Neurobiological Predictors of Hand Grip Strength as a Global Health Marker: Methodological Foundations and Interpretable Brain-Behaviour Prediction in Large-Scale Neuroimaging

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Vera Aurelia Komeyer
aus Eichstätt

Düsseldorf, Januar 2026

aus dem Institut für Systemische Neurowissenschaften
Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Berichterstatter:

1. Prof. Dr. Simon B. Eickhoff

2. Prof. Dr. Markus Kollmann

Tag der mündlichen Prüfung:

Table of Contents

1 Summary I

2 Zusammenfassung II

3 List of Abbreviations IV

4 Introduction 1

4.1 Motor performance as a window into brain and general health 3

4.1.1 Motor performance as a fundamental output of integrated body systems 3

4.1.2 Hand grip strength as a reliable, versatile and widely used health marker 5

4.2 Neuroanatomical and physiological underpinnings of hand grip strength 6

4.2.1 Phylogenetic layering and the evolution of grasping 6

4.2.2 Cortical control 7

4.2.3 Structural connectomics and white matter efficiency 8

4.2.4 Subcortical involvements 9

4.2.5 Extra-motor system contributors 10

4.2.6 HGS as a system-level read-out of multi-system integrity 11

4.3 Brain-behaviour predictive modeling using large-scale, observational neuroimaging data 11

4.3.1 Opportunities and limitations of large, observational neuroimaging data for brain-behaviour research 12

4.3.2 Brain behaviour predictive modeling: promise and current challenges 13

4.4 Confounding in brain-behaviour prediction: nuisance or biological reality? 14

4.5 Aims and scope of this thesis 17

5 Empirical work 19

5.1 Manuscript 1: Overview of Challenges in Brain-Based Predictive Modelling: Toward Meaningful Predictive Insights 19

5.2 Manuscript 2: How causal inference tools can support debiasing of machine learning models for meaningful brain-based predictions 20

5.2 Manuscript 3: Hand grip strength as a behavioral read-out of distributed but specific system-level brain integrity: A large-scale multi-modal machine learning study 21

6 Discussion 22

6.1 Integrative summary of findings across studies 22

6.1.1 From methodological challenge to biological insight 22

6.1.2 Conceptual synthesis: Why is HGS such a versatile marker? 23

6.2 Methodological contributions to brain–behaviour predictive modelling 24

6.2.1 From prediction performance to neuroscientific meaning 24

6.2.2 Confounding as a reflection of biological entanglement 24

6.2.3 Multicollinearity-aware interpretation as a prerequisite for system neuroscience 25

6.3 Neural systems supporting the prediction of HGS 26

6.3.1 A system-level perspective on HGS 26

6.3.2 Many-to-one mapping and strategic flexibility in HGS 28

6.3.3 Hand grip strength, aging, and the lifespan–healthspan gap 29

6.4 Limitations, implications, and future directions 30

6.5 Conclusion 31
8 References 32
9 Publications 45
10 Author Declaration 47
11 Appendix 48

1 Summary

Aging populations worldwide are widening the gap between lifespan and healthspan, underscoring the need for early, scalable markers of organismal health and intervention targets before overt clinical decline. Hand grip strength (HGS) has emerged as a highly reliable and low-cost predictor of system-wide factors such as frailty, cognitive decline, and mortality. Despite its simplicity and clear musculoskeletal determinants, explaining its system-wide predictive value requires a deeper understanding of underlying brain-level neurobiological architectures, which is currently lacking. This thesis addresses this gap by investigating generalizable neural predictors of HGS using machine learning (ML) with large-scale, multi-modal neuroimaging data, grounded in methodological foundations for interpretable brain-behaviour prediction.

A critical review of methodological constraints and potential mitigation strategies in observational neuroimaging-based ML studies was conducted to promote more reliable and generalizable brain-behaviour predictions and interpretations (Study 1). Such studies can be hampered by pitfalls including data leakage, site-effects in multi-site datasets, misleading post-hoc model interpretations arising from feature multicollinearity, and model bias due to confounding. Strict out-of-sample evaluation and clustering-based interpretation to deal with feature multicollinearity were identified as suitable solutions. To support principled confounder selection, a theoretically informed but empirically pragmatic 3-step approach was developped (Study 2). The proposed approach integrates methdology from causal inference - domain knowledge, directed acyclic graphs, and respective graph rules - with associative data-driven modeling.

Building on these foundations, a comprehensive, interpretable multi-modal predictive workflow revealed generalizable, system-level neuroimaging predictors of HGS in a large, healthy cohort from the UK Biobank (Study 3). Across modeling approaches, microstructural integrity – particularly in ascending medial lemniscus, thalamic radiations, and associative white-matter pathways – as well as subcortical gray matter volume (GMV), mainly in the anterior globus pallidus, emerged as relevant contributors. In contrast, cortical structural measures and functional imaging features contributed little to predictive performance. Collectively, these findings position HGS as a behavioural readout of the brain's capacity to coordinate, and integrate information across motor, sensory, cognitive, and motivational systems, rather than a purely peripheral muscle measure or an isolated motor output.

In sum, this thesis establishes a framework for neurobiologically interpretable large-scale brain-behaviour prediction and applies it to elucidate why HGS functions as a powerful marker of global health. By intergating methodological rigor with system-level neuroimaging, it demonstrates how simple behavioural phenotypes can serve as informative windows into the functioning and integrity of distributed neural architectures. Future work should determine whether identified HGS-linked neural signatures provide better or earlier prognostic and interventional value than the behavioural measure itself.

2 Zusammenfassung

Die weltweit alternde Bevölkerung vergrößert die Kluft zwischen Lebensdauer und Gesundheitsspanne und unterstreicht damit die Notwendigkeit von frühzeitigen, skalierbaren Markern für die Gesundheit des Organismus, sowie Interventionsziele vor einem offensichtlichen klinischen Verfall. Griffkraft hat sich hierbei als zuverlässiger und kostengünstiger Prädiktor für systemweite Faktoren wie Gebrechlichkeit, kognitiven Verfall und Mortalität herausgestellt. Trotz ihrer Einfachheit und klaren muskuloskelettalen Determinanten erfordert die Erklärung ihres systemweiten Vorhersagewerts ein tieferes Verständnis der zugrundeliegenden neurobiologischen Architekturen auf Gehirnebene, was derzeit fehlt. Diese Arbeit befasst sich mit dieser Lücke, indem sie verallgemeinerbare, neuronale Prädiktoren für Griffkraft unter Verwendung maschinellen Lernens (ML) mit groß angelegten, multimodalen neuronalen bildgebenden Daten untersucht und gleichzeitig auf methodischen Grundlagen für interpretierbare des Gehirn-Verhalten Vorhersagen basiert.

Es wurde eine kritische Überprüfung der methodischen Einschränkungen und potenziellen Verhinderungsstrategien in observationalen, auf neuronaler Bildgebung basierenden ML-Studien durchgeführt, um zuverlässigere und verallgemeinbarere Vorhersagen und Interpretationen des Gehirn-Verhaltens zu fördern (Studie 1). Solche Studien können durch Datenlecks, Standorteffekte in Datensätzen mit mehreren Standorten, irreführende post-hoc Modellinterpretationen aufgrund von Multikollinearität von Featuren, sowie Modellverzerrungen aufgrund von Kovariaten behindert werden. Als geeignete Lösungen wurden eine strenge out-of-sample Evaluierung und eine Clustering basierte Interpretation zur Behandlung der Multikollinearität von Featuren identifiziert. Um eine korrekte Auswahl von Kovariaten zu unterstützen, wurde ein theoretisch fundierter, aber empirisch pragmatischer 3-Stufen-Ansatz entwickelt (Studie 2). Der vorgeschlagene Ansatz integriert Methoden aus der kausalen Inferenz – Domänenwissen, gerichtete azyklische Graphen und entsprechende Graphregeln – mit assoziativer datengesteuerter Modellierung.

Auf dieser Grundlage gab ein umfassender interpretierbarer, multimodaler prädiktiver ML-Arbeitsablauf Einblicke in verallgemeinerbare und systemische neuronaler Prädiktoren für Griffkraft in einer großen, gesunden Kohorte der UK Biobank (Studie 3). Über alle Modellierungsansätze hinweg erwiesen sich die mikrostrukturelle Integrität – insbesondere im aufsteigenden medialen Lemniscus, der Thalamusstrahlung und assoziativen Bahnen weißer Substanz – sowie das Volumen der subkortikalen grauen Substanz, hauptsächlich im anterioren Globus Pallidus, als relevante Einflussfaktoren. Im Gegensatz dazu trugen kortikale Strukturmaße und funktionelle Bildgebungsmerkmale nur wenig zur Vorhersageleistung bei. Zusammengenommen positionieren diese Ergebnisse Griffkraft als Verhaltensindikator für die Fähigkeit des Gehirns, Informationen über motorische, sensorische, kognitive und motivationale Systeme hinweg zu koordinieren und zu integrieren, und nicht als rein peripheres Muskelmaß oder isolierte motorische Leistung.

Zusammenfassend lässt sich sagen, dass diese Arbeit einen Rahmen für neurobiologisch interpretierbare groß angelegte Gehirn-Verhalten Vorhersagen schafft und diese nutzt, um zu erklären, warum Griffkraft als aussagekräftiger Indikator für allgemeine Gesundheit fungieren kann. Durch die Integration methodischer Genauigkeit mit neuronaler Bildgebung auf Systemebene zeigt sie, wie einfache Verhaltensphänotypen als informative Einblicke in die Funktionsweise und Integrität verteilter neuronaler Architekturen dienen können. Zukünftige Arbeiten sollten klären, ob identifizierte Griffkraft-bezogene neuronale Signaturen einen besseren oder früheren prognostischen und interventionellen Wert bieten als die Verhaltensmessung selbst.

3 List of Abbreviations

AD	Alzheimer's Disease
aGP	anterior Globus Pallidus
ATR	Anterior Thalamic Radiation
CNS	Central Nervous System
CST	Corticospinal Tract
CV	Cross-Calidation
DAG	Directed Acyclic Graph
DWI	Diffusion-Weighted Imaging
FA	Fractional Anisotropy
GMV	Gray Matter volume
GPi	Globus Pallidus internus
HGS	Hand Grip Strength
ISOVF	Isotropic Volume Fraction
LGN	Lateral Grasping Network
M1	Primary Motor Cortex
MCP	Middle Cerebellar Peduncle
MD	Mean Diffusivity
ML	Machine Learning
MO	Diffusion Tensor Mode
OD	Orientation Dispersion Index
pGP	posterior Globus Pallidus
rs-fMRI	resting-state functional MRI
RST	Reticulospinal Tract
S1	Primary Somatosensory Cortex
SLF	Superior Longitudinal Fasciculus
SMA	Supplementary Motor Area
SMG	Supramarginal Gyrus
sMRI	structural MRI
STN	Subthalamic Nucleus
STR	Superior Thalamic Radiation
UKB	UK Biobank
vPMC	ventral Premotor Cortex
WMH	White Matter Hyperintensities

4 Introduction

Populations worldwide are aging (World Health Organization, 2025), leading to a growing emphasis not only on extending lifespan, but on promoting healthspan - the ability to live longer lives in good physical, cognitive, and functional health. However, health- and lifespan to date often exhibit a pronounced gap (Garmany & Terzic, 2025; Gianfredi et al., 2025). While improved sanitation, nutrition and advances in medicine have substantially increased longevity, aging is accompanied by gradual declines across multiple physiological systems, ultimately increasing vulnerability for frailty, disability, and disease (Gianfredi et al., 2025). A central challenge in aging research is therefore identification of early, reliable markers that reflect the integrity of underlying biological systems before overt clinical decline becomes apparent.

Health is an inherently multi-faceted construct and difficult to capture with a single measure. Multiple biological systems like physical capacity, cognitive function, and neurological integrity jointly determine an individual's ability to maintain daily functioning and resilience to disease. These domains interact in complex ways, and their trajectories of change and decline are highly heterogeneous across individuals. This heterogeneity complicates both the assessment of current health status and the prediction of future deterioration. In this context, there is a need for simple, scalable, and robust behavioral phenotypes that capture the integrated functioning of multiple biological systems.

Motor performance has emerged as a class of such integrative behavioral phenotypes (Marín-Jiménez et al., 2022). It is commonly defined as the execution of tasks requiring coordinated muscle activity (Spirduso & MacRae, 1990). Motor function as the underlying concept, involves complex physiological processes and requires the integration of multiple systems including neuromuscular, musculoskeletal, cardiopulmonary, neural motor, and sensory-perceptual systems (Barnes et al., 2025). Motor performance includes voluntary control of both fine and gross motor functions including dexterity, strength, balance, locomotion, and endurance and requires the interaction of multiple body systems, including the nervous, muscular, cardiovascular, and sensory-perceptual systems (Barnes et al., 2025). It hence serve as an umbrella term encompassing a wide range of motor outcomes. Measures of motor function are easy to obtain in a non-invasive manner, show robust associations with diverse health outcomes, but function itself consistently declines with advancing age (Buchman et al., 2007; Hunter et al., 2016; Marín-Jiménez et al., 2022; Spirduso & MacRae, 1990). Importantly, the associations extend to the cognitive domain. Impairments in fine motor function (visuomotor integration, manual dexterity) and multisensory processing (hearing, vision, vestibular, proprioception, olfaction) have been linked to cognitive dysfunction (Sayyid et al., 2024) and altered patterns in gross motor function, such as gait speed, may precede mild cognitive impairment (Buracchio et al., 2010) and show sensitivity to Alzheimer's disease (AD) pathology (Q. Tian et al., 2017, 2018; Verghese et al., 2002). Measurable declines in motor performance begin as early as midlife (~40 years) and accelerate

after approximately 65 years of age, highlighting their potential utility as preclinical indicators of health deterioration (Hunter, 2025; Spirduso & MacRae, 1990).

Among motor performance measures, hand grip strength (HGS) has received particular attention as a remarkably robust and versatile indicator of a large battery of focal and global health-related outcomes, spanning overall muscular strength, morbidity, cognition, and mortality (Chai et al., 2024; Gale et al., 2007; Ling et al., 2010; Rijk et al., 2016; Sasaki et al., 2007). Beyond reflecting peripheral musculoskeletal function, HGS depends on the integrity of distributed neural systems involved in motor planning, execution, and sensorimotor integration (Bello et a., 2021; Pruves et al., 2001; Surgent et al., 2023; Monte et al., 2003; Davis et al., 2022; Richardson et al., 2017; Borich et al., 2016; Bolognini et al., 2017) (Bello et al., 2021; Bolognini et al., 2016; Borich et al., 2015; Carson, 2018; Davis et al., 2022; Purves, Augustine, Fitzpatrick, Hall, et al., 2001; Richardson et al., 2016a; Surgent et al., 2023), as well as on central processes related to attention, motivation, and fatigue (Cass et al., 2024; J. Firth et al., 2018; Ganipineni et al., 2023; Rinne et al., 2018). As such, HGS captures variance that is shared across motor, cognitive, and neurological domains, making it a sensitive proxy for both motor performance and broader brain and body health. Building on previous evidence, this thesis investigates the neural basis of motor performance - operationalized through HGS - as a marker and predictor of general (e.g. frailty, risk for falls, morbidity etc.), mental (e.g. depression, psychosis) and cognitive (e.g. different types of dementia but also memory, attention, executive functioning) health and overall brain integrity (Duchowny et al., 2022; Ganipineni et al., 2023; Jiang, Westwater, et al., 2022; Shang et al., 2021; H. B. Ward et al., 2025). Thus, investigating the relationship between HGS and brain integrity can enable the joint assessment of aging processes and vulnerability to cognitive and neuropsychiatric disorders.

The use of machine learning (ML) methods together with recently available large-scale population neuroimaging data is necessary to capture the complex, multivariate relationships between HGS and brain integrity. ML approaches enable modeling of high-dimensional neural features, nonlinear effects, and interactions that cannot be adequately addressed with traditional univariate analyses. Furthermore, when properly implemented, ML models learn generalizable patterns, thereby enabling robust population-level inference and prediction. However, they also introduce unavoidable methodological challenges arising from biological coupling, high data dimensionality, and population heterogeneity, including confounding by age, sex, and general health status, as well as further mitigation of pitfalls in predictive modeling. This thesis directly addresses these challenges by proposing and implementing conceptually sound approaches to obtaining generalizable models, rigorously handling confounding effects, and neurobiologically interpreting model behavior.

The following sections develop these considerations in greater depth by situating motor performance in general and HGS in particular within current models of brain and general health and introducing respectively relevant neuroanatomical and physiological underpinnings.

4.1 Motor performance as a window into brain and general health

Motor behaviour constitutes the primary effector output of the central nervous system (CNS), serving as a functional interface between internal physiological integrity and environmental interaction (Wolpert & Flanagan, 2001). Moving beyond traditional musculoskeletal interpretations, contemporary research characterizes motor performance as a systemic readout of fundamental biological processes, including the integrity of distributed neural architectures (Zapparoli et al., 2022). This perspective positions motor performance as a phenotypic window onto cross-domain functional declines that may precede overt clinical symptoms (Marín-Jiménez et al., 2022). Furthermore, the inherent plasticity of motor control systems (Roth & Ding, 2024; Sadowski, 2008; Sanes & Donoghue, 2000) suggests that motor performance is not merely a prognostic or diagnostic marker but also a potent interventional target (Bolognini et al., 2009; Inoue & Nishimune, 2023). Given the relevance of neural architectures, enhancing our understanding of motor-brain interdependencies is thus critical for developing rehabilitative strategies aimed at maintaining or restoring systemic health across the lifespan (Garmany & Terzic, 2025; Hunter, 2025; Zhang et al., 2025).

4.1.1 Motor performance as a fundamental output of integrated body systems

Motor behavior represents one of the most fundamental outputs of biological systems. Even seemingly simple movements require a coordinated interaction of multiple components, including central and peripheral neural circuits, sensory feedback pathways, musculature, and metabolic and cardiovascular support. Thereby, effective motor performance is the culmination of a complex, integrated system involving high-level cognition, sensory feedback processing, and precise neural execution (Gibson & Noakes, 2004; Leisman et al., 2016; Riemann & Lephart, o. J.; Shi & Feng, 2022; Spampinato & Celnik, 2021; Winter et al., 2022). From a neuroanatomical perspective, motor output is organized across multiple hierarchical levels. Cortical regions, specifically the primary motor (M1), premotor, and supplementary motor areas, are involved in movement planning, selection, and initiation (Halsband et al., 1994; He et al., 1993, 1995; Purves, Augustine, Fitzpatrick, Hall, et al., 2001; Roland et al., 1980). M1 thereby receives and integrates input from a range of cortical and subcortical regions and is the final cortical processing site for voluntary motor commands, before they descend to the spinal cord (Stinear et al., 2009). Subcortical structures and circuits, especially the basal ganglia act through gate-keeping mechanisms that select or inhibit motor programs or prepare upper motor neuron circuits for initiation of movements (Purves, Augustine, Fitzpatrick, Hall, et al., 2001). Beyond, the motor-thalamus serves as strategic node integrating subcortical inputs to facilitate action initiation and speed (Bosch-Bouju et al., 2013; Dacre et al., 2021; Takahashi et al., 2021). Cerebellar circuits support coordination, timing and error correction necessary for fluid execution (Purves, Augustine, Fitzpatrick, Hall, et al., 2001), by modifying activtiy patterns of upper motor neurons (Purves, Augustine, Fitzpatrick, Katz, et al., 2001a). Descending pathways, such as the corticospinal tract (CST) transmit signals from the primary and secondary motor cortices to the brain stem and spinal cord (Javed et al.,

2018). This feed-forward system is continuously refined by somatosensory and proprioceptive afferents (Chakrabarty & Martin, 2011; Moreno-López et al., 2016), such as the medial lemniscus, which ensures adaptive calibration, correction and adjustment to environmental requirements (Navarro-Orozco & Bollu, 2018).

Beyond the involvement of isolated motor structures and circuits, the Scaffolding Theory of Maturation, Cognition, Motor Performance, and Motor Skill Acquisition (Klotzbier & Schott, 2025) posits that motor behaviour interacts with a variety of other brain outputs. According to this theory, motor and cognitive processes do not operate in isolation but develop, interact and decline bidirectionally across the lifespan, i.e. a decline in one domain often predicts decline in the other (Basile & Sardella, 2021). Vice versa, a variety of physical fitness regimens have been shown to improve cognition in aging individuals, as well as to delay the onset of dementia in the cognitively impaired (Ahlskog et al., 2011; Berryman et al., 2013; Guiney & Machado, 2013; Guo et al., 2017; Voelcker-Rehage & Niemann, 2013). On the neuronal level this is supported by broad evidence that motor behaviour also includes the recruitment of extra-motor structures and networks both in healthy subjects and in motor recovery, for example after stroke (e.g. Guo et al., 2017; Johnson et al., 2017; Lam et al., 2018; Mattos et al., 2023; Park et al., 2011; Rezaei et al., 2025). Alterations in motor performance may arise from dysfunction in any of the motor, but also extra-motor levels. This makes motor behaviour a sensitive, early indicator of system-wide neuronal change.

In the context of senescence, motor performance is uniquely vulnerable to age-related alterations. Aligning with the system-level informativeness of motor control, longitudinal evidence suggests that declines in motor velocity, coordination, and manual strength often precede measurable impairments in episodic memory or executive function (Beauchet et al., 2014; Camicioli et al., 1998). Additionally, it can serve as a predictor of cognitive decline, dementia and functional performance before overt clinical symptoms become apparent (Bekena et al., 2025; Fugiel et al., 2025; Xie et al., 2025). Crucially, these motoric changes are not solely attributable to age-related sarcopenia or peripheral physiology. Instead, among others, they involve and reflect reduced and more variable synaptic inputs that drive motor neuron activation, fewer and larger motor units (Hunter et al., 2016), reductions in corticospinal excitability, degeneration and altered biophysical characteristics of motor neurons (e.g. Clark, 2019) as well as alterations in neural micro-structure, specifically reduced white matter tract integrity (Oschwald et al., 2021). The motor system thus exhibits sensitivity to early-stage neurodegeneration and cerebrovascular pathology that may remain undetected by standard cognitive batteries (Buchman & Bennett, 2011). This shows that motor control is deeply integrated with neurocognitive processes of aging. Accordingly, Zapparoli et al. (2022) argue that “any neurocognitive model of aging not considering the motor system is, ipso facto, incomplete”.

Overall, motor assessments offer an ecologically valid, non-invasive phenotype of brain health, whereby motor behaviour acts as a functional readout of the brain's "scaffolding" capacity - the ability to recruit compensatory neural resources in response to age-related structural degradation. This positions

motor behaviour as a systemic primary domain for investigating the trajectories of healthy and pathological aging.

4.1.2 Hand grip strength as a reliable, versatile and widely used health marker

While motor performance is inherently multidimensional - spanning domains of dexterity, strength, postural control, balance, and coordination - the utility of motoric markers in large-scale neuroimaging and epidemiological research necessitates a robust and scalable operationalization. Within this landscape, maximal HGS has emerged not merely as a measure of localized isometric force, but also as a high-fidelity proxy for global neuromuscular integrity and systemic biological processes (Bohannon, 2019). HGS hereby refers to maximum force production in one hand as administered in a power grip, a movement that consists of flexing the fingers to allow to grasp heavy objects for their large displacement or to hold on something and maintain our body position. In contrast, precision grip targets the manipulation of small objects with a high level of detail (Landsmeer, 1962; Long et al., 1970; Quattrocchi et al., 2024). Recent evaluations suggest that simpler indices, such as absolute maximum or average HGS, provide the most reliable parameters for capturing this underlying construct (Chai et al., 2024). The superior psychometric properties of HGS additionally make it a prominent tool in clinical and population-based cohorts. HGS protocols are highly standardized, cost-effective, and demonstrate exceptional test-retest reliability across diverse demographic strata (Bohannon, 2019; Gell et al., 2024; Roberts et al., 2011).

Physiologically, HGS is influenced by structural and functional integrity of the upper limb's musculoskeletal system, directly linked to lean muscle mass in the forearms and hands (Abe & Loenneke, 2015; Lawman et al., 2016; Vaishya et al., 2024). However, absolute mass is often less critical than muscle quality, which can be compromised by myosteatosis (intramuscular fat infiltration) or shifts in fibre type, leading to functional declines even when mass is preserved (Wen et al., 2023). Additionally, low HGS is one of the components of diagnosis of sarcopenia (Fielding et al., 2011), a generalised muscle disorder associated with increased likelihood of falls, fractures, and physical disability (Cruz-Jentoft et al., 2019; Hunter, 2025). This muscular capacity is intrinsically tied to bone structure, where higher HGS was found to be associated with better bone mineral density (Amante-da-Rosa-Cardoso et al., 2025; Song et al., 2022). Furthermore, force transmission is mediated by joint and connective tissue health, as for example seen through reduced HGS in individuals with arthritis (Dedeoğlu et al., 2013; Haugen et al., 2021). Anthropometric and biomechanical factors further influence HGS, where proprioceptive joint position sense, HGS, and anthropometric measures such as forearm length or wrist circumference correlate (Abalay et al., 2024). From this perspective, HGS can be viewed as a measure of peripheral motor capacity reflecting the structural integrity of the musculoskeletal system.

However, while musculoskeletal factors can constrain maximal HGS, they do not fully account for the high degree of variance in HGS across the lifespan. In fact, muscle force generated during HGS

assessments in older adults is around half of what would be expected if the skeletal musculature itself were fully activated, a discrepancy mainly attributable to age-related neural deficits (Clark, 2019; McGrath et al., 2020; Shinohara et al., 2003). Neuromuscular function at the spinal level can partially contribute to these neural influences. However, peripheral musculoskeletal and motor components mainly represent the execution phase of a complex command chain and cannot provide explanations for its broader systemic associations. A key clinical and scientific value of HGS lies in its profound predictive validity. Low HGS is consistently associated with an increased risk of all-cause mortality, frailty syndromes, physical disability, morbidity and cardiovascular events (Bohannon, 2019; Dudzińska-Griszek et al., 2017; Reeve IV et al., 2018; Syddall et al., 2017). Notably, the prognostic power of HGS can surpass that of traditional clinical markers. For instance, the PURE study demonstrated that HGS is a stronger predictor of all-cause and cardiovascular mortality than systolic blood pressure, with every 5-kg decrement in strength corresponding to a significantly higher hazard ratio for early death regardless of socioeconomic context (Leong et al., 2015). This systemic sensitivity is further evidenced by the relationship between HGS and cellular aging. Weaker grip strength is significantly associated with accelerated DNA methylation age - an epigenetic marker of biological vs. chronological age - indicating that HGS reflects the rate of physiological erosion at a molecular level (Peterson et al., 2023). The utility of HGS extends into the cognitive domain, where it can serve as a sensitive indicator of neurodegenerative risk as lower HGS has been linked to a heightened risk of transitioning from mild cognitive impairment to Alzheimer's disease (Fritz et al., 2017; McGrath et al., 2020). By mirroring these physical, cognitive, and epigenetic dimensions, HGS functions as a robust, non-invasive phenotype of biological resilience and non-fragility. These multi-systemic associations and the widespread evidence that measures of HGS have in predicting future adverse health events (McGrath et al., 2020) cannot be explained by musculoskeletal or peripheral neuromuscular function alone. Therefore, they provide the impetus for investigating the specific neural architectures that support HGS at the brain-level.

4.2 Neuroanatomical and physiological underpinnings of hand grip strength

The characterisation of HGS as a systemic health marker is fundamentally rooted in the involvement of an expansive, multi-level neural architecture. Unlike task-specific learned motor sequences, the generation of maximal isometric force (capacity to produce force with a voluntary muscle contraction that maintains the muscle's length (Gallagher et al., 2000)) represents a system-wide recruitment of distributed but specific motor and modulatory resources, requiring the synchronized activation of spinal, supraspinal, subcortical and cortical circuits.

4.2.1 Phylogenetic layering and the evolution of grasping

Grasping constitutes a phylogenetically conserved, foundational motor capacity that emerges early in human development, predating the emergence of cortically dominated voluntary control. In humans, this is evidenced by the palmar grasp reflex, a vestigial mechanism mediated predominantly by

spinal and brainstem circuits rather than higher-order cortical planning (Capute & Accardo, 1996; Futagi et al., 2012; Marques De Moraes et al., 2017). This primitive ability matures into a complex voluntary motor pattern as supraspinal centres myelinate and develop sophisticated control (Capute & Accardo, 1996), i.e. volitional grip strength in adulthood reflects the cortical appropriation of this evolutionarily older substrate (Stephens-Sarlós et al., 2025; Zafeiriou, 2004). This layered organization and developmentally and evolutionary persistence suggests a robust neural architecture of HGS, robust to focal cortical damage by prioritizing efficiency and redundancy over isolated specialization.

4.2.2 Cortical control

At the cortical level, HGS generation is governed by a distributed network where the M1 acts as the principal executive hub (Bello et al., 2021). Activity in M1 scales linearly with force output, encoding force magnitude through population firing rates and force-dependent oscialltory dynamics (Purves, Augustine, Fitzpatrick, Hall, et al., 2001). A defining feature of maximal HGS is its dependence on intracortical disinhibition. To achieve peak force, the CNS must temporarily suppress the tonic inhibitory signals within M1 that typically prevent muscular overstrain. Research using transcranial magnetic stimulation demonstrates that maximal power grips involve a significant reduction in short interval intracortical inhibition compared to precision grips (Duval et al., 2024; Federico & Perez, 2017; Tazoe & Perez, 2017). This intracortical inhibition-excitation balance is crucial as only a transient reduction of intracortical inhibition enables increased corticospinal output, supporting maximal force production (Ding et al., 2019; Ferreiro De Andrade & Conforto, 2018; Ni et al., 2007).

On a network level, M1 is part of the lateral grasping network (LGN)**,** which is additionally comprised of the ventral premotor cortex (vPMC) and the supramarginal gyrus (SMG) (Surgent et al., 2023). The LGN contributes to the generation of grip force through the integration of polymodal sensory, contextual and motor information (Borra et al., 2017; Surgent et al., 2023). Beyond the vPMC, secondary motor areas in general, including the premotor cortex and supplementary motor areas (SMA) play a role in planning, scaling, coordinating and modulating grip force, whereby higher task difficulty or force demands are mirrored by higher activity in these regions (Kulwattho et al., 2025). SMA regions are also involved in inhibitory-excitatory balance. During high-force contractions, inter-hemispheric functional connectivity between bilateral SMAs modulates the motor drive, a mechanism that becomes increasingly critical during fatigue or in the presence of CST compromise (N. S. Ward et al., 2007; Welniarz et al., 2019). Bidirectional interactions between M1 and primary somatosensory cortex (S1) further shape force scaling and grip stability by integrating afferent feedback into ongoing motor commands. Impairment of this interaction, for example through cortical deafferentiation was seen to not just impair precision but result in reduced maximal force (Borich et al., 2015; Davis et al., 2022; Monzée et al., 2003; Richardson et al., 2016b).

4.2.3 Structural connectomics and white matter efficiency

The translation of cortical commands into physical force depends strongly on the structural and functional integrity of various white matter tracts. The descending CST originates from different cortical areas, including core-motor but also extra-motor areas (Lemon, 2008). In its most classic motor function, it serves as the major pathway conveying force-related commands from the cortex to spinal motor neurons through the excitation and inhibition of motoneurons and the descending control of afferent inputs (Lemon, 2008; Tazoe & Perez, 2017). As seen in injury studies, such as after stroke, the integrity of the CST influences how M1 and secondary regions adapt during increased force demands, underlying its essential involvement in power grip execution (N. S. Ward et al., 2006, 2007). The reticulospinal tract (RST) forms another relevant descending motor pathway, reaching from the pontine reticular formation of the brainstem to the spinal cord. It provides a relatively diffuse, high-capacity descending signal for gross force production, such as the power grip (Baker & Perez, 2017; Glover & Baker, 2022). The RST can partially compensate for corticospinal deficits, albeit with reduced selectivity, i.e. carrying simpler uniform signals more suited for gross specification of movement (Baker & Perez, 2017; Glover & Baker, 2022; Zaaimi et al., 2012). This redundancy together with the scaling of HGS in the pontine reticular nuclei of the brainstem (Danielson et al., 2024) reinforces the evolutionary robustness of grip.

Beyond descending fibres, effective grip strength relies on the integrity of ascending and associative white matter systems that support sensorimotor state estimation, cortical excitability, and cognitive-motivational integration. Ascending pathways, including the dorsal column-medial lemniscus and its thalamic projections, specifically the superior thalamic radiation (STR), relay afferent feedback essential for stabilizing force output and preventing maladaptive central inhibition (Bolognini et al., 2016; Purves, Augustine, & Fitzpatrick, 2001). Such sensory feedback is essential for skilled sensorimotor behaviour and functional motor output (Asan et al., 2022; Nowak & Hermsdörfer, 2006). Beyond classic ascending pathways, long-range association fibres - such as the superior longitudinal fasciculus (SLF) and cingulum bundle - enable the integration of attentional, executive, and sensory information. This provides a structural substrate for the well-established association between HGS and higher cognitive function, but recent studies also show that association fibres play a role in proprioception and state estimation (Chilvers et al., 2022), required for grip force execution.

Effective signal transmission through microstructurally coherent tracts underlies successful execution of power grip. While forming a critical bottleneck particularly during development and aging (Surgent et al., 2023) this coherence is crucial throughout the lifespan. For example diffuse white matter pathology, such as small-vessel disease (detectable on MRI scans through white matter hyperintensities (WMH)) (Prins & Scheltens, 2015) is already observed in healthy subjects from age 40-45 on. Clear evidence exists that WMHs precede cognitive decline and are increasingly recognized to be involved in the aetiology of AD (Brickman et al., 2015; Debette et al., 2019; Prins & Scheltens, 2015; Wardlaw et al., 2015). Even in mild states they could be associated with physical disability, possibly through reduced speed, fine motor coordination and muscular strength (Sachdev, 2005). Additionally, stronger grip was

associated with reduced WMHs in people with major depressive disorder (J. A. Firth et al., 2020). This broad influence of white matter tract integrity on various levels supports the system-level perspective on HGS.

4.2.4 Subcortical involvements

On the subcortical level, the basal ganglia function as a critical modulator of force amplitude through a functional segregation of planning and execution. Functional neuroimaging indicates that the anterior basal ganglia, specifically the caudate nucleus and anterior putamen, are primarily involved in the preparatory phase, where they scale their activity based on the predictability of the required force (Wasson et al., 2010). This, however, was mainly seen in precision grip force (Prodoehl et al., 2009). In contrast, posterior nuclei, including the subthalamic nucleus (STN) and the globus pallidus internus (GPi) were seen to function as gate-keepers. Forming part of the basal ganglia-thalamocortical circuits, those nuclei exhibit activity that scales linearly with total force amplitude and rate of force development (Prodoehl et al., 2009; Spraker et al., 2007; Vaillancourt et al., 2004). Pathophysiological evidence from Parkinson's disease suggests that the loss of dopaminergic input to the basal ganglia results in force overshoot and impaired sensorimotor integration (Nowak & Hermsdörfer, 2006). Patients often exhibit an inability to finely tune the grip-load ratio, relying on excessive maximal force to compensate for the temporal delays in force recruitment. This compensatory strategy underscores the basa ganglia's role in optimizing energetic efficiency during motor output (Nowak & Hermsdörfer, 2006).

The thalamus modulates the flow of information between basal ganglia-cerebellar loops and the cortex, acting as a pivotal integration hub within motor control circuits (Bosch-Bouju et al., 2013). Through this integrative function, thalamic involvement supports the execution of maximum grip force as targeted motor program, in which force amplitude is constrained by predictive internal models rather than arising from indiscriminate maximal muscle recruitment (Opri et al., 2019). Thalamic activity has been shown to increase with force predictability, functioning as part of a pathway that inhibits premature motor programs while facilitating the selected force level (Wasson et al., 2010). Beyond thalamic contributions, cerebellar circuits do not directly encode force magnitude but contribute to temporal precision and error correction of high-force contractions.

The reliance on subcortical circuits shifts across the lifespan and following neurological disruption. In aging populations, a compensatory hyper-activation of the thalamus, putamen and cerebellum has beend observed (Noble et al., 2011). This increased subcortical recruitment likely represents a neural strategy to maintain force stability in the presence of age-related degradation of the CST and M1 excitability (Noble et al., 2011). On the other hand side, Ejaz and Xu et al. (2018) found that post-stroke mirror movements in the non-paretic hand (unintended movements that appear in the passive hand when the active hand voluntarily moves) are caused by the upregulation of a bilaterally organized subcortical system, presumably via the reticulospinal system (Ejaz et al., 2018). Vice versa, after subcortical stroke, impaired functional integrity of the corticospinal system was found to be

associated with higher recruitment of secondary motor networks, albeit being less efficient at generating motor output (N. S. Ward et al., 2006). This highlights the motor system's ability for reorganization at different hierarchies and the important involvement of non-cortical systems in the functional reorganization of the motor system.

4.2.5 Extra-motor system contributors

Beyond classical motor structures and circuits, maximal grip strength shows involvement of extra-motor systems involved in arousal, motivation, effort and attention allocation, and interoceptive monitoring. Functional connectivtiy studies, for example, have linked higher HGS to increased segregation within the salience/ventral attention network, driven particularly by strong intra-network connectivtiy of the right anterior insula to the left posterior insula and right midcingulate/medial parietal cortex, with those charateristics also being linked to cognitive function (Chong et al., 2024). This observation aligns with the role of the insular cortex as a critical hub in various processes, including interoception and the awareness of body (including hand) movements (Craig, 2009).

Structural imaging studies further support the involvement of extra-motor systems. Widespread increase in GMV in temporal cortices as well as subcortical regions is associated with stronger HGS. These associations are partly mediated by mental health measures, which themselves correlate with GMV in these regions (Jiang, Westwater, et al., 2022). More generally, the execution of physical force involves an implicit cost-benefit analysis in which perceived effort (cost) is weighed against expected reward. This subjective valuation of effort involves neural circuits within the orbitofrontal cortex (Padoa-Schioppa & Conen, 2017). At the subcortical level, the ventral striatum plays a central role in motivational processes for goal-directed behaviour and also acts as a limbic-motor interface by modulating activtiy in M1 and premotor areas with direct effect on motor output (Suzuki & Nishimura, 2022).

Furthermore, noradrenergic projections from the locus coeruleus can enhance cortical excitability and signal-to-noise ratio, facilitating higher force output under increased arousal (Aston-Jones & Cohen, 2005). Dopaminergic systems influence motor vigor and willingness to exert effort, with dopaminergic deficiency leading to reduced grip force even in the absence of primary motor deficits (Salamone et al., 2016). Serotonergic modulation further affects motoneuron excitability and central fatigue thresholds (Perrier et al., 2013). While these represent selected examples, this extra-motor perspective on HGS converges with observations that reduced attention, motivation, or affective drive – as observed in depressive states – can result in diminished grip strength independent of muscle mass or corticospinal integrity (Cass et al., 2024; J. Firth et al., 2018; J. A. Firth et al., 2020; Ganipineni et al., 2023; Rinne et al., 2018). Collectively, it indicates that HGS should be studied through a broad range of multi-modal brain measures, as complementarily captured by structural, functional and diffusion MRI, in order to adequately characterize its neural underpinnings and systemic significance.

4.2.6 HGS as a system-level read-out of multi-system integrity

HGS relies on processes spanning wide sprectrum of muscoloskeletal, neuromuscular, and neuronal hierarchies and mechanisms. This breadth allows to conceptualize HGS as a phenotypic readout of the multidimensional biological system comprising both body and brain. Importantly, this system should be understood as a whole, characterized by complex, bidirectional influences and interactions in health and disease, with HGS functioning as a readout, i.e. a measure and window into the current state of the system. Additionally, age-related declines to both the muscular and nervous systems each contribute to reductions in HGS, providing a joint account for its association with health outcomes that are metabolically and neurologically driven (McGrath et al., 2020). However, this account remains incomplete, as it does not sufficiently explain brain-level neuronal processes, required to understand the broad and systemic predictive value of HGS.

From a neuronal perspective, HGS reflects the successful convergence of an evolutionarily conserved core (brainstem and spinal circuits supporting foundational grasping), connectomic capacity (signal transmission via white matter microstructural coherence), motor and premotor cortical output scaling (gain-control and recruitment within the LGN), sensorimotor integration (force scaling and grip stability through afferent feedback integration), subcortical and cerebellar modulations (basal ganglia & thalamus: facilitation of appropriate activation patterns, suppression of competing actions, stabilization of motor programs, cerebellum: timing, error correction), and cognitive-attentional-motivational drive (reduced attention, motivation, or affective drive can diminish HGS). Accordingly, HGS does not function as a localized motor measure but rather as a global stress test of the brain's capacity to mobilize integrated neural resources toward a unified physiological goal. This mechanistic interdependence provides a coherent explanation for the robust associations observed between HGS and system-wide health outcomes, spanning cardiovascular and cerebrovascular health (e.g. via WMHs), cognitive function (e.g. via attention, motivation, affect), and ultimately all-cause mortality. As such, HGS offers a unique, non-invasive window into the functional and structural "scaffolding" of the healthy, impaired or aging brain, positioning it as a fundamental metric in the study of organismal multi-system integrity.

4.3 Brain-behaviour predictive modeling using large-scale, observational neuroimaging data

The distributed, multi-level nature of HGS has important implications for its neuroscientific investigation. Suitable approaches must acknowledge its system-level, multi-modal and integrative nature and be sensitive to detect multivariate, distributed but specific patterns. The neural mechanisms underlying successful power grip execution are robust while remaining flexible, reflecting both conserved motor control principles and adaptive capacity. Accordingly, neuronal contributors to HGS must generalize beyond small, homogeneous samples resulting in robust, population-level patterns, while at the same time remaining sensitive to individual-level differences to detect flexible, potentially sub-group-specific neuronal strategies. ML techniques applied to large-scale, observational, population-based neuroimaging offers this necessary methodological capability.

4.3.1 Opportunities and limitations of large, observational neuroimaging data for brain-behaviour research

Over the past decade, neuroimaging research has increasingly shifted toward large-scale, population-based cohorts, such as the UK Biobank (UKB; Miller et al., 2016), which combine multi-modal brain imaging with extensive phenotypic, demographic, and health-related data in a large number of participants. This development has fundamentally expanded the scope of brain-behaviour research by enabling the investigation of subtle neural effects that are typically inaccessible in smaller samples.

A primary strength of large neuroimaging cohorts lies in their substantial sample sizes, which increase statistical power and enhance population-level generalizability. Large sample sizes facilitate the detection of small effect sizes - commonly observed in brain-behaviour associations within healthy populations - by reducing sampling error and attenuating the influence of random fluctuations (Smith & Nichols, 2018). Consequently, estimates of population parameters become more precise and reliable. Moreover, population-based sampling strategies improve external validity, allowing inferences that extend beyond narrowly defined or clinically enriched cohorts and thereby supporting conclusions at the level of the general population. An additional advantage of such cohorts is the availability of multi-modal neuroimaging data, such as structural MRI (sMRI), diffusion-weighted imaging (DWI), and resting-state functional MRI (rs-fMRI). This breadth of imaging data is often unavailable in smaller studies or data collections that are typically restricted to a single modality. The concurrent assessment of brain structure, white-matter integrity, and functional organization enables system-level investigations of neural substrates underlying behaviour in general, and HGS in particular.

Despite these advantages, large observational neuroimaging datasets are subject to several fundamental limitations. Most notably, their non-interventional nature precludes direct mechanistic inference. Unlike randomized controlled trials, observational studies cannot isolate causal brain-behaviour relationships. Furthermore, observed associations may reflect shared variance driven by non-neuronal or confounding factors rather than specific neural mechanisms. This limitation, however, is not unique to large observational cohorts but is inherent to all non-experimental designs, including many traditional neuroimaging studies with smaller samples. Furthermore, population-based cohorts typically consist of clinically unspecific, predominantly healthy individuals, resulting in restricted variance in both neural and behavioural measures. This can reduce signal-to-noise ratios for phenotypes of interest (Andrade, 2013). This issue may be exacerbated by the asynchronous acquisition of brain and behavioural data, for example when behavioural measurements are obtained outside the scanner, in contrast to task-based or activation studies. Consequently, even biologically meaningful neural contributions may explain only a small proportion of the observable behavioural variance. In addition, although a wide range of phenotypes is assessed, individual constructs are often measured with limited depth or specificity, and imaging protocols are typically constrained by relatively short acquisition times. From a statistical perspective, large samples increase the likelihood that trivial effects reach

conventional significance thresholds. Such statistically significant results however may lack scientific, clinical, or practical relevance (Smith & Nichols, 2018a).

Collectively, these characteristics indicate that large sample size alone does not guarantee neuroscientifically informative insights. Rather, the combination of weak neural signals, high-dimensional imaging features, and pervasive non-neuronal influences necessitates analytical frameworks that are capable of extracting multivariate patterns while rigorously assessing their robustness, generalizability, and interpretability (J. Chen, Patil, et al., 2023; Doshi-Velez & Kim, 2017; Molnar, 2020). At the same time, the scale of these datasets enables methodological approaches that are not feasible in smaller samples, particularly brain-behaviour predictive modelling using ML.

4.3.2 Brain behaviour predictive modeling: promise and current challenges

The emergence of large-scale, observational neuroimaging cohorts has created unprecedented opportunities for applying ML approaches to the study of brain-behaviour relationships at the population level (J. Chen, Patil, et al., 2023; Singh et al., 2022). In contrast to classical inferential frameworks that primarily quantify associations within a given dataset, predictive modelling aims to learn patterns that generalize beyond the training sample and are applicable at the individual level. This shift in focus aligns closely with the goal of identifying reproducible neural signatures of behaviour. However, empirical applications of predictive modelling in neuroimaging have revealed substantial methodological and conceptual challenges that must be addressed to enable meaningful neuroscientific interpretation (Chekroud et al., 2024; Gell et al., 2024; Li et al., 2022; Wilkinson et al., 2020).

Large observational datasets fundamentally alter the feasibility of multivariate brain-behaviour predictive models (Marek et al., 2022). Sample sizes on the order of thousands to tens of thousands substantially alleviate constraints imposed by the curse of dimensionality (Berisha et al., 2021), which are particularly severe in high-dimensional neuroimaging data and further exacerbated by the integration of multiple imaging modalities. Within this regime, predictive modelling approaches become practically viable, allowing the exploitation of distributed information across large feature spaces. Crucially, ML models can capture multivariate and interacting neural patterns that are inaccessible to univariate analyses. A central promise of brain-behaviour predictive modelling lies in its explicit emphasis on generalization. Rather than evaluating effects solely within-sample, predictive performance is assessed on independent or previously unseen data. This property is not only of interest for clinical applications but also for research in healthy populations, where the goal is to elucidate general neurobiological principles underlying behaviour. Moreover, predictive models move beyond inference on group averages and enable individual-level predictions. When combined with explanation extraction methods, such as feature importance weights and Shapley values, ML models can yield interpretable insights into the underlying brain–behavior associations. Feature importance estimates explain a model's behavior by making its internal functioning more transparent, thereby enabling the detection of spurious or undesirable predictive strategies on the one hand and supporting domain-specific interpretation on the

other. This can be informative in translational applications such as biomarker discovery, sub-typing or risk stratification, and it also advances our understanding of the neuronal mechanisms underlying measures such as HGS in healthy populations. HGS presumably relies on a neural architecture that permits multiple neurological strategies to achieve a similar behavioral output. Gaining insights at the individual level is therefore valuable for capturing such heterogeneous and flexible strategies, which may be particularly engaged in the presence of subtle impairments or compensatory processes.

In principle, predictive modelling provides a powerful framework for identifying population-level, generalizable brain-behaviour relationships and for uncovering distributed multivariate neural signatures that remain invisible to univariate analyses. These advantages have driven its rapid adoption across cognitive neuroscience, neuroimaging, psychiatry, and neuroepidemiology. However, empirical findings have often fallen short of these expectations (Chekroud et al., 2024; Kapoor & Narayanan, 2022). For many behavioural, cognitive, and health-related phenotypes, predictive performance remains modest, and models frequently fail to generalize across samples, cohorts, or acquisition sites (Arbabshirani et al., 2017). This gap between conceptual promise and empirical reality has motivated increasing scrutiny of predictive modelling practices in neuroimaging research.

Multiple interrelated factors contribute to these limitations. Neuroimaging features typically exhibit low signal-to-noise ratios, with neural measures explaining only a small proportion of variance in behavioural outcomes, particularly in non-clinical populations. Simultaneously, imaging-derived features are high-dimensional and strongly multicollinear, reflecting both biological organization and measurement properties of imaging modalities. This complicates model fitting, making it challenging to obtain generalization models and hinders straightforward neuroscientific interpretation by increasing uncertainty in feature importance estimates. In addition, methodological shortcomings (e.g. data leakage, inappropriate cross-validation (CV) schemes, biased model evaluation) can inflate apparent predictive performance, obscure true generalization ability, and limit validity of neurobiological insights (Bengio & Grandvalet, 2004; Demšar & Zupan, 2021; Sasse et al., 2025; Varoquaux, 2018). The implications of these challenges extend beyond limited predictive accuracy. Overestimated performance can foster unwarranted confidence in weak models and yield limited or even misleading insights into underlying neurobiology, particularly when unstable feature importance patterns are overinterpreted as evidence for neural mechanisms. Addressing these challenges is therefore critical not only for improving predictive performance, but also for establishing brain-behaviour predictive modelling as a reliable tool for neuroscientific discovery.

4.4 Confounding in brain-behaviour prediction: nuisance or biological reality?

Confounding represents one of the most pervasive and conceptually challenging issues in brain-behaviour predictive modelling (Alfaro-Almagro et al., 2021; Benkarim et al., 2021; Chyzhyk et al., 2022; Hamdan et al., 2023a; Rao et al., 2017), with large datasets further amplifying the problem due to their high sensitivtiy to atifactual associations (Smith & Nichols, 2018b). Within much of the

neuroimaging literature, confounders are primarily treated as technical nuisances that can be addressed through standardized preprocessing pipelines, heuristic covariate selection, or correlation-based definitions of unwanted variance. From this perspective, confounding variables are assumed to reflect extraneous influences that should be statistically controlled for, often without explicit justification regarding their selection or the implications of different adjustment strategies. While this pragmatic approach offers procedural simplicity, it risks oversimplifying the problem of confounding and is insufficient for many neurobiological research questions.

Brain structure and function are embedded within a broader biological system and do not operate in isolation; but are a part of interdependent and causally intertwined processes encompassing brain, body and environment. Specifically, in the context of HGS, variables such as age, sex, body composition, health status, and lifestyle factors influence both neural measures and the behavioural outcome through shared developmental, physiological, and environmental pathways. Consequently, some variables that are routinely labelled as confounders in neuroimaging research may represent unwanted extraneous variance, meaningful biological signal, or a combination of both, and it is often unclear whether—and to what extent—these categories can be meaningfully separated. Failure to account for the underlying causal structure of the system under study risks flawed interpretations. Therefore, treatment of these variables requires causally informed methodological approaches (Elwert & Winship, 2014; Pearl, 2000, 2009; Pearl & Mackenzie, 2018; Rohrer, 2018; Tönnies et al., 2022; VanderWeele, 2019).

As a global system-level marker, HGS exemplifies the interdependent nature of variables within biological systems rather than isolated brain–behaviour relationships. It reflects the integrated functioning of musculoskeletal, neuronal, and cardiovascular systems, as well as age-related and lifestyle-dependent processes – all of which are either influenced by or do influence brain structure and function. Predictive models that aim for identification of neuronal underpinnings of HGS can achieve substantial performance by exploiting non-neural measures that covary with both brain measures and motor output. Insights derived from such models may therefore primarily capture confounding effects, e.g. due to demographics or peripheral physiology, rather than neural mechanisms. Although these associations can be biologically valid, they do not directly server the primary objective of elucidating neural contributions to behaviour. At the individual level, reliance on unintended sources of information may mask meaningful variability in neural organization, thereby reducing the explanatory value of predictions for both basic and translational neuroscience.

Meaningful neurobiological inference therefore requires appropriate confounder handling guided by explicit causal reasoning about relationships between variables (Elwert & Winship, 2014; Pearl, 2009; Tönnies et al., 2022; Wysocki et al., 2022). Additionally, for gaining meaningful neurobiological insights from predictive models requires accounting for multicollinearity inherent in neural representations. Taken together with the further challenges of predictive modelling with large-scale, observational neuroimaging data, these considerations underscore that methodological rigor is not

an end in itself, but a necessary prerequisite for addressing substantive neuroscientific questions. Accordingly, the methodological work in the first two studies of this thesis establishes the necessary conceptual foundation required for the empirical investigations that follow. In sum, large scale observational data and ML approaches offer the opportunity to deepen our understanding of behavioural phenotypes such as HGS, which rely on distributed, system-level neuronal architectures—but only when they are implemented and interpreted with methodological rigor.

4.5 Aims and scope of this thesis

Brain–behaviour predictive modelling using large-scale neuroimaging data holds substantial promise for advancing systems neuroscience and neurobiomedical research but faces significant challenges in achieving reliable, generalizable, and interpretable findings. This thesis targets neurobiologically interpretable brain-behaviour neuroimaging-based predictive modelling, addressing key challenges arising from low signal-to-noise ratios, feature multicollinearity, and pervasive non-neuronal confounding. Through a sequential approach, it progresses from a critical synthesis of methodological pitfalls and possible solutions in neuroimaging-based machine learning (Study 1), to solution-oriented development of a causally informed framework for confounder selection and adjustment (Study 2), and finally the empirical application of these methods together with multicollinearity-aware model interpretation to identify system-level neural signatures of HGS in a large, healthy middle-to-older-aged cohort (Study 3). Beyond methodological advances, conceptually, the thesis demonstrates that HGS reflects the integrity of distributed sensorimotor transmission and subcortical control systems, providing a neurobiological explanation for its role as a global marker of brain and organismal health.

In *Study 1*, I systematically reviewed challenges in brain-based predictive modelling to promote more reliable and generalizable findings in precision psychiatry. Specifically, I discussed ubiquitous mistakes in data handling (e.g. data leakage, i.e. the inadequate use of to-remain-hidden test data during training), limitations of generalization-estimates with CV and performance inflation in small samples. Additionally, I targeted the biasing impact of third variables (confounders, colliders i.a.) including a summary of mitigation strategies and the issue of site-specific effects in multisite datasets including different harmonization methods. Lastly, study 1 is concerned with post-hoc model interpretation methods and how they can enhance transparency, while stressing the importance of consideration of feature multicollinearities and contextualization of feature importances with the model's actual performance and domain knowledge.

In *Study 2*, I zoomed in on the challenge of confounding influences in brain-based predictive models. Concretely, I investigated how causal inference tools can support the debiasing of ML models to achieve more meaningful, reliable and generalizable brain-based predictions in neurobiomedicine. I addressed how conventional practices of defining confounders often rely on heuristics or correlations alone, which risks confusing them with colliders or mediators and ultimately leads to models exploiting spurious associations rather than genuine biological mechanisms. To counter this, I theoretically justified, conceptually proposed and empirically illustrated a three-step approach for confounder selection and adjustment with the aim to be pragmatically integrable in neuroimaging studies. The framework uses established concepts from the causal inference literature and makes them applicable and usable for the broader field of neurobiomedical research, thereby trying to integrate knowledge between disciplines and at the same time solve a prominent problem in neurobiomedical

(observational) studies. Concretely, it involves a domain-knowledge-driven causal analysis formalized in a directed Acyclic Graph (DAG), the application of graph-theoretical rules to distinguish different types of third variables and identify a minimal sufficient set of confounding variables to adjust for (deconfounders), and the statistical evaluation of and adjustment for these deconfounders. As a related challenge that limits rigorous debiasing of ML models, I further discussed the limitations of common linear feature residualization methods and explored the potential of adopting double machine learning as an alternative confounder adjustment strategy, while clarifying that even with appropriate deconfounding, the models remain fundamentally associative and do not equate to causal inference without additional strong justifications.

In Study 3, I aimed to identify neuronal underpinnings of individual differences in HGS by leveraging large-scale, multi-modal neuroimaging data in a middle-to-old, healthy cohort from the UKB. Specifically, I employed thorough model evaluation, comparison and selection with rigorous confounder control (applying the previously developed framework) to generate generalizable, individual-level predictions of HGS based on structural MRI, resting-state functional MRI, and a variety of diffusion tensor imaging features in nine uni-modal and two multi-modal settings. Multicollinearity aware model interpretation (clustering-based feature importance analysis) of best models in an out-of-sample prediction, offering a comprehensive picture of the most relevant multi-modal neuroimaging predictors of HGS as operationalization of wider motor performance and general (physical and brain) health. The study found that explanatory power was concentrated in a limited set of white-matter pathways integrity and structure of pallidal nuclei rather than being uniformly distributed across the brain. Across sexes and modelling strategies, the microstructural integrity of the ascending somatosensory medial lemniscus emerged as the dominant predictor, with additional contributions from basal ganglia (mainly anterior globus pallidus) and thalamic structure as well as cerebellar and thalamocortical tracts. Both cortical structure and functional features played a minor role. These results suggest that HGS reflects the integrity of distributed sensorimotor-cognitive integration circuits, explaining its value as a global marker of overall health, frailty, cognition and mortality and identifying the underlying reasons for why it is a universal and versatile marker (despite appearing as a rather simple motor execution), namely because it directly reflects brain health, specifically integrity of wide-ranging (mainly white matter) pathways and effectiveness/success of wide-ranging/system-level signal transfer and information integration across brain systems.

In sum, the thesis aimed to establish a systematic framework from global challenges in brain-based predictive modelling, via methodologically solution development, to their application, serving the main goal of identifying neural explanations of HGS to understand and explain why and how it serves as a global health marker.

5 Empirical work

5.1 Manuscript 1: Overview of Challenges in Brain-Based Predictive Modelling: Toward Meaningful Predictive Insights

Komeyer, V., Nieto, N., Eickhoff, S. B., Raimondo, F.[+], & Patil, K. R.[+] (2025). Overview of Challenges in Brain-based Predictive Modeling: Towards meaningful predictive insights. *Biological Psychiatry*.

Impact Factor (2025): 9.0

5-year Impact Factor (2025): 10.4

Own contributions according to CRediT

- Conceptualization
- Visualization
- Validation (literature review)
- Writing – original draft
- Writing – review and editing

The following contributions do not apply to this type of article (Review): data curation, formal analysis, investigation, methodology, software.

Review

Biological Psychiatry

Overview of Challenges in Brain-Based Predictive Modeling: Toward Meaningful Predictive Insights

Vera Komeyer, Nicolás Nieto, Simon B. Eickhoff, Federico Raimondo, and Kaustubh R. Patil

ABSTRACT

Predictive analytics based on machine learning (ML) and artificial intelligence is a powerful tool enabling precision psychiatry and providing insights into brain-behavior relationships. However, given the mixed results observed in the field so far, making meaningful progress requires careful consideration of several key challenges to ensure the validity of models and findings, including overfitting, confounding biases, site effect harmonization, and interpretability, among others. First, we highlight limitations of cross-validation, a ubiquitous ML strategy used to prevent overfitting and obtain generalization estimates, emphasizing the risk of performance inflation and the need for independent validation. Next, we introduce different types of so-called third variables that can influence the examination of a brain-behavioral relationship of interest in different ways, using causal inference principles. We emphasize the biasing impact of confounding variables on ML models and summarize common mitigation strategies. We then discuss site-specific effects in multisite datasets, reviewing different harmonization strategies to reduce unwanted variability and site-specific noise. Finally, we explore post hoc model interpretation methods to enhance model transparency while cautioning against misinterpretation. By integrating rigorous result validation, confounder control, and interpretability techniques, researchers can ensure that ML models produce more reliable and generalizable findings and avoid spurious associations.

https://doi.org/10.1016/j.biopsych.2025.09.003

Predictive machine learning (ML) models using neuroimaging data that reliably forecast clinical outcomes and individual risk profiles can facilitate personalized and effective psychiatric treatments. Such models can assist conventional symptom-based assessments by providing objective, data-driven diagnostic tools, thereby addressing a longstanding need for objective biomarkers in psychiatry (1–5). Improved data acquisition, large-scale data sharing, and advanced analytics together have raised expectations for future advancements (6–9). Recent research has demonstrated the capability of predictive models to uncover complex and subtle patterns in neuroimaging data supporting diverse tasks including diagnosis, prediction of treatment response, and disease subtyping for diverse psychiatric conditions (10–16). Furthermore, neuroimaging-based ML models have helped develop and validate new symptom-based scores that capture individual differences better than traditional diagnoses, supporting more personalized assessment and treatment in mental health (2,17,18).

However, some studies have questioned the robustness and generalizability of ML results (19,20). In addition to addressing data reliability (21,22), it is crucial to tackle key challenges such as data biases including confounding effects, harmonization of multisite datasets, and methodological issues in model evaluation and interpretation. In this perspective article, we aim to deepen the understanding of these challenges and associated methodological caveats and provide guidance where possible.

Supervised ML learns patterns from measured data, where input features (X), such as imaging-derived phenotypes, are used to predict a target variable (Y), such as a clinical diagnosis. Its strength lies in detecting subtle multivariate relationships often missed by conventional analytical methods. Crucially, the main objective of a predictive model is to achieve accurate predictions on new unseen data—referred to as out-of-sample generalization. Cross-validation (CV) is commonly used for estimating generalization performance. However, it can produce biased and unstable results, particularly leading to overoptimistic performance estimates when sample sizes are limited (20,23). Therefore, the output of a CV process must be treated carefully, as we will elaborate in CV Is an Estimation. Biases in data can lead to biased models and misleading insights, compromising both clinical decision making and scientific insights. Understanding Third-Variable Effects in Biomedical Research addresses these biases through the lens of third variables, introducing key causal concepts to identify and manage them. Special emphasis is placed on confounding variables, commonly encountered type of third variables, particularly relevant in observational data (24–27). ML models typically perform better with large datasets (28), but combining data from multiple sites can introduce site effects—biases from systematic differences in

ISSN: 0006-3223

acquisition (29). Data harmonization helps address this (30,31), but Effects of Site and Data Harmonization highlights key caveats when applying commonly used data harmonization methods within ML pipelines. Finally, Post Hoc Model Interpretation details the need to carefully consider model characteristics and the validity of post hoc interpretations to ensure meaningful insights (32)—key to clinical adoption of predictive analytics. The key challenges and potential mitigation strategies are summarized in Table 1.

CV IS AN ESTIMATION

Much like the process of conducting clinical trials in drug development requires extensive testing on independent patient populations—often involving years of research, regulatory approval, and significant financial investment—collecting new datasets to validate ML models can also be lengthy and resource intensive. To circumvent this obstacle, the widely adopted methodology to evaluate a model's out-of-sample generalization is to partition the data into training and testing sets, emulating unseen data. This is known as CV.

CV comes with several challenges and limitations that have been extensively discussed in the literature, including unreliable estimations of variance (33), overoptimistic and unstable performance estimates due to small sample size (23), overfitting (34), concerns related to data averaging (35), and selective reporting of findings (36). A fundamental limitation of CV, however, has not been emphasized—it provides an estimation rather than measuring a model's true performance. While CV is often thought to estimate how a particular model is expected to perform on new unseen samples, it only estimates the average performance of models trained on different but equally sized overlapping subsets of the available data (37).

Drawing on the analogy with clinical trials in drug development, CV estimates can be thought of as results obtained during the early phases of such trials. Like drugs, ML models are intended for real-world deployment once proven effective. However, as with pharmaceuticals—where approximately 90% of candidates ultimately fail, and even 41% fail during phase III despite earlier success (38)—ML models often encounter similar challenges. A model may demonstrate high accuracy under CV but still fail when applied beyond the controlled development setting and data, highlighting the limitations of early-phase evaluation (39).

Building an ML model encompasses selecting from a diverse pool of data processing steps and learning algorithms that can be parametrized and combined in a myriad of ways. This flexibility often leads to iterative refinement to obtain high CV accuracy. This iterative process, even when unintentional or carried out by different research teams, exacerbates performance overestimation: As models are repeatedly tuned and retested on the same data, they become increasingly tailored to idiosyncrasies of the sample rather than learning robust, generalizable patterns (40), which might even lead to false positives (41). Overestimated performance can lead to false confidence in the model's ability to uncover meaningful and general biological or behavioral associations, ultimately skewing conclusions and limiting reproducibility and real-world applicability. Readers should be mindful of the inherent limitations of CV and consider the broader context in which the results were obtained, including sample size, origin of the data, and previous findings in the literature. In addition to every researcher following good practices, the responsibility lies with the reader to critically assess whether the reported findings are robust, generalizable, and meaningful within their specific domain of application.

We recommend using nested CV for unbiased error estimates, with comparisons being restricted to a preselected set of candidate workflows. To avoid data leakage, all preprocessing should be strictly performed within training folds (42). Statistical model comparisons should rely on appropriate paired tests (43), and all tested models and hyperparameters should be reported transparently. Finally, results should be interpreted carefully, keeping in mind that CV estimates reflect the average performance of the modeling procedure rather than the exact error of the final fitted model. The latter should be validated on an independent test dataset to confirm generalizability.

UNDERSTANDING THIRD-VARIABLE EFFECTS IN BIOMEDICAL RESEARCH

Predictive models in psychiatry aim to either elucidate neurobiological mechanisms or support clinical decision making. Both goals require generalizable models. However, third variables (Z) can influence the relationship of interest between features (X, e.g., brain imaging measures) and outcomes (Y, e.g., clinical phenotypes), potentially hindering generalizability. In biomedical and psychological research, where biological, behavioral, and environmental factors are tightly interwoven, third-variable effects are often unavoidable, as has been demonstrated in large-scale observational datasets such as the UK Biobank (44,45).

Third variables can act as confounders, colliders, or mediators, each affecting the feature-target relationship differently and therefore requiring distinct handling (Figure 1). Correlation-based criteria alone cannot distinguish between these types as all could produce the same correlation with both X and Y (46). Instead, cause-effect reasoning, often aided by directed acyclic graphs (DAGs), is needed for distinction (47,48).

A confounder is a common cause of both X and Y, biasing their relationship and the respective predictive model if not controlled for [confounder bias, Simpson's paradox (49)]. For example, early childhood trauma may confound the relationship between hippocampal volume (e.g., stress-induced increased glucocorticoid exposure reducing synapto- and neurogenesis) (50) and depression risk (e.g., via higher likelihood of unhealthy lifestyles) (51). In contrast, a collider is a common effect of X and Y. Controlling for a collider induces a spurious X-Y association, biasing the predictive model [collider bias, Berkson's paradox (52)] [e.g., (53,54)]. For example, depression can act as a collider in studies of serotonin receptor function (e.g., using positron emission tomography) and cortisol levels because depressive symptoms can result from both reduced 5-HT_{1A} receptor binding (55) and hypercortisolemia (hypothalamic-pituitary-adrenal axis dysfunction) (56). A mediator lies on the indirect causal path from X to Y, transmitting part of the effect (46). For example,

Table 1. Overview of the Different Challenges, Examples, Suggestions, and Recommended References

Challenge	Example	Recommendation	Key References
Model Evaluation			
Biased and Unstable CV Estimates	Performing k-fold CV in small samples (e.g., N = 100) or performing leave-one-out CV can inflate performance.	Use sample sizes that lead to reasonably sized inner CV data splits (e.g., if N = 100, in a 10-fold inner and outer CV, there would only be 1 sample left for testing in the inner CV; this is not reasonably sized). Avoid leave-one-out CV.	Varoquaux *et al.* (23)
Cherry-Picking of CV Results	Reporting results from 1) one (the best) CV fold, 2) train performances, or 3) the best-looking error metric.	Report performance mean and SD across folds. Report test set errors. Use multiple error metrics.	Komiyama and Maehara (36) Demsar *et al.* (111)
Confounding/Third Variables			
Biased and Biologically Misleading Models Due to Confounding	A model falsely attributes structural brain changes to schizophrenia when in reality these changes are (partially) driven by aging or long-term medication use.	Use DAGs to systematize variable relationships around the research question of interest to select proper adjustment variables, e.g., through the so-called backdoor criterion. Avoid using default research question agnostic confounders such as age or sex but communicate informed decisions transparently.	Pearl *et al.* (58) Wysocki *et al.* (46) Komeyer *et al.* (27) Rohrer *et al.* (59) Pearl and Mackenzie (61) VanderWeele *et al.* (60) Pearl *et al.* (58)
Incorrect Adjustment for a Collider	A variable is identified as confounder based on correlations but is actually a collider, so its adjustment introduces bias.	Use DAGs to make variable relationships transparent and identify appropriate deconfounding variables.	Wysocki *et al.* (46)
Adjustment for a Default Set of Variables	Default adjustment for demographics, such as age and sex, without further consideration of variable relationships.	Use literature and domain knowledge to arrive at relevant variables and model their relationships using a DAG.	Pearl and Mackenzie (61) Komeyer *et al.* (27)
Data Harmonization			
Separate Train-Test Splits When Harmonizing Data to Avoid Data Leakage	Original proposed ComBat finds its parameters on the whole datasets, which is only compatible with classical statistical analysis, but not with ML studies, where separated train and test sets are needed.	When integrating in ML pipelines, use newer versions of ComBat that allow separation of train-test, such as neuroHarmonize, harmonizer, and ComBat-MEGA.	Fortin *et al.* (76) Marzi *et al.* (90) Radua *et al.* (31) Hu *et al.* (84)
Expected Nonlinear Covariate Effects	Biological information, for example related to age, often presents nonlinear effects that traditional ComBat cannot model.	Allows for estimation of more complex covariate effects. neuroHarmonize allows for this flexible covariate effect estimation.	Fortin *et al.* (76)
Site-Target Relationship	Data acquired at each site may have different proportions of classes, for example patients and control participants. ComBat-based methods may require test labels to correctly harmonize without removing relevant information.	Estimate the degree of site-target dependence and use leakage-free harmonization models such as PrettyHarmonize, which do not need test targets to correctly harmonize.	Nieto *et al.* (92)
Estimating Number of Images per Site to Train the Harmonization Models	There is a minimum number of images for each site that are needed to correctly estimate the parameters of the models.	The required N is a function of the number of sites, number of features, and intrinsic characteristics of the problem. The Mahalanobis distance was proposed to quantify the multivariate site effect and estimate the minimum N. For ComBat, between 20 and 30 samples per site are needed. ComBat-based methods are recommended with a low number of images, in contrast to DL-based models, which require more data.	Parekh *et al.* (112)

Table 1. Continued

Challenge	Example	Recommendation	Key References
Harmonization on Unseen Sites	We aim to harmonize new data that were acquired in a new site that was not used for training the harmonization model.	ComBat is not able to harmonize data from sites that were not included at training time. NeuroHarmony relies on IQMs instead of site ID; thus, it can be applied to any image where the IQM can be extracted, and the obtained IQMs are in the range of the training images. DL methods can also harmonize data from unseen data.	García-Dias *et al.* (113) Abbasi *et al.* (93)
Uncompleted Effect of Site Removal	Effects of site can be due to complex interaction in the data. Some of the methods may partially remove the effects of site and lead to bias estimations.	Evaluate harmonization models to validate its capacity to remove the effect of site. Alternatively, leave-one-site-out CV can help to evaluate robustness and generalization of the models.	Solanes *et al.* (29)
Uneven Classification Prediction Across Sites	When most of the data come from 1 site, ML can underperform in smaller sites.	Report several metrics and perform separate metrics for each site. Calculate multisite-specific metrics.	Solanes *et al.* (74)
Covariance Harmonization	ComBat can only correct mean and variance but cannot correct covariance.	CovBat is recommended in those cases, as it is specifically designed to harmonize mean, variance, and covariance.	Chen *et al.* (91)
Not Possible to Access Raw Data From All the Sites	Privacy-preserving scenarios are common in medical applications. Access to raw data is not always possible.	Distributed ComBat demonstrated similar performance as ComBat without direct access to the raw data.	Chen *et al.* (114)
Repeated Measurements for Each Participant	When monitoring neurodegenerative diseases, such as Parkinson's, several images from the same participant may be acquired.	When repeated measures are available, LongitudinalComBat is recommended.	Beer *et al.* (115)
Result Interpretation			
Feature Importance Misinterpretation	SHAP shows age as the most important feature in a depression classifier—is this meaningful?	Contextualize feature importance (e.g., age may be a confounder). Use domain-specific knowledge to interpret. Do not confuse true to the model with true to the data, and do not confuse feature importance with causal explanations.	Chen *et al.* (116) Molnar (103)
Overstated Importance (Meaningless Explanations)	Gray matter volume is the most important feature. Model accuracy is almost at chance level.	Assess model's performance using multiple complementary error metrics (e.g., AUROC and balanced accuracy). Report model accuracy alongside interpretations. Contextualize interpretation of the model's accuracy.	Molnar (103)

AUROC, area under the receiver operating characteristic curve; CV, cross-validation; DAG, directed acyclic graph; DL, deep learning; IQM, image quality metric; ML, machine learning; SHAP, Shapley Additive exPlanations.

cortisol levels may mediate (indirect path) the direct effect of amygdala hyperactivity (e.g., measured through functional magnetic resonance imaging [MRI]) on depressive symptoms (56,57). Whether to adjust for mediators depends on whether the partial direct effect (amygdala hyperactivity→depressive symptoms, control for mediator) or the full effect including the indirect pathway via cortisol (do not control) is being sought (Figure 1).

To build valid models, researchers must account for third-variable types and mitigate bias accordingly. While simplified DAGs (Figure 1) illustrate basic principles, real-world neurobiological applications often involve complex interdependencies between many variables. To ensure unbiased predictive models, confounders must be controlled, while making sure not to control for colliders. However, in practice, confounder selection often lacks transparency or is based on default variables (e.g., demographics), increasing the risk of inadvertent effects, e.g., through collider adjustment.

A 3-step approach can help identify which variables to correct for. First, using literature-derived and clinical knowledge to build a DAG around the relationship of interest can clarify variable roles (types of third variables) and communicate assumptions transparently. Specifically, this DAG aids in identifying confounding pathways and therefore a correct set of variables to control for in the second step, for example through tools such as the backdoor criterion. The backdoor criterion originates in the causal literature (58) and states that to estimate a causal effect of variable X on outcome Y, it is necessary to block the so-called backdoor paths, i.e., paths

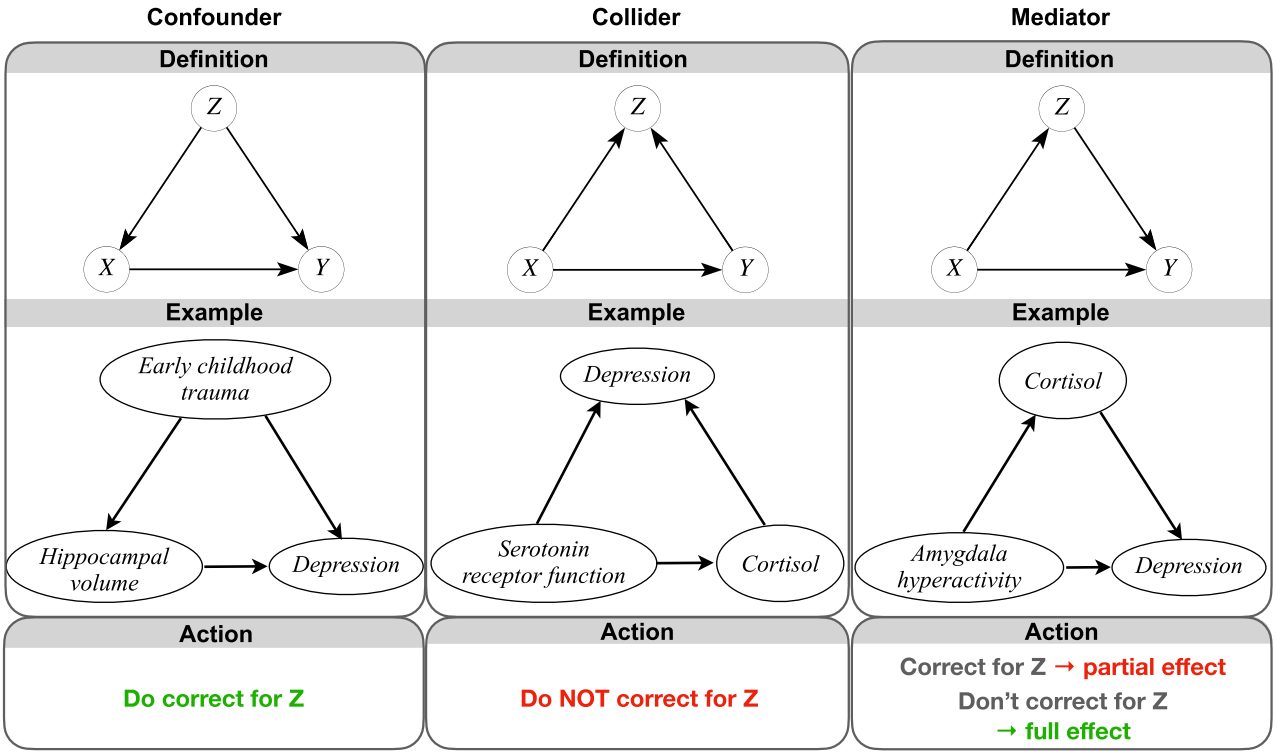


Figure 1. Definition of the different types of third variables, confounder, collider, and mediator using a directed acyclic graph. Each theoretical definition is supported by a simple biomedical example, and a recommendation for handling the respective type of third variable (action) is given.

with noncausal flow of information. Transferred to the field of associative predictive modeling, this translates to adjusting for variables that lie on indirect paths between X and Y with incoming arrows to X in the previously specified DAG (step 1). Online tools such as Dagitty (https://www.dagitty.net/dags.html) can support identifying biasing paths [see e.g., (27,48,58–61) for in-depth information]. Third, the identified variables should be checked for their statistical association with both X and Y, after which the actual model adjustment process can follow standard approaches from the ML literature [e.g., (26,62,63)].

Failing to adjust for confounders can lead ML models to learn spurious associations, capturing dataset-specific artifacts rather than neurobiologically meaningful patterns. This makes proper confounder selection and handling essential. In causal inference and treatment-outcome modeling, confounders are explicitly adjusted for to estimate causal effects, whereas in ML, they need to be considered to avoid unwanted bias or shortcuts that degrade generalizability of models and results. For example, in psychiatric research, age, medication use, and comorbidities can result in misleading associations between imaging-derived features and diagnostic labels. For example, a model may erroneously attribute structural brain changes to, e.g., schizophrenia when, in reality, these changes are (partially) driven by aging or long-term medication use. Likewise, comorbid conditions (e.g., anxiety, substance use disorder) can influence both brain imaging features and psychiatric diagnoses, making it difficult to disentangle disorder-specific neural signatures from overlapping, but distinct, effects.

Once identified, several post hoc confounder mitigation methods exist, each with implications and tradeoffs. Residualization removes confounder influences by often univariate linear regression of the confounder on features or target with subsequent residualization (true minus predicted feature/target) [e.g., (64)] but may leave nonlinear or multivariate confounding unaddressed and can even leak confounding signal into features or target (65). Matching balances distributions of confounding variables across groups or classes, thereby conditioning on the confounders and mitigating bias [e.g., (66)]. However, matching is data inefficient as unmatched samples are discarded and becomes increasingly complex with multiple confounders. Matching should not be confused with stratified CV [e.g., *StratifiedKFold* (67)], which ensures comparable distributions (e.g., of the target) between data splits but does not break confounder-feature/target associations, making it ineffective for mitigating confounding bias [e.g., (68)]. Including confounders as features/covariates can improve model performance but can reduce generalizability if confounder distributions shift across datasets (64) and does not give insights into feature-target mechanisms because these will be distorted by the confounders' influence. Post hoc tests [e.g., partial and full confounder tests (69)] can help assess model reliance on confounders but do not correct for them.

Thoughtful investigation and integration of third-variable structures is essential for unbiased models that allow for valid, generalizable, and interpretable ML research in psychiatry. Unbiased models require deliberate confounder control,

while improper adjustments (e.g., for colliders) must be avoided. Although confounder control may reduce apparent performance, it yields more meaningful and replicable results. Once appropriate confounders are identified, established confounder control strategies can be applied. DAGs provide a transparent framework for highly interwoven biomedical data to identify and justify adjustment variables by clarifying causal roles. Therefore, future work should move beyond default confounders (e.g., age, sex) toward question-specific, DAG-informed adjustments. Making this a standard practice in brain-based association studies will promote more generalizable models with greater psychiatric and clinical relevance.

EFFECTS OF SITE AND DATA HARMONIZATION

The acquisition of brain imaging data has expanded, greatly driven by advances in neuroimaging technologies. Detecting brainwide associations requires large samples for statistical and predictive analyses (28,70). Open science initiatives have facilitated this by making numerous datasets publicly available (71). The use of multisite data may also improve generalization by capturing biologically and demographically diverse samples (72). However, because collecting datasets is time- and resource intensive, pooling data from multiple sites has become common practice. While this approach offers great potential for advancing empirical neuroscience and predictive modeling, it also introduces new challenges, because systematic differences across sites can bias the resulting models.

Site-related data variability can stem from 2 main sources: acquisition differences and population differences. Variability due to acquisition is primarily driven by factors such as differences in scanners, imaging protocols, acquisition parameters, target definitions, or measurement procedures, none of which are related to biological signals (73). When this unwanted variability only affects the features (X) it is referred to as effects of site (Figure 2A) and can introduce bias into research outcomes if not properly identified or inadequately addressed (29,74). For example, MR images acquired from 2 scanners from the same manufacturer with the same parameters can differ (75), which extends to imaging-derived features (76) and nuances in image processing pipelines (77). Additionally, target can be also influenced by site due to sampling bias, differences in acquisition instruments, or different target definitions at each site/study (20). For example, there are different criteria in the Alzheimer's disease stages in the available datasets (78–80). These cases of different target definitions are also present in schizophrenia (20) and depression (81).

Beyond measurement-related variance, each site may recruit diverse populations, introducing sampling bias tied to the site's location and demographic reach. This may result in differences in diets, genetics, environmental factors, or socioeconomic factors, which are correlated with brain characteristics (82). Different sites may also recruit different target distributions, e.g., healthy control participants recruited at one site and patients at another.

If site affects the features, its role as a confounder depends on whether it also influences the target (Figure 2). When site does not affect the target, it acts as systematic noise, masking the biological signal (Figure 2A). In this case, removing site effects from features can improve the signal-to-noise ratio, thereby aiding meaningful learning and robustness (83). However, if site also influences the target, it becomes a confounder, enabling models to achieve high accuracy by exploiting nonbiological site information (Figure 2B) (see previous section).

Harmonization methods aim to eliminate site-specific variability while preserving signals of interest. When properly applied, they enhance statistical power, generalizability, and interpretability (73,84–87). These methods can be broadly classified into statistical approaches, primarily based on ComBat, and deep learning (DL) methods. It should be noted that data harmonization has different meanings across fields. For example, in psychology it refers to aligning different textual expressions to a common, semantically equivalent form (88). These fall outside the scope of this discussion. Finally, although data harmonization can provide appealing advantages, it may not be beneficial or even detrimental for some tasks (89).

ComBat is a commonly used harmonization method in statistical analyses and is the core of other proposed methods. ComBat was developed for genomics and later adapted for neuroimaging (76). Initially, ComBat estimates its parameters using the entire dataset, which is appropriate for statistical analysis but conflicts with ML principles—specifically, it violates the separation between training and test data, leading to data leakage (42,90). Extensions of ComBat allow for train-test separation (31,76,90). Some of the most prominent of ComBat's limitations are that it assumes that all features are in the same range, sites have similar numbers of images (at least 20), and the variance is equally distributed across sites. Additionally, it cannot correct features covariance (91), and it cannot be applied to data from an unseen site. Fortunately, several methods have been proposed to overcome these limitations (see Table 1). Additionally, the method struggles if site is a confounder (Figure 2B), as it assumes that any nonshared variance across sites is undesirable and could remove target variance, unless the target is preserved by specifying it as a covariate. However, this requires knowing the target value at test time, introducing leakage and precluding real-world application (92). Alternatively, normative modeling has also been proposed for harmonization, with the main difference being that the site effects are not estimated and removed, but the data are normalized instead (73).

DL-based harmonization methods offer a more flexible data-driven approach (93) by leveraging different ML architectures (94–97). DL approaches do not explicitly make assumptions about the nature of the site effects and can be applied at the image or feature level and can harmonize data from unseen sites. However, they require a substantial amount of training data. Finally, phantoms or traveling subjects allow training harmonization models without mixing biological and site variance (98), but it is inefficient and costly (99,100).

In summary, data harmonization has become a fundamental step in large-scale neuroimaging analysis. While acquisition sites mainly affect the features through instrumental factors, it can also affect the targets. It is essential to adapt the harmonization approach to the specific context of the study; in classical statistical studies, harmonization should

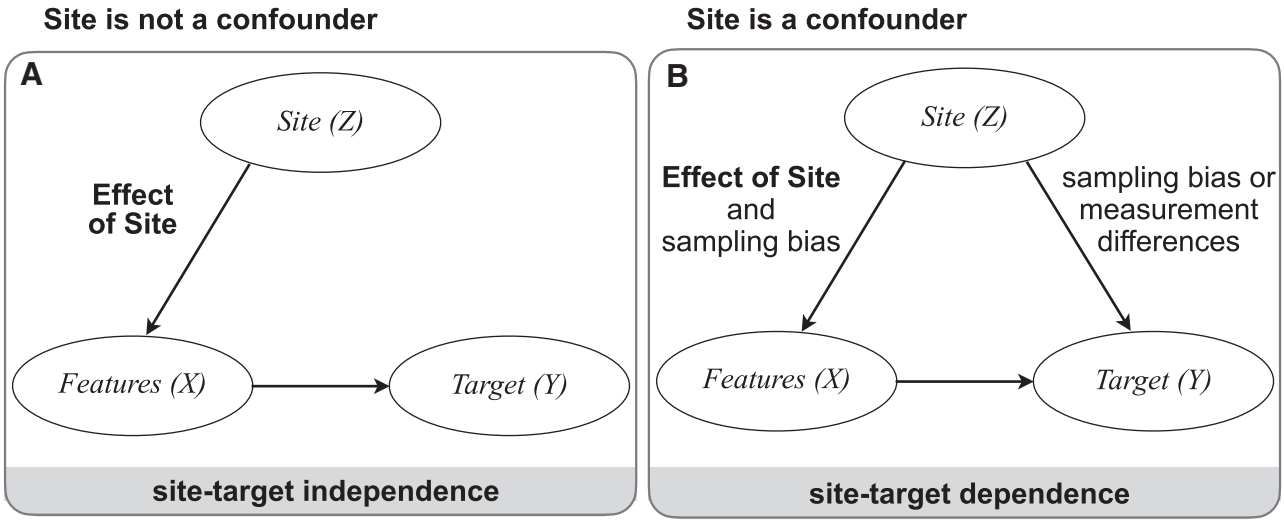


Figure 2. Different possible scenarios of the impact of site on features and target. Depending on the influence of site on the target, site does not act as a confounder **(A)** or does act as a confounder **(B)**.

prioritize the removal of site-specific biases to increase statistical power. In ML applications, extra considerations must be taken to correctly integrate harmonization methods in ML pipelines to avoid data leakage, unintentional removal of relevant signal, and ungeneralizable results. Choosing an appropriate harmonization method critically depends on the problem at hand, the data type (structural, functional or diffusion MRI), and the overarching research question. Decisions such as harmonizing on voxel or feature level, availability of traveling subjects or longitudinal data, and application on unseen sites are just a few examples of considerations important to identify a suitable harmonization method. While basic recommendations are presented in Table 1, please see (73,84,93) for a detailed discussion.

POST HOC MODEL INTERPRETATION

In clinical and research applications, high predictive accuracy alone is insufficient; understanding how models arrive at their predictions is equally important. A model must be evaluated beyond predictive accuracy (101), because it may incorrectly rely on site-specific factors or spurious third-variable associations, such as predicting improved myocardial infarction recovery in smokers due to age-related confounding (i.e., smokers are younger with a higher chance of recovery) (102). In psychiatry, an example of a common pitfall is the misattribution of brain-related differences to a disorder rather than medication effects. Model interpretability can help detect such wrong associations. Beyond determining what a model predicts, model interpretation can clarify why it reached that conclusion, ensuring meaningful and clinically sound findings (103). By examining the model's decision process, researchers can check whether it is consistent with biological and clinical knowledge before interpreting them as novel biomarkers.

Interpretability can be achieved either by model-specific (by design) or model-agnostic methods (103). By-design interpretability uses inherently interpretable algorithms (e.g., decision trees, linear models) as they rely on internal representations learned by the selected algorithm. For example, decision trees reveal exactly how decisions were made, or linear regression weights, particularly when Haufe-transformed (104,105), provide reliable feature importances. While informative, model-specific approaches restrict algorithm choice and may be too simple for capturing complex data patterns, potentially compromising predictive performance. Nonetheless, they can be a suitable choice if they align with the assumed variable relationships in the data. Model-agnostic interpretation, in contrast, can be applied to any model and must always be performed post hoc. In practice, psychiatric research questions likely will benefit from more complex models; thus, we will focus on challenges related to model-agnostic interpretation.

Consider a study yielding a successful model that accurately predicts unseen data. The next step is to determine how the model used the features to predict the target variable, i.e., to uncover the internal rules governing the model's decision-making process. Model-agnostic approaches quantify the contribution of each feature to the model's prediction by performing predictions on perturbed input data and comparing the outcomes (106). For example, permutation feature importance (107) randomly permutes a feature (or set of features) to break potential relationships with the target, compares performances between original (unpermuted) and permuted features, and the diverging loss in performance is considered the feature's contribution to the model's prediction. A widely adopted alternative is Shapley Additive exPlanations (SHAP) (108), based on Shapley values (109), which decomposes the predictions into additive feature contributions. SHAP captures the relative importance of a feature in driving a model's decisions and is applicable to a variety of model types by providing respectively suitable explainers [e.g., Molnar (103)].

Nonetheless, 2 caveats are critical when interpreting insights obtained through post hoc model interpretation. First,

feature importance is always assessed in a multivariate context, i.e., the interpretation depends on the relationships and interactions among all variables. A feature may appear important because it provides information not captured by any other variable. However, this does not necessarily imply that the feature is informative about the target variable in isolation. Conversely, a feature may appear unimportant because its information is shared with other variables (multicollinearity). In brain-related and psychiatric research, multicollinearities are common (e.g., among neighboring brain areas or between age, medication use, and illness duration). In such cases, interpreting feature groups rather than individual features is more meaningful. Usage of Owen values instead of SHAP values can be a suitable solution (110). Second, and critically, these methods estimate the contribution of features to a model's decision but do not assess whether the decision itself is correct. A feature may strongly influence a prediction, yet the prediction could still be incorrect. The insights from explanation methods are inherently tied to the model's predictive performance, reflecting only the aspects of the domain that contribute beyond chance-level predictions. As feature importance reliability is correlated with prediction accuracy (104), feature importance results should always be presented alongside performance metrics and not be interpreted as clinically meaningful when performance is near chance level.

Taken together, model interpretation is as essential as achieving high accuracy when applying ML approaches. Importantly, interpretation methods should be applied only after confirming that the model performs above chance level. Unlike hypothesis testing with defined significance thresholds, no standard defines how much above chance is enough to be considered successful. This ambiguity, in combination with the multivariate nature of predictive models, means that as readers, we must interpret feature importances while carefully considering the model's limitations (also see CV Is an Estimation). Generally, model-derived feature importances should be related to existing domain knowledge to avoid overinterpretating spurious results.

CONCLUSIONS

While ML provides powerful tools for neuroimaging-based decision making in psychiatry, evaluating the generalizability and reliability of such models demands rigorous scrutiny. CV can inflate performance estimates, and findings may be distorted by confounding variables, site-specific biases, or spurious correlations that misrepresent true brain-behavior associations. Importantly, the patterns uncovered by ML models reflect the statistical structure of the data, not necessarily causal or biologically meaningful mechanisms. To draw valid inferences, researchers must adopt robust methodological practices such as identifying and controlling for key confounders, applying proper harmonization techniques, using independent validation datasets, and predefining analysis pipelines.

Tools such as SHAP can support interpretation by highlighting which features contribute to predictions, but these outputs must be considered in light of the model's overall performance—evaluated using clinically relevant metrics such as accuracy, sensitivity, specificity, area under the receiver operating characteristic curve, and precision—and the context in which the data were collected. Where appropriate, decision-curve analysis and confusion matrices can help assess the practical utility of model-guided decisions in psychiatric settings.

Ultimately, we underline that ML should not be viewed as a shortcut to understanding complex brain-behavior associations but rather as an approach that, when used carefully, can generate testable hypotheses and clinically relevant insights. We encourage both researchers and clinicians to interpret ML findings with a critical eye, weighing methodological transparency, validation rigor, and clinical plausibility to ensure that conclusions reflect meaningful and generalizable relationships, not statistical artifacts.

ACKNOWLEDGMENTS AND DISCLOSURES

This work was supported by the Helmholtz Imagining grant BrainShapes (Grant No. ZT-I-PF-4-062 [to KRP]); the Multi-Omics Data Science project was funded from the program Profilbildung 2020 (Grant No. PROFILNRW-2020-107-A [to SBE]), an initiative of the Ministry of Culture and Science of the State of North Rhine-Westphalia; the H2020 Research Infrastructures (Grant No. EBRAIN-Health 101058516 [to SBE]); the Deutsche Forschungsgemeinschaft Collaborative Research Centre CRC1451 (Project No. 431549029 [to SBE]) on motor performance project B05; and the Universitätsklinikum Düsseldorf, Forschungskommission funded project VoxNorm [to KRP].

We thank Dr. Athena Demertzi for their insightful feedback and assistance in improving the clarity and relevance of this article to better align with the target readership of *Biological Psychiatry*.

The authors report no biomedical financial interests or potential conflicts of interest.

ARTICLE INFORMATION

From the Institute of Neuroscience and Medicine, Brain and Behavior, Forschungszentrum Jülich, Jülich, Germany (VK, NN, SBE, FR, KRP); Institute for Systems Neuroscience, Medical Faculty, Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany (VK, NN, SBE, FR, KRP); Department of Biology, Faculty of Mathematics and Natural Sciences, Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany (VK); and Institute of Diagnostic and Interventional Radiology, University Hospital Düsseldorf, Düsseldorf, Germany (VK).

FR and KRP contributed equally to this work.

Address correspondence to Federico Raimondo, Ph.D., at f.raimondo@fz-juelich.de.

Received Apr 1, 2025; revised Aug 29, 2025; accepted Sep 7, 2025.

REFERENCES

1. Wolfers T, Buitelaar JK, Beckmann CF, Franke B, Marquand AF (2015): From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. Neurosci Biobehav Rev 57:328–349.
2. Chen ZS, Kulkarni PP, Galatzer-Levy IR, Bigio B, Nasca C, Zhang Y (2022): Modern views of machine learning for precision psychiatry. Patterns (N Y) 3:100602.
3. Rutledge RB, Chekroud AM, Huys QJ (2019): Machine learning and big data in psychiatry: Toward clinical applications. Curr Opin Neurobiol 55:152–159.
4. Lucasius C, Ali M, Patel T, Kundur D, Szatmari P, Strauss J, Battaglia M (2025): A procedural overview of why, when and how to use machine learning for psychiatry. Nat Mental Health 3:8–18.
5. Bzdok D, Meyer-Lindenberg A (2018): Machine learning for precision psychiatry: Opportunities and challenges. Biol Psychiatry Cogn Neurosci Neuroimaging 3:223–230.

6. Milham MP, Craddock RC, Son JJ, Fleischmann M, Clucas J, Xu H, *et al.* (2018): Assessment of the impact of shared brain imaging data on the scientific literature. Nat Commun 9:2818.
7. Singh NM, Harrod JB, Subramanian S, Robinson M, Chang K, Cetin-Karayumak S, *et al.* (2022): How machine learning is powering neuroimaging to improve brain health. Neuroinformatics 20:943–964.
8. Giehl K, Mutsaerts H-J, Aarts K, Barkhof F, Caspers S, Chetelat G, *et al.* (2024): Sharing brain imaging data in the Open Science era: How and why? Lancet Digit Health 6:e526–e535.
9. Chen J, Patil KR, Yeo BTT, Eickhoff SB (2023): Leveraging machine learning for gaining neurobiological and nosological insights in psychiatric research. Biol Psychiatry 93:18–28.
10. Abraham A, Milham MP, Di Martino A, Craddock RC, Samaras D, Thirion B, Varoquaux G (2017): Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example. Neuroimage 147:736–745.
11. Wilkinson J, Arnold KF, Murray EJ, van Smeden M, Carr K, Sippy R, *et al.* (2020): Time to reality check the promises of machine learning-powered precision medicine. Lancet Digit Health 2:e677–e680.
12. Chen J, Patil KR, Weis S, Sim K, Nickl-Jockschat T, Zhou J, *et al.* (2020): Neurobiological divergence of the positive and negative schizophrenia subtypes identified on a new factor structure of psychopathology using non-negative factorization: An international machine learning study. Biol Psychiatry 87:282–293.
13. Chen J, Müller VI, Dukart J, Hoffstaedter F, Baker JT, Holmes AJ, *et al.* (2021): Intrinsic connectivity patterns of task-defined brain networks allow individual prediction of cognitive symptom dimension of schizophrenia and are linked to molecular architecture. Biol Psychiatry 89:308–319.
14. Gallo S, El-Gazzar A, Zhutovsky P, Thomas RM, Javaheripour N, Li M, *et al.* (2023): Functional connectivity signatures of major depressive disorder: Machine learning analysis of two multicenter neuroimaging studies. Mol Psychiatry 28:3013–3022.
15. Winter NR, Blanke J, Leenings R, Ernsting J, Fisch L, Sarink K, *et al.* (2024): A systematic evaluation of machine learning-based biomarkers for major depressive disorder. JAMA Psychiatry 81:386–395.
16. Van Essen DC, Ugurbil K, Auerbach E, Barch D, Behrens TEJ, Bucholz R, *et al.* (2012): The Human connectome Project: A data acquisition perspective. Neuroimage 62:2222–2231.
17. Chen J, Müller V, Hoffstaedter F, Nickl-Jockschat T, Derntl B, Kogler L, *et al.* (2020): Linking schizophrenia symptom dimensions to neuro-cognitive processes by multivariate pattern prediction. Biol Psychiatry 87:S408–S409.
18. Wen J, Antoniades M, Yang Z, Hwang G, Skampardoni I, Wang R, Davatzikos C (2024): Dimensional neuroimaging endophenotypes: Neurobiological representations of disease heterogeneity through machine learning. Biol Psychiatry 96:564–584.
19. Omidvarnia A, Sasse L, Larabi DI, Raimondo F, Hoffstaedter F, Kasper J, *et al.* (2024): Individual characteristics outperform resting-state fMRI for the prediction of behavioral phenotypes. Commun Biol 7:771.
20. Chekroud AM, Hawrilenko M, Loho H, Bondar J, Gueorguieva R, Hasan A, *et al.* (2024): Illusory generalizability of clinical prediction models. Science 383:164–167.
21. Milham MP, Vogelstein J, Xu T (2021): Removing the reliability bottleneck in functional magnetic resonance imaging research to achieve clinical utility. JAMA Psychiatry 78:587–588.
22. Gell M, Eickhoff SB, Omidvarnia A, Küppers V, Patil KR, Satterthwaite TD, *et al.* (2024): How measurement noise limits the accuracy of brain-behaviour predictions. Nat Commun 15:10678.
23. Varoquaux G (2018): Cross-validation failure: Small sample sizes lead to large error bars. Neuroimage 180:68–77.
24. Li J, Bzdok D, Chen J, Tam A, Ooi LQR, Holmes AJ, *et al.* (2022): Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. Sci Adv 8:eabj1812.
25. Görgen K, Hebart MN, Allefeld C, Haynes J-D (2018): The Same Analysis Approach: Practical protection against the pitfalls of novel neuroimaging analysis methods. Neuroimage 180:19–30.
26. Rao A, Monteiro JM, Mourao-Miranda J, Alzheimer's Disease Initiative (2017): Predictive modelling using neuroimaging data in the presence of confounds. Neuroimage 150:23–49.
27. Komeyer V, Eickhoff SB, Rathkopf C, Grefkes C, Patil KR, Raimondo F (2024): Correct deconfounding enables causal machine learning for precision medicine and beyond. medRxiv https://doi.org/10.1101/2024.09.20.24314055.
28. Schulz M-A, Bzdok D, Haufe S, Haynes J-D, Ritter K (2024): Performance reserves in brain-imaging-based phenotype prediction. Cell Rep 43:113597.
29. Solanes A, Gosling CJ, Fortea L, Ortuño M, Lopez-Soley E, Llufriu S, *et al.* (2023): Removing the effects of the site in brain imaging machine-learning—Measurement and extendable benchmark. Neuroimage 265:119800.
30. Saponaro S, Giuliano A, Bellotti R, Lombardi A, Tangaro S, Oliva P, *et al.* (2022): Multi-site harmonization of MRI data uncovers machine-learning discrimination capability in barely separable populations: An example from the ABIDE dataset. Neuroimage Clin 35:103082.
31. Radua J, Vieta E, Shinohara R, Kochunov P, Quidé Y, Green MJ, *et al.* (2020): Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA. Neuroimage 218:116956.
32. Ehsan U, Passi S, Liao QV, Chan L, Lee I-H, Muller M, Riedl MO (2024): The who in XAI: How AI background shapes perceptions of AI explanations. In: Proceedings of the Chi Conference on Human Factors in Computing Systems. New York, NY: ACM, 1–32.
33. Bengio Y, Grandvalet Y (2004): No unbiased estimator of the variance of K-fold cross-validation. J Mach Learn Res 5:1089–1105.
34. Ng AY (1997): Preventing "overfitting" of cross-validation data. In: Proceedings of the Fourteenth International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann, 245–253.
35. Giles CL, Lawrence S (1997): Presenting and analyzing the results of AI experiments: Data averaging and data snooping. In: Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence. Providence, RI: AAAI Press, 362–367.
36. Komiyama J, Maehara T (2018): A simple way to deal with Cherry-picking. arXiv https://doi.org/10.48550/arXiv.1810.04996.
37. Bates S, Hastie T, Tibshirani R (2024): Cross-validation: What does it estimate and how well does it do it? J Am Stat Assoc 119:1434–1445.
38. Sun D, Gao W, Hu H, Zhou S (2022): Why 90% of clinical drug development fails and how to improve it? Acta Pharm Sin B 12:3049–3062.
39. Paleyes A, Urma R-G, Lawrence ND (2023): Challenges in deploying machine learning: A survey of case studies. ACM Comput Surv 55:1–29.
40. Beyer L, Hénaff OJ, Kolesnikov A, Zhai X, van den Oord A (2020): Are we done with ImageNet? arXiv https://doi.org/10.48550/arXiv.2006.07159.
41. Thompson WH, Wright J, Bissett PG, Poldrack RA (2020): Dataset decay and the problem of sequential analyses on open datasets. eLife 9:e53498.
42. Sasse L, Nicolaisen-Sobesky E, