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-Fizgerald[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

*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. While several statistical procedures have been proposed to test correspondence between maps, there remains no consensus on the correct framing of a null hypothesis or a suitable testing approach. We propose a simple yet powerful permutation-based testing procedure for assessing similarities between two modalities using subject-level data. Our proposed method is similar to traditional permutation procedures in that it involves randomly permuting subjects to generate a null distribution. However, it differs from other recently proposed methods that have involved spherical rotations of the cortical surface or spatial autocorrelation-preserving "surrogate" maps of the brain, which depend on strong and potentially unrealistic statistical assumptions. To address these issues, we first demonstrate in simulated data that our method is conservative in terms of type I error and has high power. Next, we illustrate that our method performs well for assessing intermodal relationships from multimodal magnetic resonance imaging data from the Philadelphia Neurodevelopmental Cohort. The proposed test rejects the null hypothesis for modalities for which there is known interdependence in structure (cortical thickness and sulcal depth) but not in cases where an association would not be predicted biologically (cortical thickness and activation on the n-back working memory task). In contrast to previous methods, our approach does not depend on strong statistical assumptions other than the independence of subjects. Notably, our method is the most flexible for analyzing intermodal correspondence within subregions of the brain and has the greatest potential to be used for generalizable statistical inference.

1 Introduction

Neuroimaging studies often seek to understand similarities across modalities. However, methods underlying these studies' findings continue to vary greatly, often ignore spatial autocorrelation, and lack statistical rigor. Naive approaches to the so-called "correspondence problem" have included simply visualizing two maps next to each other for comparison, without incorporating quantitative measures or statistical tests. Another approach involves the use of parametric p-values; for example, fitting a linear regression or estimating the correlation across locations of the cerebral 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).

New methods have emerged in recent neuroimaging literature to address the correspondence problem by incorporating random rotations of the cortical surface to generate null models of spatial alignment. A precursor to this approach, now known as the *spin test*, used 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 disregard of the medial wall in spherical rotations, its lack of flexibility for incorporating subject-level data by accommodating only two (group-level) maps, and that it cannot be used to test for correspondence in small regions of the brain.

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 improves on 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 important 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. In this paper, we will consider how this may be a limiting factor when analyzing large-scale functional networks, where subregions may be small, irregularly shaped, or spatially disconnected.

The spin test, BrainSMASH, and other related methods recently described by Markello and Misic (2020) involve testing a null hypothesis of intermodal correspondence conditional on observed group-level data. Importantly, these methods are not 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). Thus, careful consideration of the plausibility of this assumption is warranted. In this paper, our goal is to introduce an alternative but complementary approach that does not require such strong statistical assumptions.

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. Our approach considers subject-level imaging data from two modalities and leverages independence of subjects to implement

a permutation-based test that does not involve any alteration to the underlying spatial structure of our data. 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, as opposed to brain maps derived from different individuals.

2 Methods

2.1 Null hypothesis

We consider 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 goal is to test the null hypothesis that the distribution of the within-subject intermodal correspondence 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 other words, we would like to evaluate whether we gain more information about $X_i(\mathbf{v})$ from $Y_i(\mathbf{v})$ (i.e., two images belonging to the same subject) as opposed to from $Y_j(\mathbf{v})$ (for $i \neq j$).

We express the null hypothesis of equality between the distributions of within- and between-subject correspondence as:

$$H_0 : \psi(X_i(\mathbf{v}), Y_i(\mathbf{v})) \stackrel{d}{=} \psi(X_i(\mathbf{v}), Y_j(\mathbf{v})) \qquad i \neq j, \tag{1}$$

where $\psi()$ represents a measure of intermodal similarity between X and Y across all locations indexed by $\mathbf{v} = [1, \ldots, V]$, either for the same subject or for different subjects (when $i \neq j$). To test H_0, we propose a permutation-based procedure. In doing this, we consider a test of equality in the means of the distributions referenced in (1). Rejecting the null hypothesis of equality between means implies rejecting (1).

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 questions about intermodal correspondence. Table 1 outlines these three null hypotheses, which we will reference throughout the paper as $H_{0(spice)}$, $H_{0(b\text{-}smash)}$, and $H_{0(spin)}$.

