



Does size matter? The relationship between predictive power of single-subject morphometric networks to spatial scale and edge weight

Pradeep Reddy Raamana¹ · Stephen C. Strother^{1,2} · for the Australian Imaging Biomarkers, Lifestyle flagship study of ageing, for The Alzheimer's Disease Neuroimaging Initiative

Received: 6 February 2019 / Accepted: 24 August 2020 / Published online: 18 September 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract

Network-level analysis based on anatomical, pairwise similarities (e.g., cortical thickness) has been gaining increasing attention recently. However, there has not been a systematic study of the impact of spatial scale and edge definitions on predictive performance, which is necessary to obtain a clear understanding of their relative performance. In this study, we present a histogram-based approach to construct subject-wise weighted networks that enable a principled comparison across different methods of network analysis. We design several weighted networks based on three large publicly available datasets and perform a robust evaluation of their predictive power under four levels of separability. An interesting insight generated is that changes in nodal size (spatial scale) have no significant impact on predictive power among the three classification experiments and two disease cohorts studied, i.e., mild cognitive impairment and Alzheimer's disease from ADNI, and Autism from the ABIDE dataset. We also release an open source python package called *graynet* to enable others to leverage the novel network feature extraction algorithms presented here. These techniques and toolbox can also be applied to other modalities due to their domain- and feature-agnostic nature) in diverse applications of connectivity research. In addition, the findings from the ADNI dataset are replicated in the AIBL dataset using an open source machine learning tool called *neuropredict*.

Keywords Cortical thickness · Graph theory · Early prognosis · Mild cognitive impairment · Alzheimer · Model comparison · Histogram distance · Magnetic resonance imaging

Data used in the preparation of this article was obtained from: (1) Alzheimer Disease Neuroimaging Initiative (ADNI) and (2) the Australian Imaging Biomarkers and Lifestyle flagship study of ageing (AIBL) funded by the Commonwealth Scientific and Industrial Research Organization (CSIRO) which was made available at the ADNI database (www.loni.usc.edu/ADNI). The ADNI and AIBL researchers contributed data but did not participate in analysis or writing of this report.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s00429-020-02136-0>) contains supplementary material, which is available to authorized users.

✉ Pradeep Reddy Raamana
praamana@research.baycrest.org

¹ Rotman Research Institute, Baycrest Health Sciences, Toronto, ON, Canada

² Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada

Introduction

Network-level analyses have become one of the dominant techniques to process and analyze different neuroimaging modalities, including functional MRI (task- and resting-state fMRI), and diffusion MRI. One of the most routine network analyses performed is the extraction of individual connectivity matrices e.g. to characterize the structure and function of the brain, as well as to develop markers of dysfunction in various brain disorders. Owing to their broad applicability and success, similar approaches have been developed in the structural MRI (sMRI) also (Raamana et al. 2015). Translation of such powerful techniques to the sMRI, and a systematic evaluation of their methodological robustness, would help assess clinical utility, esp. in the development of computer-aided diagnostic (CAD) techniques for deadly brain disorders like the Alzheimer's disease (AD) (Alzheimer's Association 2017).

Although there has been great progress in the last few decades in accurately characterizing AD as well as its progression (Weiner et al. 2015, 2017), its translation to improvement of clinical trials continues to be a great challenge (Cummings et al. 2014). For any preventive or disease-modifying therapies to succeed, early prognosis is key. Towards this goal, diverse regional and network-level analyses of features derived from different neuroimaging modalities such as sMRI (Cuingnet et al. 2011; Bron et al. 2015; Duchesne et al. 2008; Raamana et al. 2014a, b; Dyrba et al. 2015), positron emission tomography (PET) (Dukart et al. 2011; Herholz et al. 2002; Matthews et al. 2016) and resting-state fMRI (Hojjati et al. 2017; Abraham et al. 2017) have been developed and are showing great promise in identifying differences between health and disease in the early stages, as well as establishing how they correlate with cognitive measures (Alexander-Bloch et al. 2013; Tijms et al. 2013). Multimodal predictive modeling methods typically demonstrate higher prognostic accuracy (Sui et al. 2011; Arbabshirani et al. 2017) in many applications, owing to their training based on multiple sets of rich and complementary information related to disease. However, recent efforts in building more sophisticated machine learning strategies produced unimodal sMRI methods rivaling the state-of-the-art multimodal approaches (Weiner et al. 2017). Although multi-modal approaches tend to be more sensitive in general and offer richer insight, the practical advantages of sMRI being non-invasive, cost-effective and widely-accessible in the clinic, make sMRI-based CAD methods for early prognosis highly desirable.

Cortical thickness is a sensitive imaging biomarker that can be easily derived from sMRI to diagnose AD. However, its sensitivity to identify the prodromal subjects (such as mild cognitive impairment (MCI)) at risk of progressing to AD is limited (Cuingnet et al. 2011). Network-level analysis of cortical thickness and gray matter features demonstrated its potential to provide novel insights or improve predictive power (Raamana et al. 2015), and is gaining in popularity (Evans 2013; Wen et al. 2011; Reid and Evans 2013; Lerch et al. 2006). Thickness network features offer complementary information compared to the underlying fiber density (Gong et al. 2012), are shown to be disrupted in AD (Kim et al. 2016) and have been shown to have potential for early prognosis of AD (Raamana et al. 2015; Wee et al. 2012; Dai et al. 2012; Kim et al. 2016), as well as for subtype discrimination (Raamana et al. 2014a, b), outperforming the non-network raw-thickness features.

Network analysis studies in cortical thickness range from.

1. group-wise studies building networks based on group-wise covariance/correlation in cortical thickness (Evans 2013; He and Chen 2007; Lerch et al. 2006), which may be used to characterize the properties of these networks

(such as small-worldness) as well as provide useful insight into network-level changes between two diagnostic groups e.g. healthy controls (CN) and Alzheimer's disease (AD),

2. studies building individual subject-wise graphs based on within-subject ROI-wise (pairwise) similarity metrics (Raamana et al. 2015; Tijms et al. 2012; Wee et al. 2012; Dai et al. 2012; Kim et al. 2016) to enable predictive modeling. These studies resulted in disease-related insights into network-level imaging biomarkers and improved accuracy for the early prognosis of AD. However, these studies are based on distinctly different parcellation schemes of the cortex, vastly different ways of linking two different regions in the brain, and datasets differing in size and demographics.

Insights derived from various brain network studies showed considerable variability in reported group differences (Tijms et al. 2013), and widely accepted standards for network construction are yet to be established (Stam 2014). There have been recent efforts into understanding the importance and impact of graph creation methods, sample sizes and density (van Wijk et al. 2010; Phillips et al. 2015). However, these studies have been restricted to the choice of group-wise correlation methods to define the edges, or limited to understanding the group-wise differences in selected graph measures. But such important methodological analyses have not been performed in the context of building individual subject-wise predictive modeling. Hence, there is no clear understanding of the impact of different choices in subject-wise network construction and their relative predictive performance.

Given their potential for the development of accurate early prognosis methods (Raamana et al. 2015; Raamana et al. 2014a, b) demonstrated by outperforming non-network raw-thickness features, and the wide-accessibility of sMRI, thickness-based networks deserve a systematic study in terms of

1. how does the choice of edge weight or linking criterion (correlation He and Chen 2007), similarity Raamana et al. 2015) affect the performance of the predictive models? See Table 4 for more details.
2. how does the scale of parcellation (size and number of cortical ROIs) affect the predictive performance?

These questions, analyzed in our systematic study, can reveal important tradeoffs of this emerging theme of research. In this study, we present a methodological comparison of six different ways of constructing thickness-based, subject-wise networks and present classification results under varying levels of separability. We start with the classic CAD problems i.e. discriminating AD from CN,

and mild cognitive impairment (MCI) subjects (converting to AD in ~18 months) from CN in the ADNI dataset. In order to test whether the results from this methodological study generalize to different datasets, diseases and separabilities, we also study the Australian Imaging, Biomarker & Lifestyle Flagship Study of Aging (AIBL) and the Autism Brain Imaging Data Exchange (ABIDE) datasets. Based on these three large publicly available datasets, we show that the predictive power of single-subject morphometric networks, based on cortical thickness features, is insensitive to spatial scale or edge weight. This is an important finding given we were not only able to replicate these results on an independent dataset, but also replicate them in the presence of a different disease and in a different age group.

Methods

In this section, we describe the datasets we study in detail, along with a detailed description of the preprocessing and the associated methods.

ADNI dataset

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

We downloaded baseline T1 MRI scans ($n=671$) from the ADNI dataset (Jack et al. 2008), which has quality-controlled Freesurfer parcellation (version 4.3) of the cortical surfaces provided in the ADNI portal (Fischl and Dale 2000; Fischl et al. 2002). The parcellation and cortical thickness

values downloaded were carefully visually inspected by the first author PRR for errors in geometry and range. This QC process was rigorous to include a large number of cross-section slices with contours of pial and white surfaces overlaid on the sMRI image in all 3 views. We have also employed external surface views that facilitate easy inspection and identification of any topological defects as well as anatomical accuracy of the Freesurfer labels as a whole. When noticeable errors were found, we eliminated those ($n=24$) subjects, and no manual editing and corrections were performed. The thickness features from the remaining subjects for the control (CN) and AD groups (effective $n=647$) comprised the first set of subjects for our analysis in this study. The second set of subjects with a slightly lower level of separability (MCI subjects converting to AD in 18 months, denoted by MC1c) were chosen to match the benchmarking study (Cuingnet et al. 2011) as closely as possible (to enable comparison to the many methods included) based on the availability of their FS parcellation from ADNI and our quality control results. The demographics for the two sets are listed in Table 1.

AIBL dataset

In order to study whether the results from the ADNI dataset in this study generalize to another independent dataset, we have downloaded the Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing (AIBL) dataset (Ellis et al. 2009), which contained similar (but not identical) patient groups and diagnostic categories. The downloaded subjects were processed with Freesurfer v6.0. The number of Alzheimer's subjects we could download from AIBL (denoted by AD2) at baseline were $n=64$, and we randomly selected 100 healthy controls (CN4) for this study, whose subject IDs are shared in the Appendix. The resulting cortical parcellations were visually quality controlled by PRR with VisualQC (v0.4.1) (Raamana 2018; Raamana and Strother 2018b). This QC process was rigorous to include a large number of cross-sectional slices

Table 1 ADNI I Demographics

Diagnostic group	<i>N</i>	Females	Age	MMSE ^a
Dataset 1: ADNI				
Healthy controls (CN1)	224	109	75.79 (4.99)	29.11 (1.01)
Alzheimer's disease (AD)	188	89	75.22 (7.49)	23.29 (2.04)
Dataset 2: ADNI				
Healthy controls (CN2, for MCI) ^b	159	85	76.07 (5.33)	29.17 (0.98)
MCI converters to AD in 18 months (MC1c)	76	33	74.67 (7.35)	26.47 (1.86)

All statistics here are displayed in mean (SD) format.

^aADNI: Only MMSE is significantly different between CN1 and AD with $p < 0.05$.

^bADNI: Controls and MCI converters are chosen to match the benchmark samples presented in (Cuingnet et al. 2011) as closely as possible allowing for exclusions due to quality control.

with contours of pial and white surfaces overlaid on the sMRI image in all 3 views, with the ability to zoom in to the voxel-level to ensure anatomical accuracy of the pial and white surfaces. In addition, the VisualQC interface presents 6 views of the external surface view of pial surface which facilitates easy inspection and identification of any topological defects as well as anatomical accuracy of the Freesurfer labels as a whole. This QC process was employed to remove subjects with inaccurate parcellations or any other errors that render them unusable for analyses in this study. We would like to note that VisualQC is the most comprehensive QC tool for Freesurfer parcellations, and hence this process may be sensitive to catching the parcellation errors compared to that on the non-interactive tool employed on the ADNI dataset. This resulted in a usable subset of 51 AD2 and 80 CN3 subjects. Demographics of the subjects analyzed are presented in Table 2.

Motivated by intention to improve reproducibility and maximize the value of this study by employing open source tools when possible, we have employed graynet (Raamana and Strother 2017, 2018) to compute the network-level features, and neuropredict (Raamana 2017) to evaluate their predictive utility. While this change of software libraries would add another interesting level of robustness check for the results presented here, we must note that this change in technology stack may lead to some differences in numerical estimates e.g. in AUC estimates when comparing across different datasets e.g. ADNI vs. AIBL. However, given the technology employed is the same for a given dataset, the performance estimates within the dataset will be perfectly commensurable for posthoc statistical analyses.

Table 2 AIBL demographics

Diagnostic group	<i>N</i>	Females	Age	MMSE
Healthy controls (CN4)	80	44	73 (6.99)	29 (1.26)
Alzheimer's disease (AD2)	51	32	73 (7.37)	22 (5.4)

All statistics here are displayed in median (SD) format.

Table 3 ABIDE I demographics

Dataset 3: ABIDE						
Diagnostic Group	<i>N</i>	Females	Age	FIQ*	PIQ	VIQ*
Healthy controls (CN3)	100	17	17.27 (7.68)	109.10 (12.35)	105.64 (12.74)	111.89 (13.52)
Autism (AUT)	100	9	15.82 (5.93)	103.49 (14.68)	104.57 (14.68)	101.36 (15.86)

*FIQ and VIQ are significantly different between CN3 and AUT with $p < 0.05$

FIQ Full IQ standard score, VIQ Verbal IQ standard score, PIQ Performance IQ standard score

ABIDE dataset

In order to study whether the conclusions drawn from the ADNI dataset generalize to a very different disease cohort, we obtained the Freesurfer parcellations (version 5.1) from the Autism Brain Imaging Data Exchange (ABIDE) preprocessed dataset made available freely on the ABIDE website (Craddock et al. 2013). A random subset of cortical parcellations ($n = 227$) have been visually inspected by PRR for errors in geometry estimation and value ranges (using the same in-house tools and processed used on the ADNI dataset) to eliminate any subjects showing even a mild chance of failure. From the passing subjects ($n = 226$), we randomly selected 200 subjects (100 samples per diagnostic group) whose demographics are presented in Table 3 and the subject IDs are listed in the Appendix. Owing to the random selection, they come from multiple sites, which is akin to the ADNI dataset used in this study. Previous research (Abraham et al. 2017) showed that the site heterogeneity has little or no impact on the predictive accuracy of network-level features derived from task-free fMRI data. The distribution of the sites represented in this study are shown in Appendix D.

Feature extraction

In the following sections, we describe the steps involved in the extraction of weighted networks based on T1 MRI scans of the different subjects in the two independent datasets.

Alignment and dimensionality reduction

Cortical thickness features studied here were obtained from the Freesurfer parcellations (gray and white matter surfaces). They were then resampled to the *fsaverage* atlas and smoothed at $fwhm = 10$ mm to reduce the impact of noise. This is achieved by Freesurfer `-qcache` processing option, which registers each of the subjects to the *fsaverage* atlas (provided with Freesurfer) to establish vertex-wise correspondence across all the subjects.

Cortical subdivision

In order to avoid the curse of dimensionality and to reduce the computational burden, the atlas has been subdivided using a surface-based, patch-wise parcellation technique originally presented in (Raamana et al. 2015). This technique is based on Freesurfer parcellation which consists of 34 ROIs per hemisphere, which vary in size (number of vertices) greatly. In order to obtain uniform sized patches, we subdivide each of these ROIs into smaller patches, while respecting the anatomical boundaries of each ROI. Here, we use an adaptive version wherein the patch-size is controlled by number of vertices (denoted by m = vertices/patch), instead of choosing a globally fixed number of patches (say 10) per Freesurfer APARC label regardless of its size (which can vary widely resulting in vastly different patch sizes within the same subject). As we change m , the subdivision of the cortical labels is performed purely on the existing mesh, and neither the geometrical parcellation itself nor the vertex density are modified. Here, m can be taken as the size of the graph node (imagine the node as a small patch within different Freesurfer labels). Alternatively, m can be seen as the spatial scale of the graph analysis, whose impact is being assessed for different values of m . When m is small (say 100), this results in large number (273) of total patches (sum of number of patches for each aparc label) across the whole cortex, whereas it results in only 68 patches when it is very high ($m = 10,000$), as such a large patch covered the full extent of all the 68 Freesurfer APARC labels currently defined on fsaverage cortical parcellation. We have analyzed the following values of $m = 1000, 2000, 3000, 5000,$ and $10,000$, which resulted in the following total number of non-overlapping patches in the whole cortex: 273, 136, 97, 74 and 68, respectively.

Network computation

Construction of thickness networks in their early form were based on group-wise correlations (He and Chen 2007). Our previous publications based on cortical thickness (Raamana et al. 2015; Raamana et al. 2014a, b) and other interesting studies on gray matter density (Tijms et al. 2012; Wee et al. 2012) extend the earlier approaches to individualized subject-wise network extraction methods. Many of these previous studies relied on summarizing the thickness distribution in a given ROI (e.g. using mean within the entire Freesurfer label as in Tijms et al. 2012) or within a patch [Freesurfer label subdivided further as in Raamana et al. (2015)], before constructing the networks. Although such approaches reduce the dimensionality and provide us with smooth features, they do not utilize the rich description and variance of the distribution of features. Moreover, studies thus far computed characteristic features from a binary network (by applying an optimized threshold Raamana et al. 2015) or using a vector representation of weighted graphs (vector of distances in the upper triangular part of the edge weight matrix, as they are symmetric Tijms et al. 2012). Here, to enable a principled comparison across the different edge weights (and to avoid the optimization of an arbitrary threshold required to binarize the edge weight matrix), we study weighted-networks only, whose derivation is described below.

Histogram WEighted NETworks (HiWeNet)

In this section, we describe the method employed in constructing the HIstogram WEighted NETworks (HiWeNet) based on cortical thickness. First, to improve the robustness of the features, 5% outliers from both tails of the distribution of cortical thickness values are discarded from each patch at a given scale m (see Appendix for more information). The residual distribution is converted into a histogram by

Table 4 Variety of edge weights analyzed in this study

Type of base representation	Type of edge weight metric	Acronym	Mathematical definition
Summarized (median/mean in a patch)	Similarity (Raamana et al. 2015)	MD	$ M_i - M_j $
	exp(similarity)	EMD	$e^{-\frac{(\mu_i - \mu_j)^2}{2(\sigma_i + \sigma_j)}}$
Raw distribution	Wilcoxon ranksum statistic	RS	Ranksum test statistic
Normalized histogram	Histogram correlation	HCOR	$\rho(h_i, h_j)$
	χ^2 statistic	CHI2	$2 \sum_{k=1}^N \frac{h_i(k) - h_j(k)}{h_i(k) + h_j(k)}^2$
	Histogram intersection	HINT	$\frac{\sum_{k=1}^N \min(h_i(k), h_j(k))}{\sum_{k=1}^N h_i(k)}$

This table presents the list of edge weights compared in this study. Notation: each patch is indexed by i or j . For patches i and j , M , μ and σ are the median, mean and standard deviation of the within-patch distribution of vertex-wise thickness values; h_i is the normalized histogram of a given distribution. N is the number of bins in the histogram, which is fixed at $N = 100$ bins.; ρ is the Pearson correlation coefficient between two vectors of equal length.

binning into uniformly spaced $n = 100$ bins. Then the histogram counts are normalized using

$$h_i(k) = \frac{h_i(k)}{\sum_{k=1}^N h_i(k)}$$

for $k = 1 : N$, where h_i is the histogram of patch i . This method (illustrated further in Fig. 1) enables the computation of the pairwise edge-weight (distance between the histograms, denoted by EW) for the two patches i and j irrespective of the number of vertices in the two patches. A variety of histogram distances as listed in Table 4 are studied in this paper to analyze their impact on predictive power.

To analyze the relative benefit of HiWeNet, we compare the histogram-based methods to three commonly used inter-nodal weights based on descriptive summary statistics (denoted as MD, EMD and RS in Table 4). Once the edge weight matrix is computed (which is symmetric), we extract the upper-triangular part of the matrix and vectorize it (of length $n*(n-1)/2$, where n is the number of patches on the cortex for a given number of vertices/patch m). The vectorized array of edge weights (VEW) forms the input to the classifier. Each element of VEW corresponds to a unique edge in the connectivity matrix of pairwise edges. In addition, in Appendix C, we present and analyze the performance of an alternative network-representation method, based on grand mean histogram.

Note on test–retest reliability

The reliability of this network approach developed in *HiWeNet* (pairwise distances between ROIs) boils down to the reliability of the method to measure cortical thickness at the vertex-level, as the remaining parts of the algorithm

are deterministic. Several studies have previously shown that cortical thickness estimation (and Freesurfer as a tool) have high test–retest reliability (Han et al. 2006; Iscan et al. 2015) and that the brain-behavior relationships e.g. between cortical thickness and cognitive performance are stable across different sessions, scanner platforms and field strengths (Dickerson et al. 2008). In addition, given our choice of employing distance between thickness distributions over relatively large patches (1000 vertices or more), small changes in thickness (e.g. 0.2 mm) would be absorbed into the distance calculations, and hence are unlikely to change the results presented herein.

Comparison of predictive utility

In this section, we describe the procedure and techniques used to evaluate and compare the predictive power of multiple variations of the network-level features. Thanks to the relatively large sample sizes, particularly for ADNI and ABIDE, we could employ a repeated nested split-half cross-validation (CV) scheme, with 50% reserved for training, to maximize the sizes of training and test sets. Moreover, in each iteration of CV, all the methods are trained and assessed on the *exact* same training and test sets, to “pair” the performance estimates. This technique is shown to produce reliable and stable estimates of differences in predictive performance across different methods (Dietterich 1998; Burman 1989; Demšar 2006), instead of pooling multiple sets of performance distributions estimated separately on different training and test sets for each method independently. This setup allows us to compare large numbers of methods and their variants simultaneously within each dataset.

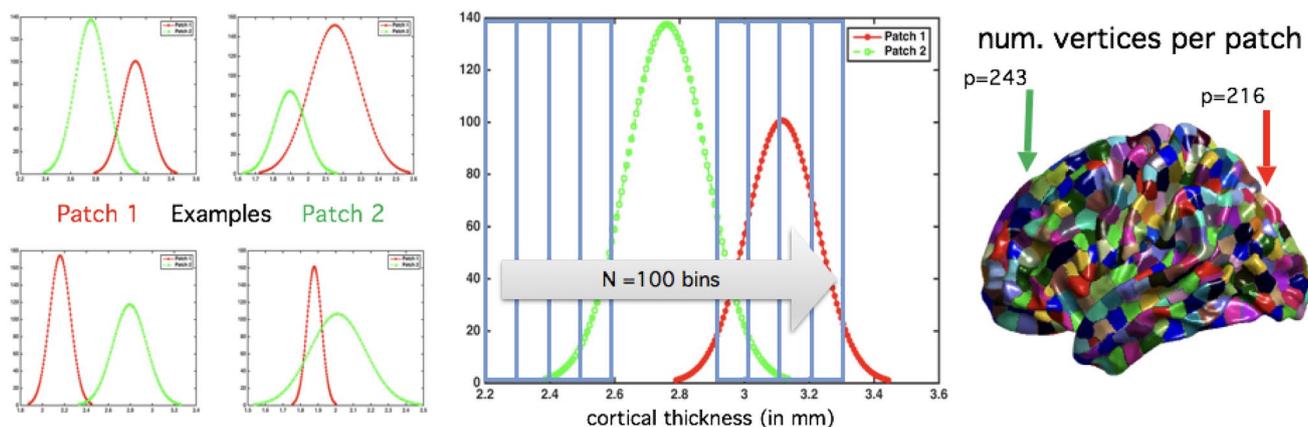


Fig. 1 Construction of histogram-distance weighted networks (HiWeNet) based on cortical thickness features using edge-weight calculations (applicable to HCOR, CHI2 and HINT metrics in Table 4). The four smaller subpanels on the left show typical distributions of cortical thickness values for four random pairs of patches (in

green and red) in a given subject (shown on cortical visualization on right). They demonstrate the means and shape of these distributions can vary substantially as you traverse across different pairs of cortical patches. The large panel in the middle illustrates the type of binning used to construct the histogram from each patch

Cross-validation scheme

The comparison scheme employed is comprised of the following steps—for a schematic, see Fig. 7 in Raamana et al. (2015):

1. Repeated split-half cross-validation scheme, with class-sizes stratified in the training set (RHsT) (Raamana et al. 2015), to minimize class-imbalance. This scheme is repeated $N=200$ times, to obtain the N paired estimates of classification performance.
2. In each CV run,
 - a. feature selection (from vectorized array of edge weights, VEW) on one split (training set of size N_{train}) is performed based on t-statistic based ranking (based on group-wise differences in the training set only), selecting only the top $N_{\text{train}}/10$ elements. The frequency of selection of a particular element (which is an edge in the cortical space) over different CV trials by the t-statistic ranking is an indication of its discriminative utility, and will be visualized to obtain better insight into the process.
 - b. Support vector machine (SVM) is chosen as the classifier to discriminate the two groups in each experiment. SVM is optimized in an additional inner split-half CV applied to the training set via a grid search. We have employed the following ranges of values in the grid search for the margin control parameter $C=10^p$; $p=-3:5$ and the kernel bandwidth $=2^q$; $q=-5:4$.
 - c. The optimized SVM is applied on the second split (test) to evaluate its performance.
3. The process in Step 2 is repeated $N=200$ times (Varoquaux et al. 2016; Raamana et al. 2015) to obtain 200 independent estimates for each method being compared.
4. In this study, we measure the performance by area under the predictive receiver operating characteristic (ROC) curve (denoted by AUC), whose distributions for different methods are shown in Fig. 6.

The results in this study were produced using Raamana's programming library implemented in Matlab based on the built-in statistics and machine learning toolbox.

Open source software

Most of the computational code applied on the ADNI and ABIDE datasets had been implemented in Matlab. In order to enable other researchers to utilize the methods presented here easily without having to pay for expensive Matlab

licenses, we have re-implemented them in python following the best practices of open science. Moreover, we have processed the AIBL dataset using the open source alternatives, and showed that our results replicate on an independent dataset, despite differences in the following 3 layers of software.

Quality control via VisualQC

To achieve higher rigor as well as ease of use, we have developed an interactive version of Freesurfer QC tool, which is available as part of the VisualQC package at github.com/raamana/VisualQC. This tool, applied on AIBL dataset, is more sensitive in detecting parcellation errors compared to the in-house Matlab tools and other existing protocols applied on the ADNI and ABIDE datasets (Raamana et al. 2020).

Feature extraction via graynet

The core HiWeNet algorithm has been implemented in Python and is publicly available at this URL: <https://github.com/raamana/hiwenet> (Raamana and Strother 2017). We have also published the original Matlab code for the computation of adjacency matrices used for this study, within the *hiwenet* package.

Further, to make this research even more accessible, we have implemented the entire workflow of morphometric network extraction as a seamless pipeline called *graynet*, implemented entirely in Python (Raamana and Strother 2018). Using this tool would enable those without much software engineering experience to simply run Freesurfer and then run *graynet* to get started with morphometric network analyses. This frees them from the hassle of assembling complicated data, implementing graph theoretical operations and managing the pipeline following the best practices, which can be a barrier to many laboratories with limited computational and software expertise. In addition, we employed this tool to process the AIBL dataset, and show that patterns in performance comparison across different weights are retained compared to those of the original Matlab toolbox.

Evaluating predictive utility via neuropredict

In order to enable a much wider audience (those without access to a Matlab license or its expensive statistical learning toolboxes (each to, or those who do not have the necessary programming skills or machine learning expertise) utilize a comprehensive performance evaluation tool, we have also built an open source tool called *neuropredict* (Raamana 2017) at github.com/raamana/neuropredict. Once the researchers run Freesurfer successfully, they can run *graynet* (Raamana and Strother 2018), which produces the necessary single-subject morphometric networks. The

outputs from graynet in turn serve as direct input to neuropredict, which runs the cross-validation scheme described in the above section to produce a comprehensive report on their predictive power. In addition, we employed this tool to evaluate the performance of network features from the AIBL dataset (CN4 vs. AD2), matching the techniques and specific optimizations to the extent possible. This showed that our findings replicated compared to that of the original Matlab toolbox, which validates neuropredict as a useful open source alternative.

Results and discussion

Within-group networks

To obtain better insight into the topology of the networks defined above, it is helpful to visualize seed-based networks and analyze their connections. A common approach to this end involves picking the posterior cingulate gyrus (core hub of the default mode network, DMN) as the seed and analyzing its connections in healthy controls, and esp. how they change for different edge metrics. The seed-based network visualizations are produced for each edge weight method separately for $m = 2000$, identical to the network construction method described in the Methods section: compute histogram-distance between the thickness distribution of the seed and all the other ROIs, averaging this edge weight across all the healthy subjects, and retaining only the strongest edges (top 5%).

To make the comparison across the three datasets easy, they are grouped for each metric e.g. for median difference (MD), the comparison is shown below for healthy controls. From this figure, we can clearly see a pattern resembling the

default mode network, in healthy controls from all the three samples. This is consistent with the results reported in previous structural covariance studies (Spreng and Turner 2013; Evans 2013; Spreng et al. 2013; Power et al. 2011) (Fig. 2).

To get a sense of how these networks change with different EW metric, we show two other networks corresponding to HCOR and CHI2 metrics below (each figure is labeled with the metric and summary statistic being displayed e.g. HCOR mean). This HCOR network loses resemblance to the DMN (e.g. loss of edges to superior frontal, banks of the superior temporal sulcus, frontal pole, fusiform), and the edge weight distribution varies widely across the three samples. However, the CHI2 network resembles the DMN pattern seen in MD network well, suggesting the similarity of the two networks (Figs. 3, 4).

Group-wise differences

To illustrate the differences between the proposed methods of computing edge weights, we compute the distributions of vertex-wise mean thickness values for CN1 and AD separately. We then visualize them in the form of a matrix of pairwise edge weights at $m = 2000$, as shown in Fig. 5a,b. Each row (say node i) in a given edge-weight matrix (from one group say CN1 in Fig. 5a) here refers to the pairwise edge weights w.r.t remaining nodes $j, j = 1:N$. As the differences are subtle and spatially distributed, for easy comparison between the two classes, we visualize the arithmetic differences between the two classes in Fig. 5c.

The visualizations in Fig. 5c offer useful insight into the group-wise differences between CN1 and AD, and across different edge weight distances. However, visual differences do not imply differences in predictive power of features extracted these networks of weights. Hence, it is important

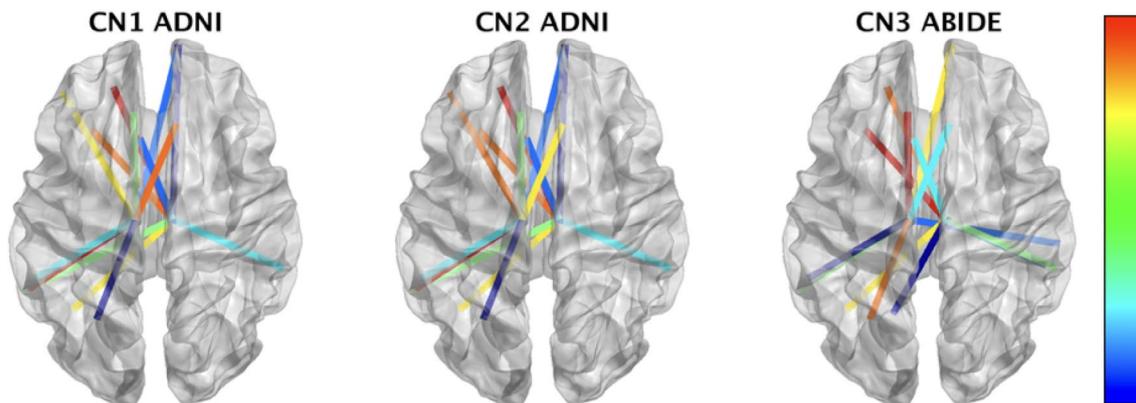


Fig. 2 Seed-based connectivity networks for the MD metric ($m = 2000$), showing average weights across each healthy control sample from the three datasets (as labeled). The colors on the edges represent the edge weight using a jet colormap (with blues indicating

the weaker and reds indicating stronger weights). From this figure, we can clearly see a pattern resembling the default mode network, in healthy controls from all the three samples

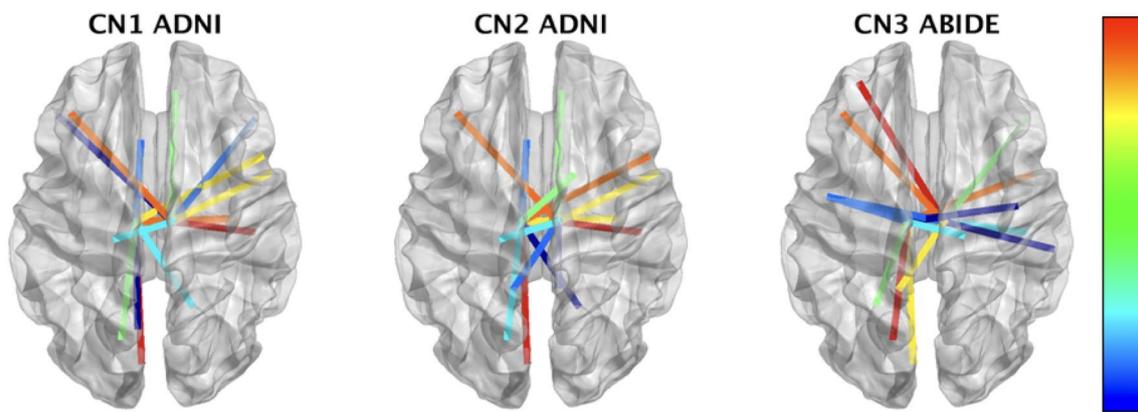


Fig. 3 Network showing edge weights (mean across samples) derived via HCOR metric. Layout of the figure is the same as above for the MD network

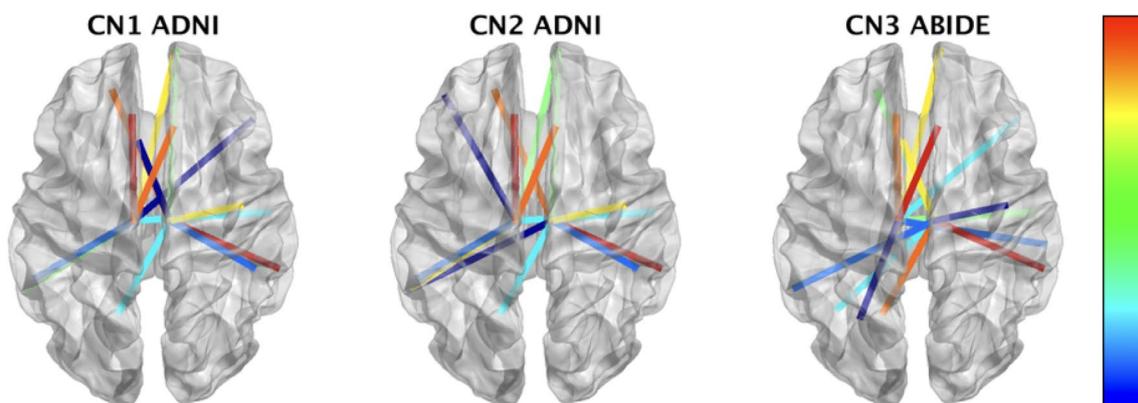


Fig. 4 Network showing edge weights (mean across samples) derived via CHI2 metric. Layout of the figure is the same as above for MD

to assess their predictive utility in discriminating AD from CN1.

Predictive utility

The RHsT cross-validation scheme is employed for each of the three classification experiments from two independent datasets i.e. CN1 vs. AD, CN2 vs. MCIc and CN3 vs. AUT. The performance distributions for the different combinations are shown in Fig. 6.

Focusing on the top panel (CN1 vs. AD), there are numerical differences in performance among different methods at fixed scale (m). However, the pattern remains similar across different spatial scales. The MD, EMD, CHI2 and HINT methods are consistently outperforming, numerically speaking, the RS and HCOR methods across different values of m . Broadly speaking, the patterns of change in AUC in Fig. 6 within each panel as we move from left to right (going over different combinations) are quite similar to the rest, although at a different median baseline (at AUC=0.87 for CN1 vs.

AD, at AUC=0.75 for CN2 vs. MCIc and at AUC=0.6 for CN3 vs. AUT).

Statistical significance testing

In order to assess the statistical significance of differences among this large set of methods, we performed a nonparametric Friedman test (Dietterich 1998) comparing the performance of the 30 different classifiers (6 methods at 5 spatial scales) simultaneously, for each of the three experiments separately. The results from post-hoc Nemenyi test (Demšar 2006) are visualized in a convenient critical difference (CD) diagram (Kourentzes 2016) as shown in Fig. 7.

The left panel in Fig. 7 shows that only the top 6 methods (with median ranks from 6.96 to 11.35) are statistically significantly different from the lowest-ranked methods, at $\alpha=0.05$, correcting for multiple comparisons. The remaining 24 methods, when compared together simultaneously, are not significantly different from each other. Similarly, the top-ranked 6 methods are not statistically significantly

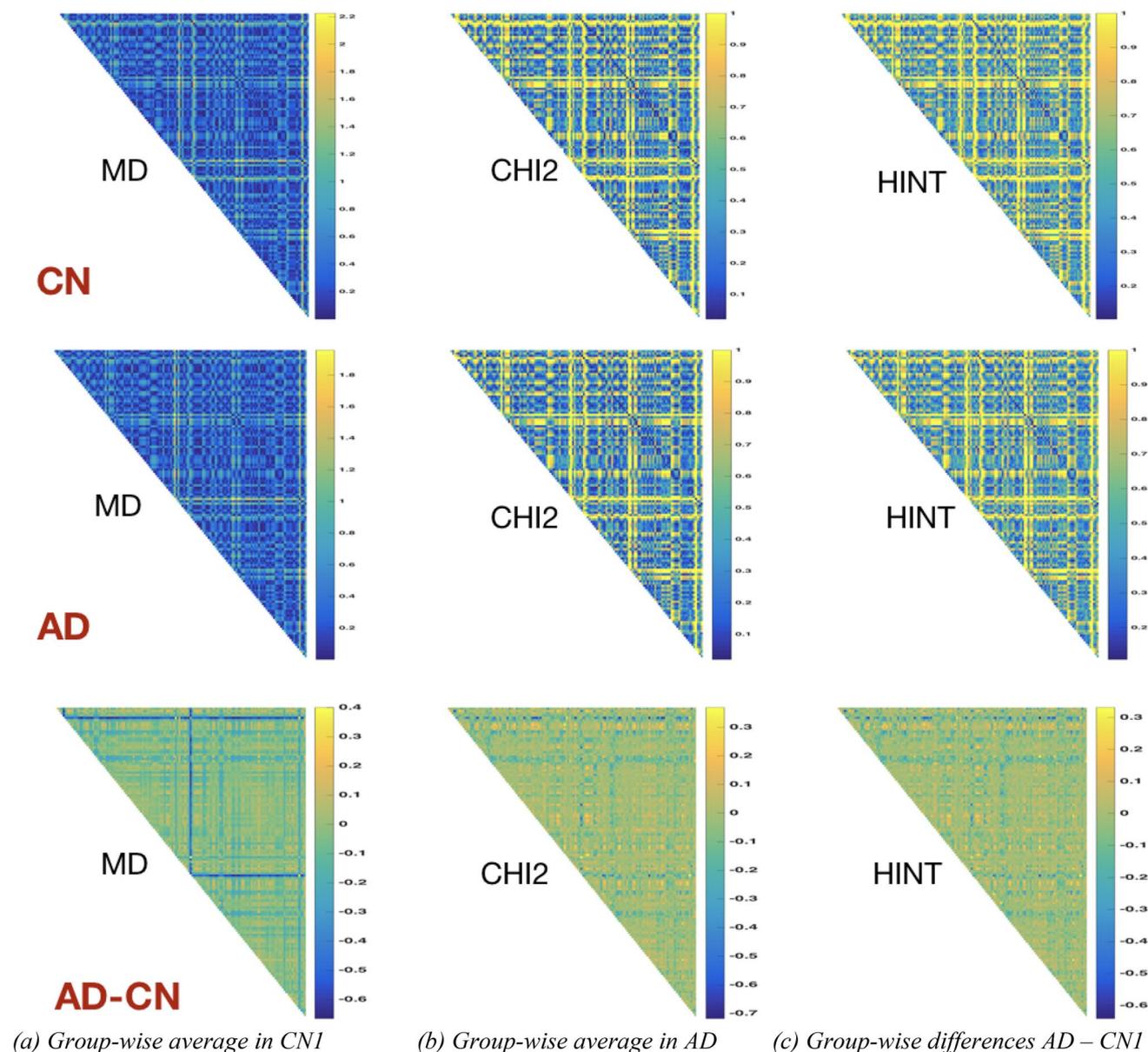


Fig. 5 Edge weights derived from group-wise average thicknesses for three definitions of edge weight. **a** healthy controls (CN1) **b** Edge weights group-wise average in Alzheimer’s disease (AD), both at $m=2000$. **c** Arithmetic differences i.e. AD–CN1. The three panels in each subfigure show the edge weights from MD, CHI2 and HINT methods as defined in Table 4. In each of the panels, we present the

upper triangular part of the edge-weight matrix (pairwise) computed using the corresponding equations in Table 4. We notice there are clear differences among the patterns in the three panels. The panels **a** and **b** appear similar at first glance, but they are sufficiently different to be observed in panel **c**

different from each other. We observe a similar pattern in the center panel (CN2 vs. MCIc), except only the top 5 are statistically significantly different from the lowest ranked methods. In the CN3 vs. AUT case, there are no significant differences at all, possibly due to rather low performance from all the methods to begin with (median AUC across methods is around 0.55).

When the comparison is made at a fixed scale m , within each experiment, the performance of the 6 different methods

(simultaneous comparison of 6 methods) for most values of m are not statistically significantly different from each other, except for $m=1000$ (CD diagrams are not shown). When the comparison is done for a fixed edge-weight metric at different values of m , the performance is not statistically significantly different for any m . Also, the top 2 methods are MD and EMD networks (based on differences in median and mean, respectively) at the highest resolution $m=1000$ and also at the lowest resolution $m=10,000$. This indicates that

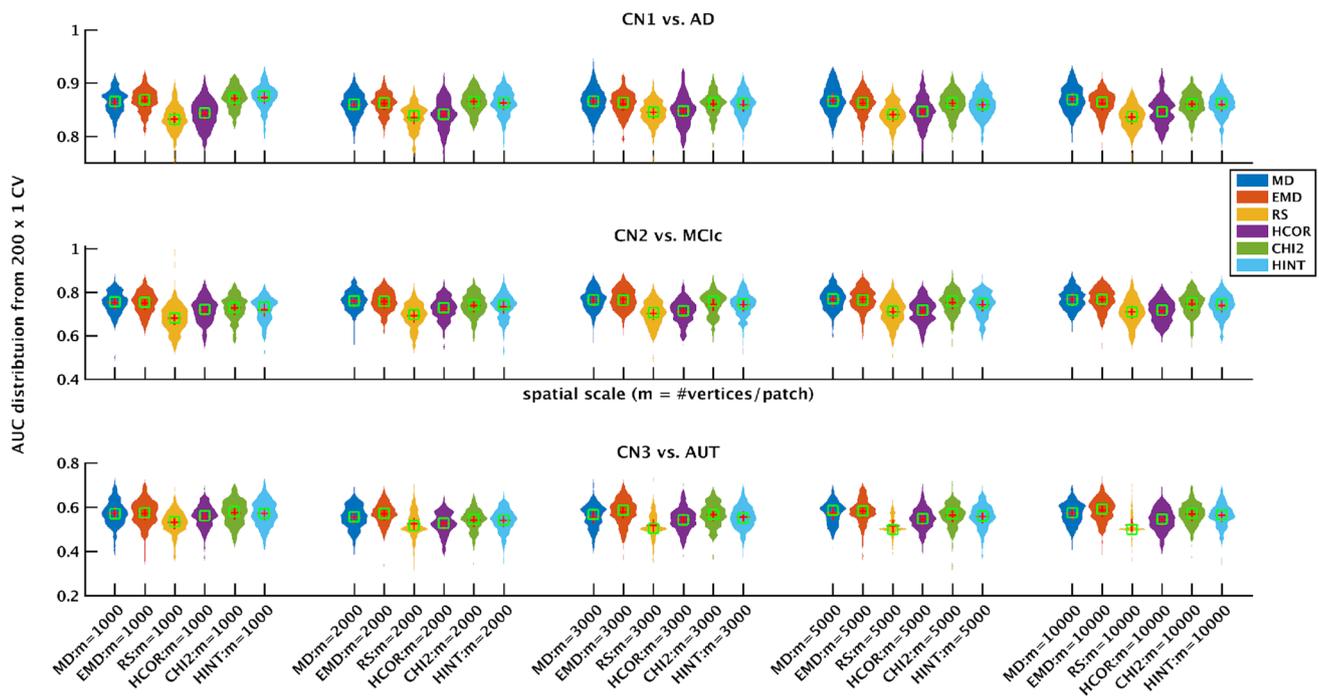


Fig. 6 Classification performance for the different network methods (different edge weight metrics at different spatial resolutions of m) in discriminating AD (top panel), MCIc (middle panel) and AUT (bottom) panel from their respective control groups under a rigorous CV scheme. The data for three experiments come from ADNI, ADNI and

ABIDE datasets, respectively (see Tables 1 and 2). The performance presented here is a distribution of AUC values from 200 randomized train/test splits of RHsT (whose median is shown with a red cross-hair symbol)

impact of the nodal size on the predictive performance of a network method may be insignificant. This result is consistent with the findings of (Zalesky et al. 2010; Evans 2013), wherein it was observed that group-wise small-worldness and scale-freeness are unaffected by spatial scale.

Most discriminative regions

As noted in our CV section earlier, our method records the frequency (across the N CV iterations) of selection (of each weighted connection in VEW) from the t-statistic based ranking method applied on the training set. This helps us gain insight into which pair-wise links have been most frequently discriminative. This pair-wise link frequency can be mapped back to individual cortical patches for intuitive visualization, identifying most discriminative regions (MDRs). One such visualization, thresholding the importance at 50% derived at $m = 2000$, is shown in Fig. 8. Each color of patch on the cortex represents a particular EW metric (labeled on the colorbar) that led to its selection, and when multiple methods selected the same region (indicating additional importance), we painted it red and labeled it “Multiple”. Note the input to the SVM classifier was a vector of edge weights (from upper-triangular part of the edge weight matrix), and hence the selection of a particular edge leads

to highlighting both the regions forming the link. Moreover, the importance of a particular node (cortical patch) could be accumulated from its multiple links, if any.

Figure 8 shows the red MDRs (identified by multiple methods as MDR) cover a large cortical area, which is not unexpected, given the changes caused by full AD are known to be widespread over the cortex. In Fig. 9, we observe the MDRs in areas consistently identified with progressive MCI or early stage AD such as middle temporal lobe, cingulate (anterior and inferior), cuneus and precuneus. Of interest here is the clear hemispheric asymmetry to the left, which can also be observed to a lesser extent in the MDRs for AD in Fig. 8. The MDRs identified in discriminating AUT from CN3 are shown in Fig. 10. They appear in the lingual, supra-marginal, post- and precentral areas, which are consistent with previous reports on Autism studying the group differences in developmental patterns of cortical thickness (Smith et al. 2016; Scheel et al. 2011), as well as found to be important in other prediction tasks (Moradi et al. 2017).

Replication in AIBL dataset

In order to test whether the results and insights from this study on ADNI would generalize to a similar dataset, we have analyzed the AIBL dataset (see Table 2). The predictive

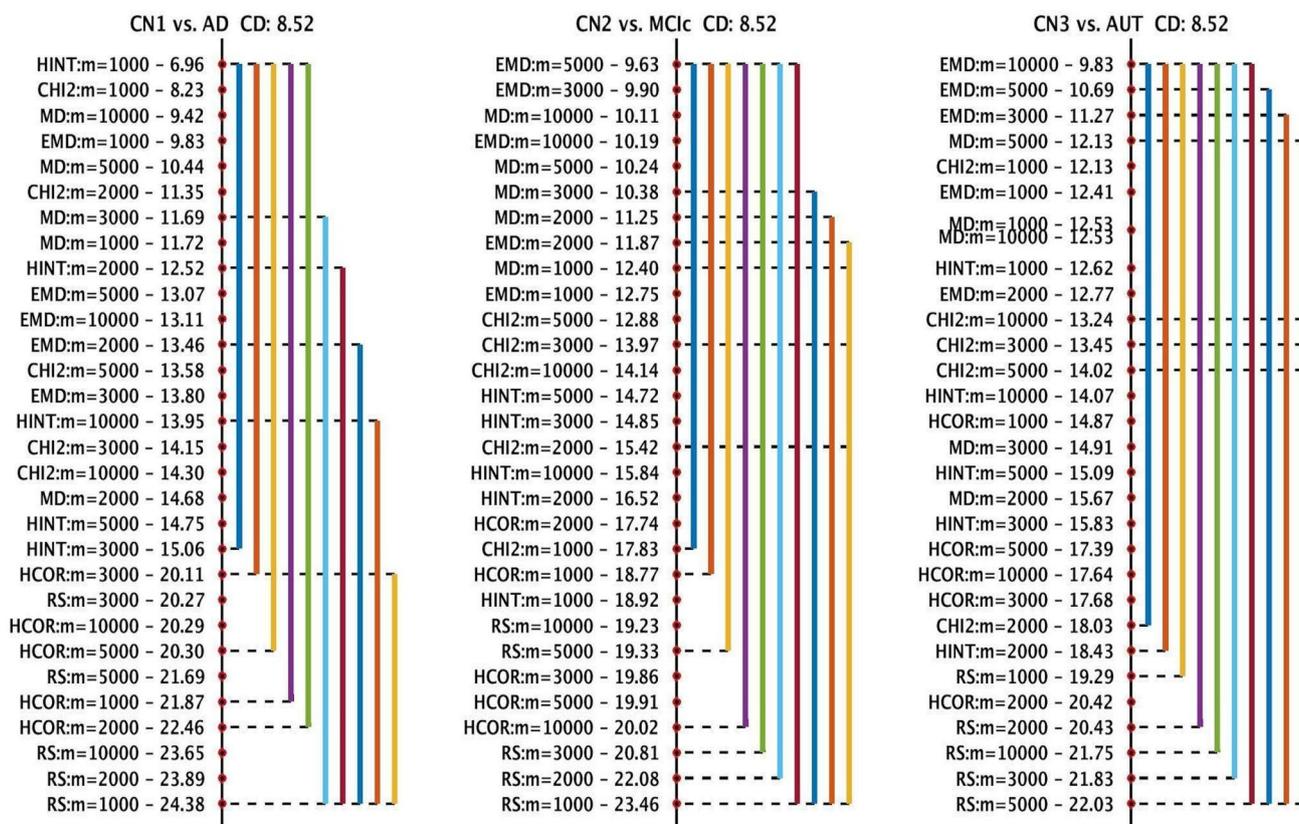


Fig. 7 Critical difference diagram comparing the ranks of different classification methods in a non-parametric Friedman test based on classification performance results from a rigorous CV evaluation method using 200 iterations of holdout. Here, smaller numerical values for rank implies higher performance. The vertical axis presents the ranks (better ranks and methods at the top, and worse ranks and methods to the bottom). The performance of any two methods are statistically significantly different from each other, if their ranks differ by at least the critical difference (CD), which is noted on top of each of the three panels. If a group of methods (annotations on the left within each panel) are connected by a line, they are not statistically significantly different from each other. Different colored lines here present groups of methods that are not significantly different from

utility for different combinations of edge weights and spatial scale (m) are shown in Fig. 11. Although there are numerical differences in performance among different methods at fixed scale (m), their pattern remains similar across different spatial scales within the same dataset, as was observed in Fig. 6 for the ADNI and ABIDE datasets. Based on posthoc statistical analyses (in the same fashion described earlier), we learn they are indeed not statistically significantly different from each other (the critical difference figure is not shown here to save space, as it is a single line connecting them all). That lack of significant differences is also true either for a fixed m (across different EW), or for a fixed EW (across different m).

each other in ranks, each one using a different method as its reference point. For example, in the leftmost panel presenting the results from CN1 vs. AD experiment, the leftmost blue line connects all the methods between the highest ranked HINT: $m=1000$ (ranked 6.96) to the HINT: $m=3000$ method (ranked 15.06), including themselves, which implies they are not statistically significantly different from each other. In the same panel, the highest-ranked HINT: $m=1000$ method is not connected to RS: $m=1000$ (least-ranked 24.38) via any of the colored lines—hence they are indeed statistically significantly different from each other (difference in ranks higher than CD). The values of $m=1000, 2000, 3000, 5000,$ and $10,000$ correspond to the following total number of non-overlapping patches in the whole cortex: 273, 136, 97, 74 and 68, respectively.

We note that the median AUC (to discriminate AD2 from CN4) across all combinations for the AIBL dataset is 0.83. This is in the typical range of NC vs. AD performance we notice in the Alzheimer’s literature, although lower than that noticed in the ADNI1 dataset of 0.87. This slight difference could be attributed to a number of factors, including a slightly different population, different feature extraction library (graynet relies on fully python-based stack), different machine learning library (scikit-learn based on libsvm vs. Matlab’s built-in SVM implementation), and most of all to a much smaller sample ($n=131$, which is only a third of the corresponding ADNI1 subset with $n=412$). That said, the patterns in performance observed in Fig. 11 i.e. lack of significant differences in performance for a fixed spatial scale

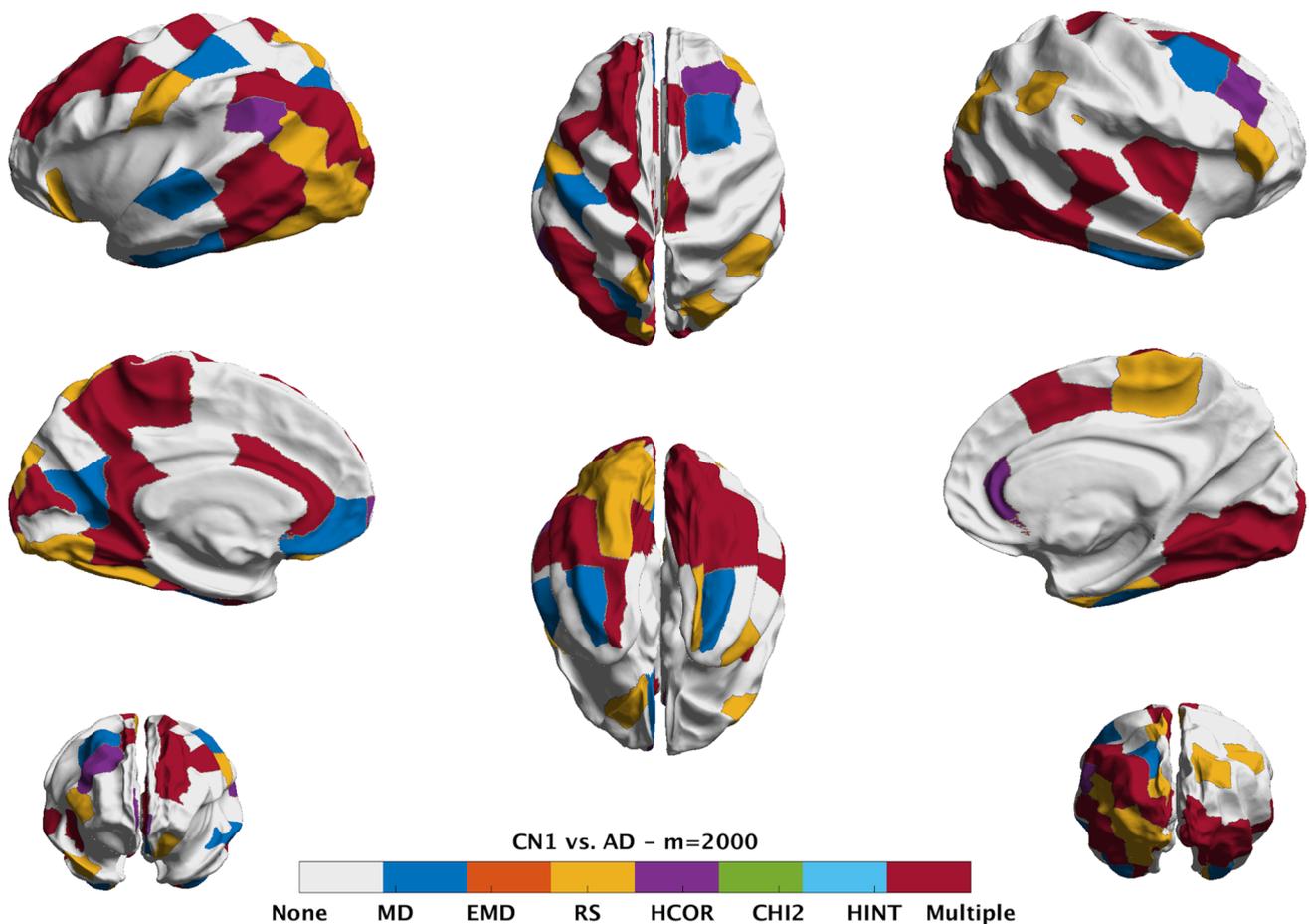


Fig. 8 Visualization of the most discriminative regions as derived from the CN1 vs. AD experiment at $m=2000$. Due to the distributed nature of the degeneration caused by AD, we expect the MDRs to span a wide area of the cortex as observed here. The color of the

patch on the cortex represents a particular EW metric (labeled on the colorbar) that led to its selection, and when multiple methods selected the same region (indicating additional importance), we painted it red and labeled it “Multiple”

(m), or for a fixed EW, replicate the main patterns from the ADNI1 dataset in the AIBL dataset.

Future directions

While we present the results from a large number of experiments ($n=90$, 6 edge weights at 5 different scales m for the 3 datasets) covering two large publicly available datasets, two disease and age groups and three different levels of separability, there is certainly room for further analysis. Future studies could consider additional histogram distances, and performing the comparison with different types of classifiers (other than SVM such as linear discriminant or random forests).

It is possible that lack of sufficiently large sample size could be a contributor to the observed lack of statistically significant differences. This might especially be the case in challenging classification experiments such as CN3 vs.

AUT. Moreover, given the multi-site nature of these large public datasets, properly accounting for the site and other relevant confounds (Raamana 2020) would be worthy of further investigation. Such a broadening of scope for the study is not only computationally very intensive, but we believe studying the above is unlikely to change the conclusions. It would be nevertheless useful to quantitatively support it.

It would also be interesting to study the impact of different atlas choices (other than *fsaverage*, such as MNI152), parcellation (such as (Destrieux et al. 2010)) and subdivision schemes (functional or geometric or multimodal) (Eickhoff et al. 2015; Glasser et al. 2016), potential neuroimaging artifacts and confounds (Churchill et al. 2015; Lerch et al. 2017), but this would be demanding not only computationally but also in expert manpower for quality control (typically unavailable). It would also be quite interesting to replicate this study in the context of differential diagnosis (Raamana et al. 2014a, b). A cross-modal comparison (Reid et al. 2015), in terms of predictive performance, with

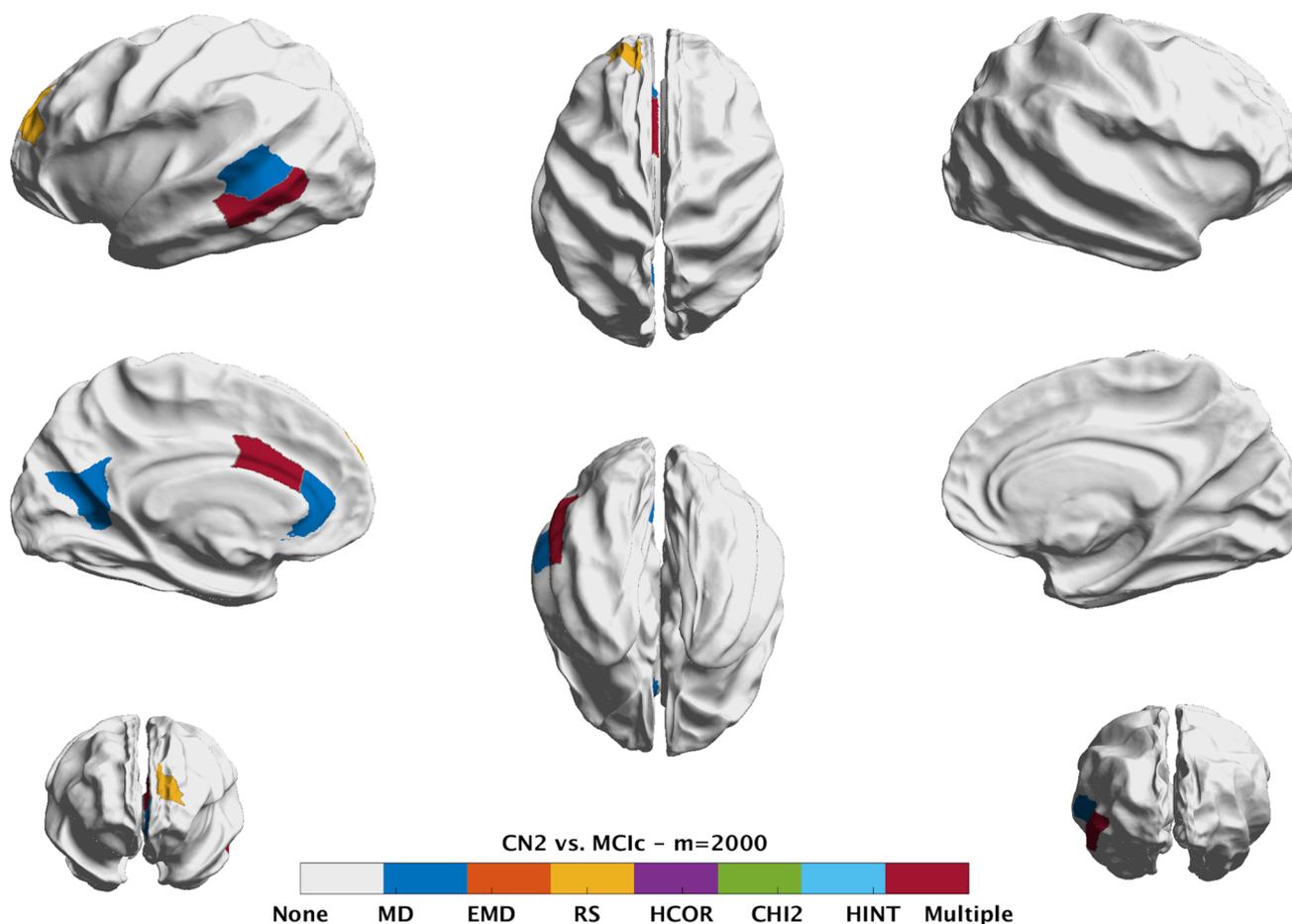


Fig. 9 Visualization of the most discriminative regions as derived from the CN2 vs. MCIc experiment at $m=2000$. MDRs in this experiment identify regions in middle temporal lobe, cingulate (anterior

and inferior), cuneus and precuneus, which are known to be associated with progressive MCI and prodromal AD

network-level features derived from modalities such as task-free fMRI would also be interesting.

Conclusion

We have studied six different ways of constructing weighted networks derived from cortical thickness features, based on a novel method to derive edge weights based on histogram distances. We performed a comprehensive model comparison based on extensive cross-validation of their predictive utility and nonparametric statistical tests. This has been studied under three separabilities (ranging from pronounced, mild to subtle differences) derived from three independent and large publicly available datasets.

Some interesting results of this study based on the single-subject classification results are:

- the simpler methods of edge weight computation such as the difference in median thickness are as predictive as the sophisticated methods relying on the richer descriptions based on complete histograms.
- within a given method, the impact of a spatial scale m on predictive performance is not significant. The most popular way of computing edge weights in group-wise analysis i.e. histogram correlation, is shown to be the least predictive of disease-status in the context of individualized prediction via HiWeNet.

We have also developed and shared multiple open source toolboxes called *graynet*, *visualqc* and *neuropredict* to enable easy reuse of the methods and best practices presented in this study.

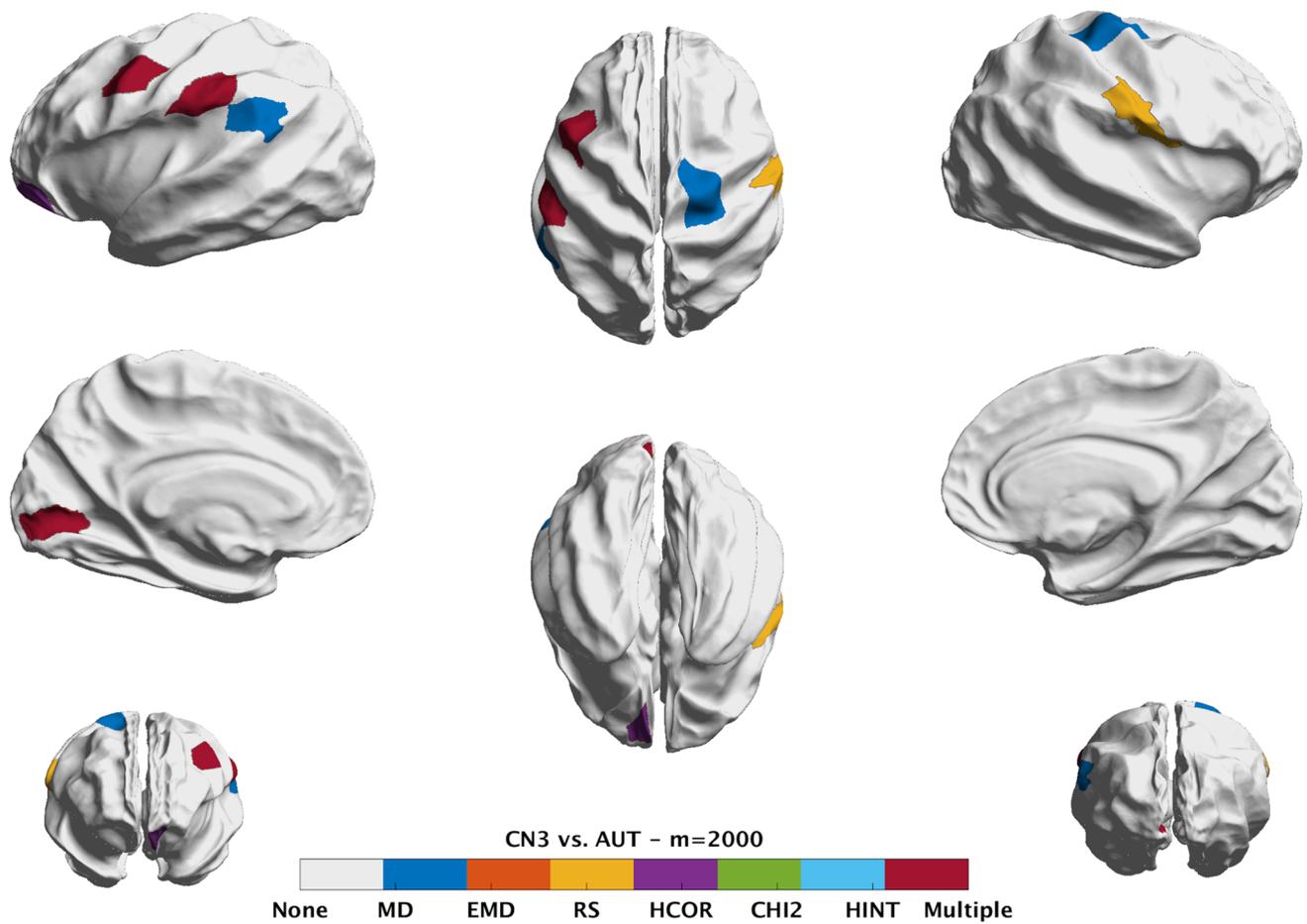


Fig. 10 Visualization of the most discriminative regions as derived from the CN3 vs. AUT experiment at $m=2000$. These regions cover the lingual, supra-marginal, post- and precentral areas

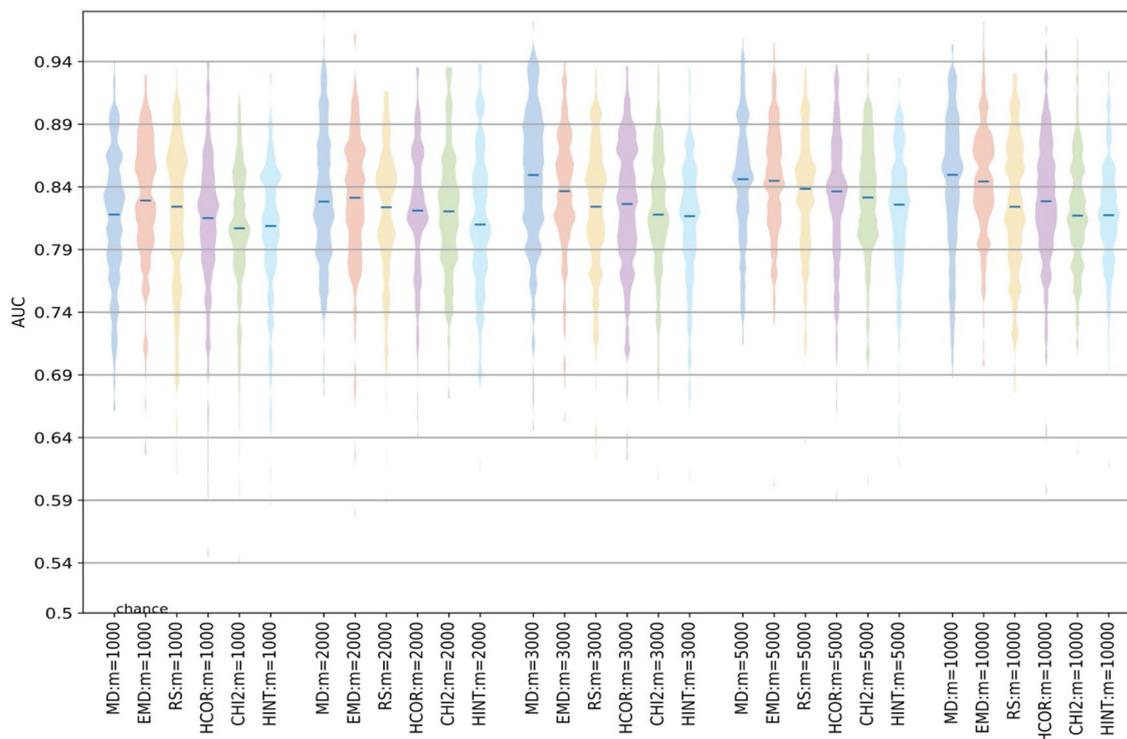


Fig. 11 Classification performance for the different network methods (different edge weight metrics at different spatial resolutions of m) in discriminating AD2 from their respective control groups (CN4) from the AIBL dataset (see Table 2). The performance presented here is a distribution of AUC values from 200 randomized train/test splits of

RHsT. It is clear that AUC for different combinations is quite similar to each other (as in Fig. 6 for ADNI and ABIDE datasets). That is also true either for a fixed m (and different edge weights) or a fixed edge weight (and different values of m)

Acknowledgements **ADNI:** Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; Euro-Immun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for NeuroImaging at the University of Southern California. **AIBL:** Data used in the preparation of this article

was obtained from the Australian Imaging Biomarkers and Lifestyle flagship study of aging (AIBL) funded by the Commonwealth Scientific and Industrial Research Organization (CSIRO) which was made available at the ADNI database (www.loni.usc.edu/ADNI). The AIBL researchers contributed data but did not participate in analysis or writing of this report. AIBL researchers are listed at www.aibl.csiro.au. **ABIDE I:** Primary support for the work by Adriana Di Martino was provided by the NIMH (K23MH087770) and the Leon Levy Foundation. Primary support for the work by Michael P. Milham and the INDI team was provided by gifts from Joseph P. Healy and the Stavros Niarchos Foundation to the Child Mind Institute, as well as by an NIMH award to MPM (R03MH096321).

Funding PRR is partly supported by CIHR (MOP 84483 to SCS), Ontario Neurodegenerative Disease Research Initiative (ONDRI) and Canadian Biomarker Integration Network for Depression (CAN-BIND), and with SCS the Ontario Brain Institute (OBI), an independent non-profit corporation, funded partially by the Ontario government. The opinions, results and conclusion are those of the authors and no endorsement by the OBI is intended or should be inferred. We also thank the financial support from the *Temery Family Foundation*.

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest.

References

- Abraham A, Milham MP, Adriana Di Martino R, Craddock C, Samaras D, Thirion B, Varoquaux G (2017) Deriving reproducible biomarkers from multi-site resting-state data: an autism-based example. *NeuroImage* 147:736–745. <https://doi.org/10.1016/j.neuroimage.2016.10.045>
- Alexander-Bloch A, Giedd JN, Bullmore E (2013) Imaging structural co-variance between human brain regions. *Nature Reviews Neuroscience* 14(5):322–36. <https://www.nature.com/doi/10.1038/nrn3465>
- Alzheimer's Association (2017) Alzheimer's disease facts and figures. *Alzheimer's Dement J Alzheimer's Assoc* 13(4):325–373. <https://doi.org/10.1016/j.jalz.2017.02.001>
- Arbabshirani MR, Plis S, Sui J, Calhoun VD (2017) Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *Neuroimage* 145(Pt B):137–65. <https://linkinghub.elsevier.com/retrieve/pii/S105381191600210X>
- Bron EE, Smits M, van der Flier WM, Vrenken H, Barkhof F, Scheltens P, Pappa JM et al (2015) Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia Challenge. *NeuroImage* 111:562–579. <https://doi.org/10.1016/j.neuroimage.2015.01.048>
- Burman P (1989) A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika* 76(3):503–514. <https://doi.org/10.1093/biomet/76.3.503>
- Churchill NW, Spring R, Afshin-Pour B, Dong F, Strother SC (2015) An automated, adaptive framework for optimizing pre-processing pipelines in task-based functional MRI. *PLoS ONE* 10(7):e0131520. <https://doi.org/10.1371/journal.pone.0131520>
- Craddock, C, Benhajali, Y, Chu C, Chouinard F, Evans A, Jakab A, Khundrakpam BS, Lewis JD, Li Q, Milham M, Yan C, Bellec P (2013) The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. *Front Neuroinform.* <https://doi.org/10.3389/conf.fninf.2013.09.00041>
- Cuingnet R, Gerardin E, Tessieras J, Auzias G, Lehéricy S, Habert M-O, Chupin M, Benali H, Colliot O, Alzheimer's Disease Neuroimaging Initiative (2011) Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. *NeuroImage* 56(2):766–781. <https://doi.org/10.1016/j.neuroimage.2010.06.013>
- Cummings JL, Morstorf T, Zhong K (2014) Alzheimer's disease drug-development pipeline: few candidates, frequent failures. *Alzheimer's Res Ther* 6(4):37. <https://doi.org/10.1186/alzrt269>
- Dai D, He H, Vogelstein JT, Hou Z (2013) Accurate prediction of AD patients using cortical thickness networks. *Int J Comput Vis* <https://www.springerlink.com/index/10.1007/s00138-012-0462-0>
- Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res JMLR* 7. <https://dl.acm.org/citation.cfm?id=1248548>
- Destrieux C, Fischl B, Dale A, Halgren E (2010) Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage* 53(1):1–15. <https://doi.org/10.1016/j.neuroimage.2010.06.010>
- Dickerson BC, Fenstermacher E, Salat DH, Wolk DA, Maguire RP, Desikan R, Pacheco J et al (2008) Detection of cortical thickness correlates of cognitive performance: reliability across MRI scan sessions, scanners, and field strengths. *NeuroImage* 39(1):10–18. <https://doi.org/10.1016/j.neuroimage.2007.08.042>
- Dietterich TG (1998) Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computat* 10(7):1895–1923. <https://doi.org/10.1162/089976698300017197>
- Duchesne S, Caroli A, Geroldi C, Barillot C (2008) MRI-based automated computer classification of probable AD versus normal controls. *IEEE Trans Med Imaging.* https://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4479633
- Dukart J, Mueller K, Horstmann A, Barthel H, Möller HE, Villringer A, Sabri O, Schroeter ML (2011) Combined evaluation of FDG-PET and MRI improves detection and differentiation of dementia. *PLoS ONE* 6(3):e18111. <https://doi.org/10.1371/journal.pone.0018111>
- Dyrba M, Barkhof F, Fellgiebel A, Filippi M, Hausner L, Hauenstein K, Kirsche T, Teipel SJ (2015) Predicting prodromal Alzheimer's disease in subjects with mild cognitive impairment using machine learning classification of multimodal multicenter diffusion- *J Neuroimaging* 25(5):738–747. <https://doi.org/10.1111/jon.12214/full>
- Eickhoff SB, Bertrand T, Gaël V, Danilo B (2015) Connectivity-based parcellation: critique and implications. *Hum Brain Map* 36(12):4771–4192. <https://onlinelibrary.wiley.com/myaccess.library.utoronto.ca/doi/10.1002/hbm.22933/full>
- Ellis KA, Bush AI, Darby D, De Fazio D, Foster J, Hudson P, Lautenschlager NT et al (2009) The Australian imaging, biomarkers and lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *Int Psychogeriatr IPA* 21(4):672–687. <https://doi.org/10.1017/S1041610209009405>
- Evans AC (2013) Networks of anatomical covariance. *NeuroImage* 80: 489–504. <https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?dbfrom=pubmed&id=23711536&retmode=ref&cmd=prlinks>
- Fischl B, Dale A (2000) Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proc Natl Acad Sci.* <https://www.pnas.org/cgi/content/abstract/97/20/11050>
- Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, van der Kouwe A et al (2002) Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron* 33(3):341–355. [https://doi.org/10.1016/S0896-6273\(02\)00569-X](https://doi.org/10.1016/S0896-6273(02)00569-X)
- Glasser MF, Coalson TS, Robinson EC, Hacker CD, Harwell J, Yacoub E, Uğurbil K et al (2016) A multi-modal parcellation of human cerebral cortex. *Nature* 536(7615):171–178. <https://doi.org/10.1038/nature18933>
- Gong G, Yong H, Zhang JC, Alan CE (2012) Convergence and divergence of thickness correlations with diffusion connections across the human cerebral cortex *NeuroImage* 59(2): 1239–1248. <https://www.sciencedirect.com/science/article/pii/S1053811911009049>
- Han X, Jorge J, David S, Andre van der K, Brian Q, Silvester C, Evelina B, et al. (2006) Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. *NeuroImage* 32(1): 180–194. <https://linkinghub.elsevier.com/retrieve/pii/S1053811906001601>
- He Y, Chen Z (2007) Small-world anatomical networks in the human brain revealed by cortical thickness from MRI. *Cerebral Cortex.* <https://cercor.oxfordjournals.org/content/17/10/2407.abstract>
- Herholz K, Salmon E, Perani D, Baron J-C, Holthoff V, Frölich L, Schönknecht P, Ito K, Mielke R, Kalbe E (2002) Discrimination between Alzheimer dementia and controls by automated analysis of multicenter FDG PET. *NeuroImage* 17(1): 302–316. <https://linkinghub.elsevier.com/retrieve/pii/S1053811902912085>
- Hojjati SH, Ebrahimzadeh A, Khazaei A (2017) Predicting conversion from MCI to AD using resting-state fMRI, graph theoretical approach and SVM. *J Neurosci.* <https://www.sciencedirect.com/science/article/pii/S0165027017300638>
- Iscan Z, Jin TB, Kendrick A, Szeglin B, Hanzhang Lu, Trivedi M, Fava M et al (2015) Test-retest reliability of freesurfer measurements within and between sites: effects of visual approval process. *Hum Brain Mapp* 36(9):3472–3485. <https://doi.org/10.1002/hbm.22856>
- Jack CR Jr, Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, Borowski B et al (2008) The Alzheimer's disease neuroimaging

- initiative (ADNI): MRI methods. *J Magn Reson Imaging JMRI* 27(4):685–691. <https://doi.org/10.1002/jmri.21049>
- Kim H-J, Shin J-H, Han CE, Kim HJ, Na DL, Seo SW, Seong J-K, Alzheimer's Disease Neuroimaging Initiative (2016) Using individualized brain network for analyzing structural covariance of the cerebral cortex in Alzheimer's patients. *Front Neurosci* 10(Sep-tember):394. <https://doi.org/10.3389/fnins.2016.00394>
- Kourentzes N (2016) ANOM and Nemenyi tests. <https://kourentzes.com/forecasting/2013/04/19/nemenyi-test/>
- Lerch JP, Worsley K, Shaw WP, Greenstein DK, Lenroot RK, Giedd J, Evans AC (2006) Mapping anatomical correlations across cerebral cortex (MACACC) using cortical thickness from MRI. *Neuroimage*. 31(3):993–1003. <https://linkinghub.elsevier.com/retrieve/pii/S1053811906000425>
- Lerch JP, Ajw van der K, Raznahan A (2017) Studying Neuroanatomy Using MRI. *Nature*. <https://www.nature.com/neuro/journal/v20/n3/abs/nn.4501.html>
- Matthews DC, Lukic AS, Andrews RD, Marendic B, Brewer J, Rissman RA, Mosconi L et al (2016) Dissociation of down syndrome and alzheimer's disease effects with imaging. *Alzheimer's Dementia* 2(2):69–81. <https://doi.org/10.1016/j.trci.2016.02.004>
- Moradi E, Khundrakpam B, Lewis JD, Evans AC, Tohka J (2017) Predicting symptom severity in autism spectrum disorder based on cortical thickness measures in agglomerative data. *NeuroImage* 144(Pt A):128–141. <https://doi.org/10.1016/j.neuroimage.2016.09.049>
- Phillips DJ, McGlaughlin A, Ruth D, Jager LR, Soldan A (2015) Graph theoretic analysis of structural connectivity across the spectrum of Alzheimer's disease: the importance of graph creation methods. *NeuroImage Clin* 7:377–390. <https://linkinghub.elsevier.com/retrieve/pii/S221315821500008X>
- Power JD, Cohen AL, Nelson SM, Wig GS, Barnes KA, Church JA, Vogel AC et al (2011) Functional network organization of the human brain. *Neuron* 72(4):665–678. <https://doi.org/10.1016/j.neuron.2011.09.006>
- Raamana PR (2017) Neuropredict: easy machine learning and standardized predictive analysis of biomarkers. <https://doi.org/10.5281/zenodo.1058993>
- Raamana PR (2018) VisualQC: assistive tools for easy and rigorous quality control of neuroimaging data. <https://doi.org/10.5281/zenodo.1211365>
- Raamana PR, Strother SC (2017) Histogram-weighted networks for feature extraction, connectivity and advanced analysis in neuroscience. *J Open Sour Softw* 2(19):380. <https://doi.org/10.21105/joss.00380>
- Raamana PR, Strother S (2018) Graynet: single-subject morphometric networks for neuroscience connectivity applications. *J Open Sour Softw* 3(30):924. <https://doi.org/10.21105/joss.00924>
- Raamana PR, Rosen H, Miller B, Weiner MW, Wang L, Beg MF (2014a) Three-class differential diagnosis among alzheimer disease, frontotemporal dementia and controls. *Front Neurol* 5(71):71. <https://doi.org/10.3389/fneur.2014.00071>
- Raamana PR, Wen W, Kochan NA, Brodaty H, Sachdev PS, Wang L, Beg MF (2014b) Novel ThickNet features for the discrimination of amnesic MCI subtypes. *NeuroImage Clin* 6:284–295. <https://www.sciencedirect.com/science/article/pii/S2213158214001417>
- Raamana PR, Weiner MW, Wang L, Beg MF (2015) Alzheimer's disease neuroimaging initiative. Thickness network features for prognostic applications in dementia. *Neurobiology Aging* 36:S91–102. <https://linkinghub.elsevier.com/retrieve/pii/S0197458014005521>
- Raamana PR, Rosen H, Miller B, Weiner MW, Wang L, Beg MF (2018) Mrivis: medical image visualization library for neuroscience in python. *J Open Sour Softw* 3(30):897. <https://doi.org/10.21105/joss.00897>
- Raamana PR (2020) Conquering confounds and covariates in machine learning with the python library confounds, version 0.1.1. <https://doi.org/10.5281/zenodo.3701528>
- Raamana PR et al (2020) Visual QC protocol for freesurfer cortical parcellations from anatomical MRI. *Biorxiv Preprint*. <https://doi.org/10.1101/2020.09.07.286807>
- Reid AT, Evans AC (2013) Structural networks in Alzheimer's disease. *Eur Neuropsychopharmacol* 23(1):63–77. <https://linkinghub.elsevier.com/retrieve/pii/S0924977X12003227>
- Reid AT, Lewis J, Bezgin G, Khundrakpam B, Eickhoff SB, McIntosh AR, Bellec P, Evans AC (2015) A cross-modal, cross-species comparison of connectivity measures in the primate brain. *Neuroimage*. <https://linkinghub.elsevier.com/retrieve/pii/S1053811915009696>
- Scheel C, Rotarska-Jagiela A, Schilbach L, Lehnhardt FG, Krug B, Vogeley K, Tepest R (2011) Imaging spe. *NeuroImage* 58(2):391–400. <https://doi.org/10.1016/j.neuroimage.2011.06.040>
- Smith E, Thurm A, Greenstein D, Farmer C, Swedo S, Giedd J, Raznahan A (2016) Cortical thickness change in autism during early childhood. *Hum Brain Mapp* 37(7):2616–2629. <https://doi.org/10.1002/hbm.23195>
- Spreng RN, Turner GR (2013) Structural covariance of the default network in healthy and pathological aging. *J Neurosci* 33(38):15226–15234. <https://www.jneurosci.org.proxy.lib.sfu.ca/content/33/38/15226.full>
- Spreng RN, Sepulcre J, Turner GR, Dale Stevens W, Schacter DL (2013) Intrinsic architecture underlying the relations among the default, dorsal attention, and frontoparietal control networks of the human brain. *J Cognit Neurosci* 25(1):74–86. https://doi.org/10.1162/jocn_a_00281
- Stam, CJ (2014) Modern network science of neurological disorders. *Nat Rev Neurosci* 15(10): 683–695. <https://www.nature.com/doi/10.1038/nrn3801>
- Sui J, Adali T, Yu Q, Chen J (2011) A review of multivariate methods for multimodal fusion of brain imaging data. *J Neurosci Methods*. <https://www.sciencedirect.com/science/article/pii/S0165027011006820>
- Tijms BM, Serié P, Willshaw DJ, Lawrie SM (2012) Similarity-based extraction of individual networks from gray matter MRI scans. *Cerebral Cortex*. 22(7):1530–1541. <https://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=21878484&retmode=ref&cmd=prlinks>
- Tijms BM, Wink AM, de Haan W, van der Flier WM, Stam CJ, Scheltens P, Barkhof F (2013) Alzheimer's disease: connecting findings from graph theoretical studies of brain networks. *Neurobiol Aging*. 34(8):2023–2036. <https://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=23541878&retmode=ref&cmd=prlinks>
- Van Wijk BC, Stam CJ, Daffertshofer A (2020) Comparing brain networks of different size and connectivity density using graph theory. *PloS One* 5(10):e13701. <https://dx.plos.org/10.1371/journal.pone.0013701>
- Varoquaux G, Raamana PR, Engemann DA, Hoyos-Idrobo A, Schwartz Y, Thirion B (2016) Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage*. <https://linkinghub.elsevier.com/retrieve/pii/S105381191630595X>
- Wee CY, Yap PT, Shen D, Alzheimer's Disease Neuroimaging Initiative (2012) Prediction of Alzheimer's disease and mild cognitive impairment using cortical morphological patterns. *Hum Brain Map*. <https://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=22927119&retmode=ref&cmd=prlinks>
- Weiner MW, Veitch DP, Aisen PS, Beckett LA, Cairns NJ, Cedarbaum J, Green RC, Harvey D, Jack CR, Jagust W, Luthman J (2015) 2014 update of the Alzheimer's disease neuroimaging initiative: a review of papers published since its inception. *Alzheimer's*

- Dementia 11(6):e1-20. <https://linkinghub.elsevier.com/retrieve/pii/S1552526014028659>
- Weiner MW, Veitch DP, Aisen PS, Beckett LA, Cairns NJ, Green RC, Harvey D et al (2017) Recent publications from the Alzheimer's disease neuroimaging initiative: reviewing progress toward improved AD clinical trials. *Alzheimer's Dementia J Alzheimer's Assoc.* <https://doi.org/10.1016/j.jalz.2016.11.007>
- Wen W, He Y, Sachdev P (2011) Structural brain networks and neuropsychiatric disorders. *Curr Opin Psychiatry* 24(3):219-25 <https://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landi&page&an=00001504-201105000-00009>
- Zalesky A, Alex F, Ian HH, Luca C, Murat Y, Christos P, Edward TB (2010) Whole-brain anatomical networks: does the choice of nodes matter?" *NeuroImage* 50(3): 970–983. <https://linkinghub.elsevier.com/retrieve/pii/S1053811909013159>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.