

1 Novel genomic loci influence patterns of structural covariance in the human 2 brain

3
4 Junhao Wen^{1,2*}, Ilya M. Nasrallah^{2,3}, Ahmed Abdulkadir², Theodore D. Satterthwaite^{2,4}, Zhijian Yang²,
5 Guray Erus², Timothy Robert-Fitzgerald⁵, Ashish Singh², Aristeidis Sotiras⁶, Aleix Boquet-Pujadas⁷,
6 Elizabeth Mamourian², Jimit Doshi², Yuhan Cui², Dhivya Srinivasan², Ioanna Skampardoni², Jiong
7 Chen², Gyujoon Hwang², Mark Bergman², Jingxuan Bao⁸, Yogasudha Veturi⁹, Zhen Zhou², Shu Yang⁸,
8 Paola Dazzan¹⁰, Rene S. Kahn¹¹, Hugo G. Schnack¹², Marcus V. Zanetti¹³, Eva Meisenzahl¹⁴, Geraldo F.
9 Busatto¹³, Benedicto Crespo-Facorro¹⁵, Christos Pantelis¹⁶, Stephen J. Wood¹⁷, Chuanjun Zhuo¹⁸, Russell
10 T. Shinohara^{2,5}, Ruben C. Gur⁴, Raquel E. Gur⁴, Nikolaos Koutsouleris¹⁹, Daniel H. Wolf^{2,4}, Andrew J.
11 Saykin²⁰, Marylyn D. Ritchie⁹, Li Shen⁸, Paul M. Thompson²¹, Olivier Colliot²², Katharina Wittfeld²³,
12 Hans J. Grabe²³, Duygu Tosun²⁴, Murat Bilgel²⁵, Yang An²⁵, Daniel S. Marcus²⁶, Pamela LaMontagne²⁶,
13 Susan R. Heckbert²⁷, Thomas R. Austin²⁷, Lenore J. Launer²⁸, Mark Espeland²⁹, Colin L Masters³⁰, Paul
14 Maruff³⁰, Jurgen Fripp³¹, Sterling C. Johnson³², John C. Morris³³, Marilyn S. Albert³⁴, R. Nick Bryan³,
15 Susan M. Resnick²⁵, Yong Fan², Mohamad Habes³⁵, David Wolk^{2,36}, Haochang Shou^{2,5}, and Christos
16 Davatzikos^{2*}, for the iSTAGING, the BLSA, the BIOCARD, the PHENOM, the ADNI studies, and the
17 AI4AD consortium

18
19 ¹Laboratory of AI and Biomedical Science (LABS), Stevens Neuroimaging and Informatics Institute, Keck School of
20 Medicine of USC, University of Southern California, Los Angeles, California, USA.

21 ²Artificial Intelligence in Biomedical Imaging Laboratory (AIBIL), Center for Biomedical Image Computing and
22 Analytics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA.

23 ³Department of Radiology, University of Pennsylvania, Philadelphia, USA.

24 ⁴Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA

25 ⁵Penn Statistics in Imaging and Visualization Center, Department of Biostatistics, Epidemiology, and Informatics,
26 Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA

27 ⁶Department of Radiology and Institute for Informatics, Washington University School of Medicine, St. Louis, USA

28 ⁷Biomedical Imaging Group, EPFL, Lausanne, Switzerland

29 ⁸Department of Biostatistics, Epidemiology and Informatics University of Pennsylvania Perelman School of Medicine,
30 Philadelphia, USA

31 ⁹Department of Genetics and Institute for Biomedical Informatics, Perelman School of Medicine, University of
32 Pennsylvania, Philadelphia, PA, USA

33 ¹⁰Department of Psychological Medicine, Institute of Psychiatry, Psychology and Neuroscience, King's College
34 London, London, UK

35 ¹¹Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, USA

36 ¹²Department of Psychiatry, University Medical Center Utrecht, Utrecht, Netherlands

37 ¹³Institute of Psychiatry, Faculty of Medicine, University of São Paulo, São Paulo, Brazil

38 ¹⁴Department of Psychiatry and Psychotherapy, HHU Düsseldorf, Germany

39 ¹⁵Hospital Universitario Virgen del Rocío, University of Sevilla-IBIS; IDIVAL-CIBERSAM, Sevilla, Spain

40 ¹⁶Melbourne Neuropsychiatry Centre, Department of Psychiatry, University of Melbourne and Melbourne Health,
41 Carlton South, Australia

42 ¹⁷Orygen and the Centre for Youth Mental Health, University of Melbourne; and the School of Psychology,
43 University of Birmingham, UK

44 ¹⁸Key Laboratory of Real Time Tracing of Brain Circuits in Psychiatry and Neurology (RTBCPN-Lab), Nankai
45 University Affiliated Tianjin Fourth Center Hospital; Department of Psychiatry, Tianjin Medical University, Tianjin,
46 China

47 ¹⁹Department of Psychiatry and Psychotherapy, Ludwig-Maximilian University, Munich, Germany

48 ²⁰Radiology and Imaging Sciences, Center for Neuroimaging, Department of Radiology and Imaging Sciences,
49 Indiana Alzheimer's Disease Research Center and the Melvin and Bren Simon Cancer Center, Indiana University
50 School of Medicine, Indianapolis

51 ²¹Imaging Genetics Center, Mark and Mary Stevens Neuroimaging and Informatics Institute, Keck School of
52 Medicine of USC, University of Southern California, Marina del Rey, California

53 ²²Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la
54 Pitié Salpêtrière, F-75013, Paris, France
55 ²³Department of Psychiatry and Psychotherapy, German Center for Neurodegenerative Diseases (DZNE), University
56 Medicine Greifswald, Germany
57 ²⁴Department of Radiology and Biomedical Imaging, University of California, San Francisco, CA, USA
58 ²⁵Laboratory of Behavioral Neuroscience, National Institute on Aging, NIH, USA
59 ²⁶Department of Radiology, Washington University School of Medicine, St. Louis, Missouri, USA
60 ²⁷Cardiovascular Health Research Unit and Department of Epidemiology, University of Washington, Seattle, WA,
61 USA
62 ²⁸Neuroepidemiology Section, Intramural Research Program, National Institute on Aging, Bethesda, Maryland, USA
63 ²⁹Sticht Center for Healthy Aging and Alzheimer's Prevention, Wake Forest School of Medicine, Winston-Salem,
64 North Carolina, USA
65 ³⁰Florey Institute of Neuroscience and Mental Health, The University of Melbourne, Parkville, VIC, Australia
66 ³¹CSIRO Health and Biosecurity, Australian e-Health Research Centre CSIRO, Brisbane, Queensland, Australia
67 ³²Wisconsin Alzheimer's Institute, University of Wisconsin School of Medicine and Public Health, Madison,
68 Wisconsin, USA
69 ³³Knight Alzheimer Disease Research Center, Washington University in St. Louis, St. Louis, MO, USA
70 ³⁴Department of Neurology, Johns Hopkins University School of Medicine, USA
71 ³⁵Glenn Biggs Institute for Alzheimer's & Neurodegenerative Diseases, University of Texas Health Science Center at
72 San Antonio, San Antonio, USA
73 ³⁶Department of Neurology and Penn Memory Center, University of Pennsylvania, Philadelphia, USA
74
75 *Corresponding authors:
76 Junhao Wen, Ph.D. – junhaowe@usc.edu
77 [2025 Zonal Ave, Los Angeles, CA 90033, United States](#)
78 Christos Davatzikos, Ph.D. – Christos.Davatzikos@pennmedicine.upenn.edu
79 3700 Hamilton Walk, 7th Floor, Philadelphia, PA 19104, [United States](#)
80
81 Word counts: 5002 words

82 **Abstract**

83 Normal and pathologic neurobiological processes influence brain morphology in coordinated
84 ways that give rise to patterns of structural covariance (PSC) across brain regions and individuals
85 during brain aging and diseases. The genetic underpinnings of these patterns remain largely
86 unknown. We apply a stochastic multivariate factorization method to a diverse population of
87 50,699 individuals (12 studies, 130 sites) and derive data-driven, multi-scale PSCs of regional
88 brain size. PSCs were significantly correlated with 915 genomic loci in the discovery set, 617 of
89 which are novel, and 72% were independently replicated. Key pathways influencing PSCs
90 involve reelin signaling, apoptosis, neurogenesis, and appendage development, while pathways
91 of breast cancer indicate potential interplays between brain metastasis and PSCs associated with
92 neurodegeneration and dementia. Using support vector machines, multi-scale PSCs effectively
93 derive imaging signatures of several brain diseases. Our results elucidate new genetic and
94 biological underpinnings that influence structural covariance patterns in the human brain.

95

96

97 **Significance statement**

98 The coordinated patterns of changes in the human brain throughout life, driven by brain
99 development, aging, and diseases, remain largely unexplored regarding their underlying genetic
100 determinants. This study delineates 2003 multi-scale patterns of structural covariance (PSCs) and
101 identifies 617 novel genomic loci, with the mapped genes enriched in biological pathways
102 implicated in reelin signaling, apoptosis, neurogenesis, and appendage development. Overall, the
103 2003 PSCs provide new genetic insights into understanding human brain morphological changes
104 and demonstrate great potential in predicting various neurologic conditions.

105 **Introduction**

106 Brain structure and function are interrelated via complex networks that operate at multiple scales,
107 ranging from cellular and synaptic processes, such as neural migration, synapse formation, and
108 axon development, to local and broadly connected circuits.¹ Due to a fundamental relationship
109 between activity and structure, many normal and pathologic neurobiological processes, driven by
110 genetic and environmental factors, collectively cause coordinated changes in brain morphology.
111 Structural covariance analyses investigate such coordinated changes by seeking patterns of
112 structural covariation (PSC) across brain regions and individuals.¹ For example, during
113 adolescence, PSCs derived from magnetic resonance imaging (MRI) have been considered to
114 reflect a coordinated cortical remodeling as the brain establishes mature networks of functional
115 specialization.² Structural covariance is not only related to normal brain development or aging
116 processes but can also reflect coordinated brain change due to disease. For example, individuals
117 with motor speech dysfunction may develop brain atrophy in Broca's inferior frontal cortex and
118 co-occurring brain atrophy in Wernicke's area of the superior temporal cortex.³ Refer to **Fig. 1C**
119 for an illustrative depiction.

120 The human brain develops, matures, and degenerates in coordinated patterns of structural
121 covariance at the macrostructural level of brain morphology.¹ However, the mechanisms
122 underlying structural covariance are still unclear, and their genetic underpinnings are largely
123 unknown. We hypothesized that brain morphology was driven by multiple genes (i.e., polygenic)
124 collectively operating on different brain areas (i.e., pleiotropic), resulting in connected networks
125 covaried by normal aging and various disease-related processes. Along the causal pathway from
126 underlying genetics to brain morphological changes, we sought to elucidate which genetic
127 underpinnings (e.g., genes), biological processes (e.g., neurogenesis), cellular components (e.g.,

128 nuclear membrane), molecular functions (e.g., nucleic acid binding), and neuropathological
129 processes (e.g., Alzheimer's disease) might influence the formation, development, and changes
130 of structural covariance patterns in the human brain.

131 Previous neuroimaging genome-wide association studies (GWAS)^{4,5} have partially
132 investigated the abovementioned questions and expanded our understanding of the genetic
133 architecture of the human brain. However, they focused on conventional neuroanatomical
134 regions of interest (ROI) instead of data-driven PSCs. In brain imaging research, prior studies
135 have applied structural covariance analysis to elucidate underlying coordinated morphological
136 changes in brain aging and various brain diseases,¹ but have had several limitations. They often
137 relied on pre-defined neuroanatomical ROIs to construct inter- and intra-individual structural
138 covariance networks. These *a priori* ROIs might not optimally reflect the molecular-functional
139 characteristics of the brain. In addition, most population-based studies have investigated brain
140 structural covariance within a relatively limited scope, such as within relatively small samples,
141 over a relatively narrow age window (e.g., adolescence²), within a single disease (e.g.,
142 Parkinson's disease⁶), or within datasets lacking sufficient diversity in cohort characteristics or
143 MRI scanner protocols. These have been imposed, in part, by limitations in both available cohort
144 size and in the algorithmic implementation of structural covariance analysis, which has been
145 computationally restricted to modest sample sizes when investigated at full image resolution.
146 Lastly, prior studies have examined brain structural covariance at a single fixed ROI
147 resolution/scale/granularity. While the optimal scale is unknown and may differ by the question
148 of interest, the highly complex organization of the human brain may demonstrate structural
149 covariance patterns that span multiple scales.^{7,8}

150 To address this gap, we modified our previously proposed orthogonally projective non-
151 negative matrix factorization (opNMF⁹) to its stochastic counterpart, sopNMF. This adaptation
152 allowed us to train the model iteratively on large-scale neuroimaging datasets with a pre-defined
153 number of PSCs (C). Non-negative matrix factorization has gained significant attention in
154 neuroimaging due to its ability to reduce complex data into a sparse, part-based brain
155 representation by projection onto a relatively small number of components (the PSCs). NMF has
156 been shown to substantially improve interpretability and reproducibility compared to other
157 unsupervised methods, such as PCA and ICA, thanks to the non-negative constraint that
158 produces parcellation-like decompositions of complex signals. Our opNMF/sopNMF approach
159 imposed an additional orthonormality constraint⁹ (*Equation 1* in **Method 1**), further enhancing
160 sparsity and facilitating clinical interpretability. In our previous work, we applied the opNMF
161 method to 934 youths ages 8–20 to depict the coordinated growth of structural brain networks
162 during adolescence – a period characterized by extensive remodeling of the human cortex to
163 accommodate the rapid expansion of the behavioral repertoire². Remarkably, this study revealed
164 PSCs that exhibited a cortical organization closely aligned with established functional brain
165 networks, such as the well-known 7-network functional parcellation proposed by Yeo et al¹⁰.
166 Notably, this alignment emerged without prior assumptions, was data-driven and hypothesis-
167 free, and potentially reflected underlying neurobiological processes related to brain development
168 and aging. Herein, we used large-scale neuroimaging data to investigate the underlying genetic
169 determinant influencing such changes in structural covariance patterns in the human brain.

170 We examined structural covariance of regional cortical and subcortical volume in the
171 human brain using MRI from a diverse population of 50,699 people from 12 studies, 130 sites,
172 and 12 countries, comprised of cognitively healthy individuals, as well as participants with

173 various diseases/conditions over their lifespan (ages 5 through 97). Herein we present results
174 from coarse to fine scales corresponding to $C = 32, 64, 128, 256, 512,$ and 1024. We
175 hypothesized that PSCs at multiple scales could delineate the human brain's multi-factorial and
176 multi-faceted morphological landscape and genetic architecture in healthy and diseased
177 individuals. We examined the associations between these multi-scale PSCs and common genetic
178 variants at different levels ($N=8,469,833$ SNPs). In total, 617 novel genomic loci were identified;
179 key pathways (e.g., neurogenesis and reelin signaling) contributed to shaping structural
180 covariance patterns in the human brain. In addition, we leveraged PSCs at multiple scales to
181 better derive individualized imaging signatures of several diseases than any single-scale PSCs
182 using support vector machines. All experimental results and the multi-scale PSCs were
183 integrated into the MuSIC (Multi-scale Structural Imaging Covariance) atlas and made publicly
184 accessible through the BRIDGEPORT (**BR**aIn knowle**DGE PORT**Al) web portal:
185 <https://www.cbica.upenn.edu/bridgeport/>. **Table 1** provides an overview of the abbreviations
186 used in the present study.

187

188 **Table 1. Abbreviations used in the present study**

Item	Abbreviation	Item	Abbreviation
Pattern of structural covariation	PSC	Independent component analysis	ICA
Genome-wide association study	GWAS	BRaIn knowleDGE PORTal	BRIDGEPORT
Orthogonal projective non-negative matrix factorization	opNMF	Multi-scale Structural Imaging Covariance	MuSIC
Stochastic orthogonal projective non-negative matrix factorization	sopNMF	Machine learning	ML
Principal component analysis	PCA	UK Biobank	UKBB
Imaging-based coordinate SysTem for AGing and NeurodeGenerative diseases	iSTAGING	Psychosis Heterogeneity Evaluated via Dimensional Neuroimaging	PHENOM
Single nucleotide polymorphism	SNP	Region of interest	ROI
Magnetic resonance imaging	MRI	Automated anatomical labeling	AAL
MUlti-atlas region Segmentation utilizing Ensembles	MUSE	Alzheimer’s disease	AD
Spatial PAtterns for REcognition	SPARE	Support vector machine	SVM

189

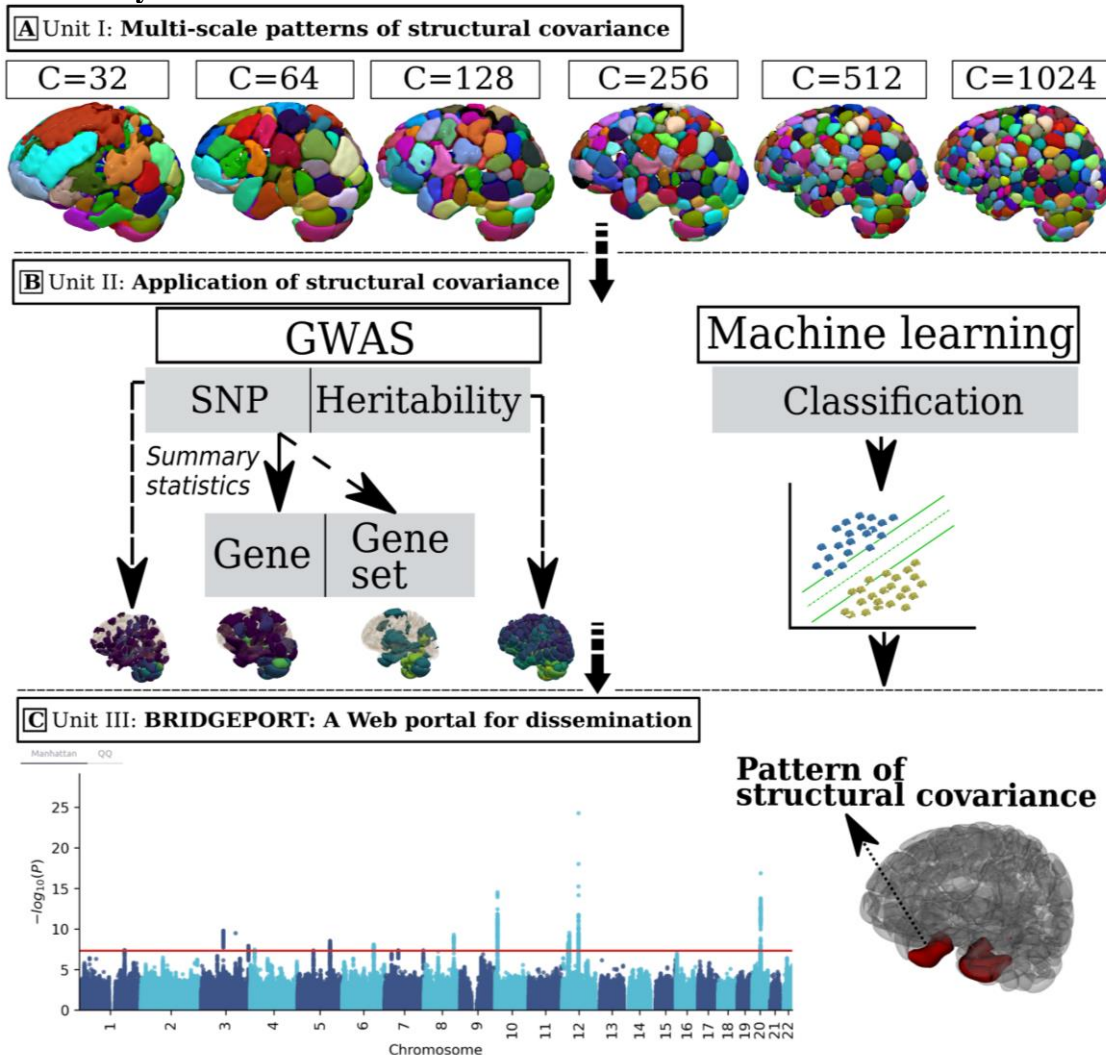
190 **Results**

191 We summarize this work in three units (I to III) outlined in **Fig. 1**. In Unit **I (Fig. 1A)**, we
192 present the stochastic orthogonally projective non-negative matrix factorization (sopNMF)
193 algorithm (**Method 1**), optimized for large-scale multivariate structural covariance analysis. The
194 sopNMF algorithm decomposes large-scale imaging data through online learning to overcome
195 the memory limitations of opNMF. A subgroup of participants with multiple disease diagnoses
196 and healthy controls (ages 5-97, training population, $N=4000$, **Method 2**) were sampled from the
197 discovery set ($N=32,440$, **Method 2**); their MRI underwent a standard imaging processing
198 pipeline (**Method 3A**). The processed images were then fit to sopNMF to derive the multi-scale
199 PSCs ($N=2003$) from the loadings of the factorization (**Method 1**). We incorporate participants
200 with various disease conditions because previous studies have demonstrated that inter-regional
201 correlated patterns (i.e., depicting a network) show variations in healthy and diseased
202 populations, albeit to a differing degree.¹¹ Multi-scale PSCs were extracted across the entire
203 population and statistically harmonized¹² (**Method 3B**). Unit **II (Fig. 1B)** investigates the
204 harmonized data for 2003 PSCs (13 PSCs have vanished in this process for $C=1024$; see **Method**
205 **1**) in two brain structural covariance analyses. Specifically, we performed *i*) GWAS (**Method 4**)
206 that sought to discover associations of PSCs at single nucleotide polymorphism (SNP), gene, or
207 gene set-level; and *ii*) pattern analysis via support vector machine (**Method 5**) to derive
208 individualized imaging signatures of several brain diseases and conditions. Unit **III (Fig. 1C)**
209 presents BRIDGEPORT, making these massive analytic resources publicly available to the
210 imaging, genomics, and machine learning communities.

211

212

213 **Figure 1: Study workflow**



214
 215 **A)** Unit I: the stochastic orthogonally projective non-negative matrix factorization (sopNMF)
 216 algorithm was applied to a large, disease-diverse population to derive multi-scale patterns of
 217 structural covariance (PSC) at different scales ($C=32, 64, 128, 256, 512,$ and 1024 ; C represents
 218 the number of PSCs). **B)** Unit II: two types of analyses were performed in this study: Genome-
 219 wide association studies (GWAS) relate each of the PSCs ($N=2003$) to common genetic variants;
 220 pattern analysis via machine learning demonstrates the utility of the multi-scale PSCs in deriving
 221 individualized imaging signatures of various brain pathologies. **C)** Unit III: BRIDGEPORT is a
 222 web portal that makes all resources publicly available for dissemination. As an illustration, a
 223 Manhattan plot for PSC ($C64-3$, the third PSC of the $C64$ atlas) and its 3D brain map are
 224 displayed.
 225

226 **Patterns of structural covariance via stochastic orthogonally projective non-negative**
 227 **matrix factorization**

228 We first validated the sopNMF algorithm by showing that it converged to the global minimum of
229 the factorization problem using the comparison population ($N=800$, **Method 2**). The sopNMF
230 algorithm achieved similar reconstruction loss and sparsity as opNMF but at reduced memory
231 demand (**Supplementary eFigure 1**). The lower memory requirements of sopNMF made it
232 possible to generate multi-scale PSCs by jointly factorizing 4000 MRIs in the training
233 population. The results of the algorithm were robust and obtained a high reproducibility index
234 (RI) (**Supplementary eMethod 2**) in several reproducibility analyses: split-sample analysis (RI
235 = 0.76 ± 0.27), split-sex analysis (RI = 0.79 ± 0.27), and leave-one-site-out analysis (RI = 0.65 -
236 0.78 for C32 PSCs) (**Supplementary eFigure 2**). We then extracted the multi-scale PSCs in the
237 discovery set ($N=32,440$) and the replication set ($N=18,259$, **Method 2**) for Unit II. These PSCs
238 succinctly capture underlying neurobiological processes across the lifespan, including the effects
239 of typical aging processes and various brain diseases. In addition, the multi-scale representation
240 constructs a hierarchy of brain structure networks (e.g., PSCs in cerebellum regions), which
241 models the human brain in a multi-scale topology.^{7,13}

242

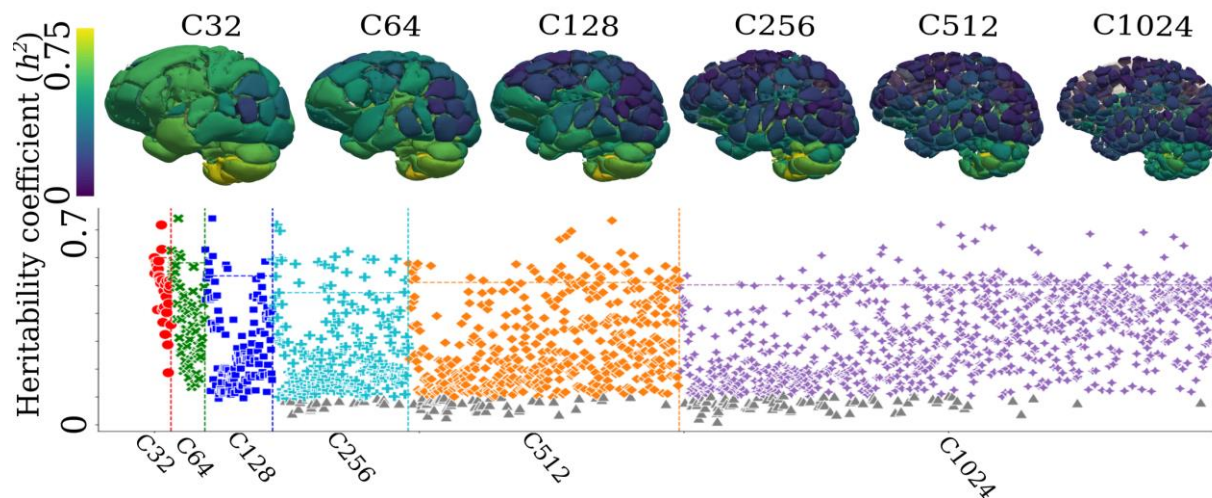
243 **Patterns of structural covariance are highly heritable**

244 The multi-scale PSCs are highly heritable ($0.05 < h^2 < 0.78$), showing high SNP-based heritability
245 estimates (h^2) (**Method 4B**) for the discovery set (**Fig. 2**). Specifically, the h^2 estimate was
246 0.49 ± 0.10 , 0.39 ± 0.14 , 0.29 ± 0.15 , 0.25 ± 0.15 , 0.27 ± 0.15 , 0.31 ± 0.15 for scales $C=32, 64, 128,$
247 $256, 512$ and 1024 of the PSCs, respectively. The Pearson correlation coefficient between the two
248 independent estimates of h^2 was $r = 0.94$ (p-value $< 10^{-6}$, between the discovery and replication
249 sets) in the UK Biobank (UKBB) data. The scatter plot of the two sets of h^2 estimates is shown in
250 **Supplementary eFigure 3**. The h^2 estimates and p-values for all PSCs are detailed in

251 **Supplementary eFile 1** (discovery set) and **eFile 2** (replication set). Our results confirm that brain
252 structure is heritable to a large extent and identify the spatial distribution of the most highly
253 heritable regions of the brain (e.g., subcortical gray matter structures and cerebellum regions).¹⁴

254

255 **Figure 2: Patterns of structural covariance are highly heritable in the human brain.**



256

257 Patterns of structural covariance (PSCs) of the human brain are highly heritable. The SNP-based
258 heritability estimates are calculated for the multi-scale PSCs at different scales (C). PSCs
259 surviving Bonferroni correction for multiple comparisons are depicted in color in the Manhattan
260 plots (gray otherwise). Each PSC's heritability estimate (h^2) was projected onto the 3D image
261 space to show a statistical map of the brain at each scale C . The dotted line indicates each scale's
262 top 10% of most heritable PSCs.

263

264 **617 novel genomic loci of patterns of structural covariance**

265 We discovered genomic locus-PSC pairwise associations (**Method 4C, Supplementary**

266 **eMethod 5**) within the discovery set and then independently replicated these associations on the

267 replication set. We found that 915 genomic loci had 3791 loci-PSC pairwise significant

268 associations with 924 PSCs after Bonferroni correction (**Method 4G**) for the number of PSCs (p -

269 value threshold per scale: $10.3 > -\log_{10}[p\text{-value}] > 8.8$) (**Supplementary eFile 3, and Fig. 3A**).

270 Our results showed that the formation of these PSCs is largely polygenic; the associated SNPs

271 might play a pleiotropic role in shaping these networks.

272 Compared to previous literature, out of the 915 genomic loci, the multi-scale PSCs
273 identified 617 novel genomic loci not previously associated with any traits or phenotypes in the
274 GWAS Catalog¹⁵ (**Supplementary eFile 4, Fig. 3B**, query date: April 5th, 2023). These novel
275 associations might indicate subtle neurobiological processes that are captured thanks to the
276 biologically relevant structural covariance expressed by sopNMF. The multi-scale PSCs
277 identified many novel associations by constraining this comparison to previous neuroimaging
278 GWAS^{12,13} using T1w MRI-derived phenotypes (e.g., regions of interest from conventional brain
279 atlases) (**Fig 3B, Supplementary eTable 3, eFile 5, 6, and 7**).

280 Our UKBB replication set analysis (**Method 4H**) demonstrated that 3638 (96%) exact
281 genomic locus-PSC associations were replicated at nominal significance ($-\log_{10}[\text{p-value}] > 1.31$),
282 2705 (72%) of which were significant after correction for multiple comparisons (**Method 4G**, -
283 $\log_{10}[\text{p-value}] > 4.27$). We present this validation in **Supplementary eFile 8** from the replication
284 set. The summary statistics, Manhattan, and QQ plots derived from the combined population
285 ($N=33,541$) are presented in BRIDGEPORT.

286 In addition to the abovementioned replication analyses, we also performed several
287 sensitivity analyses (**Supplementary eFigure 4a**). We used the GWAS results (233 significant
288 SNPs in 5 genomic loci) of the first PSC in C32 (C32_1) from the UKBB discovery set to
289 demonstrate this. First, we replicated all the 233 significant SNPs in 5 genomic loci both at the
290 nominal level ($-\log_{10}[\text{p-value}] > 1.31$), and the Bonferroni corrected p-value threshold ($-\log_{10}[\text{p-}$
291 $\text{value}] > 3.67$) using the combined discovery and replication sets ($N=33,541$) (**Supplementary**
292 **eFigure 4b**), the 20,438 participants with all ancestries in the discovery set (**Supplementary**
293 **eFigure 4c**), and the 16,743 participants in the discovery set with four additional imaging-related
294 covariates (3 parameters for the brain position in the lateral, longitudinal, and transverse

295 directions, and 1 parameter for the head motion from fMRI) (**Supplementary eFigure 4d**).

296 While replicating the results in 2386 participants with non-European ancestries, we only

297 replicated 41 SNPs (17.6%), passing the nominal significant threshold (**Supplementary eFigure**

298 **4e**). Finally, only 14 SNPs (6.4%) were replicated when replicating the results using 1481 whole-

299 genome sequencing (WGS) data from ADNI consolidated by the AI4AD consortium¹⁶

300 (**Supplementary eFigure 4f**). The low replication rates in other ancestries and independent

301 disease-specific populations are expected due to population stratification, disease-specific

302 effects, and reduced sample sizes. This further emphasizes the urge to enrich and diversify

303 genetic research with non-European ancestries and disease-specific populations.

304

305

306

307

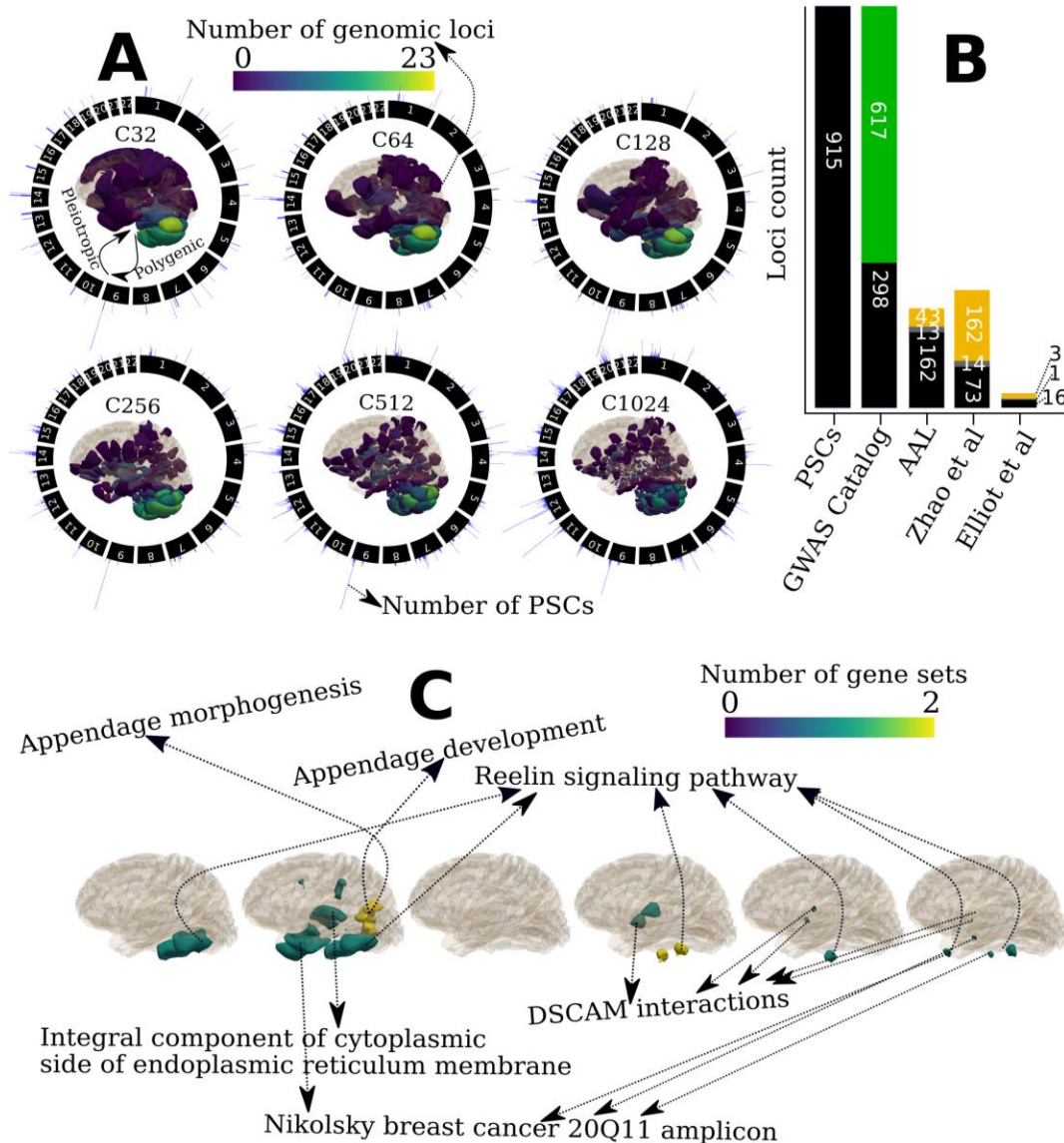
308

309

310

311

312 **Figure 3: Patterns of structural covariance highlight novel genomic loci and pathways that**
 313 **shape the human brain.**



314
 315 **A)** Patterns of structural covariance (PSC) in the human brain are polygenic: the number of
 316 genomic loci of each PSC is projected onto the image space to show a statistical brain map
 317 characterized by the number (*C*) of PSCs. In addition, common genetic variants exert pleiotropic
 318 effects on the PSCs: circular plots showed the number of associated PSCs (histograms in blue
 319 color) of each genomic loci over the entire autosomal chromosome (1-22). The histogram was
 320 plotted for the number of PSCs for each genomic locus in the circular plots. **B)** Novel genomic
 321 loci revealed by the multi-scale PSCs compared to previous findings from the GWAS Catalog,¹⁵
 322 T1-weighted MRI GWAS^{4,5}, and the AAL atlas regions of interest. The green bar indicates the
 323 617 novel genomic loci not previously associated with any clinical traits in GWAS Catalog; the
 324 black bar presents the loci identified in other studies that overlap (grey bar for loci in linkage
 325 disequilibrium) with the loci from our results; the yellow bar indicates the unique loci in other
 326 studies. **C)** Pathway enrichment analysis highlights six unique biological pathways and

327 functional categories (after Bonferroni correction for 16,768 gene sets and the number of PSCs)
328 that might influence the changes of PSCs. DSCAM: Down syndrome cell adhesion molecule.

329

330

331 **Gene set enrichment analysis highlights pathways that shape patterns of structural**

332 **covariance**

333 For gene-level associations (**Method 4D**), we discovered that 164 genes had 2489 gene-PSC

334 pairwise associations with 445 PSCs after Bonferroni correction for the number of genes and

335 PSCs (p-value threshold: $8.6 > -\log_{10}[\text{p-value}] > 7.1$) (**Supplementary eFile 9**).

336 Based on these gene-level p-values, we performed hypothesis-free gene set pathway

337 analysis using MAGMA¹⁷(**Method 4E**): a more stringent correction for multiple comparisons

338 was performed than the prioritized gene set enrichment analysis using *GENE2FUN* from FUMA

339 (**Method 4F** and **Fig. 4**). We identified that six gene set pathways had 18 gene set-PSC pairwise

340 associations with 17 PSCs after Bonferroni correction for the number of gene sets and PSCs

341 ($N=16,768$ and C from 32 to 1024, p-value threshold: $8.54 > -\log_{10}[\text{p-value}] > 7.03$) (**Fig. 3C**,

342 **Supplementary eFile 10**). These gene sets imply critical biological and molecular pathways that

343 might shape brain morphological changes and development. The reelin signaling pathway

344 regulates neuronal migration, dendritic growth, branching, spine formation, synaptogenesis, and

345 synaptic plasticity.¹⁸ The appendage morphogenesis and development pathways indicate how the

346 anatomical structures of appendages are generated, organized, and progressed over time, often

347 related to the cell adhesion pathway. These pathways elucidate how cells or tissues can be

348 organized to create a complex structure like the human brain.¹⁹ In addition, the integral

349 component of the cytoplasmic side of the endoplasmic reticulum membrane is thought to form a

350 continuous network of tubules and cisternae extending throughout neuronal dendrites and

351 axons.²⁰ The DSCAM (Down syndrome cell adhesion molecule) pathway likely functions as a

352 cell surface receptor mediating axon pathfinding. Related proteins are involved in hemophilic
353 intercellular interactions.²¹ Lastly, Nikolsky et al.²² defined genes from the breast cancer 20Q11
354 amplicon pathway that were involved in the brain might indicate the brain metastasis of breast
355 cancer, which is usually a late event with deleterious effects on the prognosis.²³ In addition,
356 previous findings^{24,25} revealed an inverse relationship between Alzheimer's disease and breast
357 cancer, which might indicate a close genetic relationship between the disease and brain
358 morphological changes mainly affecting the entorhinal cortex and hippocampus (PSC: C128_3
359 in **Fig. 4**).

360

361 **Illustrations of genetic loci and pathways forming two patterns of structural covariance**

362 To illustrate how underlying genetic underpinnings might form a specific PSC, we showcased
363 two PSCs: C32_4 for the superior cerebellum and C128_3 for the hippocampus-entorhinal
364 cortex. The two PSCs were highly heritable and polygenic in our GWAS using the entire UKBB
365 data (**Fig. 4**, $N=33,541$). We used the FUMA²⁶ online platform to perform *SNP2GENE* for
366 annotating the mapped genes and *GENE2FUNC* for prioritized gene set enrichment analyses
367 (**Method 4F**). The superior cerebellum PSC was associated with genomic loci that can be
368 mapped to 85 genes, which were enriched in many biological pathways, including psychiatric
369 disorders, biological processes, molecular functions, and cellular components (e.g., apoptotic
370 process, axon development, cellular morphogenesis, neurogenesis, and neuro differentiation).
371 For example, apoptosis – the regulated cell destruction – is a complicated process that is highly
372 involved in the development and maturation of the human brain and neurodegenerative
373 diseases.²⁷ Neurogenesis – new neuron formation – is crucial when an embryo develops and

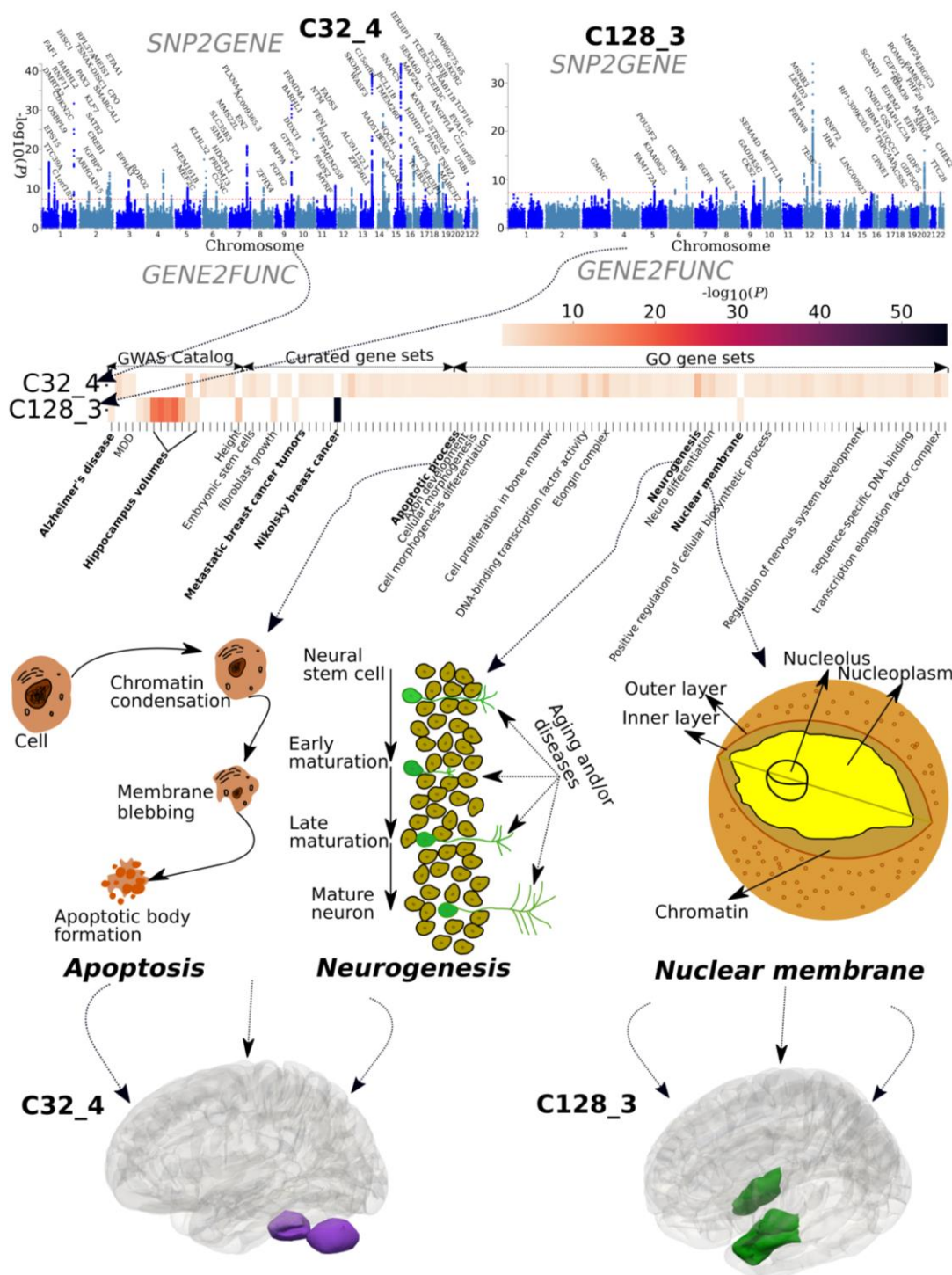
374 continues in specific brain regions throughout the lifespan.²⁸ All significant results of this
375 prioritized gene set enrichment analysis are presented in **Supplementary eFile 11.**