2.2 Implementing a permutation-based test of $H_{0(spice)}$

Having observed the subject-level pairs $X_i(\mathbf{v}), Y_i(\mathbf{v})$, we can 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 $\widehat{\psi}(X_i(\mathbf{v}), Y_i(\mathbf{v}))$ is now the sample intermodal Pearson correlation for the ith subject, across all V locations; that is,

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

although other measures of intermodal correspondence can easily be substituted. We then generate a null distribution for A_0 by randomly shuffling the Y maps across subjects. Then, for 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, illustrated in Figure 1, are fundamentally different from what the authors of the spin test and BrainSMASH refer to as permutations when they spatially re-align the data.

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

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

where we add 1 to the denominator to account for the identity permutation. We reject the null hypothesis for $p_{spice} < \alpha$, the nominal type I error rate. If the null hypothesis is true, we expect A_k to be similar in magnitude to A_0; thus, p_{spice} would be relatively large and we would fail to reject $H_{0(spice)}$. In contrast, under the alternative hypothesis, we expect A_k to rarely be greater in magnitude than A_0 and would reject $H_{0(spice)}$.

2.3 Simulations

We conduct simulation studies to evaluate the performance of our method under the null and alternative hypotheses. For the ith subject, we observe two spatial functions $X_i(\mathbf{v})$ and $Y_i(\mathbf{v})$, which have a one-to-one correspondence at all locations, indexed by $\mathbf{v}$. Subject-level bi-modal data are simulated by adding and multiplying random noise to the red ($M_1(\mathbf{v})$) and green ($M_2(\mathbf{v})$) color channels extracted from a colorful two-dimensional image, as illustrated in Figure 2. We then let $X_i(\mathbf{v}) = a_i \times a_{i1} \times M_1(\mathbf{v}) + E_{i(1)}(\mathbf{v})$ and $Y_i(\mathbf{v}) = a_i \times a_{i2} \times M_2(\mathbf{v}) + E_{i(2)}(\mathbf{v})$ for $i = 1, ..., n$.

The X and Y modalities for the same subject share a_i (distributed $N(1, \sigma_a^2)$); thus, the null hypothesis (1) is true when $\sigma_a^2 = 0$, 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 subject- and color channel-specific coefficients a_{i1} and a_{i2} are independent and both distributed as $N(1, 1)$. The $p \times q$ matrices $E_{i(1)}(\mathbf{v})$ and $E_{i(2)}(\mathbf{v})$ consist of elements that are independent and identically distributed as $N(0, \sigma_e^2)$, contributing additional noise specific to the subject, color channel, and location (i.e., pixel).

We consider sample sizes $n = 25$, 50, and 100, σ_a^2 ranging from 0 to 3 (in increments of 0.15), and σ_e^2 of either 0.5, 1.5, or 3.0. From 5000 simulations of each combination of n and σ_e^2 (when $\sigma_a^2 = 0$) we 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.

2.4 Application in the Philadelphia Neurodevelopmental Cohort

We assess the performance of our proposed method in 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 provided 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.

Our tests of intermodal correspondence will consider cortical thickness and sulcal depth—a pair of surfaces 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 the current study, we consider a subset of $n = 789$ subjects with all three image types (cortical thickness, sulcal depth, and n-back) meeting quality control criteria. This subset of 789 subjects 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.

Figure 3 shows example maps of left hemisphere sulcal depth and cortical thickness, which we consider in various tests of intermodal correspondence. In particular, group-level data (Figure 3A) are used as input for BrainSMASH and the spin test (where the medial wall is removed before analyses involving the spin test), whereas subject-level data (Figure 3B) are considered in the SPICE test. These visualizations were generated using the R package fsbrain (Schäfer 2020b).

2.4.1 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.

2.4.2 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 (Table 1) agree with one another. Given the developmental focus of the dataset, 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.

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 (2020), 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.

2.4.3 Tests of correspondence within cortical parcellation 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.

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 conservative at approximately 5% for all sample sizes and for all values of the pixel-level variance parameter, σ_e^2. Under the alternative hypothesis (when $\sigma_a^2 > 0$), lower σ_e^2, higher σ_a^2 (between-subject variance), and larger sample size correspond to higher 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 2. The SPICE method, BrainSMASH, and spin test all produce results that are consistent with what one would expect biologically: all three methods reject their respective null hypotheses when testing sulcal depth versus cortical thickness (Table 2(a)), and none of the methods indicate significant correspondence between the n-back and cortical thickness maps (Table 2(b)). For BrainSMASH, we also verify that the surrogate maps' variogram surrogate cortical thickness maps appear valid (Figure A1).

