two (group-level) maps, and that it cannot be used to test correspondence in small regions of the brain. Another limitation of the formal support for the spin test presented in Alexander-Bloch et al. (2018) is its ad hoc treatment of the medial wall, which is an artifact of generating a topologically spherical surface that does not represent neuroanatomical features (Dale et al. 1999).

More recently, Burt et al. (2020) proposed Brain Surrogate Maps with Autocorrelated Spatial Heterogeneity (BrainSMASH), using generative modeling to obtain null models seeking to preserve the spatial autocorrelation structure of the observed data. Burt et al.'s approach greatly improves on several aspects of the spin test in its flexibility to incorporate volumetric data, its inclusion of the medial wall, and its ability to test correspondence in small regions of the brain. Still, BrainSMASH has some limitations. When applied to subregions of the brain, one can only account for spatial autocorrelation among locations within each subregion. For instance, when testing intermodal correspondence between two maps of the left hemisphere, spatial autocorrelation among locations within that hemisphere are modeled, but correlation with locations in the right hemisphere are not. Furthermore, BrainSMASH models spatial, but not functional, relationships between locations throughout the brain.

The spin test, BrainSMASH, and other related methods recently described by Markello and Misic (2021) involve testing null hypotheses of intermodal correspondence conditional on observed group-level data (Burt et al. 2020; Alexander-Bloch et al. 2018; Vázquez-Rodríguez et al. 2019; Baum et al. 2020; Cornblath et al. 2020; Váša et al. 2018; Burt et al. 2018; Wael et al. 2020). As illustrated in Figure 1, averaging images across subjects precludes evaluation of inter-individual heterogeneities.

Importantly, these previous methods may not be well-suited to account for dataset-to-dataset variability without strong assumptions about stationarity of the covariance structure. Burt et al. (2020) use a variogram-matching model that assumes stationarity and normality—assumptions that have been subject to criticism in the context of neuroimaging studies (Ye et al. 2015; Eklund et al. 2016). While these methods are currently the most sophisticated in the field, their underlying assumptions may give rise to inferential missteps, with high false positive rates and substantial variability in results (Markello and Misic 2021). Thus, careful consideration of the plausibility of these assumptions is warranted, and the appeal of a method without strong assumptions motivates our current work.

In this paper, we propose the simple permutation-based intermodal correspondence (SPICE) test, a novel procedure that is intuitive, easy to implement, and does not depend on strong statistical assumptions. This test considers a setting with subjects $i = 1, \ldots, n$ for whom we observe two imaging modalities each: $X_i(\mathbf{v})$ and $Y_i(\mathbf{v})$ at locations (e.g., pixels, voxels, or vertices) indexed by $\mathbf{v} = [1, \ldots, V]$. Our null hypothesis states that the distribution of within-subject intermodal similarity between $X_i(\mathbf{v})$ and $Y_i(\mathbf{v})$ (i.e., two modalities belonging to the same subject) is no different than the distribution of between-subject intermodal correspondence—for example, between $X_i(\mathbf{v})$ and $Y_j(\mathbf{v})$ $(i \neq j)$. In prior studies, the Pearson correlation has been widely used to measure intermodal correspondence (Alexander-Bloch et al. 2018; Burt et al. 2020). While in this paper we adopt this commonly used statistic, in the Discussion we propose alternative measures that we hope to explore in future work.

The intuition behind the SPICE test is that, if there is a genuine anatomical correspondence between two modalities, this correspondence should be greater in brain maps derived from the same individual than from different individuals. To test the SPICE null hypothesis, we leverage subject-level data and do not require altering the spatial structure or alignment of the observed data, in contrast to BrainSMASH and the spin test. Since we will compare our proposed method to BrainSMASH and the spin test, it is important to highlight that their underlying null hypotheses are fundamentally different, though they may be used to answer related— but nevertheless different—questions about intermodal correspondence. Figure 1 outlines these three null hypotheses alongside the different types of data required for each test. We will reference the three null hypotheses throughout the paper as $H_{0(spice)}$, $H_{0(b\text{-}smash)}$, and $H_{0(spin)}$.

2 Methods

2.1 The SPICE test

Formally, we express the SPICE test's null hypothesis as:

$$H_{0(spice)} : \psi(X_i(\mathbf{v}), Y_i(\mathbf{v})) \overset{d}{=} \psi(X_i(\mathbf{v}), Y_j(\mathbf{v})) \qquad i \neq j, \tag{1}$$

where $\psi()$ represents a measure of intermodal similarity between X and Y across all locations indexed by $\mathbf{v}$, either for the same subject or for different subjects (when $i \neq j$). Adopting the Pearson correlation to measure intermodal correspondence (which has been widely used in previous work), the test of equality of means (1) is a test of whether within-subject intermodal correlation is higher than the between-subject intermodal correlation.

To test this null hypothesis, we first obtain the following empirical estimate of the mean of the left hand side of (1):

$$A_0 = \widehat{E}[\psi(X_i(\mathbf{v}), Y_i(\mathbf{v}))] = \frac{1}{n}\sum_{i=1}^{n} \widehat{\psi}(X_i(\mathbf{v}), Y_i(\mathbf{v})), \tag{2}$$

where, for the ith subject, the Pearson correlation across all V locations in the image is:

$$\widehat{\psi}(X_i(\mathbf{v}), Y_i(\mathbf{v})) = \frac{\left(\frac{1}{V}\sum_{v=1}^{V} X_i(v)Y_i(v)\right) - \left(\frac{1}{V}\sum_{v=1}^{V} X_i(v)\right)\left(\frac{1}{V}\sum_{v=1}^{V} Y_i(v)\right)}{\sqrt{\frac{1}{V}\sum_{v=1}^{V}\left(X_i(v) - \frac{1}{V}\sum_{v=1}^{V} X_i(v)\right)^2}\sqrt{\frac{1}{V}\sum_{v=1}^{V}\left(Y_i(v) - \frac{1}{V}\sum_{v=1}^{V} Y_i(v)\right)^2}}. \tag{3}$$

After estimating the mean within-subject intermodal correspondence (A_0), we generate a null distribution for A_0 by randomly shuffling the Y maps across subjects. Then, for the kth permuted sets of map pairs ($k = 1, ..., K$ permutations), we re-calculate (2) as A_k, providing draws from an empirical null distribution of A_0. These random permutations are illustrated in Figure 2.

A p-value is estimated by counting the number of permutations for which the magnitude of A_k exceeds that of A_0:

$$p_{spice} = \frac{\sum_{k=0}^{K} I(|A_0| \leq |A_k|)}{K+1}, \tag{4}$$

where we add 1 to the denominator to account for the identity permutation (Phipson and Smyth 2010). We reject the null hypothesis for $p_{spice} < \alpha$, the nominal type I error rate. Under the null hypothesis, p_{spice} is a random variable distributed as Uniform(0,1), and we would reject $H_{0(spice)}$ in only approximately 5% of cases. Under the alternative hypothesis, we expect A_k to rarely be greater in magnitude than A_0 and would reject $H_{0(spice)}$.

2.2 Simulations and application in the Philadelphia Neurodevelopmental Cohort

We assess the performance of our proposed method using neuroimaging data from the Philadelphia Neurodevelopmental Cohort (PNC). The PNC is a research initiative at the University of Pennsylvania consisting of 9498 subjects between the ages of 8 and 21 who received care at the Children's Hospital of Philadelphia and consented (or received parental consent) to participate. A subset of 1601 subjects were included in multimodal neuroimaging studies and were scanned in the same scanner at the Hospital of the University of Pennsylvania, per protocols described briefly below and in Appendix A1.

For the current study, we first consider correspondence between cortical thickness and sulcal depth—a pair of imaging measurements for which we expect to reject the null hypothesis due to well-established interdependence in structure throughout multiple brain regions shown in research from the last two decades (Vandekar et al. 2016; Vandekar et al. 2015; Sowell et al. 2004; Shaw et al. 2008), as well as in much earlier studies of neuroanatomy from nearly a century ago (Von Economo 1929). Second, we consider cortical thickness and n-back, a measure of working memory function (described below), to demonstrate how our method performs in a situation where intermodal correspondence would typically not be expected.

In this study, we include a subset of $n = 789$ subjects with all three image types (cortical thickness, sulcal depth, and n-back) meeting quality control criteria. This subset also excluded individuals with existing health conditions, who were taking psychoactive medications, had a history of psychiatric hospitalization, or other abnormalities that could impact brain structure or function.

2.2.1 Acquisition and processing of multimodal imaging data

Relevant protocols for the acquisition and pre-processing of both structural and functional magnetic resonance imaging (MRI) data are described in Appendix A1. In brief, all subjects underwent MRI scans in the same Siemens TIM Trio 3 tesla scanner with a 32-channel head coil and the same imaging sequences and parameters (Satterthwaite et al. 2014; Satterthwaite et al. 2016). Cortical reconstruction of T1-weighted structural images was implemented using FreeSurfer (version 5.3). We also consider the fractal n-back task sequence acquired during functional MRI (fMRI) acquisition sequences, which involved presenting each subject with a series of stimuli (geometric pictures on a screen) with instructions to press a button if the current stimulus matches the nth previous one. In our present analysis we consider maps that represent the percentage change in activation between the 2-back and 0-back sequences, which has been previously shown to isolate executive network function (Ragland et al. 2002; Satterthwaite et al. 2013). All measurements were resampled to the fsaverage5 atlas, consisting of $V = 10,242$ vertices in each hemisphere for every subject.

2.2.2 Simulations

To evaluate the performance of the SPICE test under the null and alternative hypotheses, we conduct several simulation studies. Using average cortical thickness, sulcal depth, and n-back maps from the PNC, we simulate subject-level multimodal imaging data by adding and multiplying random noise and signal to mean maps, as shown in Figure 3. While we generate three hypothetical imaging modalities per simulated subject, to test the null hypothesis of intermodal correspondence we consider only two such modalities at a time (hence Figure 3 illustrates the simulation set-up when mean cortical thickness and mean sulcal depth maps are used). The null or alternative hypothesis is true depending on σ_a^2, the variance of the simulated signal (a_i term in Figure 3), and the null hypothesis is more difficult to reject for larger values of σ_e^2, the variance of the simulated noise (elements of $E_{i1}(\mathbf{v})$ or $E_{i2}(\mathbf{v})$ in Figure 3).

For the ith subject, we simulate $X_i(\mathbf{v})$ and $Y_i(\mathbf{v})$, which have a one-to-one correspondence at all locations in the fsaverage5 atlas space, indexed by $\mathbf{v} = [1, ..., 10, 242]$. We define $X_i(\mathbf{v}) = a_i \times M_1(\mathbf{v}) + E_{i(1)}(\mathbf{v})$ and $Y_i(\mathbf{v}) = a_i \times M_2(\mathbf{v}) + E_{i(2)}(\mathbf{v})$, where $M_1(\mathbf{v})$ is the mean cortical thickness map, $M_2(\mathbf{v})$ is either the mean sulcal depth map (in which $\text{Cor}(M_1(\mathbf{v}), M_2(\mathbf{v})) = -0.15$) or the mean n-back map ($\text{Cor}(M_1(\mathbf{v}), M_2(\mathbf{v})) = -0.04$). We consider two sets of simulations, where $M_1(\mathbf{v})$ is mean cortical thickness in both cases, and $M_2(\mathbf{v})$ is mean sulcal depth in one setting and mean n-back in the other. The purpose of these two separate settings is to consider the degree to which population-level modality similarity (e.g., correlation between $M_1(\mathbf{v})$ and $M_2(\mathbf{v})$) impacts the power of our test.