376 For the hippocampus-entorhinal cortex PSC, we mapped 45 genes enriched in gene sets
377 defined from GWAS Catalog, including Alzheimer's disease and brain volume derived from
378 hippocampal regions. The hippocampus and medial temporal lobe have been robust hallmarks of
379 Alzheimer's disease.²⁹ In addition, these genes were enriched in the breast cancer 20Q11
380 amplicon pathway²² and the pathway of metastatic breast cancer tumors³⁰, which might indicate
381 a specific distribution of brain metastases: the vulnerability of medial temporal lobe regions to
382 breast cancer,²³ or highlight an inverse association between Alzheimer's disease and breast
383 cancer.²⁴ Lastly, the nuclear membrane encloses the cell's nucleus – the chromosomes reside
384 inside – which is critical in cell formation activities related to gene expression and regulation. To
385 further support the overlapping genetic underpinnings between this PSC and Alzheimer's
386 disease, we calculated the genetic correlation ($r_g = -0.28$; p-value=0.01) using GWAS summary
387 statistics from the hippocampus-entorhinal cortex PSC (i.e., 33,541 people of European ancestry)
388 and a previous independent study of Alzheimer's disease³¹ (i.e., 63,926 people of European
389 ancestry) using LDSC.³² All significant results of this prioritized gene set enrichment analysis
390 are presented in **Supplementary eFile 12.**

391

392

393 **Figure 4: Illustrations of multiple genetic loci and pathways shaping specific patterns of**
 394 **structural covariance**



395 We demonstrate how underlying genomic loci and biological pathways might influence the
 396 formation, development, and changes of two specific PSCs: the 4th PSC of the C32 PSCs
 397 (C32_4) that resides in the superior part of the cerebellum and the 3rd PSC of the C128 PSCs
 398 (C128_3) that includes the bilateral hippocampus and entorhinal cortex. We first performed
 399 *SNP2GENE* to annotate the mapped genes in the Manhattan plots and then ran *GENE2FUNC* for
 400

401 the prioritized gene set enrichment analysis (**Method 4F**). The mapped genes are input genes for
402 prioritized gene set enrichment analyses. The heat map shows the significant gene sets from the
403 GWAS Catalog, curated genes, and gene ontology (GO) that survived the correction for multiple
404 comparisons. We selectively present the schematics for three pathways: apoptosis, neurogenesis,
405 and nuclear membrane function. Several other key pathways are highlighted in bold, and the 3D
406 maps of the two PSCs are presented.

407

408 **Multi-scale patterns of structural covariance derive disease-related imaging signatures**

409 We used the multi-scale PSCs from a diverse population to derive imaging signatures that reflect
410 brain development, aging, and the effects of several brain diseases. We investigate the added
411 value of the multi-scale PSCs as building blocks of imaging signatures for several brain diseases
412 and risk conditions using linear support vector machines (SVM) (**Method 5**).³³ The aim is to
413 harness machine learning to drive a clinically interpretable metric for quantifying an individual-
414 level risk to each disease category. To this end, we define the signatures as SPARE-X (Spatial
415 PAtterns for REcognition) indices, where X is the disease. For instance, SPARE-AD captures the
416 degree of expression of an imaging signature of AD-related brain atrophy, which has been shown
417 to offer diagnostic and prognostic value in prior studies.³⁴

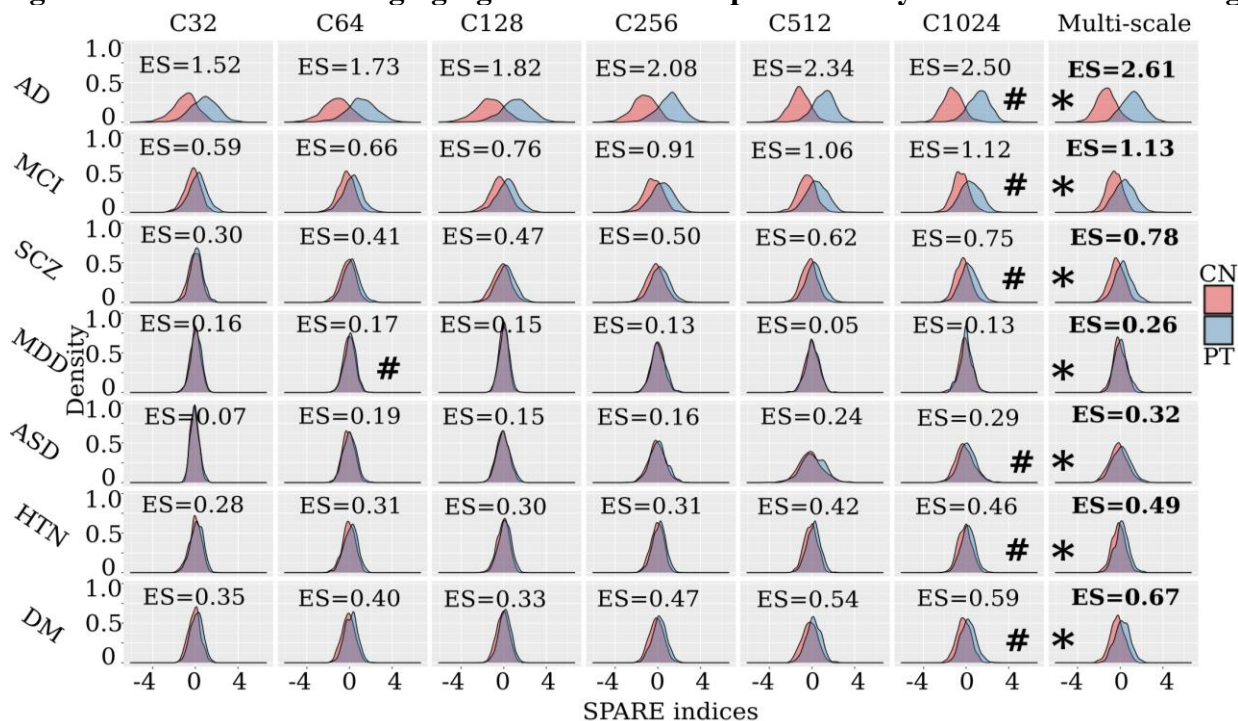
418 The most discriminative indices in our samples were SPARE-AD and SPARE-MCI (**Fig.**
419 **5, Supplementary eTable 4a** and **eFigure 5**). C=1024 achieved the best performance for the
420 single-scale analysis (e.g., AD vs. controls; balanced accuracy: 0.90 ± 0.02 ; Cohen's d : 2.50).
421 Multi-scale representations derived imaging signatures that showed the largest effect sizes to
422 classify the patients from the controls (**Fig. 5**) (e.g., AD vs. controls; balanced accuracy:
423 0.92 ± 0.02 ; Cohen's d : 2.61). PSCs obtained better classification performance than both AAL
424 (e.g., AD vs. controls; balanced accuracy: 0.82 ± 0.02 ; Cohen's d : 1.81) and voxel-wise regional
425 volumetric maps (RAVENS)³⁵ (e.g., AD vs. controls; balanced accuracy: 0.85 ± 0.02 ; Cohen's d :
426 2.04) (**Supplementary eTable 4a** and **eFigure 5**). Our classification results were higher than

427 previous baseline studies^{36,37}, which provided an open-source framework to objectively and
 428 reproducibly evaluate AD classification. Using the same cross-validation procedure and
 429 evaluation metric, they reported the highest balanced accuracy of 0.87 ± 0.02 to classify AD from
 430 healthy controls. Notably, our experiments followed good practices, employed rigorous cross-
 431 validation procedures, and avoided critical methodological flaws, such as data leakage or double-
 432 dipping (refer to critical reviews on this topic elsewhere^{36,38}).

433 To test the robustness of these SPARE indices, we performed leave-one-site-out analyses
 434 for SPARE-AD using the combined 2003 PSCs from all scales (**Supplementary eTable 4b**).
 435 Overall, holding the ADNI data out as independent test data resulted in a lower balanced
 436 accuracy (0.88 ± 0.02) compared to the other cases for AIBL (0.95 ± 0.02) and PENN data
 437 (0.95 ± 0.02). The mean balanced accuracy (0.91 ± 0.02) aligns with the nested cross-validated
 438 results using the full sample (**Fig. 5**).

439

440 **Figure 5: Individualized imaging signatures based on pattern analysis via machine learning.**



441

442 Imaging signatures (SPARE indices) of brain diseases, derived via supervised machine learning
443 models, are more distinctive when formed from multi-scale PSCs than single-scale PSCs. The
444 kernel density estimate plot depicts the distribution of the patient group (blue) in comparison to
445 the healthy control group (red), reflecting the discriminative power of the diagnosis-specific
446 SPARE (imaging signature) indices. We computed Cohen's d for each SPARE index between
447 groups to present the effect size of its discrimination power. * represents the model with the
448 largest Cohen's d for each SPARE index to separate the control vs. patient groups; # represents
449 the model with the best performance with single-scale PSCs. Our results demonstrate that the
450 multi-scale PSCs generally achieve the largest discriminative effect sizes (ES) (**Supplementary**
451 **eTable 4a**). As a reference, Cohen's d of ≥ 0.2 , ≥ 0.5 , and ≥ 0.8 , respectively, refer to small,
452 moderate, and large effect sizes.
453

454 **BRIDGEPORT: bridging knowledge across patterns of structural covariance, genomics,**
455 **and clinical phenotypes**

456 We integrated our experimental results and the MuSIC atlas into the BRIDGEPORT online web
457 portal. This online tool allows researchers to interactively browse the MuSIC atlas in 3D, query
458 our experimental results via variants or PSCs, and download the GWAS summary statistics for
459 further analyses. In addition, we allow users to search via conventional brain anatomical terms
460 (e.g., the right thalamus proper) by automatically annotating traditional anatomic atlas ROIs,
461 specifically from the MUSE atlas³⁹ (**Supplementary eTable 5**), to MuSIC PSCs based on their
462 degree of overlaps (**Supplementary eFigure 6**). Open-source software dedicated to image
463 processing,³⁹ genetic quality check protocols, MuSIC generation with sopNMF, and machine
464 learning³⁶ is also publicly available (see Code Availability for details).

465 **Discussion**

466 The current study investigates patterns of structural covariance in the human brain at multiple
467 scales from a large population of 50,699 people and, importantly, a very diverse cohort allowing
468 us to capture patterns of structural covariance emanating from normal and abnormal brain
469 development and aging, as well as from several brain diseases. Through extensive examination
470 of the genetic architecture of these multi-scale PSCs, we confirmed genetic hits from previous
471 T1-weighted MRI GWAS and, more importantly, identified 617 novel genomic loci and
472 molecular and biological pathways that collectively influence brain morphological changes and
473 development over the lifespan. Using a hypothesis-free, data-driven approach to first derive these
474 PSCs using brain MRIs, we then uncovered their genetic underpinnings and further showed their
475 potential as building blocks to predict various diseases. All experimental results and code are
476 encapsulated and publicly available in BRIDGEPORT for dissemination:
477 <https://www.cbica.upenn.edu/bridgeport/>, to enable various neuroscience studies to investigate
478 these structural covariance patterns in diverse contexts. Together, the current study highlighted
479 the adoption of machine learning methods in brain imaging genomics and deepened our
480 understanding of the genetic architecture of the human brain.

481 Our findings reveal new insights into genetic underpinnings that influence structural
482 covariance patterns in the human brain. Brain morphological development and changes are
483 largely polygenic and heritable, and previous neuroimaging GWAS has not fully uncovered this
484 genetic landscape. In contrast, genetic variants, as well as environmental, aging, and disease
485 effects, exert pleiotropic effects in shaping morphological changes in different brain regions
486 through specific biological pathways. The mechanisms underlying brain structural covariance are
487 not yet fully understood. They may involve an interplay between common underlying genetic

488 factors, shared susceptibility to aging, and various brain pathologies, which affect brain growth
489 or degeneration in coordinated brain morphological changes.¹ Our data-driven, multi-scale PSCs
490 identify the hierarchical structure of the brain under the principle of structural covariance and are
491 associated with genetic factors at different levels, including SNPs, genes, and gene set pathways.
492 These 617 novel genomic loci, as well as those previously identified, collectively shape brain
493 morphological changes through many key biological and molecular pathways. These pathways
494 are widely involved in reelin signaling, apoptotic processes, axonal development, cellular
495 morphogenesis, neurogenesis, and neuro differentiation,^{27,28} which may collectively influence the
496 formation of structural covariance patterns in the brain. Strikingly, pathways involved in breast
497 cancer shared overlapping genetic underpinnings evidenced in our MAGMA-based and
498 prioritized (*GENE2FUNC*) gene set enrichment analyses (**Fig. 3C** and **Fig. 4**), which included
499 specific pathways involved in breast cancer and metastatic breast cancer tumors. One previous
500 study showed that common genes might mediate breast cancer metastasis to the brain,²³ and a
501 later study further corroborated that the metastatic spread of breast cancer to other organs
502 (including the brain) accelerated during sleep in both mouse and human models.⁴⁰ We further
503 showcased that this brain metastasis of breast cancer might be associated with specific
504 neuropathologic processes, which were captured by PSCs data driven by Alzheimer's disease-
505 related neuropathology. For example, the hippocampus-entorhinal cortex PSC (C128_3, **Fig. 4**)
506 connected the bilateral hippocampus and medial temporal lobe – the salient hallmark of
507 Alzheimer's disease. Our gene set enrichment analysis results further support this claim: the
508 genes were enriched in the gene sets of Alzheimer's disease and breast cancer (**Fig. 4**). Previous
509 research^{24,25} also found an inverse association between Alzheimer's disease and breast cancer. In
510 addition, PSCs from the cerebellum were the most genetically influenced brain regions,

511 consistent with previous neuroimaging GWAS.^{4,5} The cerebral cortex has been thought to largely
512 contribute to the unique mental abilities of humans. However, the cerebellum may also be
513 associated with a much more comprehensive range of complex cognitive functions and brain
514 diseases than initially thought.⁴¹ Our results confirmed that many genetic substrates might
515 support different molecular pathways, resulting in cerebellar functional organization, high-order
516 functions, and dysfunctions in various brain disorders.

517 The current work demonstrates that appropriate machine learning analytics can be used to
518 shed new light on brain imaging genetics. Previous neuroimaging GWAS leveraged multimodal
519 imaging-derived phenotypes from conventional brain atlases^{4,5} (e.g., the AAL atlas). In contrast,
520 multi-scale PSCs are purely data-driven and likely to reflect the dynamics of underlying normal
521 and pathological neurobiological processes giving rise to structural covariance. The diverse
522 training sample from which the PSCs were derived, including healthy and diseased individuals of
523 a wide age range, enriched the diversity of such neurobiological processes influencing the PSCs.
524 In addition, modeling structural covariance at multiple scales (i.e., multi-scale PSCs) indicated
525 that disease effects could be robustly and complementarily identified across scales (**Fig. 5**),
526 concordant with the paradigm of multi-scale brain modeling.¹³ Imaging signatures of brain
527 diseases, derived via supervised machine learning models, were consistently more distinctive
528 when formed from multi-scale PSCs than single-scale PSCs. Multivariate learning techniques
529 have gained significant prominence in neuroimaging and have recently attracted considerable
530 attention in the domain of imaging genomics. These methods have proven valuable for analyzing
531 complex and high-dimensional data, facilitating the exploration of relationships between imaging
532 features and genetic factors. For instance, the MOSTest, a multivariate GWAS approach,
533 preserves correlation structure among phenotypes via permutation on each SNP and derives a

534 genotype vector for testing the association across all phenotypes⁴². A separate study by Soheili-
535 Nezhad et al. demonstrated that genetic components obtained through PCA or ICA applied to
536 neuroimaging GWAS summary statistics exhibited greater reproducibility than raw univariate
537 GWAS effect sizes⁴³. A recent study utilized a CNN-based autoencoder to discover new
538 phenotypes and identify numerous novel genetic signals⁴⁴. Despite the effectiveness of these
539 multivariate approaches in GWAS, they typically conduct phenotype engineering before
540 performing GWAS without explicitly incorporating imaging genetic associations during the
541 modeling process. Yang et al. recently conducted a study that employed generative adversarial
542 networks (termed GeneSGAN⁴⁵) to integrate imaging and genetic variations within the modeling
543 framework to address this limitation. By incorporating both modalities, their approach aimed to
544 capture the complexity and heterogeneity of disease manifestations.

545 MuSIC – with the strengths of being data-driven, multi-scale, and disease-effect
546 informative – contributes to the century-old quest for a "universal" atlas in brain cartography⁴⁶
547 and is highly complementary to previously proposed brain atlases. For instance, Chen and
548 colleagues⁴⁷ used a semi-automated fuzzy clustering technique with MRI data from 406 twins
549 and parcellated the cortical surface area into a genetic covariance-informative brain atlas; MuSIC
550 was data-driven by structural covariance. Glasser and colleagues⁴⁸ adopted a semi-automated
551 parcellation procedure to create a multimodal cortex atlas from 210 healthy individuals.
552 Although this method successfully integrates multimodal information from cortical folding,
553 myelination, and functional connectivity, this semi-automatic approach requires significant
554 resources, some with limited resolution. MuSIC allows flexible, multiple scales for delineating
555 macroscopic brain topology; including patient samples exposes the model to sources of
556 variability that may not be visible in healthy controls. Another pioneering endeavor is the Allen

557 Brain Atlas project,⁴⁹ whose overarching goals of mapping the human brain to gene expression
558 data via existing conventional atlases, identifying local gene expression patterns across the brain
559 in a few individuals, and deepening our understanding of the human brain's differential genetic
560 architecture, are complementary to ours – characterizing the global genetic architecture of the
561 human brain, emphasizing pathogenic variability and morphological heterogeneity.

562 Bridging knowledge across the brain imaging, genomics, and machine learning
563 communities is another pivotal contribution of this work. BRIDGEPORT provides a platform to
564 lower the entry barrier for whole-brain genetic-structural analyses, foster interdisciplinary
565 communication, and advocate for research reproducibility.^{36,50-53} The current study demonstrates
566 the broad applicability of this large-scale, multi-omics platform across a spectrum of
567 neurodegenerative and neuropsychiatric diseases.

568 The present study has certain limitations. Firstly, the sopNMF method utilized in brain
569 parcellation considers only imaging structural covariance and overlooks the genetic determinants
570 contributing to forming these structural networks, as indicated by our GWAS findings.
571 Consequently, further investigations are needed to integrate imaging and genetics into brain
572 parcellation. Additionally, it is important to note that our GWAS analyses primarily involved
573 participants of European ancestry. To enhance genetic findings for underrepresented ethnic
574 groups, future studies should prioritize the inclusion of diverse ancestral backgrounds, thereby
575 promoting a more comprehensive understanding of the genetic underpinnings across different
576 populations.

577 **Methods**

578 **Method 1: Structural covariance patterns via stochastic orthogonally projective non-** 579 **negative matrix factorization**

580 The sopNMF algorithm is a stochastic approximation built and extended based on opNMF^{9,54}.

581 We consider a dataset of n MR images and d voxels per image. We represent the data as a

582 matrix \mathbf{X} where each column corresponds to a flattened image: $\mathbf{X} = [x_1, x_2, \dots, x_n]$, $\mathbf{X} \in \mathbb{R}_{\geq 0}^{d \times n}$.

583 The sopNMF algorithm factorizes \mathbf{X} into two low-rank (r) matrices $\mathbf{W} \in \mathbb{R}_{\geq 0}^{d \times r}$ and $\mathbf{H} \in \mathbb{R}_{\geq 0}^{r \times n}$

584 under the constraints of non-negativity and column-orthonormality. Using the Frobenius norm,

585 the loss of this factorization problem can be formulated as

$$586 \quad \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2$$

$$587 \quad \text{subject to } \mathbf{H} = \mathbf{W}^T \mathbf{X}, \mathbf{W} \geq 0 \text{ and } \mathbf{W}^T \mathbf{W} = \mathbf{I} \quad (1)$$

588 where \mathbf{I} stands for the identity matrix. The columns $w_i \in \mathbb{R}^d$, $\|w_i\|^2 = 1, \forall i \in \{1..r\}$ of the so-

589 called component matrix $\mathbf{W} = [w_1, w_2, \dots, w_r]$ are part-based representations promoting sparsity

590 in data in this lower-dimensional subspace. From this perspective, the loading coefficient matrix

591 \mathbf{H} represents the importance (weights) of each feature above for a given image. Instead of

592 optimizing the non-convex problem in a batch learning paradigm (i.e., reading all images into

593 memory) as opNMF,⁹ sopNMF subsamples the number of images at each iteration, thereby

594 significantly reducing its memory demand by randomly drawing data batches $\mathbf{X}_b \in \mathbb{R}_{\geq 0}^{d \times b}$ of $b \leq$

595 n images (b is the batch size; $b=32$ was used in the current analyses); this is done without

596 replacement so that all data goes through the model once ($\lceil n/b \rceil$). In this case, the updating rule

597 can be rewritten as

$$598 \quad \mathbf{W}_{t+1} = \mathbf{W}_t \frac{(\mathbf{X}_b \mathbf{X}_b^T \mathbf{W})_t}{(\mathbf{W} \mathbf{W}^T \mathbf{X}_b \mathbf{X}_b^T \mathbf{W})_t} \quad (2)$$

599 We calculate the loss on the entire dataset at the end of each epoch (i.e., the loss is incremental
600 across all batches) with the following expression:

$$601 \quad \sum_{i=1}^{\lceil n/b \rceil} \|\mathbf{X}_{b_i} - \mathbf{W}\mathbf{W}^T \mathbf{X}_{b_i}\|_F^2 \quad (3)$$

602 We evaluated the training loss and the sparsity of \mathbf{W} at the end of each iteration. Moreover, early
603 stopping was implemented to improve training efficiency and alleviate overfitting. We
604 summarize the sopNMF algorithm in **Supplementary Algorithm 1**. An empirical comparison
605 between sopNMF and opNMF is detailed in **Supplementary eMethod 1**.

606 We applied sopNMF to the training population ($N=4000$). The component matrix \mathbf{W} was
607 sparse after the algorithm converged with a pre-defined maximum number of epochs (100 by
608 default) with an early stopping criterion. To build the MuSIC atlas, we clustered each voxel
609 (row-wise) into one of the r features/PSCs as follows:

$$610 \quad \mathbf{M}_j = \operatorname{argmax}_k (\mathbf{W}_{j,k}) \quad (4)$$

611 where \mathbf{M} is a d -dimensional vector and $j \in \{1..d\}$. The j -th element of \mathbf{M} equals k if $\mathbf{W}_{j,k}$ is the
612 maximum value of the j -th row. Intuitively, \mathbf{M} indicates which of the r PSCs each voxel belongs
613 to. We finally projected the vector $\mathbf{M} \in \mathbb{R}_{\geq 0}^d$ into the original image space to visualize each PSC
614 of the MuSIC atlas (**Fig. 1**). Of note, 13 PSCs have vanished in this process for $C=1024$: all 0 for
615 these 13 vectors.

616

617 **Method 2: Study population**

618 We consolidated a large-scale multimodal consortium ($N=50,699$) consisting of imaging,
619 cognition, and genetic data from 12 studies, 130 sites, and 12 countries (**Supplementary eTable**
620 **1**): the Alzheimer's Disease Neuroimaging Initiative⁵⁵ (ADNI) ($N=1765$); the UK Biobank⁵⁶

621 (UKBB) ($N=39,564$); the Australian Imaging, Biomarker, and Lifestyle study of aging⁵⁷ (AIBL)
622 ($N=830$); the Biomarkers of Cognitive Decline Among Normal Individuals in the Johns Hopkins
623 University⁵⁸ (BIOCARD) ($N=288$); the Baltimore Longitudinal Study of Aging^{59,60} (BLSA)
624 ($N=1114$); the Coronary Artery Risk Development in Young Adults⁶¹ (CARDIA) ($N=892$); the
625 Open Access Series of Imaging Studies⁶² (OASIS) ($N=983$), PENN ($N=807$); the Women's
626 Health Initiative Memory Study⁶³ (WHIMS) ($N=995$), the Wisconsin Registry for Alzheimer's
627 Prevention⁶⁴ (WRAP) ($N=116$); the Psychosis Heterogeneity (evaluated) via dimensional
628 Neuroimaging⁶⁵ (PHENOM) ($N=2125$); and the Autism Brain Imaging Data Exchange⁶⁶
629 (ABIDE) ($N=1220$).

630 We present the demographic information of the population under study in
631 **Supplementary eTable 1**. This large-scale consortium reflects the diversity of MRI scans over
632 different races, disease conditions, and ages over the lifespan. To be concise, we defined four
633 populations or data sets per analysis across the paper:

- 634 • Discovery set: It consists of a multi-disease and lifespan population that includes
635 participants from all 12 studies ($N=32,440$). Note that this population does not contain
636 the entire UKBB population but only our first download (July 2017, $N=21,305$).
- 637 • Replication set: We held 18,259 participants from the UKBB dataset to replicate the
638 GWAS results. We took these data from our second download of the UKBB dataset
639 (November 2021, $N=18,259$).
- 640 • Training population: We randomly drew 250 patients (PT), including AD, MCI, SCZ,
641 ASD, MDD, HTN (hypertension), DM (diabetes mellitus), and 250 healthy controls
642 (CN) per decade from the discovery set, ensuring that the PT and CN groups have
643 similar sex, study and age distributions. The resulting set of 4000 imaging data was used

644 to generate the MuSIC atlas with the sopNMF algorithm. The rationale is to maximize
645 variability across a balanced sample of multiple diseases or risk conditions, age, and
646 study protocols rather than overfit the entire data by including all images in training.

- 647 • Comparison population: To validate sopNMF compared to the original opNMF
648 algorithm, we randomly subsampled 800 participants from the training population (100
649 per decade for balanced CN and PT). For this scale of sample size, opNMF can load all
650 images into memory for batch learning.⁶⁷

651 All individual studies were approved by their local corresponding Institutional Review
652 Boards (IRB). The iSTAGING and PHENOM consortia consolidated all individual imaging and
653 clinical data; imputed genotype data were directly downloaded from the UKBB website. Data
654 from the UKBB for this project pertains to application 35148. For iSTAGING, the IRB at the
655 University of Pennsylvania (protocol number: 825722) reviewed the research proposal on
656 August 31st, 2016, and updated it on August 31st, 2022. No human subjects were recruited or
657 scanned. Existing de-identified data will be used in this mega-analysis study pooling data from
658 17 studies: BLSA, ADNI1, ADNI2, ADNI3, ACCORD-MIND, LookAhead, SPRINT,
659 CARDIA, MESA, SHIP, BIOCARD, WRAP, Penn-ADC, WHIMS-MRI, AIBL, OASIS,
660 UKBB, MESA, HANDLS. For PHENOM, the IRB at the University of Pennsylvania (protocol
661 number: 828077) reviewed the research proposal on August 19th, 2017. No human subjects were
662 recruited or scanned. Existing de-identified data will be used in this meta-analysis study pooling
663 data from 10 studies at Penn, Ludwigg-Maximilian University of Munich, Kings College-
664 London, University of Utrecht, University of Melbourne, University of Cantabria, University of
665 Sao Paulo, Xijing Hospital Shaanxi, Tianjin Anning Hospital, and Institute of Mental Health
666 Peking University.

667

668 **Method 3: Image processing and statistical harmonization**

669 **(A): Image processing.** Images that passed the quality check (**Supplementary eMethod 4**) were
670 first corrected for magnetic field intensity inhomogeneity.⁶⁸ Voxel-wise regional volumetric
671 maps (RAVENS)³⁵ for each tissue volume were then generated by using a registration method to
672 spatially align the skull-stripped images to a template in MNI-space.⁶⁹ We applied sopNMF to
673 the RAVENS maps to derive MuSIC.

674

675 **(B): Statistical harmonization of MuSIC PSCs:** We applied MuSIC to the entire population
676 ($N=50,699$) to extract the multi-scale PSCs. Specifically, MuSIC was applied to each individual's
677 RAVENS gray matter map to extract the sum of brain volume in each PSC. Subsequently, the
678 PSCs were statistically harmonized by an extensively validated approach, i.e., ComBat-GAM¹²
679 (**Supplementary eMethod 3**) to account for site-related differences in the imaging data. After
680 harmonization, the PSCs were normally distributed (skewness = 0.11 ± 0.17 , and kurtosis =
681 0.67 ± 0.68) (**Supplementary eFigure 7A and B**). To alleviate the potential violation of normal
682 distribution in downstream statistical learning, we quantile-transformed all PSCs. In agreement
683 with the literature,^{70,71} males were found to have larger brain volumes than females on average
684 (**Supplementary eFigure 7C**). Overall, the Combat-GAM model slightly improved data
685 normality across sites (**Supplementary eFigure 7E-H**). The AAL ROIs underwent the same
686 statistical harmonization procedure.

687

688 **Method 4: Genetic analyses**

689 Genetic analyses were restricted to the discovery and replication set from UKBB (**Method 2**).
690 We processed the array genotyping and imputed genetic data (SNPs). The two data sets went
691 through a "best-practice" imaging-genetics quality check (QC) protocol (**Method 4A**) and were
692 restricted to participants of European ancestry. This resulted in 18,052 participants and 8,430,655
693 SNPs for the discovery set and 15,243 participants and 8,470,709 SNPs for the replication set.
694 We reperformed the genetic QC and genetic analyses for the combined populations for
695 BRIDGEPORT, resulting in 33,541 participants and 8,469,833 SNPs. **Method 4G** details the
696 correction for multiple comparisons throughout our analyses.

697
698 **(A): Genetic data quality check protocol.** First, we excluded related individuals (up to 2nd-
699 degree) from the complete UKBB sample ($N=488,377$) using the KING software for family
700 relationship inference.⁷² We then removed duplicated variants from all 22 autosomal
701 chromosomes. We also excluded individuals for whom either imaging or genetic data were not
702 available. Individuals whose genetically identified sex did not match their self-acknowledged sex
703 were removed. Other excluding criteria were: i) individuals with more than 3% of missing
704 genotypes; ii) variants with minor allele frequency (MAF) of less than 1%; iii) variants with larger
705 than 3% missing genotyping rate; iv) variants that failed the Hardy-Weinberg test at 1×10^{-10} . To
706 adjust for population stratification,⁷³ we derived the first 40 genetic principle components (PC)
707 using the FlashPCA software⁷⁴. The genetic pipeline was also described elsewhere⁷⁵.

708
709 **(B): Heritability estimates and genome-wide association analysis.** We estimated the SNP-
710 based heritability explained by all autosomal genetic variants using GCTA-GREML.⁷⁶ We
711 adjusted for confounders of age (at imaging), age-squared, sex, age-sex interaction, age-squared-

712 sex interaction, ICV, and the first 40 genetic principal components (PC), guided by a previous
713 neuroimaging GWAS⁴. In addition, Elliot et al.⁵ investigated more than 200 confounders in
714 another study. Therefore, our sensitivity analyses included four additional imaging-related
715 covariates (i.e., brain positions and head motion). One-side likelihood ratio tests were performed
716 to derive the heritability estimates. In GWAS, we performed a linear regression for each PSC
717 and included the same covariates as in the heritability estimates using PLINK.⁷⁷

718
719 **(C): Identification of novel genomic loci.** Using PLINK, we clumped the GWAS summary
720 statistics based on their linkage disequilibrium to identify the genomic loci (see **Supplementary**
721 **eMethod 5** for the definition of the index, candidate, independent significant, lead SNP, and
722 genomic locus). In particular, the threshold for significance was set to 5×10^{-8} (*clump-p1*) for the
723 index SNPs and 0.05 (*clump-p2*) for the **candidate SNPs**. The threshold for linkage
724 disequilibrium-based clumping was set to 0.60 (*clump-r2*) for **independent significant SNPs**
725 and 0.10 for lead SNPs. The linkage disequilibrium physical-distance threshold was 250
726 kilobases (*clump-kb*). **Genomic loci** consider linkage disequilibrium (within 250 kilobases) when
727 interpreting the association results. The GWASRAPIDD⁷⁸ package (version: 0.99.14) was then
728 used to query the genomic loci for any previously-reported associations with clinical phenotypes
729 documented in the NHGRI-EBI GWAS Catalog¹⁵ (p-value $< 1.0 \times 10^{-5}$, default inclusion value of
730 GWAS Catalog). We defined a genomic locus as **novel** when it was not present in GWAS
731 Catalog (query date: April 5th, 2023).

732
733 **(D): Gene-level associations with MAGMA.** We performed gene-level association analysis
734 using MAGMA.¹⁷ First, gene annotation was performed to map the SNPs (reference variant

735 location from Phase 3 of 1,000 Genomes for European ancestry) to genes (human genome Build
736 37) according to their physical positions. The second step was to perform the gene analysis based
737 on the GWAS summary statistics to obtain gene-level p-values between the pairwise 2003 PSCs
738 and the 18,097 protein-encoding genes containing valid SNPs.

739

740 **(E): Hypothesis-free gene set enrichment analysis with MAGMA.** Using the gene-level
741 association p-values, we performed gene set enrichment analysis using MAGMA. Gene sets
742 were obtained from Molecular Signatures Database (MsigDB, v7.5.1),⁷⁹ including 6366 curated
743 gene sets and 10,402 Gene Ontology (GO) terms. All other parameters were set by default for
744 MAGMA. This hypothesis-free analysis resulted in a more stringent correction for multiple
745 comparisons (i.e., by the total number of tested genes and PSCs) than the FUMA-prioritized
746 gene set enrichment analysis (see below F).

747

748 **(F): FUMA analyses for the illustrations of specific PSCs.** In *SNP2GENE*, three different
749 methods were used to map the SNPs to genes. First, positional mapping maps SNPs to genes if
750 the SNPs are physically located inside a gene (a 10 kb window by default). Second, expression
751 quantitative trait loci (eQTL) mapping maps SNPs to genes showing a significant eQTL
752 association. Lastly, chromatin interaction mapping maps SNPs to genes when there is a
753 significant chromatin interaction between the disease-associated regions and nearby or distant
754 genes.²⁶ In addition, *GENE2FUNC* studies the expression of prioritized genes and tests for the
755 enrichment of the set of genes in pre-defined pathways. We used the mapped genes as prioritized
756 genes. The background genes were specified as all genes in FUMA, and all other parameters
757 were set by default. We only reported gene sets with adjusted p-value < 0.05.

758

759 **(G): Correction for multiple comparisons.** We practiced a conservative procedure to control
760 for the multiple comparisons. In the case of GWAS, we chose the default genome-wide
761 significant threshold (5.0×10^{-8} , and 0.05 for all other analyses) and independently adjusted for
762 multiple comparisons (Bonferroni methods) at each scale by the number of PSCs. We corrected
763 the p-values for the number of phenotypes ($N=6$) for genetic correlation analyses. We adjusted
764 the p-values for the number of PSCs at each scale for heritability estimates. For gene analyses,
765 we controlled for both the number of PSCs at each scale and the number of genes. We adopted
766 these strategies per analysis to correct the multiple comparisons because PSCs of different scales
767 are likely hierarchical and correlated – avoiding the potential of "overcorrection".

768

769 **(H): Replication analysis for genome-wide association studies.** We performed GWAS by
770 fitting the same linear regressing models as the discovery set. Also, following the same
771 procedure for consistency, we corrected the multiple comparisons using the Bonferroni method.
772 We corrected it for the number of genomic loci ($N=915$) found in the discovery set with a
773 nominal p-value of 0.05, which thereby resulted in a stringent test with an equivalent p-value
774 threshold of 3.1×10^{-5} (i.e., $-\log_{10}[\text{p-value}] = 4.27$). We performed a replication for the 915
775 genomic loci, but, in reality, SNPs in linkage disequilibrium with the genomic loci are likely
776 highly significant.

777

778 **Method 5: Pattern analysis via machine learning for individualized imaging signatures**

779 SPARE-AD captures the degree of expression of an imaging signature of AD, and prior studies
780 have shown its diagnostic and prognostic values.³⁴ Here, we extended the concept of the SPARE

781 imaging signature to multiple diseases (SPARE-X, X represents disease diagnoses). Following
782 our reproducible open-source framework³⁷, we performed nested cross-validation
783 (**Supplementary eMethod 6**) for the machine learning models and derived imaging signatures to
784 quantify individualized disease vulnerability.

785 **SPARE indices.** MuSIC PSCs were fit into a linear support vector machine (SVM) to derive
786 SPARE-AD, MCI, SCZ, DM, HTN, MDD, and ASD. Specifically, the SVM aims to classify the
787 patient group (e.g., AD) from the control group and outputs a continuous variable (i.e., the
788 SPARE indices), which indicates the proximity of each participant to the hyperplane in either the
789 patient or control space. We compared the classification performance using different sets of
790 features: i) the single-scale PSC from 32 to 1024, ii) the multi-scale PSCs by combining all
791 features (with and without feature selections embedded in the CV); iii) the ROIs from the AAL
792 atlas; and iv) voxel-wise RAVENS maps. The samples selected for each task are presented in
793 **Supplementary eTable 2.**

794 No statistical methods were used to predetermine the sample size. The experiments were
795 not randomized, and the investigators were not blinded to allocation during experiments and
796 outcome assessment.

797 **Data Availability**

798 The GWAS summary statistics corresponding to this study are publicly available on the
799 BRIDGEPORT web portal (<https://www.cbica.upenn.edu/bridgeport/>) and the MEDICINE web
800 portal (<http://labs.loni.usc.edu/medicine/>). The GWAS summary statistics used in the genetic
801 correlation analyses were fetched from the GWAS Catalog platform
802 (<https://www.ebi.ac.uk/gwas>), although each study provided the original links; The GWAS
803 Catalog platform was used to query if the SNPs identified by MuSIC were previously reported.

804 **Code Availability**

805 The software and resources used in this study are all publicly available:

- 806 • sopNMF: <https://pypi.org/project/sopnmf/>, MuSIC, and sopNMF (developed for this
807 study)
- 808 • BRIDGEPORT: <https://www.cbica.upenn.edu/bridgeport/>, (developed for this study)
- 809 • MLNI: <https://pypi.org/project/mlni/>, machine learning (developed for this study)
- 810 • MUSE: <https://www.med.upenn.edu/sbia/muse.html>, image preprocessing
- 811 • PLINK: <https://www.cog-genomics.org/plink/>, GWAS
- 812 • GCTA: <https://yanglab.westlake.edu.cn/software/gcta/#Overview>, heritability estimates
- 813 • LDSC: <https://github.com/bulik/ldsc>, genetic correlation estimates
- 814 • MAGMA: <https://ctg.cncr.nl/software/magma>, gene analysis
- 815 • GWASRAPIDD: <https://rmagno.eu/gwasrapidd/articles/gwasrapidd.html>, GWAS
816 Catalog query
- 817 • MsigDB: <https://www.gsea-msigdb.org/gsea/msigdb/>, gene sets database

818

819 **Competing Interests**

820 DAW served as Site PI for studies by Biogen, Merck, and Eli Lilly/Avid. He has received
821 consulting fees from GE Healthcare and Neuronix. He is on the DSMB for a trial sponsored by
822 Functional Neuromodulation. AJS receives support from multiple NIH grants (P30 AG010133,
823 P30 AG072976, R01 AG019771, R01 AG057739, U01 AG024904, R01 LM013463, R01
824 AG068193, T32 AG071444, and U01 AG068057 and U01 AG072177). He has also received
825 support from Avid Radiopharmaceuticals, a subsidiary of Eli Lilly (in-kind contribution of PET
826 tracer precursor); Bayer Oncology (Scientific Advisory Board); Eisai (Scientific Advisory
827 Board); Siemens Medical Solutions USA, Inc. (Dementia Advisory Board); Springer-Nature
828 Publishing (Editorial Office Support as Editor-in-Chief, Brain Imaging, and Behavior). OC
829 reports having received consulting fees from AskBio (2020) and Therapanacea (2022), having
830 received payments for writing a lay audience short paper from Expression Santé (2019), and that
831 his laboratory has received grants (paid to the institution) from Qynapse (2017-present).
832 Members of his laboratory have co-supervised a Ph.D. thesis with myBrainTechnologies (2016-
833 2019) and with Qynapse (2017-present). OC's spouse is an employee and holds stock options of
834 myBrainTechnologies (2015-present). OC has a patent registered at the International Bureau of
835 the World Intellectual Property Organization (PCT/IB2016/0526993, Schiratti J-B, Allasonniere
836 S, Colliot O, Durrleman S, A method for determining the temporal progression of a biological
837 phenomenon and associated methods and devices) (2017). ME receives support from multiple
838 NIH grants, the Alzheimer's Association, and the Alzheimer's Therapeutic Research Institute.
839 MZ serves as a consultant and/or speaker for the following pharmaceutical companies:
840 Eurofarma, Lundbeck, Abbott, Greencare, Myralis, and Elleven Healthcare.

841

842 **Authors' contributions**

843 Dr. Wen takes full responsibility for the integrity of the data and the accuracy of the data analysis.

844 *Study concept and design:* Wen, Davatzikos

845 *Acquisition, analysis, or interpretation of data:* Wen, Nasrallah, Davatzikos, Abdulkadi,

846 Satterthwaite, Dazzan, Kahn, Schnack, Zanetti, Meisenzahl, Busatto, Crespo-Facorro, Pantelis,

847 Wood, Zhuo, Koutsouleris, Wittfeld, Grabe, Marcus, LaMontagne, Heckbert, Austin, Launer,

848 Espeland, Masters, Maruff, Fripp, Johnson, Morris, Albert, Resnick, Saykin, Thompson, Li,

849 Wolf, Raquel Gur, Ruben Gur, Shinohara, Tosun-Turgut, Fan, Shou, Erus, Wolk

850 *Drafting of the manuscript:* Wen, Nasrallah, Davatzikos

851 *Critical revision of the manuscript for important intellectual content:* all authors

852 *Statistical and genetic analysis:* Wen

853 **References**

- 854 1. Alexander-Bloch, A., Giedd, J. N. & Bullmore, E. Imaging structural co-variance between
855 human brain regions. *Nat Rev Neurosci* **14**, 322–336 (2013).
- 856 2. Sotiras, A. *et al.* Patterns of coordinated cortical remodeling during adolescence and their
857 associations with functional specialization and evolutionary expansion. *Proc Natl Acad Sci*
858 *USA* **114**, 3527–3532 (2017).
- 859 3. Blank, S. C., Scott, S. K., Murphy, K., Warburton, E. & Wise, R. J. S. Speech production:
860 Wernicke, Broca and beyond. *Brain* **125**, 1829–1838 (2002).
- 861 4. Zhao, B. *et al.* Genome-wide association analysis of 19,629 individuals identifies variants
862 influencing regional brain volumes and refines their genetic co-architecture with cognitive
863 and mental health traits. *Nat Genet* **51**, 1637–1644 (2019).
- 864 5. Elliott, L. T. *et al.* Genome-wide association studies of brain imaging phenotypes in UK
865 Biobank. *Nature* **562**, 210–216 (2018).
- 866 6. Vignando, M. *et al.* Mapping brain structural differences and neuroreceptor correlates in
867 Parkinson’s disease visual hallucinations. *Nat Commun* **13**, 519 (2022).
- 868 7. Bassett, D. S. & Siebenhühner, F. Multiscale Network Organization in the Human Brain. in
869 *Multiscale Analysis and Nonlinear Dynamics* 179–204 (John Wiley & Sons, Ltd, 2013).
870 doi:10.1002/9783527671632.ch07.
- 871 8. Schaefer, A. *et al.* Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic
872 Functional Connectivity MRI. *Cerebral Cortex* **28**, 3095–3114 (2018).
- 873 9. Sotiras, A., Resnick, S. M. & Davatzikos, C. Finding imaging patterns of structural
874 covariance via Non-Negative Matrix Factorization. *NeuroImage* **108**, 1–16 (2015).

- 875 10. Thomas Yeo, B. T. *et al.* The organization of the human cerebral cortex estimated by
876 intrinsic functional connectivity. *Journal of Neurophysiology* **106**, 1125–1165 (2011).
- 877 11. Seeley, W. W., Crawford, R. K., Zhou, J., Miller, B. L. & Greicius, M. D.
878 Neurodegenerative diseases target large-scale human brain networks. *Neuron* **62**, 42–52
879 (2009).
- 880 12. Pomponio, R. *et al.* Harmonization of large MRI datasets for the analysis of brain imaging
881 patterns throughout the lifespan. *Neuroimage* **208**, 116450 (2020).
- 882 13. Betzel, R. F. & Bassett, D. S. Multi-scale brain networks. *NeuroImage* **160**, 73–83 (2017).
- 883 14. Roshchupkin, G. V. *et al.* Heritability of the shape of subcortical brain structures in the
884 general population. *Nat Commun* **7**, 13738 (2016).
- 885 15. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association
886 studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005–D1012
887 (2019).
- 888 16. Wen, J. *et al.* Genetic, clinical underpinnings of subtle early brain change along
889 Alzheimer’s dimensions. 2022.09.16.508329 Preprint at
890 <https://doi.org/10.1101/2022.09.16.508329> (2022).
- 891 17. Leeuw, C. A. de, Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-
892 Set Analysis of GWAS Data. *PLOS Computational Biology* **11**, e1004219 (2015).
- 893 18. Jossin, Y. Reelin Functions, Mechanisms of Action and Signaling Pathways During Brain
894 Development and Maturation. *Biomolecules* **10**, E964 (2020).
- 895 19. Gilbert, S. F. Morphogenesis and Cell Adhesion. *Developmental Biology*. 6th edition
896 (2000).

- 897 20. Wu, Y. *et al.* Contacts between the endoplasmic reticulum and other membranes in
898 neurons. *Proc Natl Acad Sci U S A* **114**, E4859–E4867 (2017).
- 899 21. Ly, A. *et al.* DSCAM Is a Netrin Receptor that Collaborates with DCC in Mediating
900 Turning Responses to Netrin-1. *Cell* **133**, 1241–1254 (2008).
- 901 22. Nikolsky, Y. *et al.* Genome-wide functional synergy between amplified and mutated genes
902 in human breast cancer. *Cancer Res* **68**, 9532–9540 (2008).
- 903 23. Bos, P. D. *et al.* Genes that mediate breast cancer metastasis to the brain. *Nature* **459**,
904 1005–1009 (2009).
- 905 24. Lanni, C., Masi, M., Racchi, M. & Govoni, S. Cancer and Alzheimer’s disease inverse
906 relationship: an age-associated diverging derailment of shared pathways. *Mol Psychiatry*
907 **26**, 280–295 (2021).
- 908 25. Shafi, O. Inverse relationship between Alzheimer’s disease and cancer, and other factors
909 contributing to Alzheimer’s disease: a systematic review. *BMC Neurol* **16**, 236 (2016).
- 910 26. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and
911 annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).
- 912 27. Yuan, J. & Yankner, B. A. Apoptosis in the nervous system. *Nature* **407**, 802–809 (2000).
- 913 28. Steiner, E., Tata, M. & Frisén, J. A fresh look at adult neurogenesis. *Nat Med* **25**, 542–543
914 (2019).
- 915 29. de Flores, R. *et al.* Medial Temporal Lobe Networks in Alzheimer’s Disease: Structural and
916 Molecular Vulnerabilities. *J Neurosci* **42**, 2131–2141 (2022).
- 917 30. Ginestier, C. *et al.* Prognosis and gene expression profiling of 20q13-amplified breast
918 cancers. *Clin Cancer Res* **12**, 4533–4544 (2006).

- 919 31. Kunkle, B. W. *et al.* Genetic meta-analysis of diagnosed Alzheimer’s disease identifies new
920 risk loci and implicates A β , tau, immunity and lipid processing. *Nat Genet* **51**, 414–430
921 (2019).
- 922 32. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from
923 polygenicity in genome-wide association studies. *Nat Genet* **47**, 291–295 (2015).
- 924 33. Davatzikos, C. Machine learning in neuroimaging: Progress and challenges. *NeuroImage*
925 **197**, 652–656 (2019).
- 926 34. Davatzikos, C., Xu, F., An, Y., Fan, Y. & Resnick, S. M. Longitudinal progression of
927 Alzheimer’s-like patterns of atrophy in normal older adults: the SPARE-AD index. *Brain*
928 **132**, 2026–2035 (2009).
- 929 35. Davatzikos, C., Genc, A., Xu, D. & Resnick, S. M. Voxel-based morphometry using the
930 RAVENS maps: methods and validation using simulated longitudinal atrophy. *Neuroimage*
931 **14**, 1361–1369 (2001).
- 932 36. Wen, J. *et al.* Convolutional neural networks for classification of Alzheimer’s disease:
933 Overview and reproducible evaluation. *Medical Image Analysis* **63**, 101694 (2020).
- 934 37. Samper-González, J. *et al.* Reproducible evaluation of classification methods in
935 Alzheimer’s disease: Framework and application to MRI and PET data. *NeuroImage* **183**,
936 504–521 (2018).
- 937 38. Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F. & Baker, C. I. Circular analysis in
938 systems neuroscience: the dangers of double dipping. *Nat. Neurosci.* **12**, 535–540 (2009).
- 939 39. Doshi, J. *et al.* MUSE: MUlti-atlas region Segmentation utilizing Ensembles of registration
940 algorithms and parameters, and locally optimal atlas selection. *Neuroimage* **127**, 186–195
941 (2016).

- 942 40. Diamantopoulou, Z. *et al.* The metastatic spread of breast cancer accelerates during sleep.
943 *Nature* **607**, 156–162 (2022).
- 944 41. Barton, R. A. & Venditti, C. Rapid Evolution of the Cerebellum in Humans and Other
945 Great Apes. *Curr Biol* **27**, 1249–1250 (2017).
- 946 42. van der Meer, D. *et al.* Understanding the genetic determinants of the brain with MOSTest.
947 *Nat Commun* **11**, 3512 (2020).
- 948 43. Soheili-Nezhad, S., Beckmann, C. F. & Sprooten, E. Reproducibility of Principal and
949 Independent Genomic Components of Brain Structure and Function. 2022.07.13.499912
950 Preprint at <https://doi.org/10.1101/2022.07.13.499912> (2022).
- 951 44. Patel, K. *et al.* New phenotype discovery method by unsupervised deep representation
952 learning empowers genetic association studies of brain imaging. 2022.12.10.22283302
953 Preprint at <https://doi.org/10.1101/2022.12.10.22283302> (2022).
- 954 45. Yang, Z. *et al.* Gene-SGAN: a method for discovering disease subtypes with imaging and
955 genetic signatures via multi-view weakly-supervised deep clustering. *ArXiv*
956 arXiv:2301.10772v1 (2023).
- 957 46. Eickhoff, S. B., Yeo, B. T. T. & Genon, S. Imaging-based parcellations of the human brain.
958 *Nat Rev Neurosci* **19**, 672–686 (2018).
- 959 47. Chen, C.-H. *et al.* Hierarchical Genetic Organization of Human Cortical Surface Area.
960 *Science* **335**, 1634–1636 (2012).
- 961 48. Glasser, M. F. *et al.* A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–
962 178 (2016).
- 963 49. Sunkin, S. M. *et al.* Allen Brain Atlas: an integrated spatio-temporal portal for exploring the
964 central nervous system. *Nucleic Acids Res* **41**, D996–D1008 (2013).

- 965 50. Munafò, M. R. *et al.* A manifesto for reproducible science. *Nat Hum Behav* **1**, 1–9 (2017).
- 966 51. Poldrack, R. A. *et al.* Scanning the horizon: towards transparent and reproducible
967 neuroimaging research. *Nat. Rev. Neurosci.* **18**, 115–126 (2017).
- 968 52. Routier, A. *et al.* Clinica: An Open-Source Software Platform for Reproducible Clinical
969 Neuroscience Studies. *Frontiers in Neuroinformatics* **15**, 39 (2021).
- 970 53. Wen, J. *et al.* Reproducible Evaluation of Diffusion MRI Features for Automatic
971 Classification of Patients with Alzheimer’s Disease. *Neuroinformatics* **19**, 57–78 (2021).
- 972 54. Zhirong Yang & Oja, E. Linear and Nonlinear Projective Nonnegative Matrix
973 Factorization. *IEEE Trans. Neural Netw.* **21**, 734–749 (2010).
- 974 55. Petersen, R. C. *et al.* Alzheimer’s Disease Neuroimaging Initiative (ADNI): clinical
975 characterization. *Neurology* **74**, 201–209 (2010).
- 976 56. Miller, K. L. *et al.* Multimodal population brain imaging in the UK Biobank prospective
977 epidemiological study. *Nat Neurosci* **19**, 1523–1536 (2016).
- 978 57. Ellis, K. A. *et al.* The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging:
979 methodology and baseline characteristics of 1112 individuals recruited for a longitudinal
980 study of Alzheimer’s disease. *Int. Psychogeriatr.* **21**, 672–687 (2009).
- 981 58. Soldan, A. *et al.* Relationship of medial temporal lobe atrophy, APOE genotype, and
982 cognitive reserve in preclinical Alzheimer’s disease. *Hum Brain Mapp* **36**, 2826–2841
983 (2015).
- 984 59. Resnick, S. M., Pham, D. L., Kraut, M. A., Zonderman, A. B. & Davatzikos, C.
985 Longitudinal magnetic resonance imaging studies of older adults: a shrinking brain. *J*
986 *Neurosci* **23**, 3295–3301 (2003).

- 987 60. Resnick, S. M. *et al.* One-year age changes in MRI brain volumes in older adults. *Cereb*
988 *Cortex* **10**, 464–472 (2000).
- 989 61. Friedman, G. D. *et al.* CARDIA: study design, recruitment, and some characteristics of the
990 examined subjects. *J Clin Epidemiol* **41**, 1105–1116 (1988).
- 991 62. LaMontagne, P. J. *et al.* OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive
992 Dataset for Normal Aging and Alzheimer Disease. 2019.12.13.19014902 Preprint at
993 <https://doi.org/10.1101/2019.12.13.19014902> (2019).
- 994 63. Resnick, S. M. *et al.* Postmenopausal hormone therapy and regional brain volumes: the
995 WHIMS-MRI Study. *Neurology* **72**, 135–142 (2009).
- 996 64. Johnson, S. C. *et al.* The Wisconsin Registry for Alzheimer’s Prevention: A review of
997 findings and current directions. *Alzheimer’s & Dementia: Diagnosis, Assessment & Disease*
998 *Monitoring* **10**, 130–142 (2018).
- 999 65. Chand, G. B. *et al.* Two distinct neuroanatomical subtypes of schizophrenia revealed using
1000 machine learning. *Brain* **143**, 1027–1038 (2020).
- 1001 66. Di Martino, A. *et al.* Enhancing studies of the connectome in autism using the autism brain
1002 imaging data exchange II. *Sci Data* **4**, 170010 (2017).
- 1003 67. Shalev-Shwartz, S., Singer, Y. & Ng, A. Y. Online and batch learning of pseudo-metrics. in
1004 *Proceedings of the twenty-first international conference on Machine learning* 94
1005 (Association for Computing Machinery, 2004). doi:10.1145/1015330.1015376.
- 1006 68. Tustison, N. J. *et al.* N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imaging* **29**,
1007 1310–1320 (2010).
- 1008 69. Ou, Y., Sotiras, A., Paragios, N. & Davatzikos, C. DRAMMS: Deformable Registration via
1009 Attribute Matching and Mutual-Saliency Weighting. *Med Image Anal* **15**, 622–639 (2011).

- 1010 70. Coupé, P., Catheline, G., Lanuza, E., Manjón, J. V. & Initiative, for the A. D. N. Towards a
1011 unified analysis of brain maturation and aging across the entire lifespan: A MRI analysis.
1012 *Human Brain Mapping* **38**, 5501–5518 (2017).
- 1013 71. Bethlehem, R. a. I. *et al.* *Brain charts for the human lifespan*. 2021.06.08.447489
1014 <https://www.biorxiv.org/content/10.1101/2021.06.08.447489v1> (2021)
1015 doi:10.1101/2021.06.08.447489.
- 1016 72. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies.
1017 *Bioinformatics* **26**, 2867–2873 (2010).
- 1018 73. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population
1019 stratification in genome-wide association studies. *Nat Rev Genet* **11**, 459–463 (2010).
- 1020 74. Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: principal component analysis of Biobank-
1021 scale genotype datasets. *Bioinformatics* **33**, 2776–2778 (2017).
- 1022 75. Wen, J. *et al.* Characterizing Heterogeneity in Neuroimaging, Cognition, Clinical
1023 Symptoms, and Genetics Among Patients With Late-Life Depression. *JAMA Psychiatry*
1024 (2022) doi:10.1001/jamapsychiatry.2022.0020.
- 1025 76. Yang, J., Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. GCTA-GREML
1026 accounts for linkage disequilibrium when estimating genetic variance from genome-wide
1027 SNPs. *PNAS* **113**, E4579–E4580 (2016).
- 1028 77. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based
1029 Linkage Analyses. *Am J Hum Genet* **81**, 559–575 (2007).
- 1030 78. Magno, R. & Maia, A.-T. gwasrapidd: an R package to query, download and wrangle
1031 GWAS catalog data. *Bioinformatics* **36**, 649–650 (2020).

- 1032 79. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for
1033 interpreting genome-wide expression profiles. *Proceedings of the National Academy of*
1034 *Sciences* **102**, 15545–15550 (2005).
- 1035 80. Kuhn, H. W. The Hungarian method for the assignment problem. *Naval Research Logistics*
1036 *Quarterly* **2**, 83–97 (1955).
- 1037 81. Stasinopoulos, D. M. & Rigby, R. A. Generalized Additive Models for Location Scale and
1038 Shape (GAMLSS) in R. *Journal of Statistical Software* **23**, 1–46 (2008).
- 1039 82. Klein, A. & Tourville, J. 101 Labeled Brain Images and a Consistent Human Cortical
1040 Labeling Protocol. *Frontiers in Neuroscience* **6**, 171 (2012).
- 1041 83. Nadeau, C. & Bengio, Y. Inference for the Generalization Error. in *Advances in Neural*
1042 *Information Processing Systems 12* (eds. Solla, S. A., Leen, T. K. & Müller, K.) 307–313
1043 (MIT Press, 2000).
- 1044

1045 **Acknowledgments**

1046 The iSTAGING consortium is a multi-institutional effort funded by NIA by RF1 AG054409.
1047 The Baltimore Longitudinal Study of Aging neuroimaging study is funded by the Intramural
1048 Research Program, National Institute on Aging, National Institutes of Health and by
1049 HHSN271201600059C. The BIOCARD study is in part supported by NIH grant U19-
1050 AG033655. The PHENOM study is funded by NIA grant R01MH112070 and by the PRONIA
1051 project as funded by the European Union 7th Framework Program grant 602152. Other
1052 supporting funds are 5U01AG068057, 1U24AG074855, and S10OD023495. Data were provided
1053 [in part] by OASIS OASIS-3: Principal Investigators: T. Benzinger, D. Marcus, J. Morris; NIH
1054 P50 AG00561, P30 NS09857781, P01 AG026276, P01 AG003991, R01 AG043434, UL1
1055 TR000448, R01 EB009352. AV-45 doses were provided by Avid Radiopharmaceuticals, a
1056 wholly owned subsidiary of Eli Lilly. OC has received funding from the French government
1057 under management of Agence Nationale de la Recherche as part of the "Investissements
1058 d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and reference ANR-
1059 10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6). This
1060 research has been conducted using the UK Biobank Resource under Application Number 35148.
1061 Data used in preparation of this article were in part obtained from the Alzheimer's Disease
1062 Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within
1063 the ADNI contributed to the design and implementation of ADNI and/or provided data but did
1064 not participate in the analysis or writing of this report. A complete listing of ADNI investigators
1065 can be found
1066 at: http://adni.loni.usc.edu/wpcontent/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.
1067 ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging

1068 and Bioengineering, and through generous contributions from the following: AbbVie,
1069 Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica,
1070 Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan
1071 Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its
1072 affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer
1073 Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research
1074 & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.;
1075 NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer
1076 Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.
1077 The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in
1078 Canada. Private sector contributions are facilitated by the Foundation for the National Institutes
1079 of Health (www.fnih.org). The grantee organization is the Northern California Institute for
1080 Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research
1081 Institute at the University of Southern California. ADNI data are disseminated by the Laboratory
1082 for Neuro Imaging at the University of Southern California. Dr. Wen and Dr. Davatzikos had full
1083 access to all the data in the study. They took responsibility for the integrity of the data and the
1084 accuracy of the data analysis.

1085

1086 **Novel genomic loci influence patterns of structural covariance in the human**
1087 **brain**

1088
1089 Junhao Wen^{1,2*}, Ilya M. Nasrallah^{2,3}, Ahmed Abdulkadir², Theodore D. Satterthwaite^{2,4}, Zhijian Yang²,
1090 Guray Erus², Timothy Robert-Fitzgerald⁵, Ashish Singh², Aristeidis Sotiras⁶, Aleix Boquet-Pujadas⁷,
1091 Elizabeth Mamourian², Jimit Doshi², Yuhan Cui², Dhivya Srinivasan², Ioanna Skampardoni², Jiong
1092 Chen², Gyujoon Hwang², Mark Bergman², Jingxuan Bao⁸, Yogasudha Veturi⁹, Zhen Zhou², Shu Yang⁸,
1093 Paola Dazzan¹⁰, Rene S. Kahn¹¹, Hugo G. Schnack¹², Marcus V. Zanetti¹³, Eva Meisenzahl¹⁴, Geraldo F.
1094 Busatto¹³, Benedicto Crespo-Facorro¹⁵, Christos Pantelis¹⁶, Stephen J. Wood¹⁷, Chuanjun Zhuo¹⁸, Russell
1095 T. Shinohara^{2,5}, Ruben C. Gur⁴, Raquel E. Gur⁴, Nikolaos Koutsouleris¹⁹, Daniel H. Wolf^{2,4}, Andrew J.
1096 Saykin²⁰, Marylyn D. Ritchie⁹, Li Shen⁸, Paul M. Thompson²¹, Olivier Colliot²², Katharina Wittfeld²³,
1097 Hans J. Grabe²³, Duygu Tosun²⁴, Murat Bilgel²⁵, Yang An²⁵, Daniel S. Marcus²⁶, Pamela LaMontagne²⁶,
1098 Susan R. Heckbert²⁷, Thomas R. Austin²⁷, Lenore J. Launer²⁸, Mark Espeland²⁹, Colin L Masters³⁰, Paul
1099 Maruff³⁰, Jurgen Fripp³¹, Sterling C. Johnson³², John C. Morris³³, Marilyn S. Albert³⁴, R. Nick Bryan³,
1100 Susan M. Resnick²⁵, Yong Fan², Mohamad Habes³⁵, David Wolk^{2,36}, Haochang Shou^{2,5}, and Christos
1101 Davatzikos^{2*}, for the iSTAGING, the BLSA, the BIOCARD, the PHENOM, the ADNI studies, and the
1102 AI4AD consortium

1103
1104 ¹Laboratory of AI and Biomedical Science (LABS), Stevens Neuroimaging and Informatics Institute, Keck School of
1105 Medicine of USC, University of Southern California, Los Angeles, California, USA.

1106 ²Artificial Intelligence in Biomedical Imaging Laboratory (AIBIL), Center for Biomedical Image Computing and
1107 Analytics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA.

1108 ³Department of Radiology, University of Pennsylvania, Philadelphia, USA.

1109 ⁴Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA

1110 ⁵Penn Statistics in Imaging and Visualization Center, Department of Biostatistics, Epidemiology, and Informatics,
1111 Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA

1112 ⁶Department of Radiology and Institute for Informatics, Washington University School of Medicine, St. Louis, USA

1113 ⁷Biomedical Imaging Group, EPFL, Lausanne, Switzerland

1114 ⁸Department of Biostatistics, Epidemiology and Informatics University of Pennsylvania Perelman School of Medicine,
1115 Philadelphia, USA

1116 ⁹Department of Genetics and Institute for Biomedical Informatics, Perelman School of Medicine, University of
1117 Pennsylvania, Philadelphia, PA, USA

1118 ¹⁰Department of Psychological Medicine, Institute of Psychiatry, Psychology and Neuroscience, King's College
1119 London, London, UK

1120 ¹¹Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, USA

1121 ¹²Department of Psychiatry, University Medical Center Utrecht, Utrecht, Netherlands

1122 ¹³Institute of Psychiatry, Faculty of Medicine, University of São Paulo, São Paulo, Brazil

1123 ¹⁴Department of Psychiatry and Psychotherapy, HHU Düsseldorf, Germany

1124 ¹⁵Hospital Universitario Virgen del Rocío, University of Sevilla-IBIS; IDIVAL-CIBERSAM, Sevilla, Spain

1125 ¹⁶Melbourne Neuropsychiatry Centre, Department of Psychiatry, University of Melbourne and Melbourne Health,
1126 Carlton South, Australia

1127 ¹⁷Orygen and the Centre for Youth Mental Health, University of Melbourne; and the School of Psychology,
1128 University of Birmingham, UK

1129 ¹⁸Key Laboratory of Real Time Tracing of Brain Circuits in Psychiatry and Neurology (RTBCPN-Lab), Nankai
1130 University Affiliated Tianjin Fourth Center Hospital; Department of Psychiatry, Tianjin Medical University, Tianjin,
1131 China

1132 ¹⁹Department of Psychiatry and Psychotherapy, Ludwig-Maximilian University, Munich, Germany

1133 ²⁰Radiology and Imaging Sciences, Center for Neuroimaging, Department of Radiology and Imaging Sciences,
1134 Indiana Alzheimer's Disease Research Center and the Melvin and Bren Simon Cancer Center, Indiana University
1135 School of Medicine, Indianapolis

1136 ²¹Imaging Genetics Center, Mark and Mary Stevens Neuroimaging and Informatics Institute, Keck School of
1137 Medicine of USC, University of Southern California, Marina del Rey, California

- 1138 ²²Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la
1139 Pitié Salpêtrière, F-75013, Paris, France
- 1140 ²³Department of Psychiatry and Psychotherapy, German Center for Neurodegenerative Diseases (DZNE), University
1141 Medicine Greifswald, Germany
- 1142 ²⁴Department of Radiology and Biomedical Imaging, University of California, San Francisco, CA, USA
- 1143 ²⁵Laboratory of Behavioral Neuroscience, National Institute on Aging, NIH, USA
- 1144 ²⁶Department of Radiology, Washington University School of Medicine, St. Louis, Missouri, USA
- 1145 ²⁷Cardiovascular Health Research Unit and Department of Epidemiology, University of Washington, Seattle, WA,
1146 USA
- 1147 ²⁸Neuroepidemiology Section, Intramural Research Program, National Institute on Aging, Bethesda, Maryland, USA
- 1148 ²⁹Sticht Center for Healthy Aging and Alzheimer's Prevention, Wake Forest School of Medicine, Winston-Salem,
1149 North Carolina, USA
- 1150 ³⁰Florey Institute of Neuroscience and Mental Health, The University of Melbourne, Parkville, VIC, Australia
- 1151 ³¹CSIRO Health and Biosecurity, Australian e-Health Research Centre CSIRO, Brisbane, Queensland, Australia
- 1152 ³²Wisconsin Alzheimer's Institute, University of Wisconsin School of Medicine and Public Health, Madison,
1153 Wisconsin, USA
- 1154 ³³Knight Alzheimer Disease Research Center, Washington University in St. Louis, St. Louis, MO, USA
- 1155 ³⁴Department of Neurology, Johns Hopkins University School of Medicine, USA
- 1156 ³⁵Glenn Biggs Institute for Alzheimer's & Neurodegenerative Diseases, University of Texas Health Science Center at
1157 San Antonio, San Antonio, USA
- 1158 ³⁶Department of Neurology and Penn Memory Center, University of Pennsylvania, Philadelphia, USA
- 1159
- 1160 *Corresponding authors:
1161 Junhao Wen, Ph.D. – junhaowe@usc.edu
1162 [2025 Zonal Ave, Los Angeles, CA 90033, United States](#)
1163 Christos Davatzikos, Ph.D. – Christos.Davatzikos@pennmedicine.upenn.edu
1164 3700 Hamilton Walk, 7th Floor, Philadelphia, PA 19104, [United States](#)
1165

1166 **eMethod 1: Empirical validation of sopNMF**
1167 **eMethod 2: Reproducibility index**
1168 **eMethod 3: Inter-site image harmonization**
1169 **eMethod 4: Quality check of the image processing pipeline**
1170 **eMethod 5 Definition of the index, candidate, independent significant, and SNP and**
1171 **genomic locus**
1172 **eMethod 6 Cross-validation procedure for PAML**
1173 **eFigure 1: Comparison between opNMF and sopNMF**
1174 **eFigure 2: Reproducibility of the sopNMF brain parcellation**
1175 **eFigure 3: Scatter plot for the h^2 estimates from the discovery and replication sets**
1176 **eFigure 4: Sensitivity check for the GWAS results using the discovery set in UKBB**
1177 **eFigure 5: Machine learning performance for disease classification and age prediction**
1178 **eFigure 6: Annotation of MUSE PSCs to MuSIC PSCs based on the overlap index**
1179 **eFigure 7: Summary statistics of the multi-scale PSCs of MuSIC**
1180 **eTable 1: Study cohort characteristics**
1181 **eTable 2: Clinical phenotypes and diagnoses used in machine learning classification**
1182 **eTable 3: Comparison of variants identified via MuSIC with other studies**
1183 **eTable 4: Classification balanced accuracy for disease classification and effect size of these**
1184 **imaging signatures**
1185 **eTable 5: 119 MUSE gray matter regions of interest**
1186 **eAlgorithm 1: Algorithm for sopNMF**
1187

1188 **eMethod 1: Empirical validation of sopNMF.**

1189 For the empirical validation of sopNMF, the comparison population (**Method 1** in the main
1190 manuscript) was used so that the machine's memory could be sufficient to read the entire data for
1191 opNMF. For sopNMF, different choices of batch size (i.e., BS=32, 64, 128, and 256) were
1192 tested. We hypothesized that sopNMF could approximate the optima of opNMF during
1193 optimization, i.e., resulting in similar parts-based representation, training loss, and sparsity.
1194 TensorboardX was embedded into the sopNMF framework to monitor the training process
1195 dynamically. All experiments were performed on an Ubuntu machine with a maximum RAM of
1196 32 GB and 8 CPUs. The predefined maximum number of epochs for all experiments is 50,000,
1197 and the tolerance of early stopping criteria is 100 epochs based on the training loss.

1198 We qualitatively compared the extracted PSCs and quantitatively for the training loss, the
1199 sparsity of the component matrix W , and the memory consumption for $C=20$ (number of PSCs).
1200 The 20 PSCs were spatially consistent between opNMF and sopNMF, despite that some regions
1201 were decomposed into different PSCs (i.e., the white ellipse in **eFig. 1A**). For the training loss,
1202 opNMF obtained the lowest loss (1.103×10^6), and the loss of sopNMF were 1.107×10^6 , 1.108
1203 $\times 10^6$, 1.111×10^6 and 1.210×10^6 for BS =256, 128, 64, and 32, respectively (**eFig. 1D**). For the
1204 sparsity of the component matrix, all models obtained comparable results (sparsity ≈ 0.83 , **eFig.**
1205 **1E**). The estimated memory consumptions during the training process were 28.65, 4.02, 3.81,
1206 2.60, 1.47 GB for opNMF and sopNMF (BS =256, 128, 64, and 32), respectively
1207 (**Fig. e1F**).

1208

1209 **eMethod 2: Reproducibility index.**

1210 We proposed a reproducibility index (RI) to test the reproducibility of sopNMF for brain

1211 parcellation:

1212 • We used the Hungarian match algorithm⁸⁰ to match the pairs of PSCs between two splits
1213 under the specific condition that maximizes the similarity (i.e., minimizes the cost of
1214 workers/jobs in its original formulation).

1215 • For each pair of PSCs, we calculated the inner product of the vectors (R^d), referred to as
1216 RI. This index takes values between [0, 1], with higher values indicating higher
1217 reproducibility.

1218 • For each scale C , we presented the mean/standard deviation of the RIs for all PSCs.

1219

1220

1221

1222 **eMethod 3: Inter-site image harmonization**

1223 We used an extensively validated statistical harmonization approach, i.e., ComBat-GAM,¹² to
1224 harmonize the extracted multi-scale PSCs. This method estimates the variability in volumetric
1225 measures due to differences in site/cohort-specific imaging protocols based on variances observed
1226 within and across control groups while preserving normal variances due to age, sex, and
1227 intracranial volume (ICV) differences. The model was initially trained on the discovery set and
1228 then applied to the replication set.

1229

1230 **eMethod 4: Quality check of the image processing pipeline.**

1231 Raw T1-weighted MRIs were first quality checked (QC) for motion, image artifacts, or restricted
1232 field-of-view. Another QC was performed: First, the images were examined by manually
1233 evaluating for pipeline failures (e.g., poor brain extraction, tissue segmentation, and registration
1234 errors). Furthermore, a second step automatically flagged images based on outlying values of
1235 quantified metrics (i.e., PSC values); those flagged images were re-evaluated.

1236

1237 **eMethod 5: Definition of the index, candidate, independent significant, and lead SNP and**
1238 **genomic locus.**

1239 *Index SNP*

1240 They are defined as SNPs with a p-value threshold $\leq 5e-8$ (*clump-p1*) from GWAS summary
1241 statistics.

1242 *Independent significant SNP*

1243 They are defined as the index SNPs, which are independent of each other (not in linkage
1244 disequilibrium) with $r^2 \leq 0.6$ (*clump-r2*) within 250 kilobases (non-overlapping, *clump-kb*) away
1245 from each other.

1246 *Lead SNP and genomic loci*

1247 They are defined as the independent significant SNPs, which are independent of each other with
1248 a more stringent $r^2 \leq 0.1$ (*clump-r2*) within 250 kilobases (non-overlapping, *clump-kb*) away
1249 from each other. Each of these clumps is defined as a *genomic locus*.

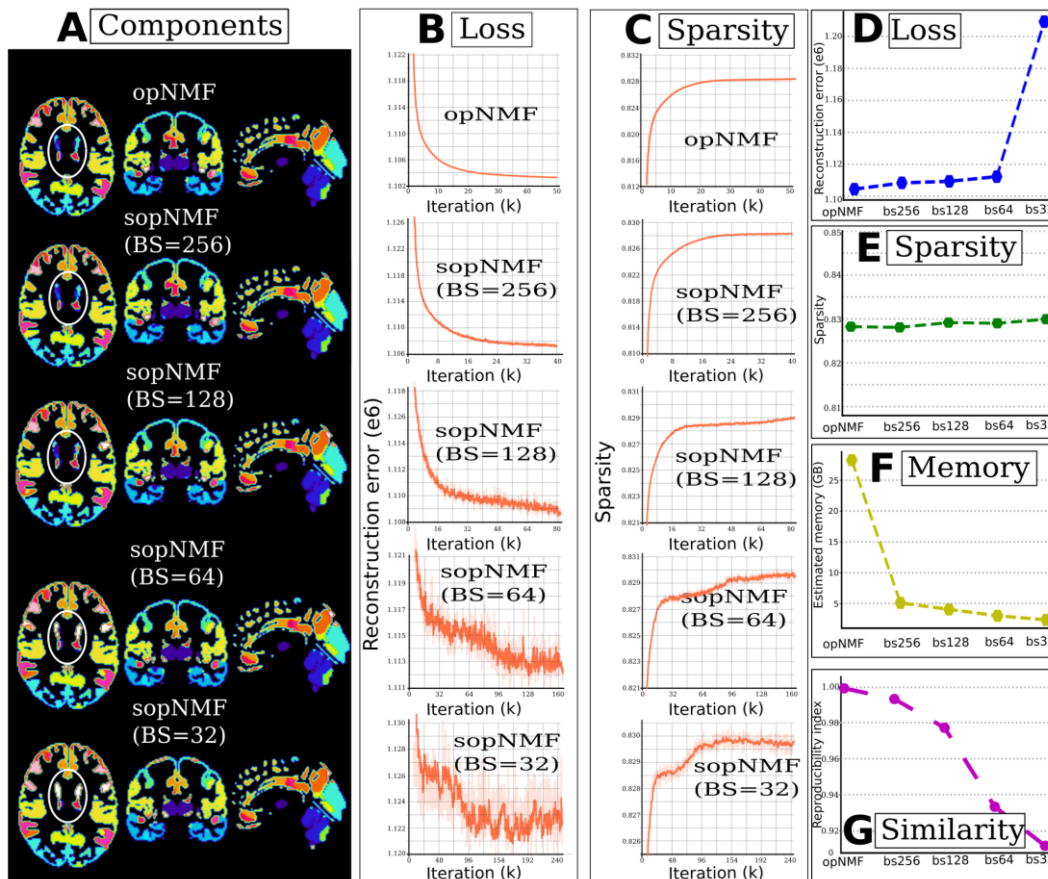
1250 *Candidate SNP*

1251 With each genomic locus, candidate SNPs are defined as the SNPs whose association p-values
1252 are smaller than 0.05 (*clump-p2*). The definitions followed instructions from FUMA²⁶ and
1253 Plink⁷⁷ software.

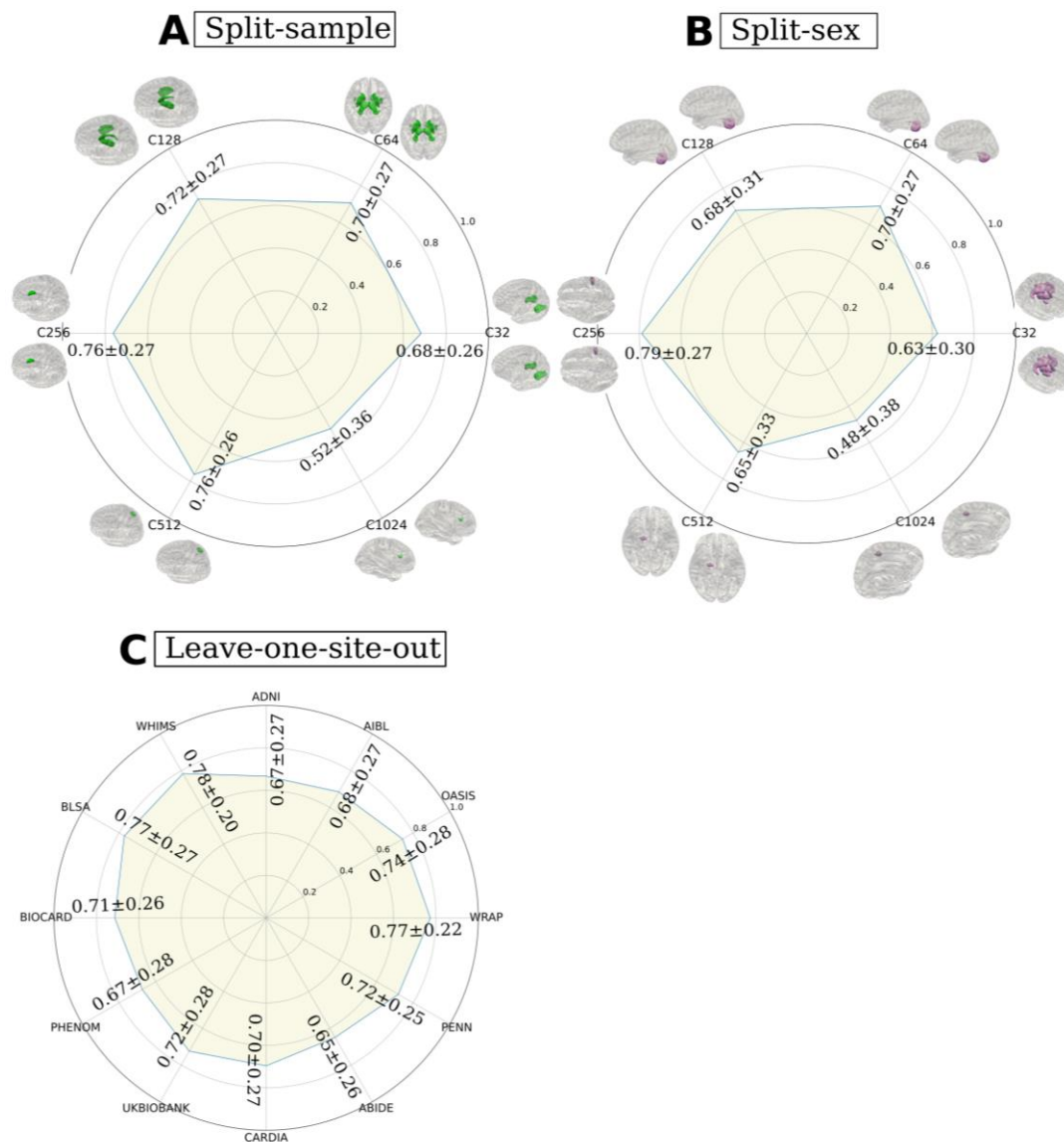
1254 **eMethod 6: Cross-validation procedure for PAML.**

1255 Nested cross-validation was adopted for all tasks following the good-practice guidelines
1256 proposed in our previous works^{36,37,53}. In particular, an outer loop was used to evaluate the task
1257 performance (250 repetitions of random hold-out splits with 80% of data for training). In
1258 contrast, an inner loop focused on tuning the hyperparameters (10-fold splits). We computed the
1259 balanced accuracy (BA) to evaluate the classification tasks. We calculated the effect size
1260 (Cohen's d) and p-value for each SPARE index to quantify its discriminative power.

1261

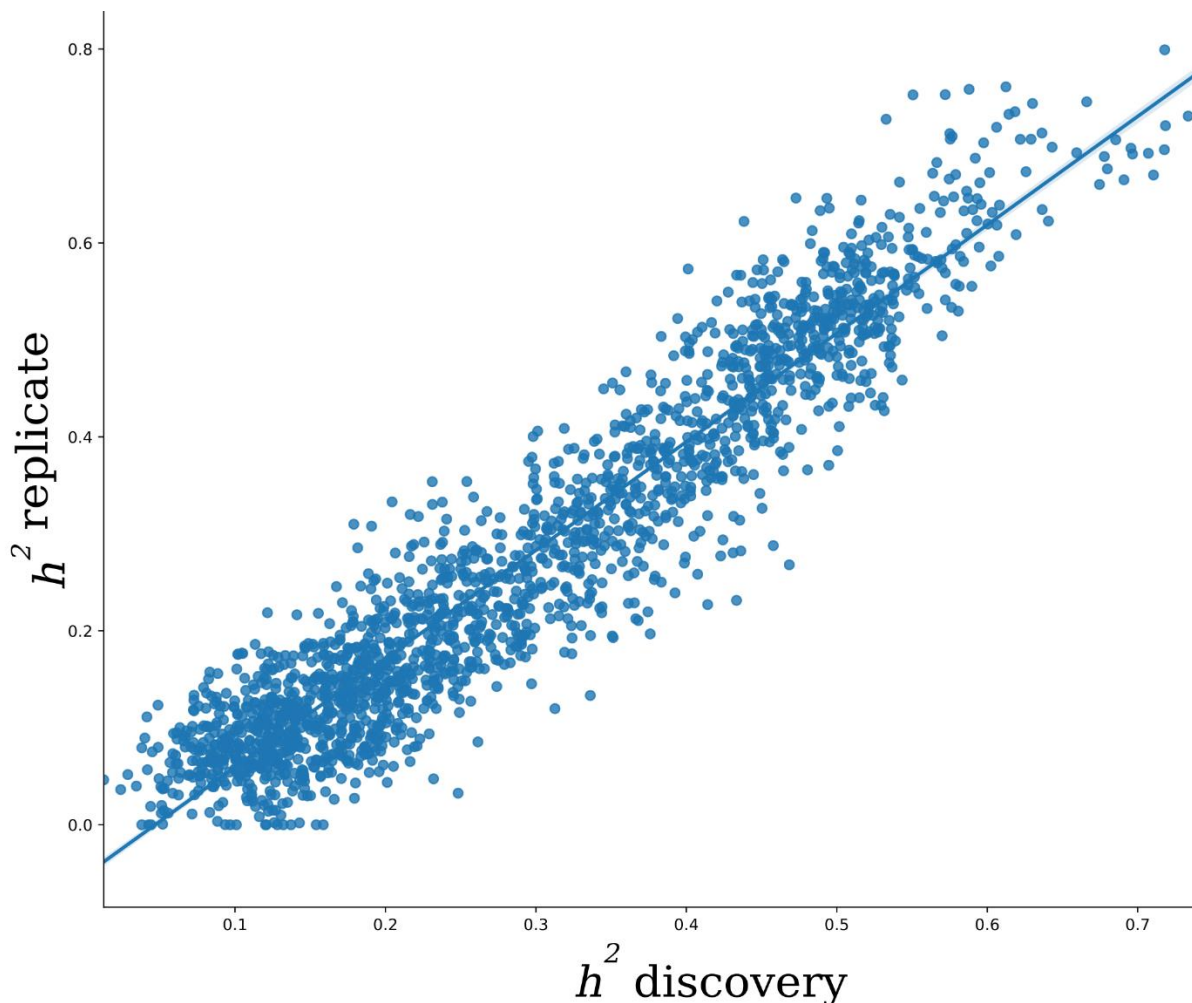


1262
1263 **eFigure 1: Comparison between opNMF and sopNMF.** (A) Qualitative evaluation: The
1264 extracted components are shown in the original image space, with each PSC displayed in a
1265 distinct color. The white ellipse indicates the region where the models diverge. Quantitative
1266 evaluation: training loss (B, D) and sparsity (C, E) demonstrated similar patterns between
1267 models, except that batch size (BS) = 32 had a larger loss than the other models. Comparing the
1268 estimated memory consumption during training across models shows significant advantages for
1269 all sopNMF models compared to opNMF.



1270
 1271 **eFigure 2: Reproducibility of the sopNMF brain parcellation.** In general, sopNMF
 1272 demonstrated high reproducibility under various conditions. For each brain PSC, the
 1273 reproducibility index (RI) was calculated (**Supplementary eMethod 2**). (A) Split-sample
 1274 analyses, where the training population ($N=4000$) was randomly split into two halves while
 1275 maintaining similar age, sex, and site distribution between groups. (B) Split-sex analyses, where
 1276 the training population was divided into males and females. Colored PSCs on the brain template
 1277 illustrate the same PSC independently derived from the two splits. (C) Leave-one-site-out
 1278 analyses for C32 PSCs., where the training populations excluding participants from each site
 1279 (BIOCARD, ADNI, WARP, AIBL, ABIDE, BLSA, OASIS, CARDIA, PHENOM, PENN,
 1280 UKBB, and WHIMS) were independently trained with sopNMF. The RI indices were compared
 1281 to the sopNMF results using the full training sample ($N=4000$).
 1282

1283



1284

1285

1286

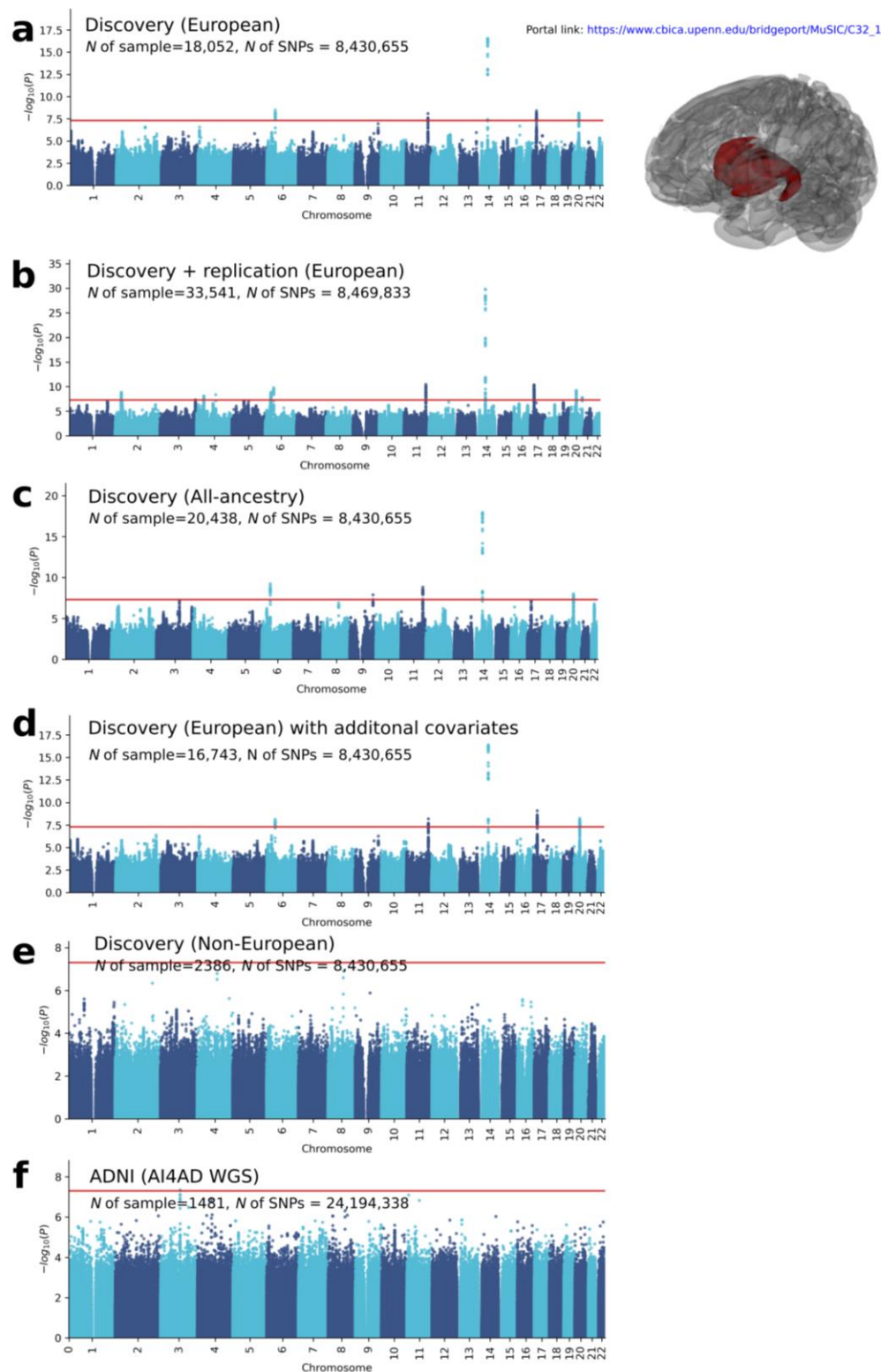
1287

1288

1289

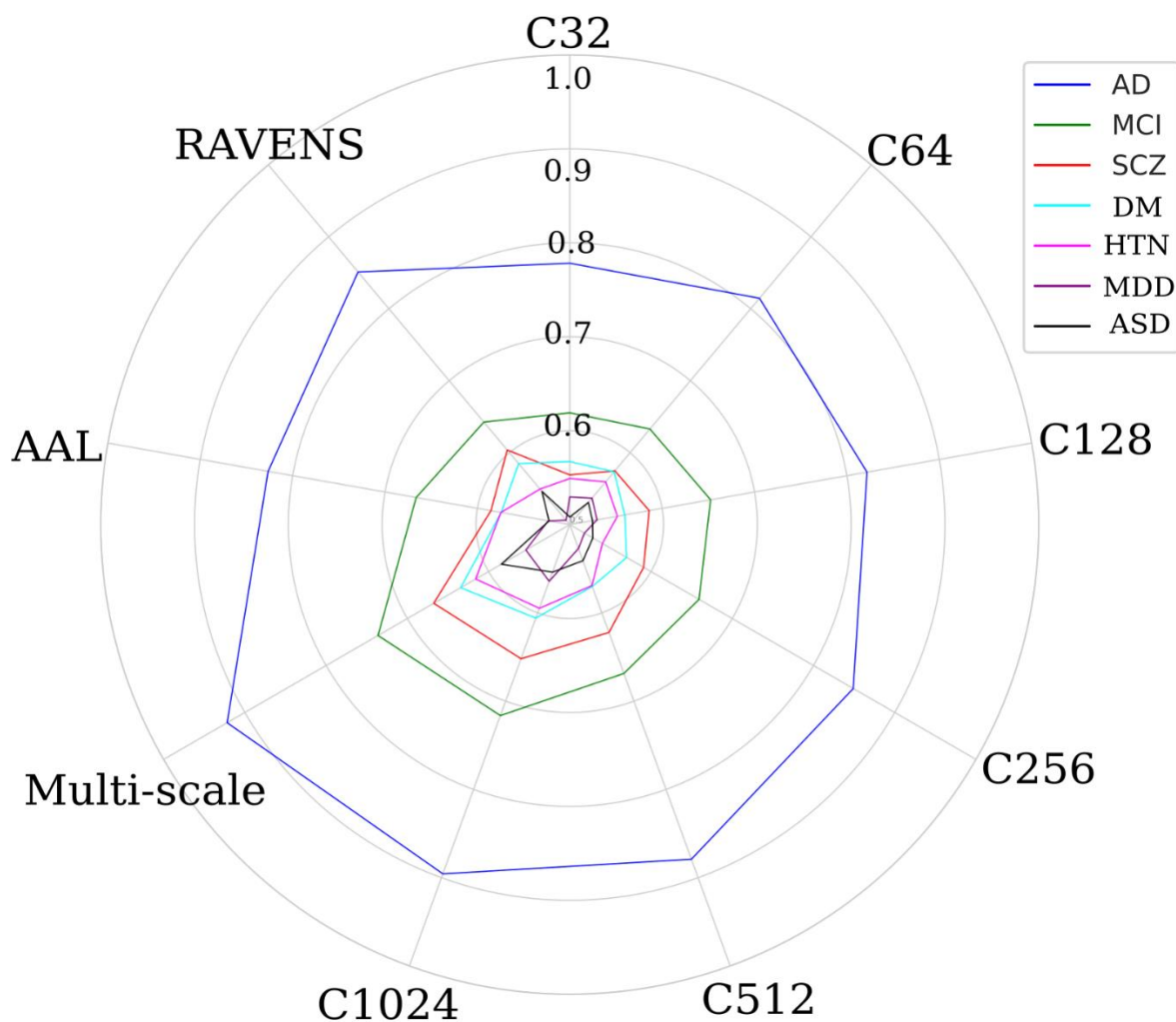
1290

eFigure 3: Scatter plot for the h^2 estimates from the discovery and replication sets. The SNP-based heritability was estimated independently for the discovery set ($N=18,052$) and replication set ($N=15,243$). In particular, the two estimates were highly correlated ($r = 0.94$, p -value $< 10^{-6}$), demonstrating a highly similar genetic architecture across different sets of UKBB data.

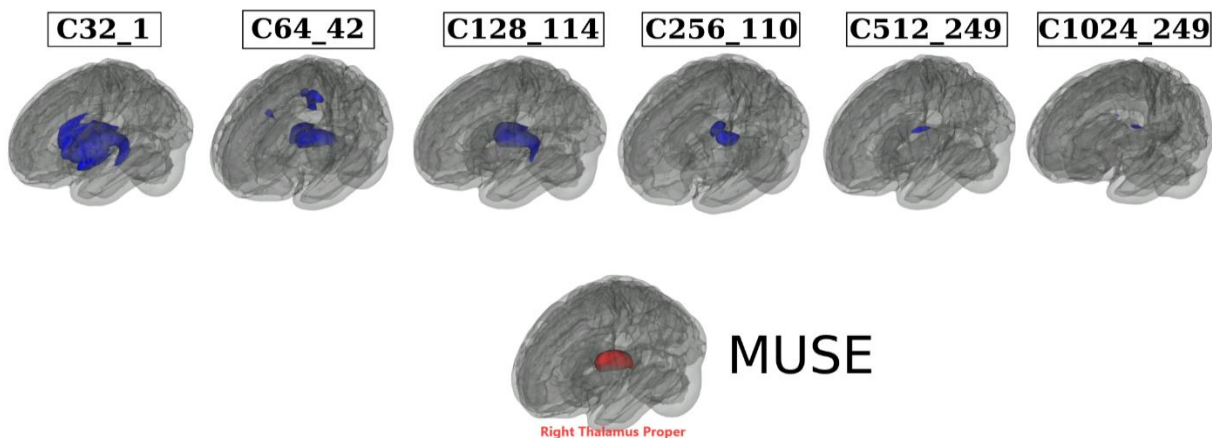


1291
1292 **eFigure 4: Sensitivity check for the GWAS results using the discovery set in UKBB.** A) The
1293 GWAS results for participants with European ancestry in the discovery set. B) The GWAS
1294 results for participants with European ancestry in the discovery and replication sets. C) The
1295 GWAS results for participants with all different ancestries in the discovery set. D) The GWAS

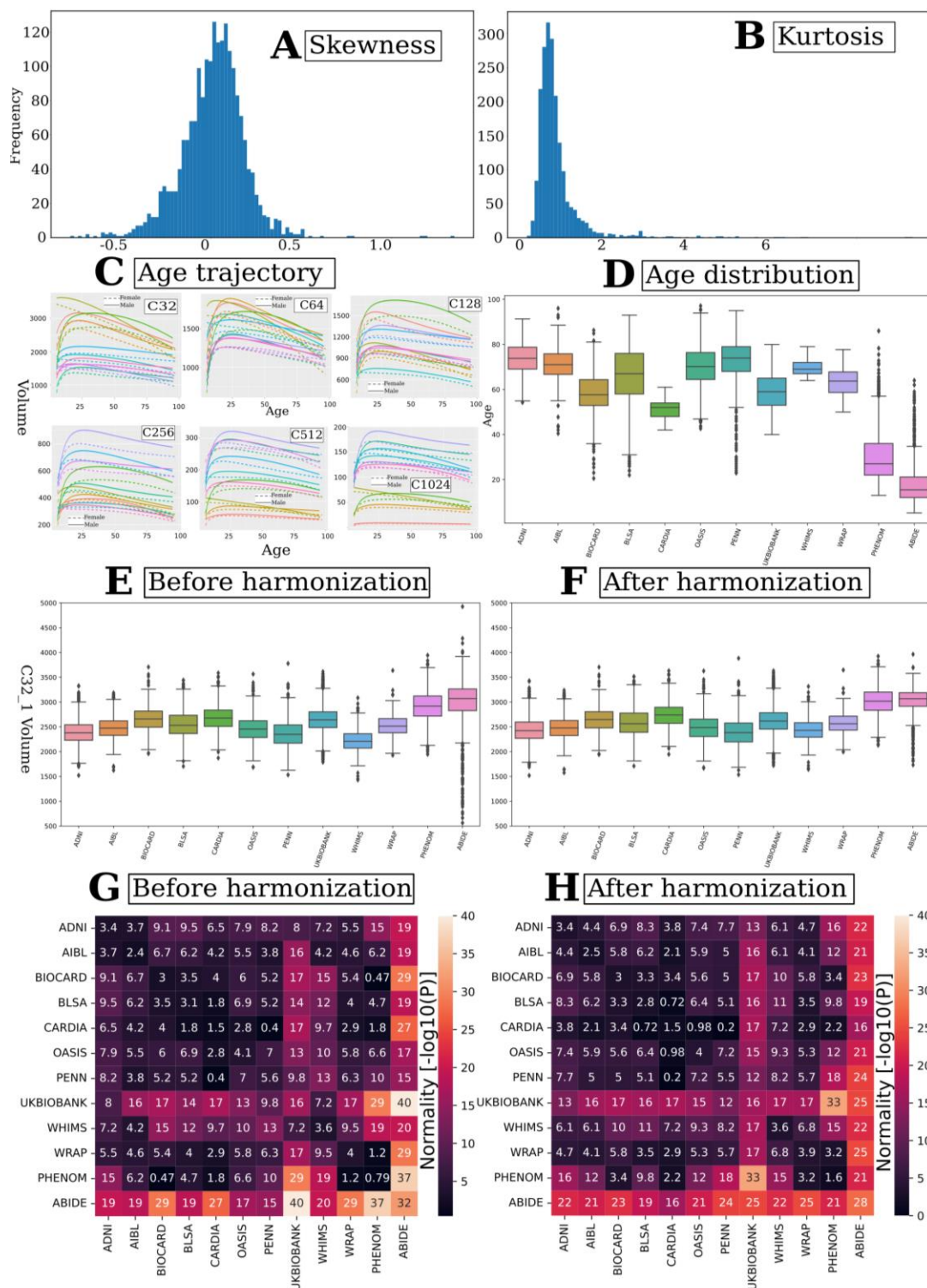
1296 results for participants with European ancestry in the discovery set by adding four additional
1297 imaging-related covariates. **E)** The GWAS results for participants with non-European ancestry in
1298 the discovery set. **F)** The GWAS results for participants with the independent ADNI WGS data.



1299
1300 **eFigure 5: Machine learning performance for disease classification.** Balanced accuracy (BA)
1301 for each classification task using different features from multi-scale MuSIC, AAL, and RAVENS
1302 (higher score better). Details are presented in **eTable 4**.
1303



1304
1305 **eFigure 6: Annotation of MUSE ROIs to MuSIC PSCs based on the overlap index.** We
1306 automatically annotated the 119 MUSE GM PSCs to the MuSIC atlases at all six scales ($C=32$,
1307 64, 128, 256, 512, and 1024). To this end, we calculated an overlap index (OI) to quantify the
1308 spatial overlaps between MUSE and MuSIC. For instance, for each MUSE PSC (**eTable 5**) vs.
1309 each of the 32 PSCs of MuSIC at $C=32$ scale, the OI equals the proportion of the number of
1310 overlap voxels and the total number of voxels in the MUSE PSC. Here we illustrate by mapping
1311 the right thalamus of MUSE to all 6 MuSIC atlases. The highest OIs are 0.82, 0.70, 0.86, 0.30,
1312 0.09, 0.05 for C32_1, C64_42, C128_114, C256_110, C512_249, and C1024_249 PSCs. This
1313 functionality is available in BRIDGEPORT:
1314 <https://www.cbica.upenn.edu/bridgeport/MUSE/Right%20Thalamus%20Proper>
1315



1316
 1317 **eFigure 7: Summary statistics of the multi-scale PSCs of MuSIC.** Multi-scale PSCs show
 1318 considerable normal distributions, i.e., symmetrical distribution (A) with a low kurtosis (B).
 1319 Moreover, we fit the Generalized Additive Model for Location, Scale, and Shape (GAMLSS)⁸¹
 1320 model (fractional polynomials with 2 degrees) to each PSC to delineate the age trajectory over
 1321 the lifespan in males (solid lines) and females (dotted lines), respectively (C). For visualization

1322 purposes, we selectively display the first 10 PSCs from each scale of the MuSIC atlases. In
1323 general, males have larger brain volumes than females. For **D-F**, we selectively showed the
1324 distribution of age (**D**) and the distribution of PSC volume before harmonization (**E**) and after
1325 harmonization (**F**) for C32_1 within each site in the discovery set. For **G** and **H**, we tested the
1326 normality of the PSC volume (C32_1) from each pair of sites using the Shapiro-Wilk test
1327 (*scipy.stats.shapiro* function) in the discovery set before (**G**) harmonization and after
1328 harmonization (**H**). A higher $-\log_{10}(P)$ indicates the data are less likely to be normally
1329 distributed. As a general trend, our statistical harmonization techniques demonstrated a slight
1330 improvement in the normality of the data. Additionally, we consistently applied normality
1331 transformations to all statistical analyses, including GWAS, to mitigate any non-normality.
1332

1333 **eTable 1. Study cohort characteristics.**

1334 The current study consists of two main populations/sets: the discovery set ($N=32,440$, including
 1335 participants from the first download of the UKBB data) and the replication set ($N=18,259$, the
 1336 second download of the UKBB data). To train the sopNMF model for MuSIC, we selected 250
 1337 patients (PT) and 250 healthy controls (CN) for each decade of the discovery set, resulting in
 1338 4000 participants in total, referred to as the training population. Age ranges from 5 to 97 years
 1339 and is shown with mean and standard deviation. Sex is displayed with the number and
 1340 percentage of female participants. Data was collected from 12 studies, 130 sites, and 12
 1341 countries. The number of sites (country) per study is detailed as follows:

- 1342 • ADNI: 63 sites (USA)
- 1343 • UKBB: 5 sites (UK)
- 1344 • AIBL: 2 sites (Australia)
- 1345 • BIOCARD: 2 sites (USA)
- 1346 • BLSA: 1 site (USA)
- 1347 • CARDIA: 3 sites (USA)
- 1348 • OASIS: 1 site (USA)
- 1349 • PENN: 1 site (USA)
- 1350 • WHIMS: 14 sites (USA)
- 1351 • WRAP 1 site (USA)
- 1352 • PHENOM: 12 sites (China, Brazil, Australia, Germany, Spain, USA, Netherlands)
- 1353 • ABIDE: 25 sites (USA, Netherlands, Belgium, Germany, Ireland, Switzerland, France)

1354 Abbreviations: CN: healthy control; AD: Alzheimer's disease; MCI: mild cognitive impairment;
 1355 SCZ: schizophrenia; ASD: autism spectrum disorder; MDD: major depressive disorder; DM:
 1356 diabetes; HTN: hypertension.