Next, we 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). Neither method provides evidence against the null hypothesis of intermodal correspondence for n-back versus cortical thickness (solid lines). 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 intermodal correspondence in the somatomotor, dorsal attention, ventral attention, limbic, frontoparietal control, and default networks. BrainSMASH similarly provides evidence against its null hypothesis ($H_{0(b\text{-}smash)}$) in tests of sulcal depth versus cortical thickness in the somatomotor (except ages 8-9 and 10-11), dorsal 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 A2, may give some insight for

understanding these discrepancies. Specifically, we observe discrepancies 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 both the SPICE test and BrainSMASH do not reject their respective null hypotheses for testing correspondence between sulcal depth and cortical thickness within the visual network. This result is consistent with those of 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).

4 Discussion

In this paper, we introduce an intuitive and powerful 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) and Burt et al. (2020), as the null hypotheses of these three methods fundamentally differ (Table 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 fully resolves five major 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) disregards the medial wall, and (iii) cannot be applied to 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 3). 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.

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.

An important extension of our work will be to incorporate nonlinear measures of intermodal correspondence. Such measures were also not considered when the spin test and BrainsSMASH were originally proposed Alexander-Bloch et al. (2018) and Burt et al. (2020), and it will be interesting to compare the three methods in settings where nonlinear associations are of interest. In addition, 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 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.

6 Acknowledgments

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.**), R01MH112847 (**T.D.S.**), 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 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.**).

7 Competing interests

The authors declare no competing interests.

8 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.

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, 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.

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.

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 and Bratislav Misic (2020). "Comparing spatially-constrained null models for parcellated brain maps". In: *BioRxiv*.

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.

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, 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.

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, 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.

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.

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.

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.

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.

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.

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: Illustration of simple permutation-based intermodal correspondence (SPICE) testing procedure. A null distribution for A_0, the 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).