Since $X_i(\mathbf{v})$ and $Y_i(\mathbf{v})$ share a_i (distributed $N(1, \sigma_a^2)$), the SPICE test's null hypothesis (1) is true when the variance of a_i, $\sigma_a^2 = 0$ because in this case, the covariation between $X_i(\mathbf{v})$ and $Y_i(\mathbf{v})$ is no stronger than it is between $X_i(\mathbf{v})$ and $Y_j(\mathbf{v})$ ($i \neq j$). The V-dimensional surfaces $E_{i(1)}(\mathbf{v})$ and $E_{i(2)}(\mathbf{v})$ consist of elements that are independent and identically distributed as $N(0, \sigma_e^2)$.

We consider sample sizes of n = 25, 50, and 100, σ_a^2 ranging from 0 to 3.0, and σ_e^2 of 0.5, 1.5, 3.0, or 6.0. While the signal-to-noise ratio varies with the magnitude of intensities at different locations in $M_1(\mathbf{v})$ and $M_2(\mathbf{v})$ (i.e., higher magnitudes of intensities at each location would multiplicatively increase the variance of the signal at those locations), the relative size of σ_a^2 and σ_e^2 translates to varying degrees of statistical power when we conduct tests of intermodal correspondence between simulated surfaces. For instance, higher variance of the signal (σ_a^2) and lower variance of the noise (σ_e^2) would produce more signal than noise, adding to the similarity between $X_i(\mathbf{v})$ and $Y_i(\mathbf{v})$ (from the same subject). In contrast, lower σ_a^2 relative to σ_e^2 would mute the dissimilarity between $X_i(\mathbf{v})$ and $Y_j(\mathbf{v})$ ($i \neq j$) (more noise than signal) and induce lower statistical power.

When $\sigma_a^2 = 0$, we repeat 5000 simulations of each combination of n, σ_e^2, and $M_1(\mathbf{v})$ and $M_2(\mathbf{v})$ to estimate type I error as the proportion of simulations for which $p_{spice} < 0.05$. In addition, from 5000 simulations of each combination of parameters where $\sigma_a^2 > 0$, we estimate power by the proportion of simulations for which $p_{spice} < 0.05$. R code for implementing the SPICE test and reproducing our simulations may be found at https://github.com/smweinst/spice_test.

2.2.3 Comparing the SPICE test with previous methods

In assessing intermodal correspondence in the PNC, we compare the performance of our method to that of the spin test and BrainSMASH. Since both the spin test and BrainSMASH require group-level surfaces and their underlying null hypotheses differ, our goal here is to assess whether the general conclusions of their respective null hypotheses (Figure 1) agree with one another. Given the developmental focus of the dataset and prior research suggesting that intermodal coupling may change throughout development (Vandekar et al. 2016), we apply each method in imaging data within seven age subgroups (group-averaged for the spin test and BrainSMASH, and subject-level for the SPICE test). We also apply each method to test for intermodal correspondence in the full group of n=789 subjects. We account for multiple comparisons by using a Bonferroni-adjusted p-value threshold for statistical significance ($0.05/16 = 0.003$) to correct for the seven age groups and one group including all subjects, times two hemispheres tested within each group.

The spin test is implemented using publicly available MATLAB code (https://github.com/spin-test). We generate 1000 null models of spatial alignment through random rotations of cortical thickness maps within each hemisphere, after removing vertices identified to be part of the medial wall. The same "spun" maps are used to test $H_{0(spice)}$ of cortical thickness versus sulcal depth as well as cortical thickness versus n-back.

BrainSMASH is implemented in Python (https://github.com/murraylab/brainsmash) on parcellated left and right hemispheres. Parcellated surfaces, according to the 1000 cortical network parcellation described by Schaefer et al. (2018), are used instead of vertex-level data due to computational challenges of computing the required pairwise distance matrix for vertex-level data. As noted by Markello and Misic (2021), using such parcellations does not appear to compromise null model performance. For each hemisphere, a 501 × 501 geodesic distance matrix (which includes one parcel for the medial wall that does not need to be removed for BrainSMASH, unlike the spin test) is constructed using the SurfDist package in Python (Margulies et al. 2016). For BrainSMASH, we generate 1000 surrogate maps whose variograms are intended to be as close as possible to the empirical variogram of the observed cortical thickness maps. Variogram estimation includes pairwise distances in the bottom 25th percentile of possible pairwise distances, in accordance with Viladomat et al. (2014) and Burt et al. (2020). Since both tests of interest involve cortical thickness, for computational efficiency, we use the same 1000 surrogate cortical thickness maps tests against both sulcal depth and n-back in each age group.

2.2.4 Testing correspondence within functional networks

Next, we test $H_{0(spice)}$ and $H_{0(b\text{-}smash)}$ and consider the same pairs of modalities as before in the left hemisphere, within canonical large-scale functional networks defined by Yeo et al. (2011) in their reported 7-network solution. Each network consists of a subset of the 1000 cortical parcellations described by Schaefer et al. (2018). For BrainSMASH, we increase the truncation threshold for pairwise distances included in variogram calculation to the 70th percentile, thus allowing variation between locations located farther apart to be incorporated in variogram calculation, following Burt et al. (2020)'s methodology for assessment of spatial autocorrelation in subregions of the brain. To adjust for multiple comparisons across the seven age groups, we use a Bonferroni-adjusted p-value threshold for statistical significance ($0.05/7 = 0.007$). Because the main interest is in studying developmental changes throughout development in each network, we adjust for comparisons across age groups, but not across the Yeo networks nor the two tests.

3 Results

3.1 Type I error and power of the SPICE method

Simulation results are shown in Figure 4. Under both the null and alternative hypotheses, our proposed method performs well and as expected. The type I error (when $\sigma_a^2 = 0$) is close to 5% for all sample sizes, for all values of the vertex-level variance parameter, σ_e^2, and for both settings involving $M_2(\mathbf{v})$ as either mean sulcal depth or mean n-back. Under the alternative hypothesis (when $\sigma_a^2 > 0$), a higher signal-to-noise ratio (lower σ_e^2, higher σ_a^2) correspond to higher power, and power decreases incrementally as the signal-to-noise ratio decreases (higher σ_e^2, lower σ_a^2). As expected, larger sample size and higher magnitude of population-level image similarity $(\mathrm{Corr}(M_1(\mathbf{v}), M_2(\mathbf{v})))$ also improve statistical power.

3.2 Results and comparison of methods in the Philadelphia Neurodevelopmental Cohort

We first consider tests of intermodal correspondence for seven age groups in the left and right hemispheres. Unadjusted p-values are provided in Table 1. The SPICE test, BrainSMASH, and spin test all produce results that are consistent with what one would expect biologically: all three methods produce small p-value when testing sulcal depth versus cortical thickness (Table 1(a)), although BrainSMASH's results in the right hemisphere for the youngest three age groups do not reach statistical significance based on our Bonferroni-adjusted p-value threshold of 0.003. None of the methods indicate significant correspondence between the n-back and cortical thickness maps (Table 1(b)). For BrainSMASH, we also verify that the surrogate maps' variogram surrogate cortical thickness maps appear valid (Figure A1).

It is clear that intermodal coupling may occur with both anatomical and developmental specificity (Vandekar et al. 2016). For example, a positive result could be driven disproportionately by specific neuroanatomical sub-systems or specific age groups, suggesting the need for post-hoc regional or network analysis tests at higher anatomical or developmental resolutions. An important attribute of the SPICE test is its applicability for post-hoc testing in this sense. We therefore consider age-stratified tests of intermodal correspondence within functional networks from Yeo et al. (2011) with the SPICE test and BrainSMASH (plotted in blue and red, respectively, in Figure 5). As expected, the p-values for SPICE and BrainSMASH tests of correspondence for n-back versus cortical thickness (solid lines) appear uniformly distributed between 0 and 1, which we anticipate under their respective null hypotheses. However, we observe some discrepancies in their results for testing sulcal depth versus cortical thickness (dotted lines) within functional networks, despite the agreement of the two methods in tests conducted within full cortical hemispheres discussed before.

Specifically, the SPICE test provides evidence against its null hypothesis ($H_{0(spice)}$) of coupling between cortical thickness and sulcal depth in the somatomotor, dorsal attention, ventral attention, limbic, frontoparietal control, and default networks, with $p_{spice} < 0.007$, the Bonferroni-adjusted threshold used for this analysis. BrainSMASH similarly provides evidence against its null hypothesis ($H_{0(b\text{-}smash)}$) for coupling between sulcal depth and cortical thickness in the somatomotor (except ages 8-9 and 10-11), dorsal attention, ventral attention, and default networks, but not in the limbic and frontoparietal control networks.

The poor fit of the network-specific surrogate variograms, plotted in Figure A3, may give some insight for understanding these discrepancies. Specifically, we observe differences between the empirical and surrogate map variograms in all age groups for the dorsal attention, ventral attention, limbic, frontoparietal control, and default networks, suggesting that the surrogate cortical thickness maps in these networks failed to preserve the spatial autocorrelation structure of the target (i.e., original cortical thickness) map, undermining the results from tests in those networks, and suggesting that this method may not be reliable in the context of smaller brain regions.

It is also notable that neither the SPICE test nor BrainSMASH reject their respective null hypotheses for testing correspondence between sulcal depth and cortical thickness within the visual network, but that the non-significant p-values for the visual network are not uniformly distributed. This result is not surprising, given previous work from Vandekar et al. (2016), who found that in primary visual regions, we observe a complex relationship between these two modalities. These authors noted that the nonlinear nature of the association between sulcal depth and cortical thickness within the primary visual cortex violates the linearity assumption of their methodology for assessing intermodal coupling. It is plausible that this nonlinearity also explains SPICE's and BrainSMASH's failure to reject $H_{0(spice)}$ and $H_{0(b\text{-}smash)}$, respectively, within the visual network, as both methods currently use a linear measurement of similarity (Pearson correlation).

Lastly, although we did not hypothesize an association between the n-back task and cortical thickness, given the age-related regional heterogeneity in both cortical thickness growth curves (Tamnes et al. 2017) and also patterns of activation in the n-back task (Andre et al. 2016), differences in coupling across age groups and neuroanatomical systems are not entirely unexpected. Speculatively, the maturation of multi-modal association areas may be related to their differential activation in working memory tasks, which will be an interesting area for future work.

4 Discussion

In this paper, we introduce an intuitive and easy-to-implement method that leverages subject-level data to test and make inference about intermodal correspondence. Our method is complementary to those proposed by Alexander-Bloch et al. (2018), Burt et al. (2020), and other related methods described by Markello and Misic (2021), as the null hypotheses of these methods fundamentally differ (Figure 1) but may all provide meaningful insights about intermodal correspondence. While the spin test and BrainSMASH give a picture of population-level spatial alignment between two maps, our method considers subject-level intermodal correspondence, depending only on the plausible assumption that subjects are independent of one another. By approaching the correspondence problem from a new angle—emphasizing inference on subject-level associations as opposed to more general patterns of spatial alignment in group-averaged data—the SPICE test addresses five limitations of the earlier approaches.