1357 ^aUKBB data were separately downloaded two times: the first was the $N=21,305$ in the discovery
 1358 set, and the second was the replication set.

1359 ^bWe define CN (healthy controls) as participants that do not have any of the diseases listed here.
 1360 These CN participants might have diagnoses of other illnesses or comorbidities (e.g., participants
 1361 from UKBB have a wide range of pathology based on ICD-10).
 1362

Study	N (50,699)	Age (5-97 year)	Sex (female/%)	CN ^b	AD	MCI	SCZ	ASD	MDD	DM	HTN
Discovery set	32,440	60.04± 14.87	16,868/52	24,980	954	1288	1094	597	1476	1093	958
ADNI	1765	73.66± 7.19	798/45	297	343	875	NA	NA	NA	NA	250
UKBB ^a	21,305	62.58± 7.48	10,101/53	18,735	1	NA	NA	NA	1476	1093	NA
AIBL	830	71.36± 6.78	471/57	625	86	115	NA	NA	NA	NA	4
BIOCARD	288	58.15± 10.54	115/60	283	1	4	NA	NA	NA	NA	NA
BLSA	1114	65.44± 14.11	589/53	729	9	11	NA	NA	NA	NA	365
CARDIA	892	51.21± 3.98	471/53	620	NA	NA	NA	NA	NA	NA	272
OASIS	983	69.92± 9.75	557/57	759	220	NA	NA	NA	NA	NA	4

PENN	807	72.63± 10.65	333/59	173	294	283	NA	NA	NA	NA	57
WHIMS	995	69.61± 3.64	995/100	986	NA	NA	NA	NA	NA	NA	6
WRAP	116	63.36± 6.06	79/68	116	NA	NA	NA	NA	NA	NA	NA
PHENOM	2125	30.21± 10.60	854/40	1031	NA	NA	1094	NA	NA	NA	NA
ABIDE	1220	17.92± 9.01	203/17	623	NA	NA	NA	597	NA	NA	NA
Replication set^a	18,259	54.70± 7.43	9742/53	NA	NA	NA	NA	NA	NA	NA	NA

1363
1364

1365 **eTable 2: Clinical phenotypes and diagnoses used in machine learning classification.**

1366 We harmonized the population of the phenotypes of interest per study definitions:

- 1367 • We combined AD and MCI patients from ADNI, PENN, and AIBL but excluded OASIS
 1368 subjects because of the different diagnostic criteria of an AD patient in OASIS.
 1369 • For several binary disease phenotypes, we used the ICD-10 diagnosis
 1370 (<https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=41270>). Note that ICD-10 diagnoses are
 1371 generally collected from the participants' medical inpatient records. We first included
 1372 diseases from the following categories:
- 1373 ○ Diseases of the blood and blood-forming organs and certain disorders involving the
 1374 immune mechanism (D-XXX, XXX represents the ID of a specific disease);
 - 1375 ○ Endocrine, nutritional, and metabolic diseases (E-XXX);
 - 1376 ○ Mental and behavioral disorders (F-XXX);
 - 1377 ○ Diseases of the nervous system (G-XXX);
 - 1378 ○ Diseases of the circulatory system (I-XXX).

1379 We then set a threshold of 75 patients for any ICD-10 diagnosis. We finally randomly
 1380 selected age and sex-matched healthy controls (excluding all patients in all diagnoses).^a
 1381 For major depressive disorder, we used the inclusion criteria from our previous work.⁷⁵

- 1382 • For cognitive scores, we included:
- 1383 ○ Tower rearranging (<https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=21004>)
 - 1384 ○ Matrix pattern (<https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=6373>)
 - 1385 ○ TMT-A (<https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=6348>)
 - 1386 ○ TMT-B (<https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=6350>)
 - 1387 ○ DSST (<https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=23324>)
 - 1388 ○ Pairs matching (<https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=399>)
 - 1389 ○ Numerical memory (<https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=4282>)
 - 1390 ○ Prospective memory (<https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=4288>)
 - 1391 ○ Reaction time (<https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=20023>)
 - 1392 ○ Fluid intelligence (<https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=20016>)

1393 AD: Alzheimer's disease; MCI: mild cognitive impairment; SCZ: schizophrenia; DM: diabetes
 1394 mellitus; MDD: major depressive disorder; HTN: hypertension; ASD: autism spectrum disorder;
 1395 CN: healthy control; PT: patient; *N*: number of participants. We decided not to harmonize
 1396 cognitive scores from different studies.

1397

Trait (ICD-10 code or ID)	Sample size (CN/PT or <i>N</i>)	Site	Trait (ICD-10 code or ID)	Sample size (CN/PT or <i>N</i>)	Site
AD	1095/723	ADNI, PENN, & AIBL	Carpal tunnel syndrome (G560)	901/901	UKBB
MCI	1273/1095	ADNI, PENN, & AIBL	Lesion of ulnar nerve (G562)	104/104	UKBB
SCZ	1031/1094	PHENOM	Lesion of plantar nerve (G576)	163/163	UKBB
DM	1093/1093	UKBB	Angina pectoris (I20)	1535/1535	UKBB
MDD ^a	1476/1476	UKBB	Acute myocardial infarction (I21)	769/769	UKBB
HTN	934/887	ADNI, BLSA & CARDIA	Chronic ischaemic heart disease (I25)	2217/2217	UKBB

ASD	623/597	ABIDE	Pulmonary embolism (I20)	351/351	UKBB
Iron deficiency anemia (D50)	1012/1012	UKBB	Cardiomyopathy (I42)	116/116	UKBB
Vitamin B12 deficiency anemia (D50)	78/78	UKBB	Paroxysmal tachycardia (I47)	320/320	UKBB
Agranulocytosis (D70)	245/245	UKBB	Heart failure (I50)	436/436	UKBB
Thyrotoxicosis (E05)	205/205	UKBB	Cerebral infarction (I63)	291/291	UKBB
Vitamin D deficiency (E55)	180/180	UKBB	Vitamin B deficiency (E53)	130/130	UKBB
Obesity (E66)	1481/1481	UKBB	Hemiplegia (G81)	111/111	UKBB
Lipoprotein metabolism disorder (E78)	3880/3880	UKBB	Facial nerve disorders (G51)	95/95	UKBB
Mineral metabolism disorder (E83)	291/291	UKBB	Tower rearranging (21004)	8412	UKBB
Volume depletion	240/240	UKBB	Matrix pattern (6373)	8501	UKBB
Delirium	92/92	UKBB	TMT-A (6348)	8599	UKBB
Alcohol abuse	341/341	UKBB	TMT-B (6350)	8599	UKBB
Tobacco abuse	863/863	UKBB	DSST (23324)	8523	UKBB
Bipolar affective disorder	77/77	UKBB	Pairs matching (399)	20945	UKBB
Phobic anxiety disorder	84/84	UKBB	Numerical memory (4282)	9323	UKBB
Multiple sclerosis	109/109	UKBB	Prospective memory (4288)	19681	UKBB
Epilepsy	250/250	UKBB	Reaction time (20023)	21258	UKBB
Migraine	508/508	UKBB	Fluid intelligence (20016)	19184	UKBB
Sleep disorders	590/590	UKBB			

1398
1399

1400 **eTable 3: Comparison of variants identified via MuSIC with other studies.** Using the AAL
1401 atlas, we found (using the same data in the current study) that 269 independent significant SNPs
1402 had 356 pairwise associations with 54 AAL brain regions. 230 out of the 269 SNPs matched with
1403 the SNPs in MuSIC. Among the 39 unmatched SNPs, 15 SNPs were in linkage disequilibrium
1404 (LD, $r^2 > 0.6$) with MuSIC SNPs (**Supplementary eFile 5**). As a second example, Zhao et al.⁴
1405 reported that 251 independent significant SNPs had 346 pairwise associations with 43 GM regions
1406 using the Mindboggle atlas on the UKBB ($N=19,629$).⁸² 129 of the 251 SNPs matched with SNPs
1407 identified by MuSIC. Among these non-matching SNPs (127), 31 were in LD with MuSIC SNPs
1408 (**Supplementary eFile 6**). Similarly, Elliot et al.⁵ ($N=8428$) discovered that 20 independent
1409 significant SNPs had 58 pairwise associations with 52 GM regions from atlases in Freesurfer and
1410 FSL software. Out of the 20 SNPs, 16 coincided with MuSIC SNPs. Among the four unmatched
1411 SNPs, 1 SNP was in LD with MuSIC SNPs (**Supplementary eFile 7**). Note that the definition of
1412 independent significant SNPs or genomic loci might slightly differ between studies.

Study/Atlas	Identified genomic loci	Matched loci	Loci in LD	Novel loci	Database	Sample size	Ancestry
MuSIC	915	NA	NA	NA	UKBB	18,052	European
AAL	218	162	13	740	UKBB	18,052	European
Zhao et al. ⁴	251	73	14	828	UKBB	19,629	European
Elliot et al. ⁵	20	16	1	898	UKBB	8428	European
GWAS Catalog	NA	298	NA	617	NA	NA	NA

1413
1414

1415 **eTable 4: Classification balanced accuracy for disease classification and effect size of these**
 1416 **imaging signatures.**

1417 Disease classification performance is presented using balanced accuracy. The mean and standard
 1418 deviation are presented. Cohen's d was computed to compare the SPARE scores between groups.
 1419 Multi-scale classification^a: All 2003 PSCs from multiple scales were fit into the classifier.
 1420 Multi-scale classification^b: PSCs from all scales were fit into the classifier with a nested feature
 1421 selection procedure (SVM-REF). The motivation is that PSCs from different scales are
 1422 hierarchical and correlated. The nested feature selection can select the features most relevant to
 1423 the specific task. We avoided any statistical comparison of the performance of machine learning
 1424 models because available statistical tests are liberal and often lead to false-positive conclusions
 1425 due to the complexity of the cross-validation procedure.⁸³

1426 a): Classification results for all subjects in all sites using a nested CV procedure

PSC	AD	d	MCI	d	SCZ	d	DM	d	HTN	d	MDD	d	ASD	d
C32	0.78± 0.02	1.52	0.62± 0.02	0.59	0.55± 0.02	0.30	0.56± 0.02	0.35	0.55± 0.02	0.28	0.52± 0.02	0.16	0.50± 0.02	0.07
C64	0.81± 0.02	1.73	0.63± 0.02	0.66	0.57± 0.02	0.41	0.57± 0.02	0.40	0.56± 0.02	0.31	0.53± 0.02	0.17	0.53± 0.02	0.19
C128	0.82± 0.02	1.82	0.65± 0.02	0.76	0.59± 0.02	0.47	0.56± 0.02	0.33	0.55± 0.02	0.30	0.52± 0.02	0.15	0.52± 0.02	0.15
C256	0.85± 0.02	2.08	0.66± 0.02	0.91	0.59± 0.02	0.50	0.56± 0.02	0.47	0.54± 0.02	0.31	0.51± 0.02	0.13	0.52± 0.02	0.16
C512	0.88± 0.02	2.34	0.67± 0.02	1.06	0.62± 0.02	0.62	0.57± 0.02	0.54	0.56± 0.02	0.42	0.52± 0.02	0.05	0.54± 0.02	0.24
C1024	0.90± 0.02	2.50	0.72± 0.02	1.12	0.65± 0.02	0.75	0.60± 0.02	0.59	0.59± 0.02	0.46	0.56± 0.02	0.13	0.55± 0.02	0.29
Multi-scale ^a	0.91± 0.02	2.54	0.72± 0.02	1.12	0.66± 0.02	0.77	0.61± 0.02	0.64	0.59± 0.02	0.47	0.55± 0.02	0.23	0.56± 0.02	0.30
Multi-scale ^b	0.92± 0.02	2.61	0.73± 0.02	1.13	0.67± 0.02	0.78	0.64± 0.02	0.67	0.61± 0.02	0.49	0.55± 0.02	0.26	0.58± 0.02	0.32
AAL	0.82± 0.02	1.81	0.66 ±0.02	0.75	0.59± 0.02	0.46	0.57± 0.02	0.32	0.57± 0.02	0.35	0.52± 0.02	0.08	0.52± 0.02	0.14
RAVENS	0.85± 0.02	2.04	0.64 ±0.02	0.74	0.60± 0.02	0.45	0.58± 0.02	0.33	0.55± 0.02	0.34	0.50± 0.02	0.05	0.54± 0.02	0.15

1427 b): The classification results of the balanced accuracy (BA) from the test data in the nested CV
 1428 and the independently left-out site for the task of AD vs. CN were assessed using all available
 1429 multi-scale PSCs^a. Three sites, namely ADNI, AIBL, and PENN, were considered for this
 1430 analysis. However, UKBB, BIOCARD, and BLSA data were excluded due to limited AD cases
 1431 (eTable 1). Similarly, data from OASIS were excluded due to discrepancies in the diagnosis
 1432 criteria for AD, as previously stated in our previous work³⁶.

Left-out site	Test BA in CV	Test BA in the left-out site
ADNI	0.90±0.02	0.88±0.02
AIBL	0.88±0.02	0.95±0.02
PENN	0.90±0.02	0.95±0.02

1434

1435 **eTable 5:** 119 MUSE gray matter regions of interest.

1436 L: Left hemisphere; R: Right hemisphere; ROI: region of interest.

MUSE ROI	MUSE ROI	MUSE ROI
Precentral gyrus (R)	Occipital fusiform gyrus (R)	Anterior insula (L)
Precentral gyrus (L)	Planum temporale (R)	Anterior orbital gyrus (R)
Accumbens area (R)	Cerebellar vermal lobules I-V	Anterior orbital gyrus (L)
Accumbens area (L)	Cerebellar vermal lobules VI-VII	Angular gyrus (R)
Amygdala (R)	Cerebellar vermal lobules VIII-X	Angular gyrus (L)
Amygdala (L)	Basal forebrain (R)	Calcarine cortex (R)
Occipital pole (L)	Basal forebrain (L)	Calcarine cortex (L)
Caudate (R)	Middle temporal gyrus (L)	Central operculum (R)
Caudate (L)	Occipital pole (R)	Central operculum (L)
Cerebellum exterior (R)	Planum temporale (L)	Cuneus (R)
Cerebellum exterior (L)	Parietal operculum (L)	Cuneus (L)
Planum polare (L)	Postcentral gyrus (R)	Entorhinal area (R)
Middle temporal gyrus (R)	Postcentral gyrus (L)	Entorhinal area (L)
Hippocampus (R)	Posterior orbital gyrus (R)	Frontal operculum (R)
Hippocampus (L)	Temporal pole (R)	Frontal operculum (L)
Precentral gyrus medial segment (R)	Temporal pole (L)	Frontal pole (R)
Precentral gyrus medial segment (L)	Triangular part of the inferior frontal gyrus (R)	Frontal pole (L)
Superior frontal gyrus medial segment (R)	Triangular part of the inferior frontal gyrus (L)	Fusiform gyrus (R)
Superior frontal gyrus medial segment (L)	Transverse temporal gyrus (R)	Fusiform gyrus (L)
Pallidum (R)	Superior frontal gyrus medial segment (L)	Gyrus rectus (R)
Pallidum (L)	Planum polare (R)	Gyrus rectus (L)
Putamen (R)	Transverse temporal gyrus (L)	Inferior occipital gyrus (R)
Putamen (L)	Anterior cingulate gyrus (R)	Inferior occipital gyrus (L)
Thalamus proper (R)	Anterior cingulate gyrus (L)	Inferior temporal gyrus (R)
Thalamus proper (L)	Anterior insula (R)	Inferior temporal gyrus (L)
Lingual gyrus (R)	Occipital fusiform gyrus (L)	Subcallosal area (R)
Lingual gyrus (L)	Opercular part of inferior frontal gyrus (R)	Subcallosal area (L)
Lateral orbital gyrus (R)	Opercular part of inferior frontal gyrus (L)	Superior frontal gyrus (R)
Lateral orbital gyrus (L)	Orbital part of inferior frontal gyrus (R)	Superior frontal gyrus (L)
Middle cingulate gyrus (R)	Orbital part of inferior frontal gyrus (L)	Supplementary motor cortex (R)
Middle cingulate gyrus (L)	Posterior cingulate gyrus (R)	Supplementary motor cortex (L)
Medial frontal cortex (R)	Posterior cingulate gyrus (L)	Supramarginal gyrus (R)
Medial frontal cortex (L)	Precuneus (R)	Supramarginal gyrus (L)
Middle frontal gyrus (R)	Precuneus (L)	Superior occipital gyrus (R)
Middle frontal gyrus (L)	Parahippocampal gyrus (R)	Superior occipital gyrus (L)
Middle occipital gyrus (R)	Parahippocampal gyrus (L)	Superior parietal lobule (R)
Middle occipital gyrus (L)	Posterior insula (R)	Superior parietal lobule (L)
Medial orbital gyrus (R)	Posterior insula (L)	Superior temporal gyrus (R)
Medial orbital gyrus (L)	Parietal operculum (R)	Superior temporal gyrus (L)
Superior frontal gyrus medial segment (R)	Posterior orbital gyrus (L)	

1437

1438 **eAlgorithm 1:** Algorithm for sopNMF.
1439 The source code of the Python implementation of sopNMF is available here:
1440 <https://github.com/anbai106/SOPNMF>

Algorithm 1: sopNMF

- **Input:** maximum number of epochs e , number of component C or r , batch size b , early stopping criteria θ (i.e., the loss without decreasing for a certain epochs) ;
- **Output:** $\mathbf{W} \in \mathbb{R}^{d \times r}$, $\mathbf{H} \in \mathbb{R}^{r \times n}$;
- **Initialization:** \mathbf{W} ;

if *not* θ or *epoch* $\neq e$ then
 for $p \leftarrow 0$ to e do
 for $i \leftarrow 0$ to t do
 Read mini-batch \mathbf{X}_{bi}
 Update \mathbf{W}_{i+1} via Eq. 2
 end
 $loss = \sum_{i=1}^{\lfloor \frac{n}{b} \rfloor} \|\mathbf{X}_{bi} - \mathbf{W}\mathbf{W}^T \mathbf{X}_{bi}\|_F^2$ (Eq.3)
 if *loss in* θ then
 | Stop
 else
 | Shuffle \mathbf{X}
 | Continue
 end
 end
else
 | Stop
end

1441
1442
1443

1444 **References**

- 1445
- 1446 1. Kuhn, H. W. The Hungarian method for the assignment problem. *Naval Research Logistics*
- 1447 *Quarterly* **2**, 83–97 (1955).
- 1448 2. Pomponio, R. *et al.* Harmonization of large MRI datasets for the analysis of brain imaging
- 1449 patterns throughout the lifespan. *Neuroimage* **208**, 116450 (2020).
- 1450 3. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and
- 1451 annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).
- 1452 4. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based
- 1453 Linkage Analyses. *Am J Hum Genet* **81**, 559–575 (2007).
- 1454 5. Samper-González, J. *et al.* Reproducible evaluation of classification methods in
- 1455 Alzheimer’s disease: Framework and application to MRI and PET data. *NeuroImage* **183**,
- 1456 504–521 (2018).
- 1457 6. Wen, J. *et al.* Convolutional neural networks for classification of Alzheimer’s disease:
- 1458 Overview and reproducible evaluation. *Medical Image Analysis* **63**, 101694 (2020).
- 1459 7. Wen, J. *et al.* Reproducible Evaluation of Diffusion MRI Features for Automatic
- 1460 Classification of Patients with Alzheimer’s Disease. *Neuroinformatics* **19**, 57–78 (2021).
- 1461 8. Stasinopoulos, D. M. & Rigby, R. A. Generalized Additive Models for Location Scale and
- 1462 Shape (GAMLSS) in R. *Journal of Statistical Software* **23**, 1–46 (2008).
- 1463 9. Wen, J. *et al.* Characterizing Heterogeneity in Neuroimaging, Cognition, Clinical
- 1464 Symptoms, and Genetics Among Patients With Late-Life Depression. *JAMA Psychiatry*
- 1465 (2022) doi:10.1001/jamapsychiatry.2022.0020.

- 1466 10. Zhao, B. *et al.* Genome-wide association analysis of 19,629 individuals identifies variants
1467 influencing regional brain volumes and refines their genetic co-architecture with cognitive
1468 and mental health traits. *Nat Genet* **51**, 1637–1644 (2019).
- 1469 11. Klein, A. & Tourville, J. 101 Labeled Brain Images and a Consistent Human Cortical
1470 Labeling Protocol. *Frontiers in Neuroscience* **6**, 171 (2012).
- 1471 12. Elliott, L. T. *et al.* Genome-wide association studies of brain imaging phenotypes in UK
1472 Biobank. *Nature* **562**, 210–216 (2018).
- 1473 13. Nadeau, C. & Bengio, Y. Inference for the Generalization Error. in *Advances in Neural*
1474 *Information Processing Systems 12* (eds. Solla, S. A., Leen, T. K. & Müller, K.) 307–313
1475 (MIT Press, 2000).
- 1476