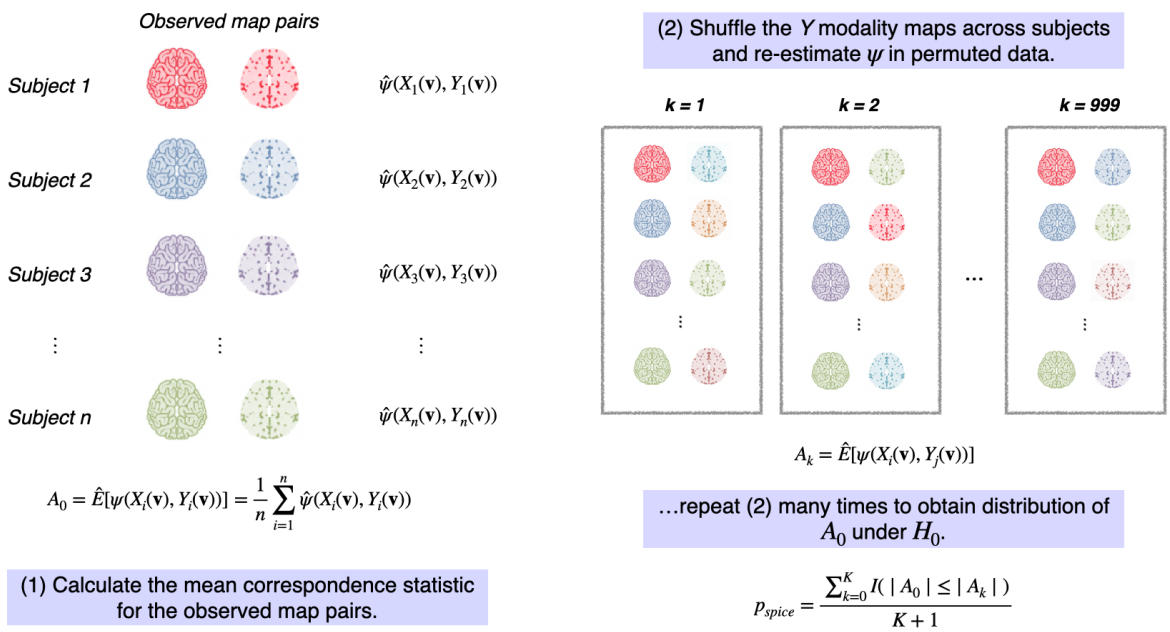

Figure 2: Illustration of red ($M_1(\mathbf{v})$) and green ($M_2(\mathbf{v})$) color channels, which are used for simulating two image modalities. Subject-level bi-modal image data are simulated as $X_i(\mathbf{v}) = a_i \times a_{i1} \times M_1(\mathbf{v}) + E_{i(1)}(\mathbf{v})$ and $Y_i(\mathbf{v}) = a_i \times a_{i2} \times M_2(\mathbf{v}) + E_{i(2)}(\mathbf{v})$, where a_i, a_{i1}, a_{i2}, and the elements of $E_{i(1)}(\mathbf{v})$ and $E_{i(2)}(\mathbf{v})$ are normally distributed, with parameters specified in Section 2.3. (Original image source: https://tinyurl.com/yy4dzaek/)

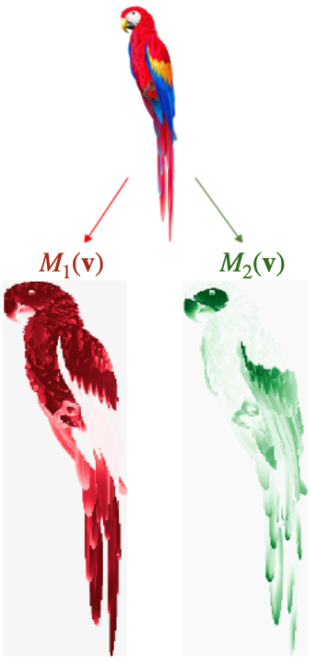

Figure 3: Example left hemisphere sulcal depth and cortical thickness maps used for (A) the spin test and BrainSMASH and (B) the SPICE test. While the spin test, BrainSMASH, and other prior methods consider group-level maps, the SPICE method leverages subject-level data. Brain map visualizations were generated using the R package fsbrain (Schäfer 2020b).

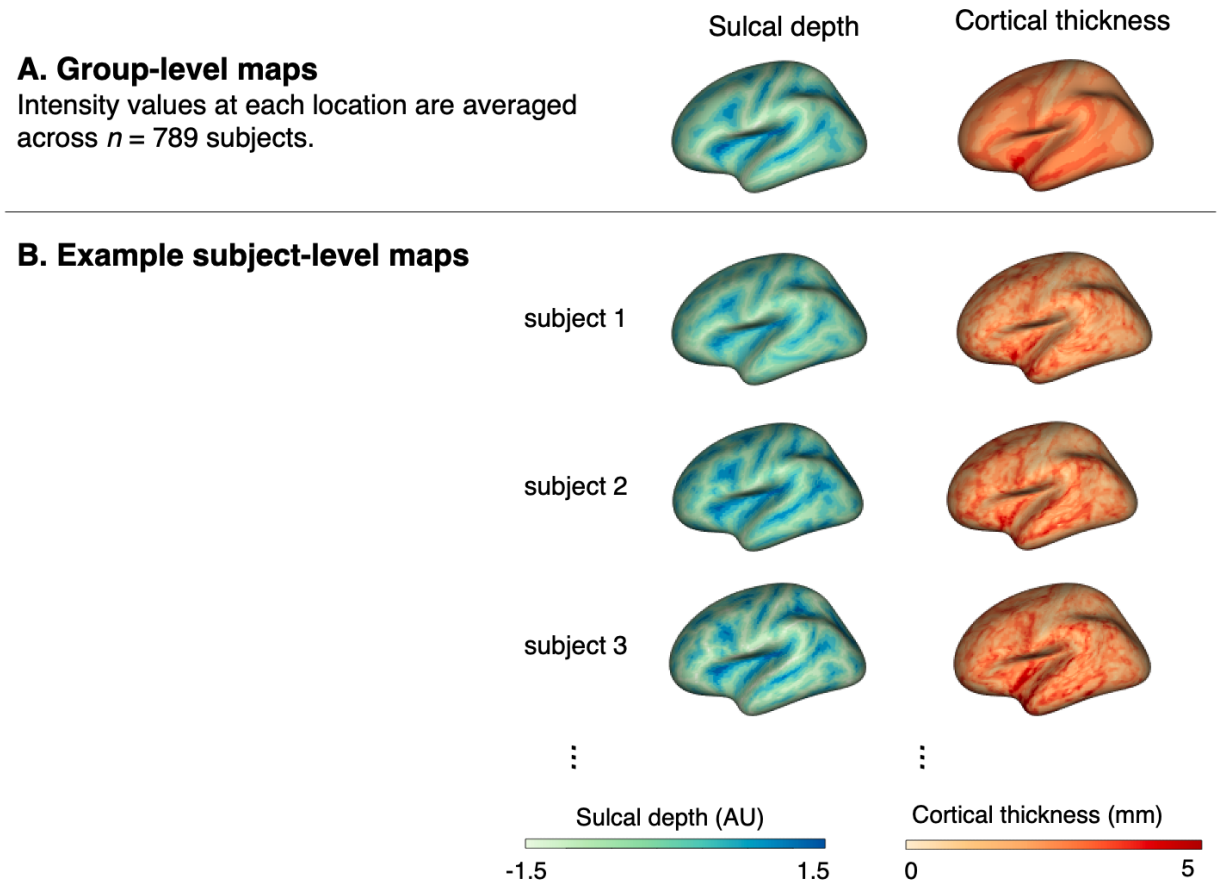

Figure 4: Power and type I error of Simple Permutation-based intermodal correspondence (SPICE) method based on 5000 simulations.

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 2) 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 pixel-level variance (σ_e^2, either 0.5, 1,5, or 3.0).

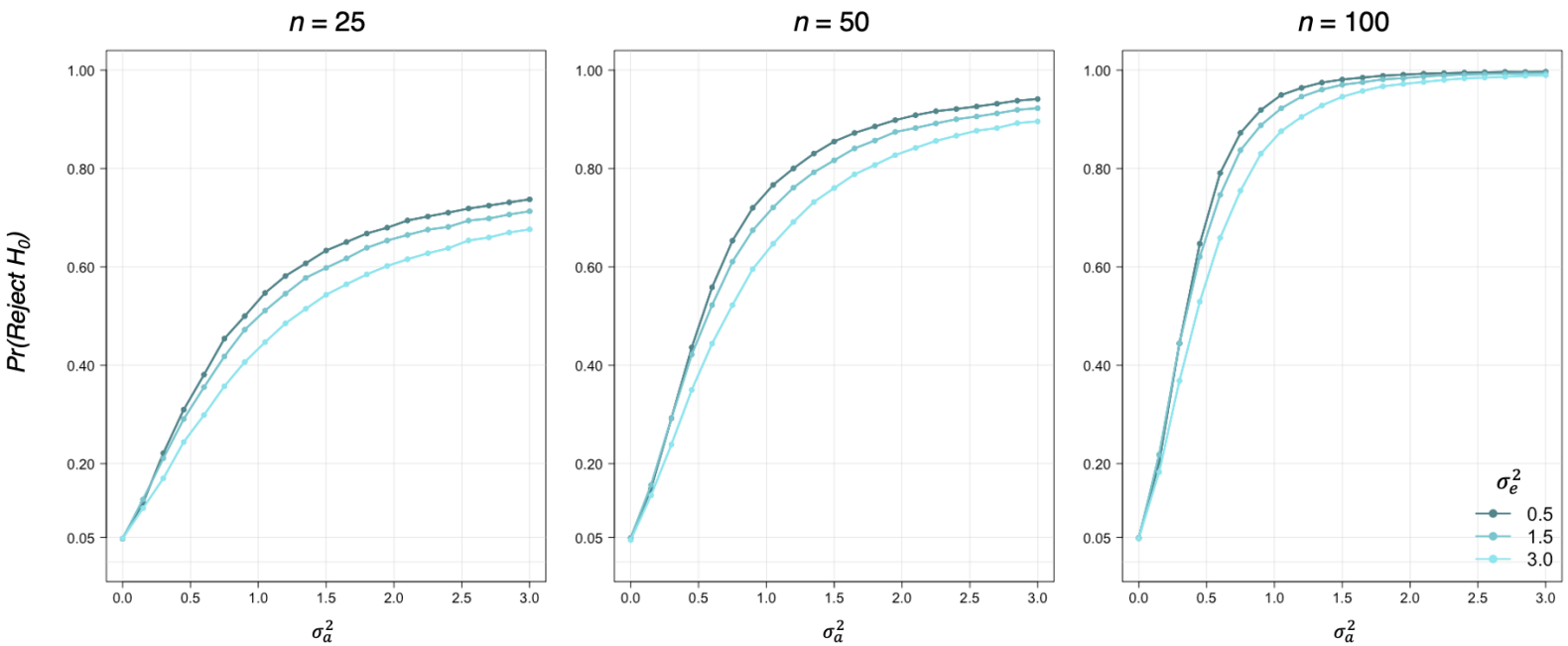

Figure 5: Unadjusted p-values from tests of intermodal correspondence within seven functional networks described by Yeo et al. 2011 in the left hemisphere for different age groups. We consider $p < 0.007$ to provide evidence against the null hypotheses (defined in Table 1), after using a Bonferroni correction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

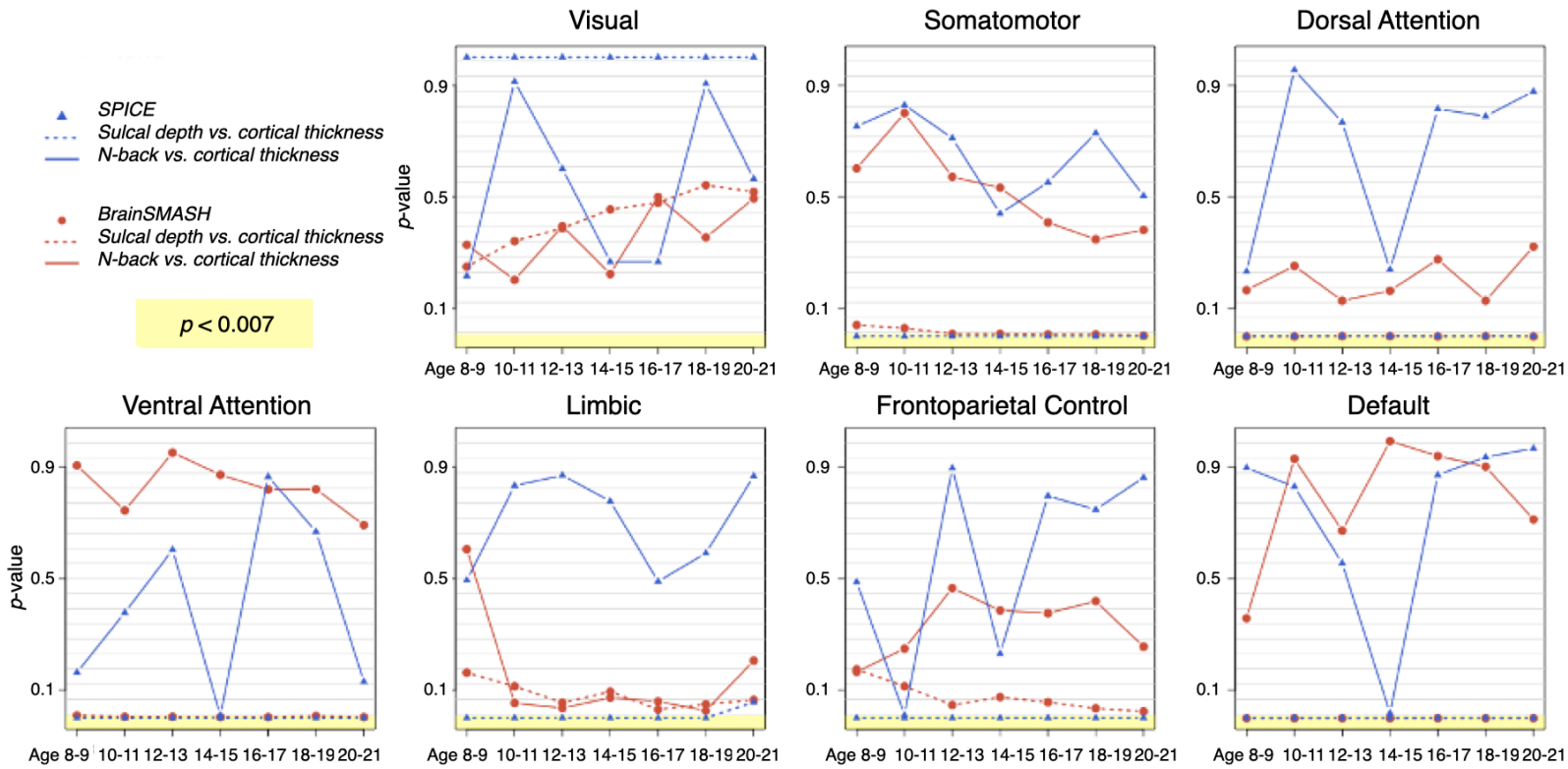

Table 1: Summary of the null hypotheses from our proposed Simple Permutation-based Intermodal Correspondence (SPICE) test, Brain Surrogate Maps with Autocorrelated Spatial Heterogeneity (BrainSMASH, Burt et al. 2020), and the spin test (Alexander-Bloch et al. 2018). The null hypotheses are inherently different, but may all be used to approach questions about intermodal correspondence.

SPICE test	$H_{0(spice)}$	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 $Y_i(\mathbf{v})$ as opposed to from $Y_j(\mathbf{v})$ $(i \neq j)$.
BrainSMASH	$H_{0(b\text{-}smash)}$	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	$H_{0(spin)}$	The observed correspondence between two group-average surface maps, $X(\mathbf{v})$ and $Y(\mathbf{v})$, provides no information about the alignment between $X(\mathbf{v})$ and $Y(\mathbf{v})$.

Table 2: 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 Table 1.

(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 $0.9 \times 0.9 \times 1$ 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 $3 \times 3 \times 3$ 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 Empirical and surrogate variograms for BrainSMASH

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.

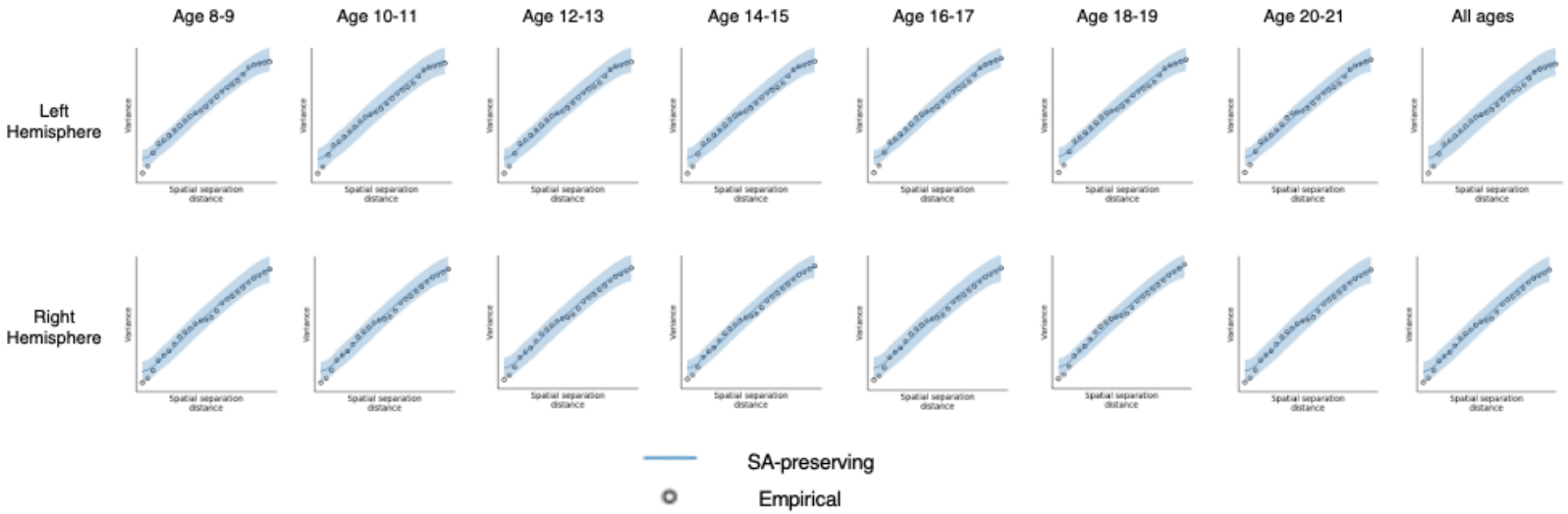
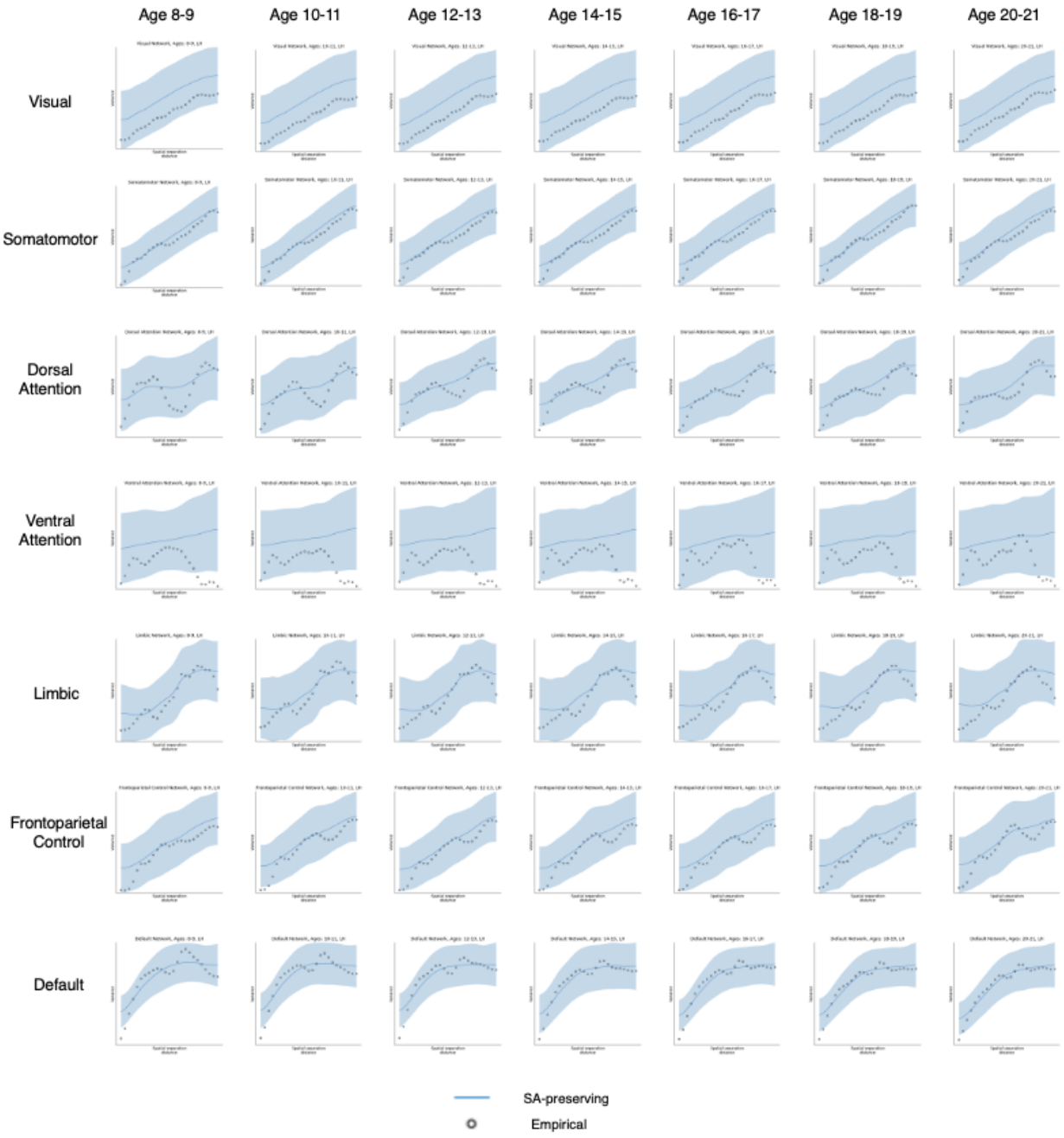
Figure A2: Empirical and surrogate (no. surrogates = 1000) variograms of parcellated cortical thickness measurements within seven functional networks, according to Yeo et al. (2011). These variograms are constructed to assess the reliability of BrainSMASH in testing for intermodal correspondence between 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.