Linking interindividual variability in brain structure to behaviour

Sarah Genon[1,2], Simon B. Eickhoff[1,2] and Shahrzad Kharabian[1]

[1]Institute of Systems Neuroscience, Heinrich Heine University Düsseldorf, Düsseldorf, Germany.

[2]Institute of Neuroscience and Medicine (INM-7), Research Centre Jülich, Jülich, Germany.

Corresponding author:

Correspondence should be addressed to Sarah Genon at the Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Centre Juelich, Germany, e-mail: s.genon@fz-juelich.de.

Abstract

"What are the brain structural correlates of interindividual differences in behaviour?" More than a decade ago, advances in structural MRI opened promising perspectives to address this question. The initial wave of research then progressively led to substantial conceptual and methodological shifts. A replication crisis has unveiled the limitations of traditional approaches, searching for associations between local measurements of neuroanatomy and behavioural variables in small samples of healthy individuals. Given these methodological issues and broadening scepticisms regarding the idea of one-to-one mappings between psychological constructs and brain regions, new perspectives emerged. These embrace the multivariate nature of structural brain-behaviour relationships and promote generalizability, but also representation of the relationships between brain structure and behavioural data by latent dimensions of interindividual variability. Here, we review the past and present of the study of structural brain-behaviour associations in healthy population and address current challenges and open questions for future investigations.

The quest of the brain structural bases of differences between people in behavioural aspects such as personality traits or intelligence has always impassioned scientists. In the last two decades, advances in structural imaging opened the door to a broad range of reports of associations between specific aspects of human behaviour and local brain morphological features in healthy populations. This literature, in turn, served as an empirical background on which further investigations and conceptual theories about interindividual variability are still building, such as the interpretation of local brain morphological differences between men and women[1] or the genetic bases of interindividual variability in psychological measures[2]. However, this empirical pillar of cognitive neuroscience has recently been shaken by a replication crisis.

Accordingly, after a brief historical perspective on previous century practices and motivations, we here first review the main developments and influential studies based on neuroimaging measures of brain structure and describe the subsequent replication crisis that has progressively emerged in the study of structural brain-behaviour associations (SBB). As main potential contributing or limiting factors, small sample size and sampling variability, but also the multicollinearity of brain voxels/vertices, as well as of behavioural variables are considered. These considerations converge with a conceptual shift from one-to-one mapping between brain region and behavioural features towards a multivariate view. In that view, we then address two perspectives. First, we consider the mapping of multiple brain variables to a specific behavioural variable using predictive approaches. Second, we describe multivariate approaches aiming to identify brain-behaviour latent dimensions by accounting for the multivariate nature of both sets of data, brain variables and behavioural variables. In the last section, we also consider the possible pitfalls and limitations in these new trends, as well as the challenges for an ideal out-of-sample replication. Finally, as a closing section, we discuss the interpretation pitfalls in line with current methodological challenges.

Mapping local brain structure to behaviour

One striking feature of humankind is our marked inter-individual variability in behavioural aspects such as personality and cognition. This observation has long driven the quest for elucidating their biological bases. Over the past two decades, neuroimaging has offered the possibility to perform brain morphometry in-vivo across samples of individuals and hence the examination of structural brain features in relation to variability in individual traits[3]. In particular, current neuroimaging techniques readily provide estimates of local grey matter volume and cortical thickness across the brain (see Box 1). As reviewed by Kanai and Rees in 2011[3], reports of structural brain-behaviour (SBB) associations quickly ranged from common cognitive functions, such as working memory[4], to social and affective traits measured with standard questionnaires, such as personality traits[2,5] or impulsivity[6] but also extended to a variety of aspects evaluated with specific instruments pertaining to dedicated theories or

research aims. As illustrated in Fig. 1, complex behavioural aspects such as political orientation[7] and the number of Facebook friends were hence related to local grey matter volume (i.e. grey matter volume in specific parts of the brain). A vast SBB literature has hence emerged since the development of MRI techniques for quantifying brain structure.

Box 1: neuroimaging estimates of brain structure

Grey matter Volume/Concentration (GMV) is generally assessed using *voxel-based morphometry (VBM)*. This technique aims to quantitatively compare the T1-weighted scans across individuals by examining the local composition of brain tissue after macroscopic differences have been discounted[8]. This is done first by segmenting each anatomical scan into different tissue types and normalizing to a template. The consequences of wrapping an individual scan to a template can then be accounted for by adjusting grey matter images with the spatial deformations parameters. By doing so, local volume of grey matter can be estimated in indivduals[9].

Cortical thickness is in contrast usually estimated using *surface-based techniques* as the cerebral cortical grey matter has a sheet-like structure. After a minimal preprocessing of the images in the volumetric space and assigning a neuroanatomical label (e.g., white matter, cortex) to each voxel, white/grey matter and grey matter/pial surfaces are delineated through an iterative process. A vertex being the place where the points of neighbouring triangles on the surface meet, the cortical thickness at a given vertex is then defined as the distance between the final white and pial surfaces at that vertex[10]. Structural properties of the vertices, such as cortical thickness, but also surface area and curvature could then be studied across a group of individuals at each vertex[11].

Beyond these macro-structural or morphological estimates of grey matter structure, myelination can also be quantified using T1/T2 ratio[12] and, in turn, be related to psychometric data, such as personality scores[13]. Additionally, quantification of white matter properties at the individual level has been facilitated by the development of diffusion MRI. After specific artefacts correction steps, techniques of diffusion tensor imaging can be used to derive local measures (i.e. at each voxel) of white matter properties from diffusion data. The most popular metrics are fractional anisotropy (FA), which reflects the degree of diffusion anisotropy, and mean diffusivity (MD), which is used as an indicator of the overall magnitude of diffusion[14]. While popular measures of grey matter, such as GMV and cortical thickness probe brain macrostructural features, popular diffusion-based metrics such as FA and MD, tap more into white mater microstructural properties. As we here focus on the study of interindividuality variability in brain morphometry features, we focused on the former measures. However, it should be noted that the relationship between interindividuality variability in white matter microstructural features and behaviour has enjoyed a board interest in the scientific literature[15]. Finally, as the acquisition of MRI images beyond T1-weighted scans are getting more common in large sample cohorts, additional features of the brain microstructure can also be derived. Examples of these features, include, but are not limited to, proxies of myelin concentration in the grey matter using quantitative multi-parameter maps (MPM)[16] and proxies of cellular cortical architecture using quantitative modelling of diffusion weighted MRI[17]. Similarly to most popular structural metrics, interindividual variability in these structural estimates have been related to age[18] and interindividual variability in psychometric data[20], respectively. Overall, the range of structural estimates that can be derived from MRI data will enable a rich, multivariate, description of interindividual variability in local brain structure that can in turn be related to interindividual variability in behaviour.

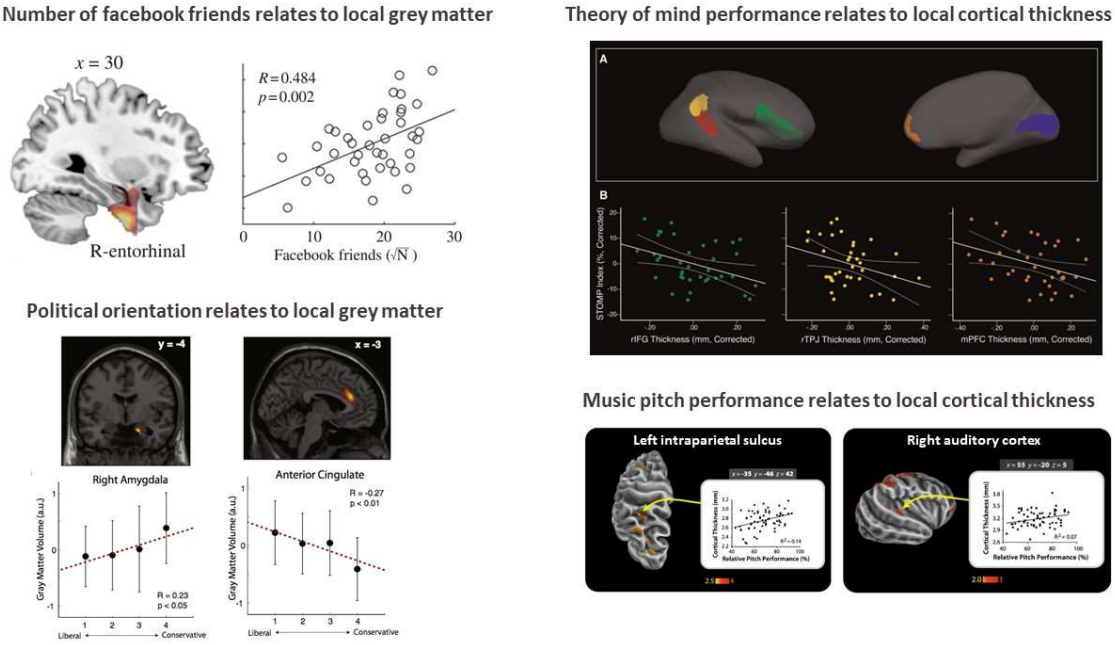

Figure 1. Examples of mapping between local brain morphology and behavioural traits/performance. Left panel: Mapping of political orientation and social network size[19] to local grey matter volume[7]. R value indicates correlation coefficient value and p value indicates the associated p-value. A linear positive relationship between the square root-transformed number of Facebook friend and grey matter volume in the right entorhinal cortex could be found (top panel). Along the same line, a positive linear relationship could be found between the degree of political orientation towards conservative and grey matter volume in the right amygdala, while a negative one could be found between the former and grey matter volume in the anterior cingulate cortex. Right panel: Mapping of theory of mind performance (STOP)[20] and musical performance[21] to local cortical thickness. A negative linear relationship could be found between performance at a spontaneous theory of mind protocol (STOMP) and cortical thickness in the right inferior frontal cortex (rIFG), the right temporo-parietal junction (rTPJ), the medial prefrontal cortex (mPFC) (top panel). A linear positive relationship could be found between the performance in processing relative pitch and cortical thickness in the left intraparietal sulcus and in the right auditory cortex.