Alexander-Bloch et al.'s spin test (i) can only be applied to surfaces with spherical representations (i.e., not volumetric data), (ii) does not specifically address the presence of the medial wall, which must be accounted for in practical implementations of the test, for example by disregarding cortical areas that overlap with the medial wall in spun maps when estimating the null distribution, and (iii) cannot be applied in small regions of the brain. While Burt et al.'s BrainSMASH can technically be applied in subregions of the brain, our findings suggest its generative null models may not be suitable when the small regions are irregularly shaped or disconnected, such as the widely used functional networks described by Yeo et al. (2011). One possible extension of BrainSMASH could incorporate nonparametric variogram estimation, as proposed by Ye et al. (2015), to account for both spatial proximity and functional connectivity in variogram estimation. Both the spin test and BrainSMASH are further limited in that they (iv) are only able to incorporate group-level data, precluding assessments of inter-subject heterogeneity (visualized in Figure 1). Finally, by conditioning on observed group-level maps, both of the previous methods would require (v) assuming covariance stationarity in order to take into account dataset-to-dataset variability or be used for generalizable statistical inference—that is, inference that formally considers the variability that results when sampling new data from the population. Without a treatment of this source of variability that does not rely on unrealistic assumptions, neither the spin test nor BrainSMASH may fully address concerns about the external validity of findings. The ability to directly address such concerns is a significant theoretical advantage offered by the SPICE test.

The SPICE test accounts for spatial autocorrelation to the extent that it is a feature of the underlying data-generating process. Spatially relevant information has already been encoded in the observed data, which renders moot any need for assumptions about stationarity. By conditioning on observed group-level data and altering its spatial structure in generating a reference null distribution, both the spin test and BrainSMASH necessitate explicitly modeling spatial autocorrelation or assuming covariance stationarity to make generalizable inference. In this sense, the SPICE test has the benefit of being readily usable for post-hoc analyses; if investigators decide to test correspondence between brain maps within subregions not considered in their primary analyses, the SPICE test would not require revisiting questionable assumptions about the spatial structure of the data when transitioning between testing more broadly or narrowly defined brain regions.

In addition to the SPICE test's advantage of having minimal statistical assumptions, it is also clear that a method which allows for inter-individual variability would be advantageous in the neuroimaging field. Inter-individual variability in coupling between imaging-derived phenotypes has been shown to be sensitive to age, sex, and disease-related changes (Vandekar et al. 2016). However, we know of no systematic exploration of coupling across possible phenotype pairs, possibly due to the lack of a clear statistical framework for such a study. An example use case for the SPICE test is to prioritize phenotypic pairs for further analysis in large multimodal imaging studies, where coupling may be related to outcomes of interest such as psychopathology and individual differences (Karcher and Barch 2021). Given that inter-individual variability from both genetic and environmental sources is known to influence imaging phenotypes across modalities (Tooley et al. 2021; Fjell et al. 2015), higher within-subject coupling is consistent with "true" biological basis of intermodal coupling that also manifests marked inter-individual variability.

The SPICE test does not replace BrainSMASH, the spin test, or other methods that specifically wish to compare group-level surfaces, as SPICE is only applicable in settings where subject-level data are available. For example, the spin test and BrainSMASH may be used to test similarities between different atlases (e.g., Yeo versus Desikan), and the SPICE test would not be applicable in such studies.

In future work, we hope to consider more comprehensive testing strategies, including the use of different correspondence metrics, since the choice of statistic should be appropriate for the modalities under consideration. For example, while the Pearson correlation has known limitations when used in sparse data such as connectivity matrices, a geodesic distance measure between two connectivity matrices may be more suitable (Venkatesh et al. 2020). In addition, since the null hypothesis for the SPICE test is framed in terms of equality in distributions of a correspondence measure (1), one may consider additional test statistics (for example, the Kolmogrov-Smirnov statistic) so that the test can explicitly consider the full distribution function, rather than focusing on the mean. Implementing the SPICE test using a Pearson correlation (or other average measurement) tests only one aspect of the distribution—testing a sufficient but not necessary condition for equivalence in distributions. A rank-based measure of similarity (e.g., the Spearman correlation coefficient) would be another option more sensitive to nonlinear associations.

Finally, since our method involves permuting subject-level maps in generating a null distribution, in future work we plan to consider implications of subject exchangeability on inference using our method. Various approaches to preserve exchangeability (e.g., defining blocks within which subjects may be considered "exchangeable") may be adapted from methods described by Winkler et al. (2014).

5 Conclusions

The SPICE, BrainSMASH, and spin methods may all support similar findings despite their different null hypotheses. However, the SPICE test is the most flexible when it comes to analyzing subregions of the brain, particularly when assumptions regarding the structure of spatial autocorrelation can pose obstacles to generating surrogate maps intended to preserve those complex structures and it is the only method that is able to consider subject-level data. Depending on the question of interest and the available data, using a combination of these three methods may be beneficial in future neuroimaging experiments to obtain a more complete picture of intermodal correspondence.

Author contributions

S.M.W.: conceptualization, data analysis, software, and writing. **S.N.V.:** conceptualization, writing/review of manuscript. **A.A.:** data processing and review/editing of manuscript. **T.M.T.:** data processing and review/editing of manuscript. **T.R.F.:** data analysis, software, and review/editing of manuscript. **R.C.G.**: data collection, data processing, and review/editing of manuscript. **R.E.G.**: data collection, data processing, and review/editing of manuscript. **A.R.**: conceptualization and review/editing of manuscript. **T.D.S.**: conceptualization, data collection, data processing, and review/editing of manuscript. **A.F.A.B.:** conceptualization, data analysis, writing/review of manuscript, and supervision. **R.T.S.:** conceptualization, data analysis, writing/review of manuscript, and supervision.

References

Alexander-Bloch, Aaron, Armin Raznahan, Ed Bullmore, and Jay Giedd (2013). “The convergence of maturational change and structural covariance in human cortical networks”. In: *Journal of Neuroscience* 33.7, pp. 2889–2899. ISSN: 0270-6474.

Alexander-Bloch, Aaron F, Haochang Shou, Siyuan Liu, Theodore D Satterthwaite, David C Glahn, Russell T Shinohara, Simon N Vandekar, and Armin Raznahan (2018). “On testing for spatial correspondence between maps of human brain structure and function”. In: *Neuroimage* 178, pp. 540–551.

Andre, Julia, Marco Picchioni, Ruibin Zhang, and Timothea Toulopoulou (2016). “Working memory circuit as a function of increasing age in healthy adolescence: A systematic review and meta-analyses”. In: *NeuroImage: Clinical* 12, pp. 940–948.

Baum, Graham L., Zaixu Cui, David R. Roalf, Rastko Ciric, Richard F. Betzel, Bart Larsen, Matthew Cieslak, Philip A. Cook, Cedric H. Xia, Tyler M. Moore, Kosha Ruparel, Desmond J. Oathes, Aaron F. Alexander-Bloch, Russell T. Shinohara, Armin Raznahan, Raquel E. Gur, Ruben C. Gur, Danielle S. Bassett, and Theodore D. Satterthwaite (Jan. 2020). "Development of structure–function coupling in human brain networks during youth". In: *Proceedings of the National Academy of Sciences* 117.1, pp. 771–778. ISSN: 0027-8424. DOI: 10.1073/PNAS.1912034117. URL: https://www.pnas.org/content/117/1/771.

Burt, Joshua B, Murat Demirtaş, William J Eckner, Natasha M Navejar, Jie Lisa Ji, William J Martin, Alberto Bernacchia, Alan Anticevic, and John D Murray (2018). "Hierarchy of transcriptomic specialization across human cortex captured by structural neuroimaging topography". In: *Nature neuroscience* 21.9, pp. 1251–1259.

Burt, Joshua B, Markus Helmer, Maxwell Shinn, Alan Anticevic, and John D Murray (2020). "Generative modeling of brain maps with spatial autocorrelation". In: *NeuroImage* 220, p. 117038. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2020.117038. URL: http://www.sciencedirect.com/science/article/pii/S1053811920305243.

Ciric, Rastko, Adon F G Rosen, Guray Erus, Matthew Cieslak, Azeez Adebimpe, Philip A Cook, Danielle S Bassett, Christos Davatzikos, Daniel H Wolf, and Theodore D Satterthwaite (2018). "Mitigating head motion artifact in functional connectivity MRI". In: *Nature protocols* 13.12, pp. 2801–2826.

Cornblath, Eli J, Arian Ashourvan, Jason Z Kim, Richard F Betzel, Rastko Ciric, Azeez Adebimpe, Graham L Baum, Xiaosong He, Kosha Ruparel, Tyler M Moore, et al. (2020). "Temporal sequences of brain activity at rest are constrained by white matter structure and modulated by cognitive demands". In: *Communications biology* 3.1, pp. 1–12.

Cui, Zaixu, Hongming Li, Cedric H Xia, Bart Larsen, Azeez Adebimpe, Graham L Baum, Matt Cieslak, Raquel E Gur, Ruben C Gur, and Tyler M Moore (2020). "Individual variation in functional topography of association networks in youth". In: *Neuron.* ISSN: 0896-6273.

Dale, Anders M, Bruce Fischl, and Martin I Sereno (1999). "Cortical surface-based analysis: I. Segmentation and surface reconstruction". In: *Neuroimage* 9.2, pp. 179–194.

Eklund, Anders, Thomas E Nichols, and Hans Knutsson (2016). "Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates". In: *Proceedings of the national academy of sciences* 113.28, pp. 7900–7905.

Fischl, Bruce, Martin I Sereno, and Anders M Dale (1999). "Cortical surface-based analysis: II: inflation, flattening, and a surface-based coordinate system". In: *Neuroimage* 9.2, pp. 195–207.

Fjell, Anders M, Håkon Grydeland, Stine K Krogsrud, Inge Amlien, Darius A Rohani, Lia Ferschmann, Andreas B Storsve, Christian K Tamnes, Roser Sala-Llonch, Paulina Due-Tønnessen, et al. (2015). "Development and aging of cortical thickness correspond to genetic organization patterns". In: *Proceedings of the National Academy of Sciences* 112.50, pp. 15462–15467.

Gordon, Evan M, Timothy O Laumann, Babatunde Adeyemo, Jeremy F Huckins, William M Kelley, and Steven E Petersen (2016). "Generation and evaluation of a cortical area parcellation from resting-state correlations". In: *Cerebral cortex* 26.1, pp. 288–303.

Gordon, Evan M, Timothy O Laumann, Babatunde Adeyemo, and Steven E Petersen (2017a). "Individual variability of the system-level organization of the human brain". In: *Cerebral cortex* 27.1, pp. 386–399.

Gordon, Evan M, Timothy O Laumann, Adrian W Gilmore, Dillan J Newbold, Deanna J Greene, Jeffrey J Berg, Mario Ortega, Catherine Hoyt-Drazen, Caterina Gratton, Haoxin Sun, et al. (2017b). "Precision functional mapping of individual human brains". In: *Neuron* 95.4, pp. 791–807.

Honey, Christopher J, Olaf Sporns, Leila Cammoun, Xavier Gigandet, Jean-Philippe Thiran, Reto Meuli, and Patric Hagmann (2009). "Predicting human resting-state functional connectivity from structural connectivity". In: *Proceedings of the National Academy of Sciences* 106.6, pp. 2035–2040. ISSN: 0027-8424.

Horn, Andreas, Dirk Ostwald, Marco Reisert, and Felix Blankenburg (2014). "The structural–functional connectome and the default mode network of the human brain". In: *Neuroimage* 102, pp. 142–151. ISSN: 1053-8119.

Jenkinson, Mark, Christian F Beckmann, Timothy E J Behrens, Mark W Woolrich, and Stephen M Smith (2012). "Fsl". In: *Neuroimage* 62.2, pp. 782–790.

Karcher, Nicole R and Deanna M Barch (2021). "The ABCD study: understanding the development of risk for mental and physical health outcomes". In: *Neuropsychopharmacology* 46.1, pp. 131–142.

Lefèvre, Julien, Antonietta Pepe, Jennifer Muscato, Francois De Guio, Nadine Girard, Guillaume Auzias, and David Germanaud (2018). "SPANOL (SPectral ANalysis of Lobes): A Spectral Clustering Framework for Individual and Group Parcellation of Cortical Surfaces in Lobes". In: *Frontiers in Neuroscience* 12, p. 354. ISSN: 1662-453X. URL: https://www.frontiersin.org/article/10.3389/fnins.2018.00354.

Margulies, Daniel S, Marcel Falkiewicz, and Julia M Huntenburg (Nov. 2016). "A cortical surface-based geodesic distance package for Python". In: *GigaScience* 5.suppl_1. ISSN: 2047-217X. DOI: 10.1186/s13742-016-0147-0-q. URL: https://doi.org/10.1186/s13742-016-0147-0-q.

Markello, Ross D and Bratislav Misic (2021). "Comparing spatial null models for brain maps". In: *Neuroimage*, p. 118052.

Muschelli, John (2018). *gifti: Reads in 'Neuroimaging' 'GIFTI' Files with Geometry Information*. R package version 0.7.5. URL: https://CRAN.R-project.org/package=gifti.

Paquola, Casey, Richard A I Bethlehem, Jakob Seidlitz, Konrad Wagstyl, Rafael Romero-Garcia, Kirstie J Whitaker, Reinder Vos De Wael, Guy B Williams, Petra E Vértes, and Daniel S Margulies (2019). "Shifts in myeloarchitecture characterise adolescent development of cortical gradients". In: *Elife* 8, e50482. ISSN: 2050-084X.

Phipson, Belinda and Gordon K Smyth (2010). "Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn". In: *Statistical applications in genetics and molecular biology* 9.1.

Ragland, J Daniel, Bruce I Turetsky, Ruben C Gur, Faith Gunning-Dixon, Travis Turner, Lee Schroeder, Robin Chan, and Raquel E Gur (2002). "Working memory for complex figures: an fMRI comparison of letter and fractal n-back tasks." In: *Neuropsychology* 16.3, p. 370.

Rosen, Adon, David R Roalf, Kosha Ruparel, Jason Blake, Kevin Seelaus, L Prayosha Villa, Rastko Ciric, Philip A Cook, Christos Davatzikos, Mark A Elliott, and others (2017). "Data-driven assessment of structural image quality". In: *BioRxiv*, p. 125161.

Satterthwaite, Theodore D, John J Connolly, Kosha Ruparel, Monica E Calkins, Chad Jackson, Mark A Elliott, David R Roalf, Ryan Hopson, Karthik Prabhakaran, Meckenzie Behr, and others (2016). "The Philadelphia Neurodevelopmental Cohort: A publicly available resource for the study of normal and abnormal brain development in youth". In: *Neuroimage* 124, pp. 1115–1119.

Satterthwaite, Theodore D, Mark A Elliott, Kosha Ruparel, James Loughead, Karthik Prabhakaran, Monica E Calkins, Ryan Hopson, Chad Jackson, Jack Keefe, Marisa Riley, and others (2014). "Neuroimaging of the Philadelphia neurodevelopmental cohort". In: *Neuroimage* 86, pp. 544–553.

Satterthwaite, Theodore D, Daniel H Wolf, Guray Erus, Kosha Ruparel, Mark A Elliott, Efstathios D Gennatas, Ryan Hopson, Chad Jackson, Karthik Prabhakaran, Warren B Bilker, Monica E Calkins, James Loughead, Alex Smith, David R Roalf, Hakon Hakonarson, Ragini Verma, Christos Davatzikos, Ruben C Gur, and Raquel E Gur (Oct. 2013). "Functional Maturation of the Executive System during Adolescence". In: *The Journal of Neuroscience* 33.41, p. 16249. DOI: 10.1523/JNEUROSCI.2345-13.2013. URL: http://www.jneurosci.org/content/33/41/16249.abstract.

Schaefer, Alexander, Ru Kong, Evan M Gordon, Timothy O Laumann, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, and B T Thomas Yeo (2018). "Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI". In: *Cerebral Cortex* 28.9, pp. 3095–3114.

Schmitt, J Eric, Michael C Neale, Liv S Clasen, Siyuan Liu, Jakob Seidlitz, Joshua N Pritikin, Alan Chu, Gregory L Wallace, Nancy Raitano Lee, and Jay N Giedd (2019). "A comprehensive quantitative genetic analysis of cerebral surface area in youth". In: *Journal of Neuroscience* 39.16, pp. 3028–3040. ISSN: 0270-6474.

Schäfer, Tim (2020a). *freesurferformats: Read and Write 'FreeSurfer' Neuroimaging File Formats.* R package version 0.1.12. URL: https://CRAN.R-project.org/package=freesurferformats.

— (2020b). *fsbrain: Managing and Visualizing Brain Surface Data.* R package version 0.3.0. URL: https://CRAN.R-project.org/package=fsbrain.

Shafiei, Golia, Ross D Markello, Carolina Makowski, Alexandra Talpalaru, Matthias Kirschner, Gabriel A Devenyi, Elisa Guma, Patric Hagmann, Neil R Cashman, Martin Lepage, M Mallar Chakravarty, Alain Dagher, and Bratislav Mišić (2020). "Spatial Patterning of Tissue Volume Loss in Schizophrenia Reflects Brain Network Architecture". In: *Biological Psychiatry* 87.8, pp. 727–735. ISSN: 0006-3223. DOI: https://doi.org/10.1016/j.biopsych.2019.09.031. URL: http://www.sciencedirect.com/science/article/pii/S0006322319317858.

Shaw, Philip, Noor J Kabani, Jason P Lerch, Kristen Eckstrand, Rhoshel Lenroot, Nitin Gogtay, Deanna Greenstein, Liv Clasen, Alan Evans, Judith L Rapoport, and others (2008). "Neurodevelopmental trajectories of the human cerebral cortex". In: *Journal of neuroscience* 28.14, pp. 3586–3594.

Sowell, Elizabeth R, Paul M Thompson, Christiana M Leonard, Suzanne E Welcome, Eric Kan, and Arthur W Toga (2004). "Longitudinal mapping of cortical thickness and brain growth in normal children". In: *Journal of neuroscience* 24.38, pp. 8223–8231.

Stoecklein, Sophia, Anne Hilgendorff, Meiling Li, Kai Förster, Andreas W. Flemmer, Franziska Galiè, Stephan Wunderlich, Danhong Wang, Sophie Stein, Harald Ehrhardt, Olaf Dietrich, Qihong Zou, Shuqin Zhou, Birgit Ertl-Wagner, and Hesheng Liu (Jan. 2020). "Variable functional connectivity architecture of the preterm human brain: Impact of developmental cortical expansion and maturation". In: *Proceedings of the National Academy of Sciences* 117.2, pp. 1201–1206. ISSN: 0027-8424. DOI: 10.1073/PNAS.1907892117. URL: https://www.pnas.org/content/117/2/1201.

Tamnes, Christian K, Megan M Herting, Anne-Lise Goddings, Rosa Meuwese, Sarah-Jayne Blakemore, Ronald E Dahl, Berna Güroğlu, Armin Raznahan, Elizabeth R Sowell, Eveline A Crone, et al. (2017). "Development of the cerebral cortex across adolescence: a multisample study of inter-related longitudinal changes in cortical volume, surface area, and thickness". In: *Journal of Neuroscience* 37.12, pp. 3402–3412.

Tooley, Ursula A, Danielle S Bassett, and Allyson P Mackey (2021). "Environmental influences on the pace of brain development". In: *Nature Reviews Neuroscience*, pp. 1–13.

Vandekar, Simon N, Russell T Shinohara, Armin Raznahan, Ryan D Hopson, David R Roalf, Kosha Ruparel, Ruben C Gur, Raquel E Gur, and Theodore D Satterthwaite (2016). "Subject-level measurement of local cortical coupling". In: *NeuroImage* 133, pp. 88–97.

Vandekar, Simon N, Russell T Shinohara, Armin Raznahan, David R Roalf, Michelle Ross, Nicholas DeLeo, Kosha Ruparel, Ragini Verma, Daniel H Wolf, Ruben C Gur, and others (2015). "Topologically dissociable patterns of development of the human cerebral cortex". In: *Journal of Neuroscience* 35.2, pp. 599–609.

Váša, František, Jakob Seidlitz, Rafael Romero-Garcia, Kirstie J Whitaker, Gideon Rosenthal, Petra E Vértes, Maxwell Shinn, Aaron Alexander-Bloch, Peter Fonagy, Raymond J Dolan, Peter B Jones, Ian M Goodyer, NSPN consortium, Olaf Sporns, and Edward T Bullmore (Jan. 2018). "Adolescent Tuning of Association Cortex in Human Structural Brain Networks". eng. In: *Cerebral cortex (New York, N.Y. : 1991)* 28.1, pp. 281–294. ISSN: 1460-2199. DOI: 10.1093/cercor/bhx249. URL: https://pubmed.ncbi.nlm.nih.gov/29088339https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5903415/.

Vázquez-Rodríguez, Bertha, Laura E Suárez, Ross D Markello, Golia Shafiei, Casey Paquola, Patric Hagmann, Martijn P van den Heuvel, Boris C Bernhardt, R Nathan Spreng, and Bratislav Misic (Oct. 2019). "Gradients of structure-function tethering across neocortex." In: *Proceedings of the National Academy of Sciences of the United States of America* 116.42, pp. 21219–21227. ISSN: 1091-6490. DOI: 10.1073/pnas.1903403116. URL: http://www.ncbi.nlm.nih.gov/pubmed/31570622http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6800358.

Venkatesh, Manasij, Joseph Jaja, and Luiz Pessoa (2020). "Comparing functional connectivity matrices: A geometry-aware approach applied to participant identification". In: *NeuroImage* 207, p. 116398.

Viladomat, Júlia, Rahul Mazumder, Alex McInturff, Douglas J McCauley, and Trevor Hastie (June 2014). "Assessing the significance of global and local correlations under spatial autocorrelation: A nonparametric approach". In: *Biometrics* 70.2, pp. 409–418. ISSN: 0006-341X. DOI: 10.1111/biom.12139. URL: https://doi.org/10.1111/biom.12139.

Von Economo, Constantin (1929). *The cytoarchitectonics of the human cerebral cortex.* H. Milford Oxford University Press.

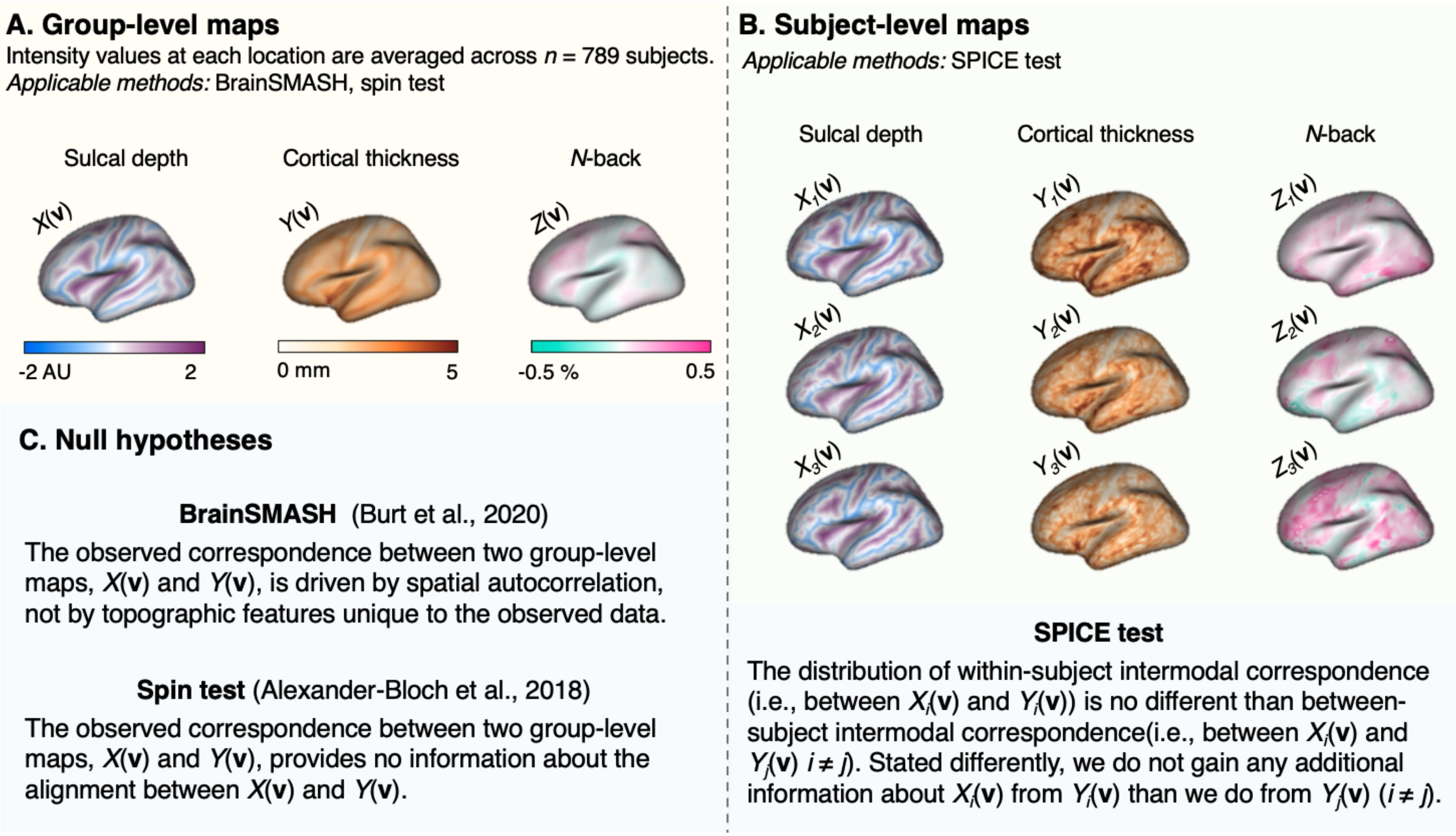
Wael, Reinder Vos de, Oualid Benkarim, Casey Paquola, Sara Lariviere, Jessica Royer, Shahin Tavakol, Ting Xu, Seok-Jun Hong, Georg Langs, Sofie Valk, et al. (2020). "BrainSpace: a toolbox for the analysis of macroscale gradients in neuroimaging and connectomics datasets". In: *Communications biology* 3.1, pp. 1–10.

Winkler, Anderson M, Gerard R Ridgway, Matthew A Webster, Stephen M Smith, and Thomas E Nichols (2014). "Permutation inference for the general linear model". In: *NeuroImage* 92, pp. 381–397. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2014.01.060. URL: http://www.sciencedirect.com/science/article/pii/S1053811914000913.

Ye, Jun, Nicole A Lazar, and Yehua Li (2015). "Nonparametric variogram modeling with hole effect structure in analyzing the spatial characteristics of fMRI data". In: *Journal of neuroscience methods* 240, pp. 101–115.

Yeo, B T, Fenna M Krienen, Jorge Sepulcre, Mert R Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L Roffman, Jordan W Smoller, Lilla Zöllei, Jonathan R Polimeni, and others (2011). "The organization of the human cerebral cortex estimated by intrinsic functional connectivity". In: *Journal of neurophysiology* 106.3, pp. 1125–1165.

Figure 1: Example left hemisphere sulcal depth, cortical thickness, and n-back maps used for (A) the spin test and BrainSMASH and (B) the SPICE test. Brain map visualizations were generated using the R package fsbrain (Schäfer 2020b). (C) Null hypotheses for previous methods and our proposed method for testing intermodal correspondence. Of the three methods, the SPICE test is the only one to incorporate subject-level data.



BrainSMASH (Burt et al., 2020)

The observed correspondence between two group-level maps, $X(\mathbf{v})$ and $Y(\mathbf{v})$, is driven by spatial autocorrelation, not by topographic features unique to the observed data.

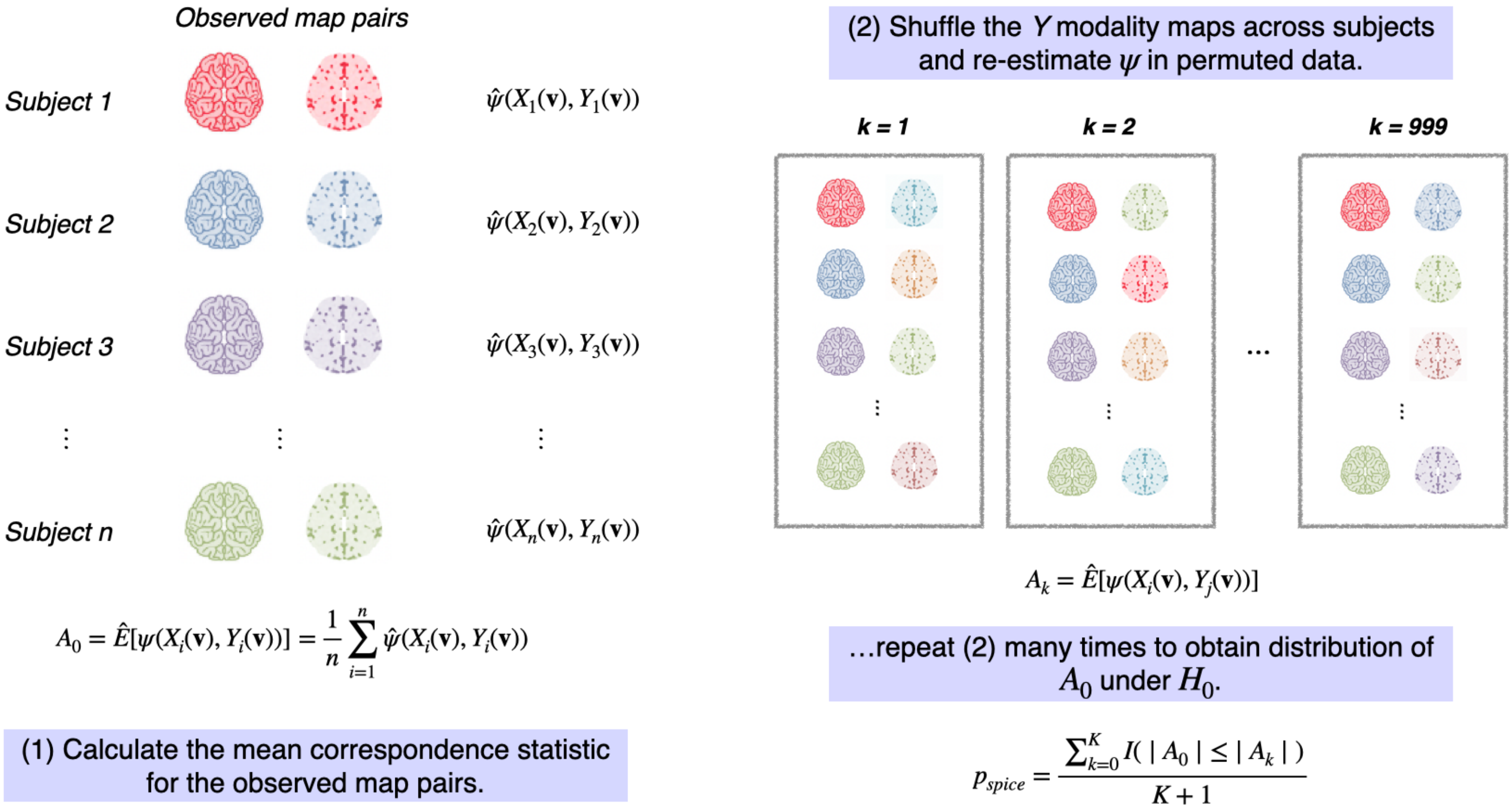
Spin test (Alexander-Bloch et al., 2018)

The observed correspondence between two group-level maps, $X(\mathbf{v})$ and $Y(\mathbf{v})$, provides no information about the alignment between $X(\mathbf{v})$ and $Y(\mathbf{v})$.

SPICE test

The distribution of within-subject intermodal correspondence (i.e., between $X_i(\mathbf{v})$ and $Y_i(\mathbf{v})$) is no different than between-subject intermodal correspondence(i.e., between $X_i(\mathbf{v})$ and $Y_j(\mathbf{v})$ $i \neq j$). Stated differently, we do not gain any additional information about $X_i(\mathbf{v})$ from $Y_i(\mathbf{v})$ than we do from $Y_j(\mathbf{v})$ ($i \neq j$).

Figure 2: Illustration of simple permutation-based intermodal correspondence (SPICE) testing procedure. A null distribution for A_0, the average within-subject correspondence statistic, is calculated by randomly permuting the Y maps $K = 999$ times and re-estimating A_k in each permuted version of the data. A p-value is estimated as in (4).



Observed map pairs
Subject 1
$\hat{\psi}(X_1(\mathbf{v}), Y_1(\mathbf{v}))$
Subject 2
$\hat{\psi}(X_2(\mathbf{v}), Y_2(\mathbf{v}))$
Subject 3
$\hat{\psi}(X_3(\mathbf{v}), Y_3(\mathbf{v}))$
Subject n
$\hat{\psi}(X_n(\mathbf{v}), Y_n(\mathbf{v}))$
$A_0 = \hat{E}[\psi(X_i(\mathbf{v}), Y_i(\mathbf{v}))] = \frac{1}{n}\sum_{i=1}^{n}\hat{\psi}(X_i(\mathbf{v}), Y_i(\mathbf{v}))$
(1) Calculate the mean correspondence statistic for the observed map pairs.
(2) Shuffle the Y modality maps across subjects and re-estimate ψ in permuted data.
k = 1
k = 2
k = 999
$A_k = \hat{E}[\psi(X_i(\mathbf{v}), Y_j(\mathbf{v}))]$
…repeat (2) many times to obtain distribution of A_0 under H_0.
$p_{spice} = \frac{\sum_{k=0}^{K} I(|A_0| \leq |A_k|)}{K+1}$

Figure 3: Illustration of bi-modal image simulation. Subject-level images are derived from average cortical thickness $(M_1(\mathbf{v}))$ and sulcal depth $(M_2(\mathbf{v}))$ data from the Philadelphia Neurodevelopmental Cohort. For illustrative purposes, the example below shows $M_2(\mathbf{v})$ as mean sulcal depth, but we also consider a simulation setting using mean n-back as the second population-level map.

Subject-level images for $i = 1, ..., n$ subjects are simulated as $X_i(\mathbf{v}) = a_i \times M_1(\mathbf{v}) + E_{i(1)}(\mathbf{v})$ and $Y_i(\mathbf{v}) = a_i \times M_2(\mathbf{v}) + E_{i(2)}(\mathbf{v})$, where a_i and the elements of $E_{i(1)}(\mathbf{v})$ and $E_{i(2)}(\mathbf{v})$ are normally distributed, with mean and variance parameters specified in Section 2.2.2. The null hypothesis is true when the variance of a_i, σ_a^2 is equal to 0. Otherwise, we expect to reject H_0, with test power varying according to other parameter values.

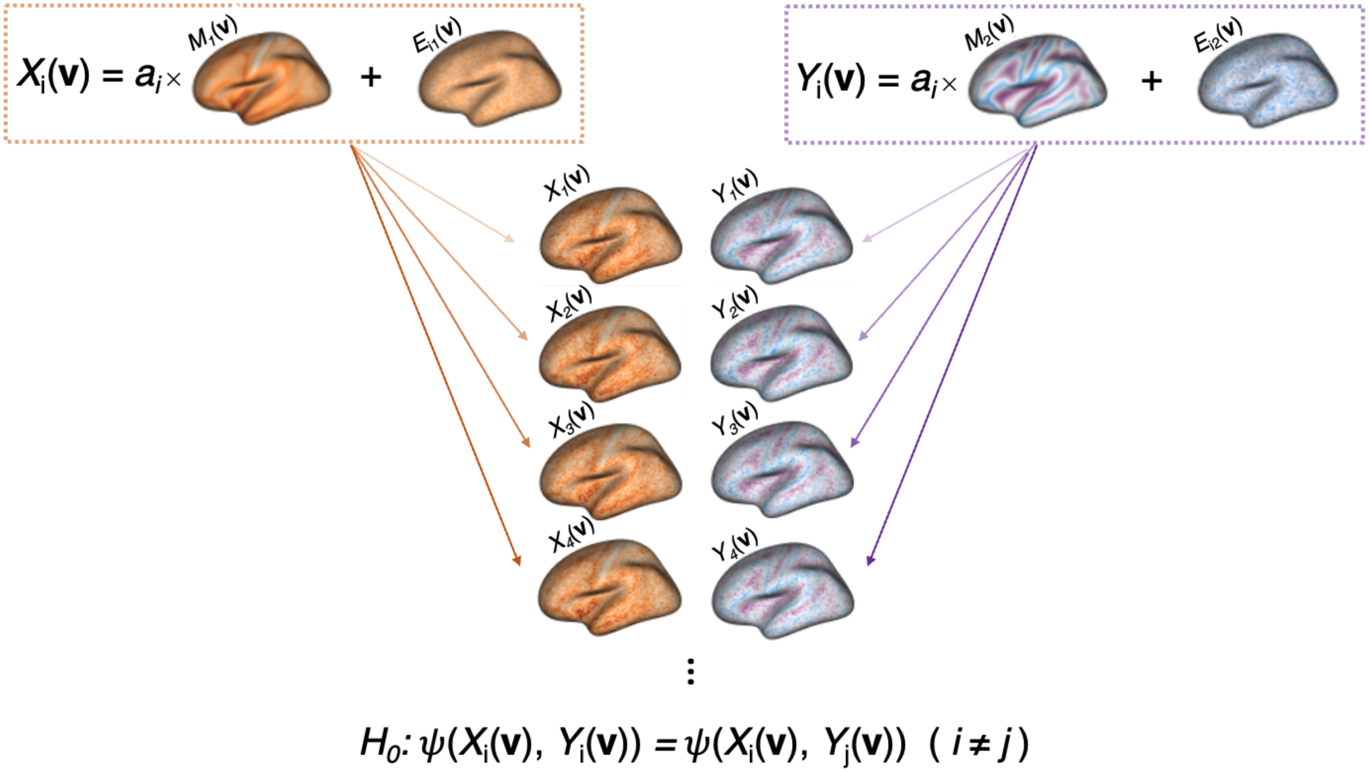
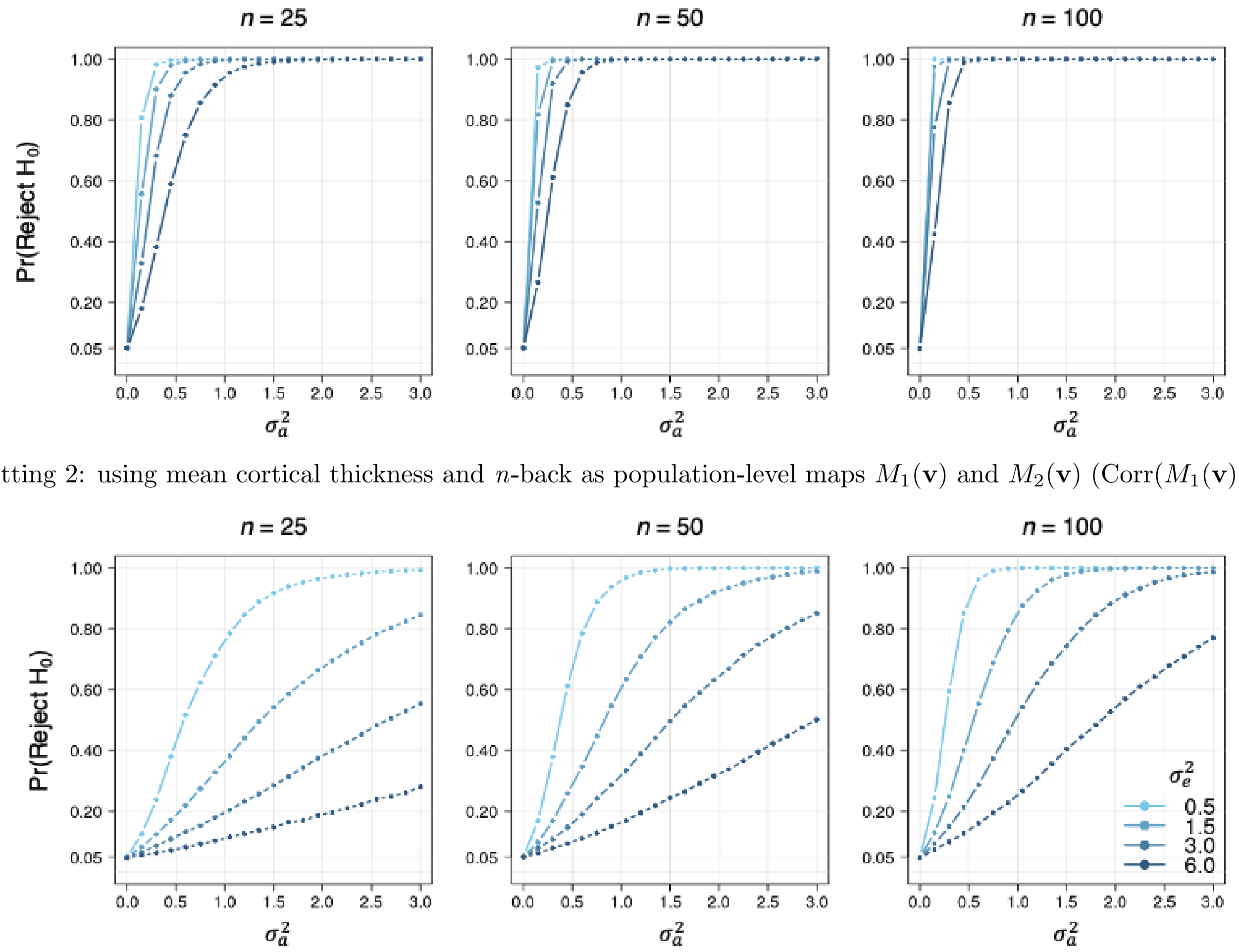


Figure 4: Power and type I error of the simple permutation-based intermodal correspondence (SPICE) test based on 5000 simulations. Each point is the rate of rejecting H_0 (based on $p_{spice} < 0.05$) from 5000 simulations of of the data (as shown in Figure 3) with unique combinations of parameters: sample size ($n = 25, 50,$ or 100), subject-level variance (σ_a^2, ranging from 0.0 to 3.0 in increments of 0.15), and vertex-level variance (σ_e^2, either 0.5, 1.5, 3.0, or 6.0). Both simulation settings involve using $M_1(\mathbf{v})$ as mean cortical thickness from a subset of 789 participants in the Philadelphia Neurodevelopmental Cohort. $M_2(\mathbf{v})$ is mean sulcal depth in setting 1 and mean n-back in setting 2.

(a) Simulation setting 1: using mean cortical thickness and sulcal depth as population-level maps $M_1(\mathbf{v})$ and $M_2(\mathbf{v})$ ($\mathrm{Corr}(M_1(\mathbf{v}), M_2(\mathbf{v})) = -0.14$).



(b) Simulation setting 2: using mean cortical thickness and n-back as population-level maps $M_1(\mathbf{v})$ and $M_2(\mathbf{v})$ ($\mathrm{Corr}(M_1(\mathbf{v}), M_2(\mathbf{v})) = -0.04$).

Figure 5: Unadjusted p-values from tests of intermodal correspondence within seven functional networks described by Yeo et al. 2011 in the left hemisphere for different age groups. We consider $p < 0.007$ to provide evidence against the null hypotheses (defined in Figure 1), after using a Bonferroni correction for comparisons across seven age groups.

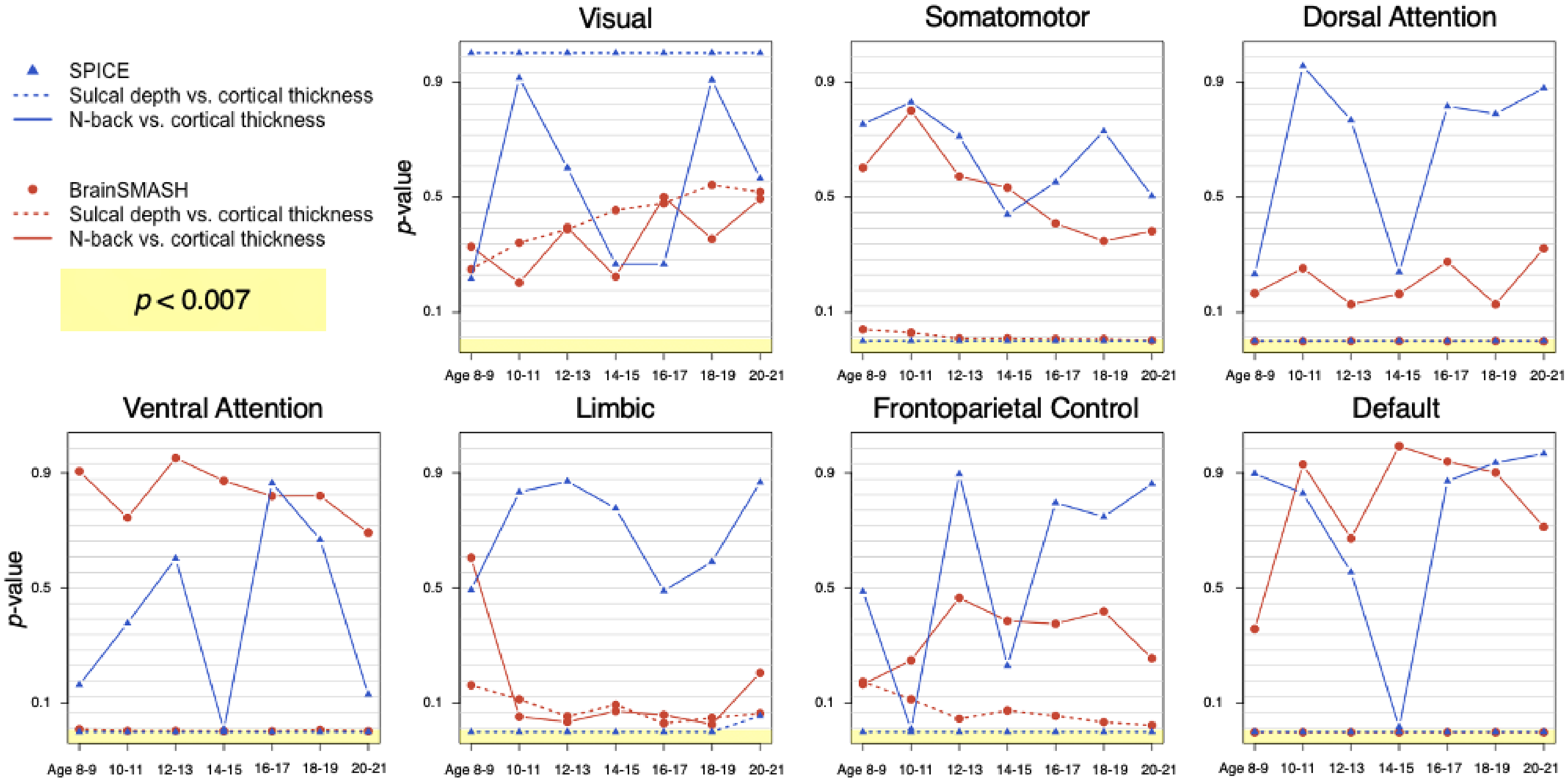


Table 1: Unadjusted p-values from tests of intermodal correspondence in the left and right hemispheres using the SPICE, BrainSMASH, and spin methods. Null hypotheses for each method are summarized in Figure 1. Additionally, Figure A2 shows the null distributions and observed test statistics used to estimate each p-value below.

(a) Sulcal depth vs. cortical thickness

Age	**8-9**	**10-11**	**12-13**	**14-15**	**16-17**	**18-19**	**20-21**	**All subjects**
	$n = 69$	$n = 95$	$n = 129$	$n = 157$	$n = 170$	$n = 147$	$n = 22$	$n = 789$
SPICE								
Left	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Right	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
BrainSMASH								
Left	0.007	0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Right	0.034	0.004	0.005	0.002	< 0.001	0.001	0.001	0.002
Spin								
Left	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Right	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

(b) N-back vs. cortical thickness

Age	**8-9**	**10-11**	**12-13**	**14-15**	**16-17**	**18-19**	**20-21**	**All subjects**
	$n = 69$	$n = 95$	$n = 129$	$n = 157$	$n = 170$	$n = 147$	$n = 22$	$n = 789$
SPICE								
Left	0.899	0.842	0.281	0.958	0.796	0.345	0.829	0.838
Right	0.222	0.879	0.784	0.513	0.752	0.254	0.468	0.470
BrainSMASH								
Left	0.726	0.974	0.907	0.932	0.798	0.771	0.948	0.914
Right	0.344	0.464	0.795	0.786	0.936	0.878	0.868	0.831
Spin								
Left	0.987	0.851	0.760	0.894	0.575	0.572	0.835	0.730
Right	0.609	0.711	0.994	0.964	0.742	0.655	0.922	0.947

Appendix

A1 Imaging protocols

A1.1 Acquisition and pre-processing of structural imaging data

All subjects underwent magnetic resonance imaging (MRI) in the same Siemens TIM Trio 3 tesla scanner with a 32-channel head coil and the same imaging sequences and parameters, described in detail by Satterthwaite et al. (2014) and Satterthwaite et al. (2016). For structural imaging, the protocol included a magnetization-prepared, rapid-acquisition gradient echo (MPRAGE) T1-weighted structural image with a voxel resolution of 0.9 × 0.9 × 1 mm. To ensure adequate quality of images included in our analysis, quality assurance was independently rated by three experienced image analysts, as described by Rosen et al. (2017).

Cortical reconstruction of T1-weighted structural images was implemented using FreeSurfer (version 5.3), which included template registration, intensity normalization, and inflation of cortical surfaces to a template. Cortical measurements were resampled to the fsaverage5 atlas, consisting of exactly $V = 10,242$ vertices in each hemisphere for every subject. We quantified cortical thickness as the minimum distance between pial and white matter surfaces (Dale et al. 1999) and sulcal depth as the height of gyri (with positive and negative values indicating outward and inward movement, respectively) (Fischl et al. 1999).

A1.2 Acquisition and pre-processing of n-back task sequence

Protocols for functional MRI (fMRI) included task-based BOLD scans using a single-shot, interleaved multi-slice, gradient-echo, echo planar imaging sequence, with a voxel resolution of 3 × 3 × 3 mm voxels and 46 slices. Preprocessing was implemented using the eXtensible Connectivity Pipeline (XCP) Engine (Ciric et al. 2018). In the current study, we consider the fractal n-back task sequence, which involves presenting subjects with a series of stimuli (geometric pictures on a screen) with instructions to press a button if the current stimulus matches the nth previous one. For example, in a 1-back sequence, participants would be instructed to recognize whether a current stimulus matches the one that appeared last; a 2-back sequence would involve recognizing whether the current stimulus matches the one that appeared two stimuli ago. A 0-back sequence is also considered, in which subjects are instructed to simply press a button each time a stimulus appears (regardless of whether it matches a previous stimulus). In our present analysis we consider maps that represent the percentage change in activation between the 2-back and 0-back sequences, which has been previously shown to isolate executive network function (Ragland et al. 2002).

We used the FEAT software tool in the FSL library to fit first-level general linear models on the *n*-back time series data (Jenkinson et al. 2012). Three first level models for each subject considered contrasts to assess change in working memory load between 1-back and 0-back, 2-back and 1-back, and 2-back and 0-back. Activation maps representing the percent change between the 2-back and 0-back sequences were projected into the fsaverage5 template with exactly 10,242 vertices per hemisphere per subject.

A1.3 Use of cortical and network parcellations from Yeo et al. (2011) and Schaefer et al. (2018)

Images converted to the fsaverage5 template consisted of exactly 10,242 vertices per hemisphere. We then identified the vertices belonging to each of the 1000 parcellations suggested by Schaefer et al. (2018) as well as each of the seven networks proposed by Yeo et al. (2011), available for download from GitHub (`https://github.com/ThomasYeoLab/CBIG/tree/master/stable_projects/brain_parcellation/Schaefer2018_LocalGlobal/Parcellations/FreeSurfer5.3/fsaverage5/label`).

A2 Supplemental figures

Figure A1: Empirical and surrogate (no. surrogates = 1000) variograms of parcellated cortical thickness measurements. These variograms are constructed to assess the reliability of the Brain Surrogate Maps with Autocorrelated Spatial Heterogeneity (BrainSMASH) method in testing for intermodal correspondence between cortical thickness and sulcal depth and cortical thickness and n-back in the left and right hemispheres (parcellations from Schaefer et al. (2018)). The horizontal axis of each figure indicates the spatial separation distance (d), and the vertical axis describes the variation between the cortical thickness measurements observed in parcels separated by distance d.

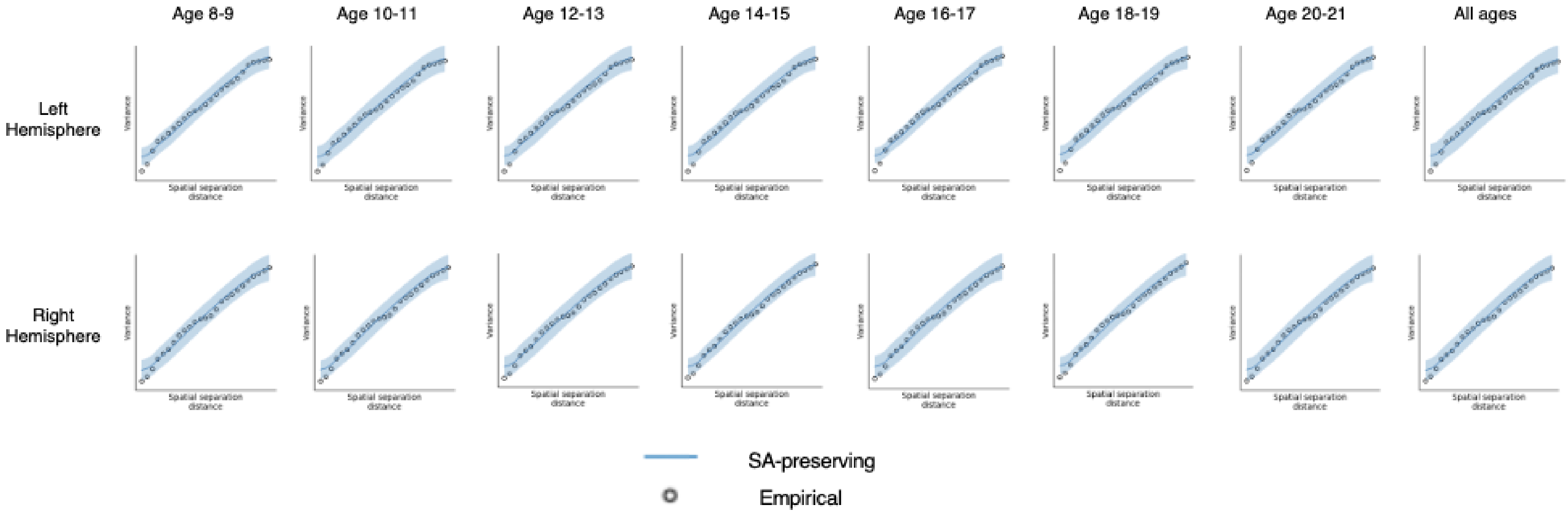
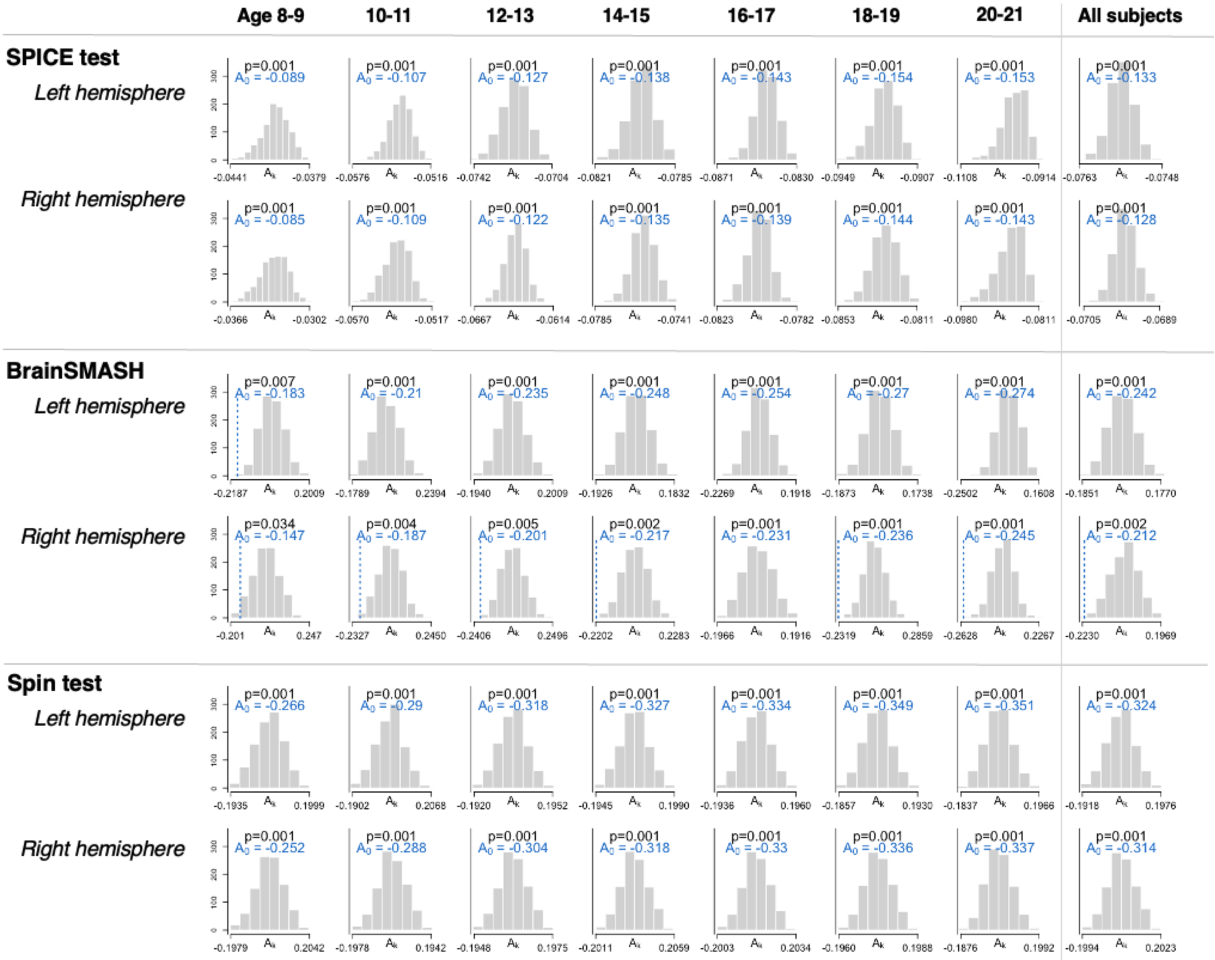


Figure A2: Null test statistic distributions corresponding to results shown in Table 1 for the SPICE test, BrainSMASH, and spin test. The observed test statistic, A_0, is plotted or written in blue. (A_0 is not plotted when it falls outside the range of the null test statistics.) Note: the observed test statistics for BrainSMASH and the spin test are not identical, even though both these methods use group-level surfaces, since the spin test removes the medial wall before calculating intermodal correspondence.

(a) Cortical thickness vs. sulcal depth (corresponds to Table 1(a)).



(b) Cortical thickness vs. n-back (corresponds to Table 1(b)).

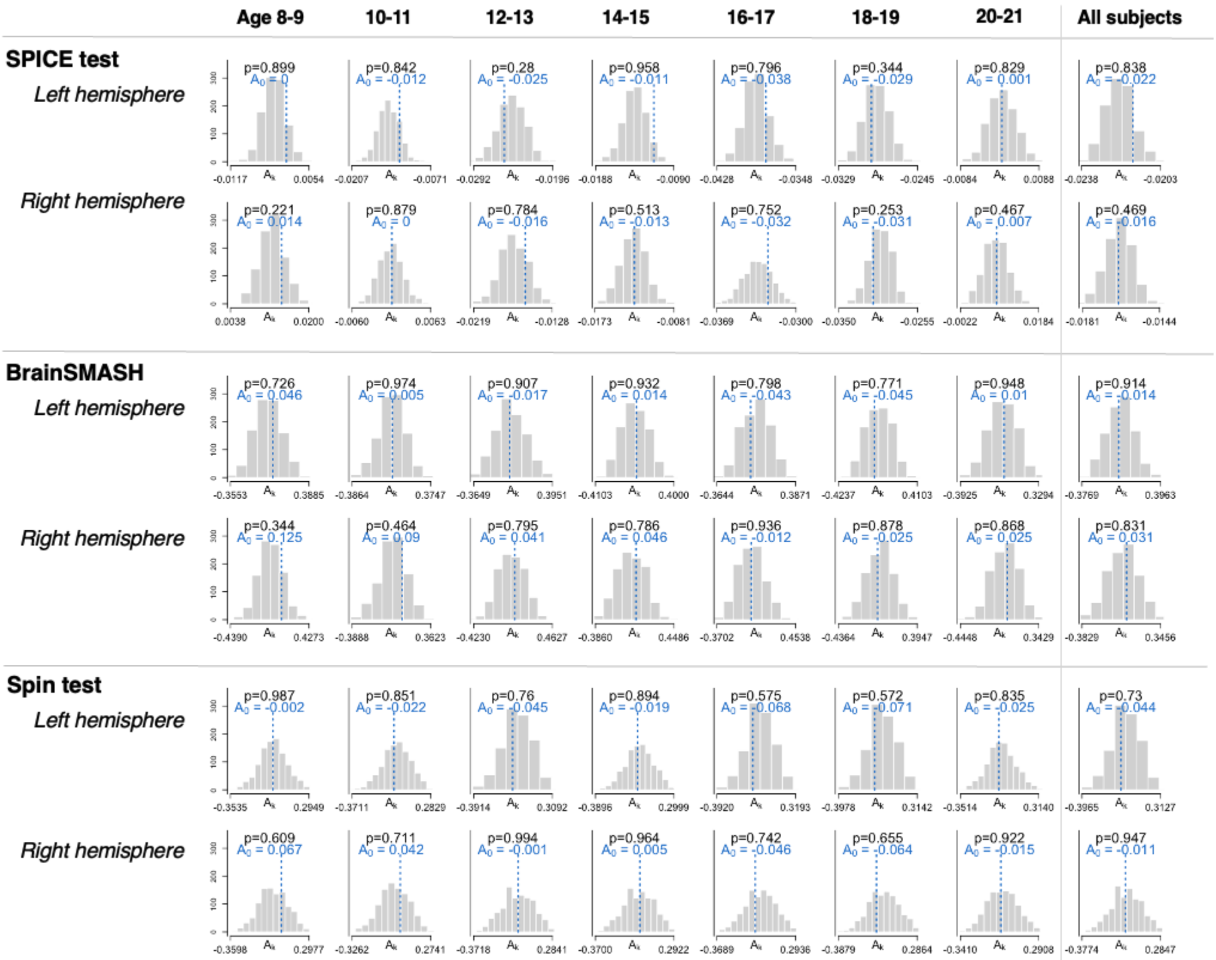
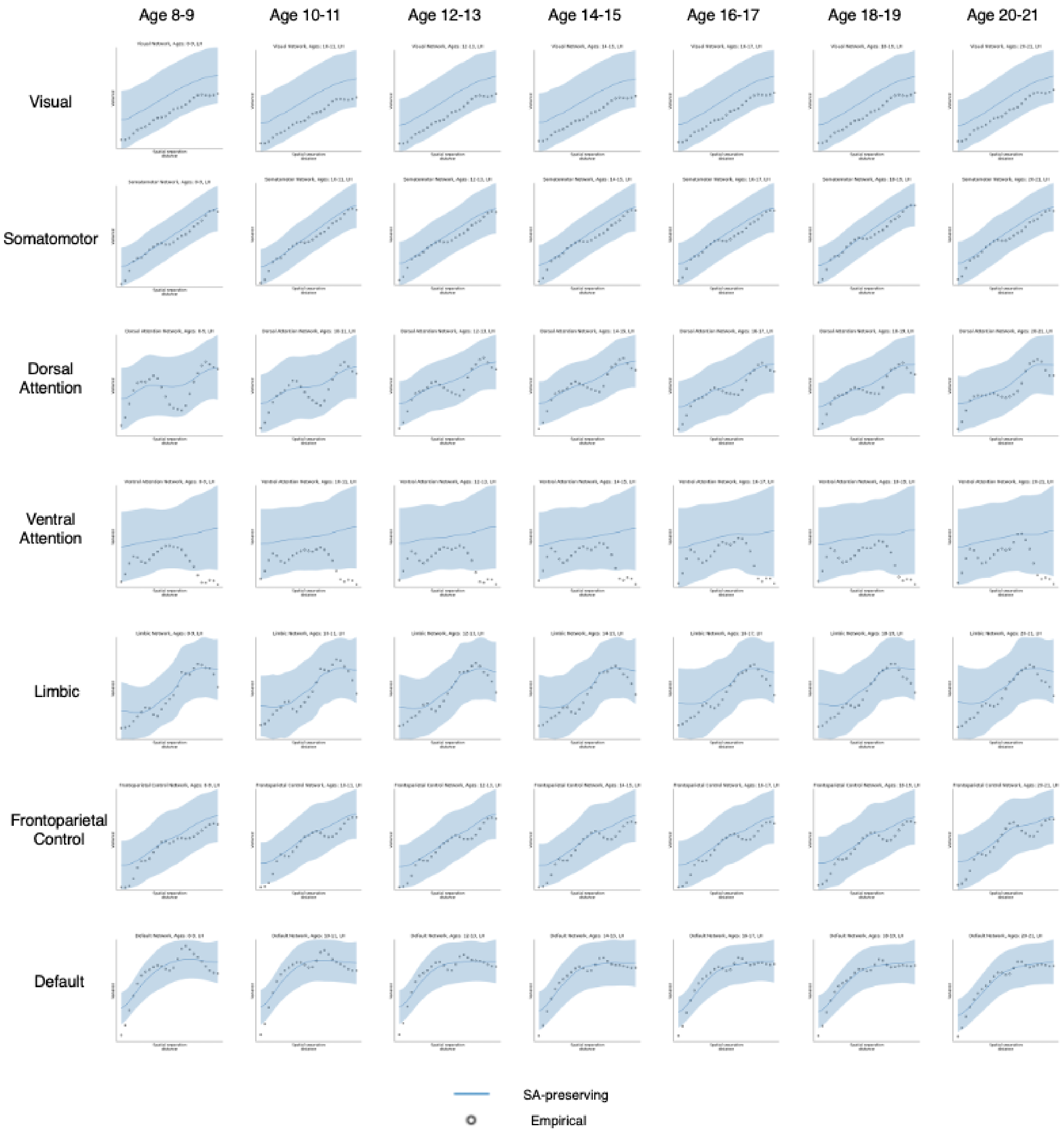


Figure A3: Empirical and surrogate (no. surrogates = 1000) variograms of parcellated cortical thickness measurements within seven functional networks, according to Yeo et al. (2011). These variograms are constructed to assess the reliability of BrainSMASH in testing for intermodal correspondence between cortical thickness and sulcal depth and cortical thickness and n-back within age-stratified groups. The horizontal axis of each plot indicates the spatial separation distance (d); the vertical axis describes the variation between cortical thickness measurements observed in parcels separated by the distance d.


Age 8-9
Age 10-11
Age 12-13
Age 14-15
Age 16-17
Age 18-19
Age 20-21
Visual
Somatomotor
Dorsal Attention
Ventral Attention
Limbic
Frontoparietal Control
Default
SA-preserving
Empirical