The replication crisis

A broad replication crisis has recently emerged in neuroscience and psychology[22-24] with some authors even suggesting that the majority of published neuroscientific literature may not be replicable[25]. In that context, the replicability of traditional SBB studies has likewise been questioned. A purely confirmatory replication study of structural brain-behaviour correlations[26] has been conducted by re-addressing the findings of several SBB studies (Fig. 2). Strikingly, for almost all the examined associations, support for the original results could not be found in the replication investigations. In fact, for most (previously

significant) relationships, confirmatory Bayesian hypothesis tests indicated evidence in favour of the null hypothesis[26]. These worrying findings were then followed by vivid discussions between the authors and others on the limitations in replication studies[26-29].

Along the same line, we highlighted that, a significant correlation between a standard measure of cognitive control and GMV in a well-powered sample could not be replicated in a second, demographically similar, sample even when using the same analysis pipeline (Fig. 2)[30]. It could be argued, though, that the focus on specific region-of-interest in these replication attempts potentially limited the replicability of observed relationships. To address this question, an extensive evaluation of the replicability of SBB associations with an exploratory and confirmatory approach across a range of psychometric variables and by computing different replicability indices drawing from a single large cohort of healthy adults[31] was then conducted. The ensuing results not only indicated that finding significant associations is relatively unlikely, but also demonstrated that associations found in a well-powered subset of subjects could hardly be replicated in a second, matched subset from the same cohort (Fig. 2). These results point towards a publication bias for associations between brain structure and behavioural measures in which null findings are likely to be very frequent, but unreported, whereas significant findings receive more attention but have very poor replicability.

It bears mentioning, that these worrisome findings primarily pertained to voxel-based morphometry (VBM), as this approach has been criticized as a relatively crude estimation of local neuroanatomy (see Box 1), which may contribute to the low replicability of probably rather delicate associations. This raises the question, whether other approaches may lead to more replicable SBB associations. Directly addressing this question, recent studies demonstrated that previous reports of association between personality traits on one hand, and cortical thickness, surface area or white matter integrity on the other, could likewise not be replicated in a large cohort[32]. These findings were corroborated by poor replicability of associations between cortical thickness estimates and a range of behavioural measurements within a high-quality large dataset of healthy young adults[33]. Thus, it appears that the replication crisis in SBB associations does not concern a specific MRI measure of brain structure, but is a general crisis that encompasses grey and white matter volumetric measurements, as well as surface-based measurements.

Altogether these findings resonate with a general context of replication concerns for neuroimaging studies (e.g.[34]). In that context, inter-scanner variability in cortical thickness estimation remains as an important factor to control for[35,36]. Furthermore, the validity and reliability of analysis software were also discussed[37,38]. Worrisome differences for specific versions of a given software have hence been reported a decade ago[37]. The low replicability of associations between cognitive factors and cortical thickness when evaluated across different cortical thickness pipelines was also demonstrated[39] (Fig. 2). However, for the recent versions of commonly used estimation pipelines, a relatively reliable thickness estimation of cortical thickness and its interindividual variabilities has been demonstrated in spite of

focal estimation failures[38]. Hence, although the users community of neuroanatomical computational tools can be reassured to some extent, many issues pertaining to inter-scanner, cross-versions and local effect of processing pipelines variability should still deserve attention in SBB studies.

Besides these concerns on the quantification of individual brain structure with MRI techniques, the psychometric properties of behavioural measures should also be carefully considered. The measurement of core concepts in differential psychology, such as intelligence and personality traits, have generally enjoyed continuous development and evaluation (e.g.[40]) aiming to improve their validity and reliability. In contrast, more research-field-specific constructs may show poorer validity, reliability and/or inadequate distribution for statistical analyses. The lack of variations in political orientation variables was, for example, one factor potentially contributing to SBB replication attempts failure (Fig. 2 top left panel)[26]. Thus, when working on complex psychological concepts and brain estimates derived from MRI techniques, the limited validity, reliability and distributions of variables from both, brain and behaviour sides, should be kept in mind as limiting factors partly contributing to replication issues in SBB studies.

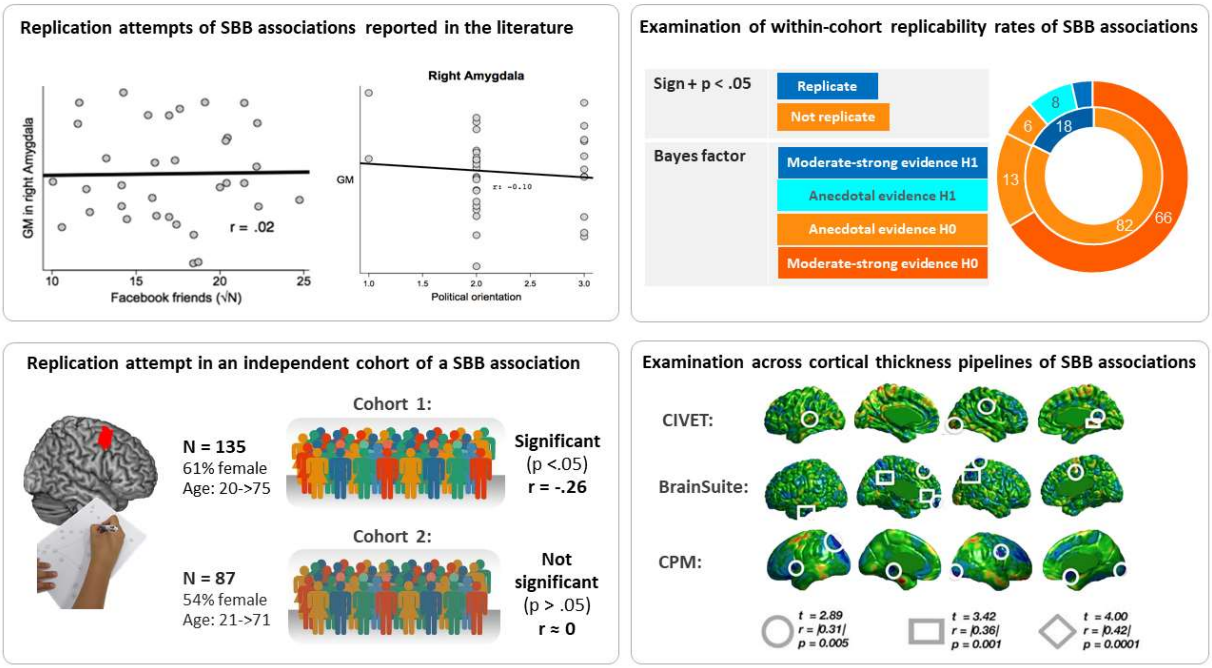

Figure 2. Poor replicability of structural brain-behavioural (SBB) associations. Top left panel: replication attempts of previous SBB, GM: grey matter, r; correlation coefficient in the replication sample[26]. While significant linear relationship between grey matter volume in the right amygdala and the square root-transformed number of friend and also the political orientation was previously evidence, replication investigations failed to evidence this relationship with a coefficient correlation value r close to 0. Bottom left panel: illustration of the replication attempt of a found association between grey matter volume in the dorsal premotor cortex and performance at the Trail Making Test in an independent cohort[30]. While a significant negative correlation was found in a first sample (cohort 1), a replication attempt in another cohort (cohort 2) did not show a significant relationship. Right top panel: within-cohort replication attempts of an association between perceptual IQ and grey matter volume (figure adapted[31]). The outer ring of the donut plot reflects Bayes factor indices while the inner

ring reflects p-value- and sign-based replication indices. While a significant association was initially found, replication attempts across 100 resampling show a high rate of replication failures when considering p-value and direction of the association (82% of replication failure shown in the inner ring) and correspondingly a high rate of moderate to strong evidence for no association (66% shown in the outer ring). Bottom right panel. Local cortical thickness associations with a working memory component score across three different cortical thickness pipelines (CIVET, BrainSuite and CPM)[39]. The significant associations based on r and t values vary in their spatial location across the different used pipelines suggesting that different investigations of associations using different analysis' choices provide different patterns of association.

In sum, the replicability of SBB associations has been questioned across a range of behavioural measurements and various popular estimates of brain structure. Importantly, this pertains to replications of findings from previously published papers[41], but also for within-study replications of findings in an independent sample and even when subsampling within a given cohort. These results should not lead however to the conclusion that association between brain structure and behavioural phenotype are per se unattainable. Rather, they suggest that the magnitude of such associations may be relatively limited and not reliably captured by current standard approaches.

Small effect sizes, power and false positives

Recent replication studies, but also explorative studies in very large datasets such as the UK BioBank, have suggested that the effect size of associations between brain structure and behavioural aspects as measured by standard psychometric tools in normal populations are likely very small[31,32], seemingly overall around 0.10 in terms of correlation coefficient[41,42]. The probability of finding an effect that is genuinely true, e.g., an association between estimates of local brain structural and behavioural measurements, that is, the power of an experiment, is directly related to both, the effect size and the sample size. Consequently, the probability of finding a true effect is relatively limited in small samples. More precisely, samples consisting of ~200-300 participants appear to have low power to identify reliable SBB-associations among healthy participants. Recent investigations actually point out that substantially bigger cohorts of ~1000 participants are required for reliably identifying SBB associations for standard cognitive tests (such as intelligence tests) and psychological scales (such as personality and psychopathology scales)[31,41]. Hence, overall, small samples show extremely low power to find a real association.

Yet, the vast majority of SBB studies forming the current literature is typically based on relatively small samples ($n<200$). It can therefore reasonably be stated that the probability of having reported spurious or inconclusive results in these studies is extremely high. While the exact factors driving spurious associations can remain as a topic of investigation, it can be assumed that report of statistically significant associations could be influenced by data dredging and related practices[43]. At the conceptual level, these issues imply that SBB findings and neuroscientific theories building on those must be taken

with caution. In turn, at the methodological level, together, small effect sizes for specific brain-behaviour associations and the false positive risk of data fishing expeditions imply that large cohorts and alternative approaches are needed to bring insight into the relationship between brain structure and behavioural aspects in healthy populations[44,45].

Multivariate brain structure-behaviour mapping

Multivariate analyses as an alternative approach

From an epistemological standpoint, the first scientific evidence of mapping between brain structure and behavioural functions in humans emerged from lesions studies[46]. By showing causal relationship between a relatively localized lesion and relatively specific behavioiral deficits, the first lesion studies hence supported the concept of a relatively specific mapping between brain regions and the respective behavioural function, such as between the hippocampus and episodic memory[47]. However, in the last decades, this one-to-one mapping initial conceptualization has been revisited in favour of a many-to-many view, in particular following the boom of functional and structural neuroimaging studies in healthy populations. Accordingly, the underlying mechanisms that give rise to the complex behavioural aspects probed by psychometric tools are nowadays thought as not modularly localized to individual brain regions, but to rather rely on distributed neural networks[48,49]. In other words, while the structural integrity of some regions appears to be needed for normal functioning in one behavioural domain based on seminal lesion studies in clinical populations, variations of performance in the range of normal functioning seem to rely on structural variations across a range of brain regions.

Yet, the search for structural correlates of behavioural measurements in healthy populations has been typically performed by capitalizing on statistical univariate approaches in which a statistical test of association with a behavioural variable is performed locally in the brain, either with a general linear model for each voxels or brain regions or with a correlation approaches with a-priori defined region(s)-of-interest. Considering the small effect size and the often assumed spatially distributed nature of these associations in the brain, univariate approaches appear particularly limited to capture complex brain-behaviour relationships. Furthermore, these approaches do not take into account the mutual dependencies between different brain voxels/vertices or regions which are readily seen in structural covariance pattern[50]. For these reasons, in exploratory studies whose aim is to identify brain structural features correlating with a given (set of) psychological variable(s), multivariate techniques offer an alternative approach taking into consideration the multivariate nature of brain data[51,52].

To consider the joint covariance of many brain regions (or voxels/vertices) with a given behavioural variable, a multiple regression approach can be used. In that view, we will describe how brain structural features can be conjointly mapped to a specific behavioural measure by using prediction techniques

taking the form of a regularized multiple regression. However, when studying brain and behaviour, it is worth considering that mutual dependencies between multiple sources of measurements exist not only between brain variables (voxels, vertices or regions), but also among behavioural measurements. From an epidemiological standpoint, collinearity between behavioural measurements can be expected in a population covariance framework, as it could be assumed that unique factors, such as age, education, or culture, influence many aspects of the behavioural phenotype conjointly (see last section). Furthermore, from a psychological standpoint, collinearity between different cognitive measures can be assumed to happen because these different measurements tap onto the underlying process/hidden constructs or processes (e.g.[53-55]). This latter consideration generally justifies the use of factorial analyses in psychological sciences to extract latent factors (also called synthetic variables)[56]. The derived synthetic behavioural variable(s) can then be mapped individually to brain structure using either univariate (such as the traditional voxel-wise GLM) or multivariate approaches. The former approach will not be addressed here as, from a conceptual standpoint, it could be assumed that synthetic variable in healthy population psychometric data reflects a broad behavioural dimension (such as "fluid intelligence") which could hence be expected to be multidetermined with regards to cognitive processes and thus would not map to specific brain regions[57,58]. The mapping of such synthetic variable to brain structural features considered conjointly can be performed using a predictive framework as described in the following section.

Predicting a behavioural variable from a brain structural multivariate pattern

A prediction framework enables mapping multiple brain structural features, considered jointly, to a single behavioural variables. Concretely, a set of brain variables are used as predictors for a target behavioural variable. As aforementioned, this variable of interest could be a measured variable, such as specific score at a questionnaire, but also a synthetic variable, such as a composite score defined based on specific theories (e.g. typically an intelligence composite score) or a loading score computed based on the data factorization. For instance, when starting from an applied clinical question or a psychological sciences standpoint, the demonstration of a relationship between a large set of brain structural features and the investigated variable could contribute to the neurobiological validation of the latter. Put it simply, if participants' score for this variable can be predicted from brain structural data, a relationship between the brain structural features and the behavioural variable can be assumed. The predictive power is generally evaluated within a machine learning framework with a cross-validation setting (Box 2). By doing so, the generalizability of the fitted brain-behavioural relationship out of the training sample is evaluated using the prediction accuracy in a validation or unseen dataset (Box 2).

Brain-based prediction of behavioural phenotype has enjoyed a vivid interest in the neuroimaging community across the recent years. Presumably for conceptual reasons (behavioural function arise from

functional interaction between brain regions), the majority of psychometric prediction studies has focused on brain resting-state functional connectivity as key features (or a combination of structural and functional connectivity features). Accordingly, the main developments and discussions in brain-based predictions that SBB can now enjoy have originally arisen from functional (and/or diffusion) studies. Nevertheless, in the last decade, a good dozen of studies have evaluated the prediction of a behavioural variable based on estimated brain structures using a cross-validation approach. These pioneer implementations of a machine learning approach for SBB associations were generally conducted in a hypothesis-driven framework to demonstrate associations between specific brain regions' structure (such as the amygdala[59] or the striatum[60,61]) and a specific cognitive[60-62], mood[59] or personality aspect[63,64]. In a related framework, the sample sizes used in the first predictive SBB studies were generally relatively limited (below 200[65], 100[59,64,66]or even 50[60,62] participants) leading to inflated effect sizes, similar to observations made in univariate studies[31]. More generally, following a global trend in the application of predictive approaches of phenotype from neuroimaging features, the prediction power as reflected by prediction accuracy metrics was overestimated due to the limited cohort size, improper cross-validation scheme affected by data leakage or double dipping[67]. Hence, the first studies in limited cohorts (<200 participants) generally reported very optimistic prediction accuracy (in terms of correlation between the predicted and observed behavioural scores) ranging from 0.40 to 0.74[59-62,64-66]. However, when sample sizes got bigger (> 200 participants), lower prediction accuracies were observed, within a range from 0.11 to 0.28[57,68,69], in validation datasets (Fig. 3).

From a basic neuroscientific standpoint, the modest prediction of behaviour from brain structure in healthy population may suggest a limited contribution of interindividual variability in brain structure to interindividual variability in behaviour. However, it should be noted that, overall, the prediction performance of behaviour based on brain structure is similar to performance achieved when instead brain functional features are used (usually functional connectivity estimates)[70]. Along the same line, despite combination of structural features with functional features in multimodal frameworks generally leads to increased predictive power than focusing on single modality in large cohorts[71-73], the predictive power remains relatively limited[74,75]. This state of the art hence highlights that predictive models of behaviour based on neuroimaging markers in healthy populations still hold their own challenges, regardless of the neurobiological aspect probed (brain structure or functional connectivity). Acknowledging these global challenges, the contribution of brain structure to the prediction of behavioural phenotype remains an important research topic. Preliminary investigations[74] suggest that using the same dataset of healthy adults, predictive models based on brain structural features may perform as well, or even better, than those based on functional features for the prediction of some behavioural scores (Fig 3.). It could also been seen from these investigations that, when prediction is based on multimodal data, structural features (such as surface area, cortical thickness and grey matter volume) tend to have higher weights than functional features in the prediction of many cognitive

measures[74]. Despite further studies are needed, brain structural features thus already appear to represent an important source of information in the study of brain-behaviour relationship. In the next section, we further discuss approaches that address this question while considering a range of behavioural variables jointly.

Mapping sets of phenotypical variables to brain structural data

A complementary view to the mapping of multivariate brain features to one behavioural variable can be offered by a "doubly-multivariate approaches" in which a set of brain variables are jointly mapped to a set of behavioural variables[76]. More concretely, covariance patterns between two sets or blocks of data, here behavioural and brain data, can be summarized along latent or hidden dimensions[76] that are linear (or non-linear) combinations of brain (such as grey matter volume in several brain regions) and behavioural variables. This doubly-multivariate approach hence enables the representation of broad patterns of interindividual variability in brain structure and behaviour in a latent space formed by emerging distinct dimensions. In addition to enabling possibly new structure-behaviour factors to emerge as a result of considering both multivariate sets of features simultaneously, from a conceptual standpoint, such approaches avoid the pitfall of focusing on a single a-priori behavioural aspect as a tree that hides the forest. Furthermore, from a data science or statistical standpoint, such approaches take into account, both at the brain and behavioural level, that different variables can represent redundant sources of variability.

Partial least squares correlation (PLSC) and the closely related canonical correlation analysis (CCA) are the most popular techniques in that view[77,78]. Concretely, these techniques maximize the association between linear combinations of brain and behavioural variables by searching for weight vectors or directions, such that the projection of the dataset(s) (e.g., the brain and behaviour) onto the obtained weight vector(s) has maximal correlation (CCA), or covariance (PLSC). The resulting profiles of weights or individual correlations for each dataset can then also be examined providing insight into the features that form the association. Hence, these approaches decompose the complex nature of brain-behaviour associations into parsimonious overlapping patterns, dissociating different aspects of brain-behaviour relationships. For example, the relationships between IQ scores and interindividual variability in "morphometric similarity networks" have been explored in a large sample of adolescents with this approach[79]. Focusing on interindividual variability in standard intelligence measures hence revealed two latent dimensions capturing interindividual variability in distinct brain systems roughly corresponding to language and cognitive control networks versus visual and memory networks.

However, generally, such doubly-multivariate approaches have been more intensively used to map brain functional connectivity to behavioural phenotype[80,81], in particular in clinical populations[82-86]. When focusing on a large healthy cohort (the Human Connectom Project, HCP[87]), such an approach revealed a main mode or dimension of covariance linking brain connectivity to demographics and behaviour[80]. Follow-up data-driven studies in the same cohort giving more attention to structural brain estimates strikingly spotted light on the substantial contribution of brain structural interindividual variability in the initially reported population mode[88,89]. In particular, relating brain cortical thickness to a range of behavioural/life style measures with a similar CCA approach in the same cohort (HCP) replicate the main significant mode or dimension, initially reported with resting-state functional connectivity, portraying a positive vs negative pole of behavioural phenotype with measures such as fluid intelligence, vocabulary, life satisfaction vs. behavioural aggression, tobacco use, cognitive failure (Fig. 3). Interestingly, the associated cortical thickness pattern showed a clear differentiation across the cortical hierarchy with positive correlations mostly at lower order sensory/motor areas and negative correlations mostly at higher-order brain regions (including the frontal, anterior temporal, and parietal cortices that encompass the most parts of the default mode network)[88]. The emergence of this pattern offered by the doubly-multivariate approach suggests that interindividual variability in the pattern of cortical thickness difference between the lower- and higher-order brain regions could be closely linked to interindividual variability in phenotype.

A similar pattern has been observed in paediatric populations when CCA was applied to a large cohort of adolescents[90,91]. Generally, across studies, higher cortical thickness in frontal regions appear to be associated with more negative life events, lower cognitive functions and more negative social behaviour/increased psychopathology in healthy young populations[88,90,91]. Altogether, the results of recent studies in large cohort using data-driven approaches hence suggest that broad patterns of interindividual variability in brain structure can reliably relate to interindividual variability in behaviour and this, to a similar, or even greater extent than interindividual variability in functional connectivity[89]. These findings further resonate with multimodal prediction modelling according to which brain structural features importantly contribute to prediction of behavioural variables[74]. However, the emerging data-driven pattern can appear relatively minimalist, when expressed along a summary positive vs. negative dimension spanning across different brain features, offering a very limited insight into brain-behavioural phenotype relationships. Alternatively, when multiple and finer dimensions are discovered, establishing their correspondence across different cohorts and their interpretation from a basic neuroscience standpoint may pose some conceptual challenges. Overall, the broader the spectrum of variables included in the doubly-multivariate model, the highest the interpretation challenge. The interpretation and neurobiological validity of such multivariate models is even obviously further complicated when highly derivative features are used as inputs. For example, graph theory-based features of brain structural mearsuments and transformed features thereof may be used as input for a

PLS analysis, or principal components (from Principal Component Analysis) are sometimes extracted from functional connectivity networks derived with Independent Component Analysis and used as input for a CCA analysis. Generally, it should be kept in mind that highly derivative frameworks with oversophisticated analyses may obfuscate our understanding from a neuroscientific conceptual standpoint and easily degenerate into poorly informative neuro-informatics methods. Thus, the exploratory nature of fully data-driven approach should still be complemented by carefully designed SBB associations studies with clear neurobiological theorization and a predictive utility evaluation. In the next section, we discuss further open general challenges for SBB association in healthy populations, with regards to their replication, their relationships with non-brain and behaviour variables, as well as their extrapolation to brain pathology, and finally their discussion beyond the informed expert scientific community.

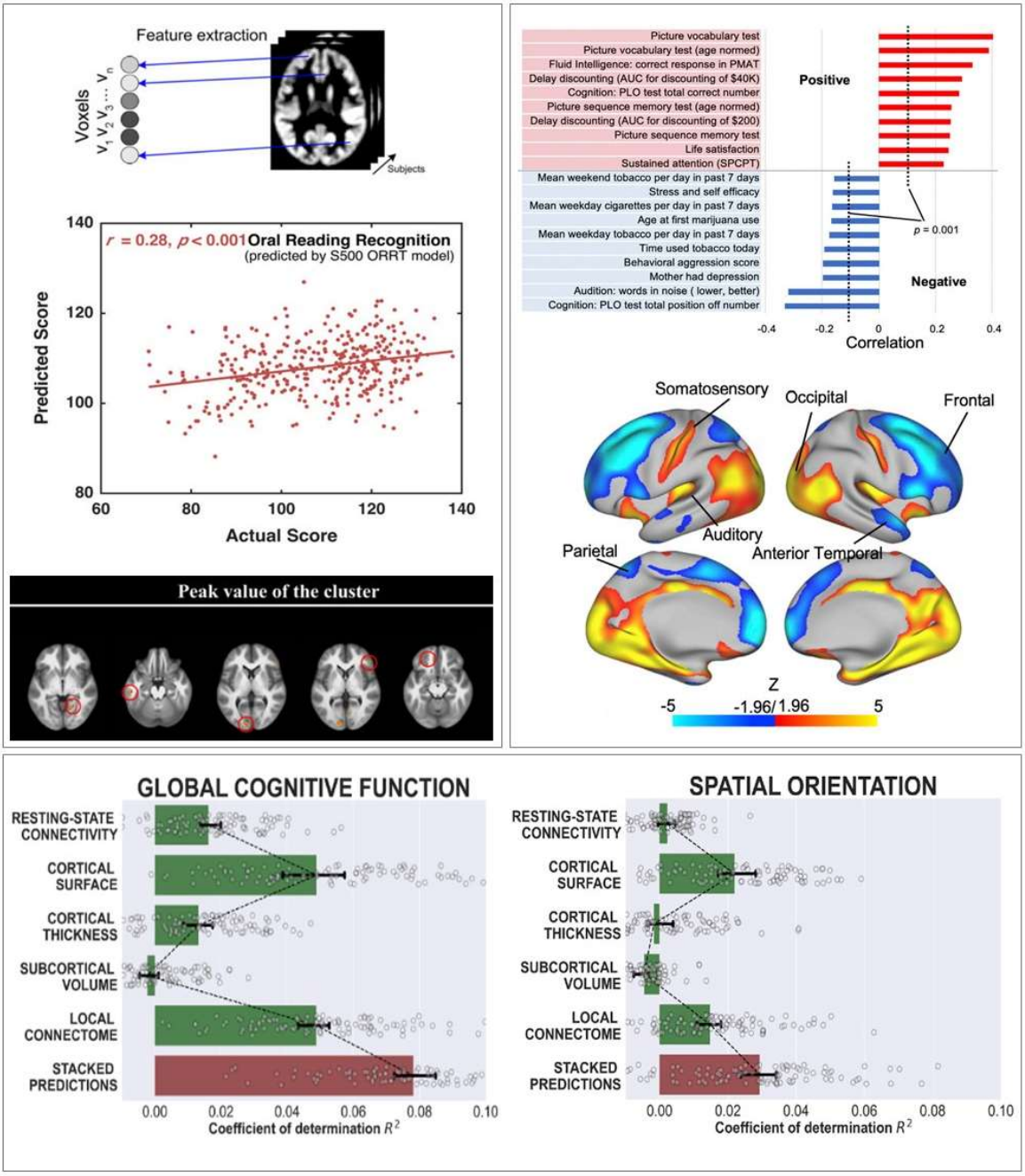

Figure 3. Machine learning/multivariate approaches to study structural brain-behaviour (SBB) associations. Top left panel: prediction of oral reading recognition performance in a validation dataset of the Human Connectom Project, r; coefficient correlation between predicted reading score and actual reading score. The top row illustrates voxel-wise grey matter volume feature extraction. In each subject, the grey matter value estimates at each voxel are used as features or predictors to predict oral reading recognition score. The middle panel shows that the predicted reading score in an unseen sample are significantly correlated with the actual reading score for those subjects. The lower row illustrates some assumed contributing regions derived from the prediction model[69]. Top right panel: illustration of the canonical mode linking interindividual variability in cortical thickness to interindividual variability in behavioural variables in the Human Connectome Project[88]. The behavioural measures most strongly correlated with the identified thickness-behaviour CCA mode are shown in the top panel. The positively correlated subject measures (red) generally reflect positive traits, whereas the negatively correlated measures (blue) pertains to more negative behavioural aspects. The Fisher's z-transformed correlations between local cortical thickness and the identified CCA mode are shown in the lower panel.

Positive correlations (red-yellow colours) were mainly observed in the lower-level sensory/motor regions, while negative correlations (blue-cyan colours) mostly appeared in the higher-order cognitive brain regions. Bottom panel: comparison of the performance of predictive models of psychometric variables using different brain structural and functional features[74]. Model performance is assessed by the coefficient of determination (the higher the coefficient of determination, the higher the prediction accuracy). For some psychometric variables (here a composite score of cognitive function and a score of spatial orientation), a predictive model based on cortical thickness show similar or higher performance than one based on functional connectivity. "Stacked predictions" refer to models combining all types of features.

Box 2: machine learning approaches in SBB

From a statistical standpoint, SBB studies as most types of neuroimaging studies fall into typical N<p problem (i.e. the number of data points is smaller than the number of variables) and multicollinearity issues calling for multivariate analyses and features selection/reduction approaches[92,93]. Large sample sizes enable proper assessment of generalizability and stability of the multivariate solutions within a cross-validated setting. This ensures that the reported patterns are not driven by sample-dependent spurious covariations (despite this does not ensure that the pattern could be replicated in a completely independent dataset). In particular, CCA/PLS approaches are particularly prone to overfitting and although the effect size may seem high when the analyses is initially run on a dataset, the associations are generally much lower in an independent dataset[94]. In other words, overfitting may give an impression that the found associations between brain and behavioural variables is much stronger than it would be in an independent hold-out dataset[94].

Therefore, in a cross-validation setting, the original sample is divided into a train and test (holdout) subsample (or set). In CCA/PLSC approaches, the multivariate patterns are derived based on the train sample and the individuals' data of the holdout set are projected to the weight vectors from this train sample. This process of randomly splitting the data into train and test sets is usually repeated for a limited number of times, e.g. 10[80,95]. Generalizability of the model is then evaluated by summarizing (e.g. averaging) out-of-sample accuracies on the holdout sets. Stability can in turn be evaluated by the similarity of the weights' profiles in the train sets across the cross-validation[80,95]. Beyond the standard versions of CCA and PLSC, variants thereof are developed for neuroimaging-behaviour dataset, aiming at reducing overfitting in high dimensional feature space, and extending the scope of analysis to address nonlinearities in the data (regularized extensions such as sparse CCA/PLSC[96] and kernel CCA[97]). In that context, the identification of the optimal parameters (hyperparameter tuning) for these extended methods also requires a cross-validated setting[95], which puts additional requirements for large sample sizes[98].

Similar concepts and constraints apply to predictive models in which multiple brain features are used to predict a behavioural variable. The generalizability of the fitted model is assessed by using a cross-validation scheme in which the behavioural scores predicted by the model are compared with the observed scores. The model is hence fitted in the training sample and tested in the holdout set by randomly splitting the dataset between train and test a certain number of times. As for specific variants of CCA/PLSC, some prediction algorithms require hyperparameter tuning, which is done through nested cross-validation, which again requires large sample sizes. When the prediction performance of the model has been demonstrated, researchers are often tempted to look at the weights assigned to individual brain features as an indication of the magnitude of the importance of the association between the feature and the behavioural variable. However, because of the multicollinearity of the features, such an approach can be dangerously misleading in a multivariate framework[99]. Accordingly, general caution should remain in the neuroscientific interpretation of brain-behaviour associations patterns by focusing

on highest contributing features in a multivariate framework, be it a predictive or correlational framework.

As discussed above, these approaches have been generally used to investigate relationship between functional connectivity and behavioural phenotype in clinical samples. For CCA/PLS approaches, this context implies that, despite sample sizes that were generally larger than in univariate studies (> 100 participants), the number of subjects relative to the number of investigated features remains relatively small, potentially leading to inflated effect size[100]. It has indeed recently been shown that for these approaches, effect sizes decrease as the sample size increases (a similar statistical phenomenon as reported for univariate analysis and predictive approaches). In other words, the effect sizes reported may often be inflated or at least overoptimistic. Furthermore, as most studies have included a range of phenotypical variables beyond psychometric data (such as life style and demographic variables), the specific magnitude of brain-behaviour associations in such multivariate approaches remains uncertain, although it has been assumed to be small to moderate (≤.30[100]). Given that most previous studies contributing to this discussion point have focused on functional connectivity or multimodal brain data, the question of the strength of associations between brain structure and behaviour in healthy populations remains open for future studies.

Open questions, challenges and interpretation pitfalls

From association in healthy to clinical utility

Historically, as described above, the study of SBB associations has been strongly influenced in its inception by early observations of associations between localized lesions and specific behavioural deficits[46]. However, later, it appears that the region-to-behaviour relationship, suggested by these studies in clinical populations, do not in its simplest form, i.e. one-to-one mapping, extend into interindividual SBB patterns in the healthy populations. Nowadays with neuroimaging techniques having spurred SBB studies in normal population, the relevance of the reverse conceptual extension can similarly be questioned. To which extent the SBB patterns discovered in healthy populations relate and can be used to better understand brain-symptom relationships in clinical population? is indeed a non-trivial question. While acknowledging that the application of machine learning techniques in neuroimaging may remain at a premature stage, when a pure data-driven approach is taken, the patterns of brain-behaviour associations revealed by these techniques do not readily echo the brain mapping literature[58]. For instance, general intelligence score in a cohort of healthy adults has been found to be best predicted by cerebellar grey matter volume[57] in an adult cohort (the enhanced NKI cohort[101]) and to cortical thickness measurements in the sensori-motor cortex in the HCP cohort[74]. While such findings would need "out-of-cohort" (see below) replication in the future, they currently highlight that the scientific path from prediction-based neuroscientific discovery in heathy population to implications for clinical populations is still long and convoluted.

Along the same line, one important question pertains to the relationship between the latent dimensions of interindividual variability emerging in healthy populations and the SBB patterns that can be found

in clinical populations. As previously mentioned, CCA/PLSC approaches have been often used to identify dimensions linking neurobiological patterns to a range of symptoms or cognitive deficits[82-86,102-105]. In that context, (sub)clinical and healthy populations were often mixed, e.g.[86,106], based on the assumption that symptom expression can be summarized on dimensions of psychopathology, which are extensions of dimensions of brain-behaviour variability in the healthy population. However, this assumption may not always hold true, which would result in artificial dimensions when pooling clinical and healthy samples together, an issue that does not only concern multivariate correlational analyses, but that could be less easily tackled in a pure multivariate data-driven framework. Attention should hence be given to the broad interpretation and implications of these latent dimensions in the future. More concretely, scientific investigations should carefully evaluate whether latent dimensions of interindividual variability reflect biological susceptibility from which brain-symptoms/brain-behavioural deficits dimensions in clinical populations represent extreme expression, or whether these normal dimensions reflect very general demographic aspects (such as general education) with limited extrapolation to brain pathology mechanisms. In the next section, we further discuss the challenges of generalizability, modelling confounds and the related interpretation issues.

Absolute out-of-sample replications as an open challenge

The replication crisis in SBB has not only pointed to the limitation of small samples in SBB and hinted at a substantial publication bias, but also highlighted the critical need for out-of-sample replication of SBB findings. This calls for replication attempt in an independent dataset, that is, an "out-of-cohort" replication. However, for practical reasons, in machine learning studies the focus is usually set on generalization within the dataset (see Box 2). This practice unfortunately does not fully prevent the statistical model to capture idiosyncrasies of the dataset from which it has been derived. This pitfall is particularly likely for psychometric data because they rely on assessment tools that have been typically developed for a specific population in a specific context, whose validity across different subpopulations may be variable[107]. Furthermore, psychometric data may be susceptible to examiner effect. Importantly, neuroimaging estimates of brain structures are also susceptible to scanner and sequences effects[36,108], hence both psychometric and neuroimaging data are particularly prone to batch effect.

In this context it is particularly worrisome that the vast majority of studies capitalizing on large healthy cohort of young adults, tend to focus on the few openly available datasets[101,109]. As can be seen from the above review, studies in healthy population, aiming for large sample size and extended behavioural phenotyping, tend to specifically rely on the HCP dataset[87] clearly biasing in the current literature. Due to limited data (and computational) resources, these studies seldomly include replication in an independent sample. Evaluating the extent to which standard and higher construct measurement consistently map onto similar structural brain patterns across different cohorts will hence depend on the

availability of additional healthy population cohorts in the future. In that regards, openly available datasets, representative of the world-wide populations, are needed. Diversity in the dataset is not only needed at the geographical level, but also in other demographic and sociocultural aspects. Such endeavours would not only enable conceptual replications, but also promote new discoveries on the factors influencing brain-behaviour relationships.

Covariance, Confounds, mediating variables and wise scientific interpretation

When aiming to understand how brain structure and behavioural phenotype are related, the explicit definition of confounds and mediating variables play a major role. The definition as either confounds or mediating variables should rely on a-priori assumptions that should be carefully formulated or evaluated. For instance, on the one hand, head morphometry (i.e. size but also shape) can be associated with specific pattern of structural estimations in some part of the brain (resulting from normalization to a standard template). On the other hand, it may also be indirectly associated to some specific patterns of behaviour (for example women have typically smaller head and are early encouraged to develop care-taking behaviour). Consequently, head and therefore brain size can be considered as a confounder with the potential to create spurious brain-behaviour associations (such as association between grey matter volume and care-taking behaviour).

Illustrating that concern, in a healthy adults' cohort, when multimodal brain components including several structural measures were correlated with a range of phenotypical variables, several components detected simple associations between brain size (encoded in grey matter density and cortical area)and gender, strength and endurance[89]. This raises the very concerning question of the extent to which the SBB associations can be spuriously influenced by general head morphometry. In that view, head morphometry should be controlled for as a confounding variable. In contrast, when a variable is assumed to underlie a brain-behaviour association based on specific neurobiological assumption, this variable should deserve specific attention and mediation approaches to disentangle its relevant role should be considered. A typical example of such case is the influence of age. If older age is assumed to go along with lower grey matter volume in some parts of the brain, which in turn results in reduced cognitive abilities, age should not be considered as a confounder. Nevertheless, clear and documented assumptions on confounders and mediators are rarely formulated and discussed in recent multivariate studies. A range of variables are often controlled for (from in-scanner movements to education) or included in the multivariate model without any a-priori and/or post-hoc conceptual considerations of the role of these variables (e.g.[74,80]). Yet, different adjustment or controlling can substantially modify the pattern of results, and sometimes, in an unexpected way. For example, different strategies for brain size adjustment may potentially result in negative associations of brain structure with behaviour and the researchers should be aware that the interpretation is conditioned by the adjustments (see e.g.[110,111]).

Relatedly, it is generally acknowledged that brain-behaviour covariance does not mean causation, even in a predictive design. Accordingly, evidence of a relationship is rarely explicitly presented as a causal relationship. Nevertheless, the communication of the message outside the scientific community, in particular by general media, who contribute to a simplified communication of the statistical results and mass dissemination, is rarely carefully considered. Scientific researchers, themselves, may foster causal interpretation through the terminology they used. While the exact neurobiological mechanisms accounting for findings of SBB associations are not clear, the terminology used to some extent implicitly conveyed the ideas of the behavioural aspects shown by individuals as "hard-wired" or "deeply biologically rooted". This can be seen by the terminology used by Kanai & Rees[3] in "the structural basis of inter-individual differences in human behaviour and cognition". But this can also be seen in the frequent use of the term "substrate" or "bases" (e.g.[112,113]), which is semantically related to "foundation", in this literature. Further illustrating a-priori assumption on directionality of phenomenon can be found on recent machine learning studies: "*A parsimonious interpretation would be that the structural difference leads to functional changes. However, we should not exclude the possibility that spontaneous brain activity underlying the resting state functional connectivity may help to shape brain structure and morphology, such as the cortical thickness*"[88]. Hence brain structural patterns are often implicitly assumed to determine functional patterns and behavioural phenotype and following that view, these "bases" are often primarily assumed to emerge from nature.

Somewhat contrasting with this implicit view, compelling evidence of a relationship between human skills and traits and brain morphometry were actually initially brought by several studies probing brain plasticity[114]. Subtler changes were hence shown following new learning/training demonstrating dynamic relationships between regular behaviour and brain structural patterns. Along the same line, current theories suggest that functional interaction between brain regions early in development may drive observed patterns of co-morphology between brain region, rather than the other way around[115,116]. Furthermore, there is now a substantial literature on the relationship between external factors and brain structure development[117-119]. Reinforcing these evidence, in a very large paediatric cohort, the variance in the sociodemographic factors was found to be shared with the cognitive and brain structural features and accordingly, the relationship between brain structure and cognition was strongly related to the sociodemographic factors[120]. Along the same line, several recent studies suggest that the relationship between brain structure and behavioural phenotype is driven or mediated by family income in developmental cohorts[121,122]. Hence, altogether, all these lines of evidence point to brain plasticity as a major aspect that should always be a-priori considered when interpreting, discussing relationships and communicating the findings outside the scientific expert community.

A vast SBB association literature has been produced across the last two decades offering additional empirical background for differential psychology and cognitive neuroscience theories. However, this empirical pillar has recently been shaken by a replication crisis. The poor replicability of SBB associations has been hence demonstrated across a range of behavioural measurements and various popular estimates of brain structure. This crisis should lead the field to reconsider our scientific approaches of SBB associations based on interindividual variability patterns in the healthy population. In particular, large cohorts are crucially needed (~1000 participants). In parallel, machine learning techniques, by taking into consideration the multivariate nature of structural brain and behavioural measurements and promoting robust association patterns, promise to offer a complementary view to the more traditional univariate approaches. These new trends have already highlighted that interindividual variability in brain structure relates to an extent, which is at least similar to the one of brain functional connectivity to interindividual variability in behavioural phenotype. However, the conceptual validity of the findings remains to be carefully evaluated with "out-of-cohort" and conceptual replication. This challenge dramatically calls for the availability of additional cohorts. In that regards, geographic diversity and socioeconomical diversity endeavours in new cohorts further hold the keys of the usefulness of the derived model and further insight into moderator factors in SBB associations. Ultimately such endeavours could contribute to better identify the role of environmental factors in SBB and develop actions for education and diseases prevention strategies.

References

1 Kiesow, H. *et al.* 10,000 social brains: Sex differentiation in human brain anatomy. *Science advances* **6**, eaaz1170 (2020).

2 Valk, S. L. *et al.* Personality and local brain structure: Their shared genetic basis and reproducibility. *NeuroImage* **220**, 117067 (2020).

3 Kanai, R. & Rees, G. The structural basis of inter-individual differences in human behaviour and cognition. *Nature Reviews Neuroscience* **12**, 231-242 (2011).

4 Colom, R., Jung, R. E. & Haier, R. J. General intelligence and memory span: Evidence for a common neuroanatomic framework. *Cognitive Neuropsychology* **24**, 867-878, doi:10.1080/02643290701781557 (2007).

5 Nostro, A. D., Müller, V. I., Reid, A. T. & Eickhoff, S. B. Correlations between personality and brain structure: a crucial role of gender. *Cerebral cortex* **27**, 3698-3712 (2017).

6 Matsuo, K. *et al.* A voxel-based morphometry study of frontal gray matter correlates of impulsivity. *Human brain mapping* **30**, 1188-1195 (2009).

7 Kanai, R., Feilden, T., Firth, C. & Rees, G. Political orientations are correlated with brain structure in young adults. *Current biology* **21**, 677-680 (2011).

8 Good, C. D. *et al.* A voxel-based morphometric study of ageing in 465 normal adult human brains. *Neuroimage* **14**, 21-36 (2001).

9 Mechelli, A., Price, C. J., Friston, K. J. & Ashburner, J. Voxel-based morphometry of the human brain: methods and applications. *Current Medical Imaging* **1**, 105-113 (2005).

10 Fischl, B. & Dale, A. M. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences* **97**, 11050-11055 (2000).

11 Greve, D. N. & Fischl, B. False positive rates in surface-based anatomical analysis. *Neuroimage* **171**, 6-14 (2018).

12 Glasser, M. F. & Van Essen, D. C. Mapping human cortical areas in vivo based on myelin content as revealed by T1-and T2-weighted MRI. *Journal of Neuroscience* **31**, 11597-11616 (2011).
13 Toschi, N. & Passamonti, L. Intra-cortical myelin mediates personality differences. *Journal of personality* **87**, 889-902 (2019).
14 Le Bihan, D. *et al.* Diffusion tensor imaging: concepts and applications. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* **13**, 534-546 (2001).
15 Forkel, S. J., Friedrich, P., Thiebaut de Schotten, M. & Howells, H. White matter variability, cognition, and disorders: a systematic review. (2020).
16 Weiskopf, N. *et al.* Quantitative multi-parameter mapping of R1, PD*, MT, and R2* at 3T: a multi-center validation. *Frontiers in neuroscience* **7**, 95 (2013).
17 Menon, V. *et al.* Microstructural organization of human insula is linked to its macrofunctional circuitry and predicts cognitive control. *Elife* **9**, e53470 (2020).
18 Carey, D. *et al.* Quantitative MRI provides markers of intra-, inter-regional, and age-related differences in young adult cortical microstructure. *Neuroimage* **182**, 429-440 (2018).
19 Kanai, R., Bahrami, B., Roylance, R. & Rees, G. Online social network size is reflected in human brain structure. *Proceedings of the Royal Society B: Biological Sciences* **279**, 1327-1334 (2012).
20 Rice, K. & Redcay, E. Spontaneous mentalizing captures variability in the cortical thickness of social brain regions. *Social Cognitive and Affective Neuroscience* **10**, 327-334 (2015).
21 Foster, N. E. & Zatorre, R. J. Cortical structure predicts success in performing musical transformation judgments. *Neuroimage* **53**, 26-36 (2010).
22 Button, K. S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* **14**, 365-376 (2013).
23 Collaboration, O. S. Estimating the reproducibility of psychological science. *Science* **349** (2015).
24 De Boeck, P. & Jeon, M. Perceived crisis and reforms: Issues, explanations, and remedies. *Psychological bulletin* **144**, 757 (2018).
25 Poldrack, R. A. *et al.* Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature reviews neuroscience* **18**, 115 (2017).
26 Boekel, W. *et al.* A purely confirmatory replication study of structural brain-behavior correlations. *Cortex* **66**, 115-133 (2015).
27 Boekel, W., Forstmann, B. U. & Wagenmakers, E.-J. Challenges in replicating brain-behavior correlations: Rejoinder to Kanai (2015) and Muhlert and Ridgway (2015). *cortex* **74**, 348-352 (2016).
28 Muhlert, N. & Ridgway, G. R. Failed replications, contributing factors and careful interpretations: Commentary on Boekel et al., 2015. *Cortex; a journal devoted to the study of the nervous system and behavior* **74**, 338-342, doi:10.1016/j.cortex.2015.02.019 (2016).
29 Kanai, R. Open questions in conducting confirmatory replication studies: Commentary on Boekel et al., 2015. *Cortex* **74**, 343-347, doi:10.1016/j.cortex.2015.02.020 (2016).
30 Genon, S. *et al.* Searching for behavior relating to grey matter volume in a-priori defined right dorsal premotor regions: lessons learned. *NeuroImage* **157**, 144-156 (2017).
31 Masouleh, S. K., Eickhoff, S. B., Hoffstaedter, F., Genon, S. & Initiative, A. s. D. N. Empirical examination of the replicability of associations between brain structure and psychological variables. *Elife* **8**, e43464 (2019).
32 Avinun, R., Israel, S., Knodt, A. R. & Hariri, A. R. Little evidence for associations between the Big Five personality traits and variability in brain gray or white matter. *NeuroImage*, 117092 (2020).
33 Kharabian, S., Eickhoff, S. B. & Genon, S. Searching for replicable associations between cortical thickness and psychometric variables in healthy adults: empirical facts. *BioRxiv* (2020).
34 Botvinik-Nezer, R. *et al.* Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* **582**, 84-88, doi:10.1038/s41586-020-2314-9 (2020).

35 Han, X. *et al.* Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. *Neuroimage* **32**, 180-194 (2006).

36 Fortin, J.-P. *et al.* Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage* **167**, 104-120, doi:https://doi.org/10.1016/j.neuroimage.2017.11.024 (2018).

37 Gronenschild, E. H. B. M. *et al.* The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements. *PLOS ONE* **7**, e38234, doi:10.1371/journal.pone.0038234 (2012).

38 Kharabian Masouleh, S. *et al.* Influence of Processing Pipeline on Cortical Thickness Measurement. *Cereb Cortex* **30**, 5014-5027, doi:10.1093/cercor/bhaa097 (2020).

39 Martínez, K. *et al.* Reproducibility of brain-cognition relationships using three cortical surface-based protocols: An exhaustive analysis based on cortical thickness. *Human Brain Mapping* **36**, 3227-3245 (2015).

40 Climie, E. A. & Rostad, K. (SAGE Publications Sage CA: Los Angeles, CA, 2011).

41 Marek, S. *et al.* Towards reproducible brain-wide association studies. *BioRxiv* (2020).

42 Miller, K. L. *et al.* Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature neuroscience* **19**, 1523 (2016).

43 Steegen, S., Tuerlinckx, F., Gelman, A. & Vanpaemel, W. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science* **11**, 702-712 (2016).

44 Albers, C. & Lakens, D. When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of experimental social psychology* **74**, 187-195 (2018).

45 Schönbrodt, F. D. & Perugini, M. At what sample size do correlations stabilize? *Journal of Research in Personality* **47**, 609-612 (2013).

46 Genon, S., Reid, A., Langner, R., Amunts, K. & Eickhoff, S. B. How to Characterize the Function of a Brain Region. *Trends Cogn Sci* **22**, 350-364, doi:10.1016/j.tics.2018.01.010 (2018).

47 Scoville, W. B. & Milner, B. Loss of recent memory after bilateral hippocampal lesions. *Journal of neurology, neurosurgery, and psychiatry* **20**, 11 (1957).

48 Poldrack, R. A. Mapping mental function to brain structure: how can cognitive neuroimaging succeed? *Perspectives on psychological science* **5**, 753-761 (2010).

49 Pessoa, L. Understanding brain networks and brain organization. *Physics of life reviews* **11**, 400-435 (2014).

50 Alexander-Bloch, A., Raznahan, A., Bullmore, E. & Giedd, J. The convergence of maturational change and structural covariance in human cortical networks. *Journal of Neuroscience* **33**, 2889-2899 (2013).

51 Habeck, C. & Stern, Y. Multivariate data analysis for neuroimaging data: overview and application to Alzheimer's disease. *Cell biochemistry and biophysics* **58**, 53-67 (2010).

52 McIntosh, A. R. & Mišić, B. Multivariate statistical analyses for neuroimaging data. *Annual review of psychology* **64**, 499-525 (2013).

53 van der Linden, D. *et al.* Overlap between the general factor of personality and emotional intelligence: A meta-analysis. *Psychological bulletin* **143**, 36 (2017).

54 Lyall, D. M. *et al.* Cognitive test scores in UK Biobank: data reduction in 480,416 participants and longitudinal stability in 20,346 participants. *PloS one* **11**, e0154222 (2016).

55 Cox, S., Ritchie, S., Fawns-Ritchie, C., Tucker-Drob, E. & Deary, I. Structural brain imaging correlates of general intelligence in UK Biobank. *Intelligence* **76**, 101376 (2019).

56 Watkins, M. W. Exploratory factor analysis: A guide to best practice. *Journal of Black Psychology* **44**, 219-246 (2018).

57 Hilger, K. *et al.* Predicting intelligence from brain gray matter volume. *Brain Structure and Function* **225**, 2111-2129 (2020).

58 Wu, J. *et al.* A Connectivity-based Psychometric Prediction Framework for Brain-behavior Relationship Studies. *Cereb Cortex* (In press).

59 Qin, S. *et al.* Amygdala subregional structure and intrinsic functional connectivity predicts individual differences in anxiety during early childhood. *Biological psychiatry* **75**, 892-900 (2014).

60 Vo, L. T. *et al.* Predicting individuals' learning success from patterns of pre-learning MRI activity. *PloS one* **6**, e16093 (2011).

61 Eldar, E., Hauser, T. U., Dayan, P. & Dolan, R. J. Striatal structure and function predict individual biases in learning to avoid pain. *Proceedings of the National Academy of Sciences* **113**, 4812-4817 (2016).

62 Chen, C., Yang, J., Lai, J., Li, H. & Yuan, J. Correlating gray matter volume with individual difference in the flanker interference effect. *PloS one* **10**, e0136877 (2015).

63 Wei, L. *et al.* Grey matter volumes in the executive attention system predict individual differences in effortful control in young adults. *Brain topography* **32**, 111-117 (2019).

64 Wang, X. *et al.* Predicting trait-like individual differences in fear of pain in the healthy state using gray matter volume. *Brain imaging and behavior* **13**, 1468-1473 (2019).

65 Wang, L., Wee, C.-Y., Suk, H.-I., Tang, X. & Shen, D. MRI-based intelligence quotient (IQ) estimation with sparse learning. *PloS one* **10**, e0117295 (2015).

66 Yang, J.-J. *et al.* Prediction for human intelligence using morphometric characteristics of cortical surface: partial least square analysis. *Neuroscience* **246**, 351-361 (2013).

67 Scheinost, D. *et al.* Ten simple rules for predictive modeling of individual differences in neuroimaging. *NeuroImage* **193**, 35-45 (2019).

68 He, Q. *et al.* Decoding the neuroanatomical basis of reading ability: a multivoxel morphometric study. *Journal of Neuroscience* **33**, 12835-12843 (2013).

69 Cui, Z., Su, M., Li, L., Shu, H. & Gong, G. Individualized prediction of reading comprehension ability using gray matter volume. *Cerebral Cortex* **28**, 1656-1672 (2018).

70 Sui, J., Jiang, R., Bustillo, J. & Calhoun, V. Neuroimaging-based individualized prediction of cognition and behavior for mental disorders and health: methods and promises. *Biological psychiatry* (2020).

71 Jiang, R. *et al.* Multimodal data revealed different neurobiological correlates of intelligence between males and females. *Brain imaging and behavior* **14**, 1979-1993 (2020).

72 Ullman, H., Almeida, R. & Klingberg, T. Structural maturation and brain activity predict future working memory capacity during childhood development. *Journal of Neuroscience* **34**, 1592-1598 (2014).

73 Wang, Y., Goh, J. O., Resnick, S. M. & Davatzikos, C. Imaging-based biomarkers of cognitive performance in older adults constructed via high-dimensional pattern regression applied to MRI and PET. *PloS one* **8**, e85460 (2013).

74 Rasero, J., Sentis, A. I., Yeh, F.-C. & Verstynen, T. V. Integrating across neuroimaging modalities boosts prediction accuracy of cognitive ability. *PLoS computational biology* **17**, e1008347 (2021).

75 Boeke, E. A., Holmes, A. J. & Phelps, E. A. Toward robust anxiety biomarkers: a machine learning approach in a large-scale sample. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* **5**, 799-807 (2020).

76 Smith, S. M. & Nichols, T. E. Statistical challenges in “big data” human neuroimaging. *Neuron* **97**, 263-268 (2018).

77 Krishnan, A., Williams, L. J., McIntosh, A. R. & Abdi, H. Partial Least Squares (PLS) methods for neuroimaging: a tutorial and review. *Neuroimage* **56**, 455-475 (2011).

78 Wang, H.-T. *et al.* Finding the needle in a high-dimensional haystack: canonical correlation analysis for neuroscientists. *NeuroImage* **216**, 116745 (2020).

79 Seidlitz, J. *et al.* Morphometric similarity networks detect microscale cortical organization and predict inter-individual cognitive variation. *Neuron* **97**, 231-247. e237 (2018).

80 Smith, S. M. *et al.* A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nature neuroscience* **18**, 1565-1567 (2015).

81 Moser, D. A. *et al.* An integrated brain–behavior model for working memory. *Molecular psychiatry* **23**, 1974-1980 (2018).

82 Kebets, V. *et al.* Somatosensory-motor dysconnectivity spans multiple transdiagnostic dimensions of psychopathology. *Biological psychiatry* **86**, 779-791 (2019).

83 Xia, C. H. *et al.* Linked dimensions of psychopathology and connectivity in functional brain networks. *Nature communications* **9**, 1-14 (2018).

84 Drysdale, A. T. *et al.* Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature medicine* **23**, 28-38 (2017).

85 Grosenick, L. *et al.* Functional and optogenetic approaches to discovering stable subtype-specific circuit mechanisms in depression. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* **4**, 554-566 (2019).

86 Mihalik, A. *et al.* Brain-behaviour modes of covariation in healthy and clinically depressed young people. *Scientific reports* **9**, 1-11 (2019).

87 Van Essen, D. C. *et al.* The WU-Minn human connectome project: an overview. *Neuroimage* **80**, 62-79 (2013).

88 Han, F., Gu, Y., Brown, G. L., Zhang, X. & Liu, X. Neuroimaging contrast across the cortical hierarchy is the feature maximally linked to behavior and demographics. *Neuroimage* **215**, 116853 (2020).

89 Llera, A., Wolfers, T., Mulders, P. & Beckmann, C. F. Inter-individual differences in human brain structure and morphology link to variation in demographics and behavior. *Elife* **8**, e44443 (2019).

90 Modabbernia, A., Janiri, D., Doucet, G. E., Reichenberg, A. & Frangou, S. Multivariate patterns of brain-behavior-environment associations in the Adolescent Brain and Cognitive Development study. *Biological psychiatry* **89**, 510-520 (2021).

91 Alnæs, D., Kaufmann, T., Marquand, A. F., Smith, S. M. & Westlye, L. T. Patterns of sociocognitive stratification and perinatal risk in the child brain. *Proceedings of the National Academy of Sciences* **117**, 12419-12427 (2020).

92 Cremers, H. R., Wager, T. D. & Yarkoni, T. The relation between statistical power and inference in fMRI. *PloS one* **12**, e0184923 (2017).

93 Mwangi, B., Tian, T. S. & Soares, J. C. A review of feature reduction techniques in neuroimaging. *Neuroinformatics* **12**, 229-244, doi:10.1007/s12021-013-9204-3 (2014).

94 Dinga, R. *et al.* Evaluating the evidence for biotypes of depression: Methodological replication and extension of. *NeuroImage: Clinical* **22**, 101796 (2019).

95 Mihalik, A. *et al.* Multiple holdouts with stability: Improving the generalizability of machine learning analyses of brain–behavior relationships. *Biological psychiatry* **87**, 368-376 (2020).

96 Hardoon, D. R. & Shawe-Taylor, J. Sparse canonical correlation analysis. *Machine Learning* **83**, 331-353 (2011).

97 Fukumizu, K., Bach, F. R. & Gretton, A. Statistical Consistency of Kernel Canonical Correlation Analysis. *Journal of Machine Learning Research* **8** (2007).

98 Helmer, M., Ji, J. L., Anticevic, A. & Murray, J. On Discovery of Brain-Phenotype Relationships: Detection, Estimation, and Prediction. *Biological Psychiatry* **87**, S207 (2020).

99 Haufe, S. *et al.* On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage* **87**, 96-110 (2014).

100 Helmer, M. *et al.* On stability of Canonical Correlation Analysis and Partial Least Squares with application to brain-behavior associations. *BioRxiv*, 2020.2008. 2025.265546 (2021).

101 Nooner, K. B. *et al.* The NKI-Rockland sample: a model for accelerating the pace of discovery science in psychiatry. *Frontiers in neuroscience* **6**, 152 (2012).

102 Avants, B. B., Cook, P. A., Ungar, L., Gee, J. C. & Grossman, M. Dementia induces correlated reductions in white matter integrity and cortical thickness: a multivariate neuroimaging study with sparse canonical correlation analysis. *Neuroimage* **50**, 1004-1016 (2010).

103 Genon, S. *et al.* Relating pessimistic memory predictions to Alzheimer's disease brain structure. *Cortex* **85**, 151-164 (2016).

104 Sui, J. *et al.* Multimodal neuromarkers in schizophrenia via cognition-guided MRI fusion. *Nature communications* **9**, 1-14 (2018).

105 Moser, D. A. *et al.* Multivariate associations among behavioral, clinical, and multimodal imaging phenotypes in patients with psychosis. *JAMA psychiatry* **75**, 386-395 (2018).

106 Ing, A. *et al.* Identification of neurobehavioural symptom groups based on shared brain mechanisms. *Nature human behaviour* **3**, 1306-1318 (2019).

107 Wasserman, J. D. & Bracken, B. A. Fundamental psychometric considerations in assessment. *Handbook of Psychology, Second Edition* **10** (2012).
108 Chen, J. *et al.* Exploration of scanning effects in multi-site structural MRI studies. *Journal of Neuroscience Methods* **230**, 37-50, doi:https://doi.org/10.1016/j.jneumeth.2014.04.023 (2014).
109 Holmes, A. J. *et al.* Brain Genomics Superstruct Project initial data release with structural, functional, and behavioral measures. *Scientific data* **2**, 1-16 (2015).
110 Habeck, C., Gazes, Y., Razlighi, Q. & Stern, Y. Cortical thickness and its associations with age, total cognition and education across the adult lifespan. *PLOS ONE* **15**, e0230298, doi:10.1371/journal.pone.0230298 (2020).
111 Peelle, J. E., Cusack, R. & Henson, R. N. A. Adjusting for global effects in voxel-based morphometry: gray matter decline in normal aging. *NeuroImage* **60**, 1503-1516, doi:10.1016/j.neuroimage.2011.12.086 (2012).
112 Dayan, E., Hamann, J. M., Averbeck, B. B. & Cohen, L. G. Brain structural substrates of reward dependence during behavioral performance. *Journal of Neuroscience* **34**, 16433-16441 (2014).
113 Valk, S. L., Bernhardt, B. C., Böckler, A., Kanske, P. & Singer, T. Substrates of metacognition on perception and metacognition on higher-order cognition relate to different subsystems of the mentalizing network. *Hum Brain Mapp* **37**, 3388-3399, doi:10.1002/hbm.23247 (2016).
114 May, A. Experience-dependent structural plasticity in the adult human brain. *Trends in cognitive sciences* **15**, 475-482 (2011).
115 Geng, X. *et al.* Structural and maturational covariance in early childhood brain development. *Cerebral Cortex* **27**, 1795-1807 (2017).
116 Zielinski, B. A., Gennatas, E. D., Zhou, J. & Seeley, W. W. Network-level structural covariance in the developing brain. *Proceedings of the National Academy of Sciences* **107**, 18191-18196 (2010).
117 Lanphear, B. P. The Impact of Toxins on the Developing Brain. *Annual Review of Public Health* **36**, 211-230, doi:10.1146/annurev-publhealth-031912-114413 (2015).
118 Farah, M. J. The neuroscience of socioeconomic status: Correlates, causes, and consequences. *Neuron* **96**, 56-71 (2017).
119 Teicher, M. H., Samson, J. A., Anderson, C. M. & Ohashi, K. The effects of childhood maltreatment on brain structure, function and connectivity. *Nature reviews neuroscience* **17**, 652 (2016).
120 Palmer, C. E. *et al.* Fluid and crystallised intelligence are associated with distinct regionalisation patterns of cortical morphology. *bioRxiv* (2020).
121 Noble, K. G. *et al.* Family income, parental education and brain structure in children and adolescents. *Nature neuroscience* **18**, 773 (2015).
122 Decker, A. L., Duncan, K., Finn, A. S. & Mabbott, D. J. Children's family income is associated with cognitive function and volume of anterior not posterior hippocampus. *Nature communications* **11**, 1-11 (2020).

Acknowledgments

The authors are supported by the Deutsche Forschungsgemeinschaft (DFG, GE 2835/2–1), the National Institute of Mental Health (R01-MH074457), the Helmholtz Portfolio Theme "Supercomputing and Modeling for the Human Brain", the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement785907 (HBP SGA2), 945539 (HBP SGA3), and The Virtual Brain Cloud (EU H2020, no. 826421).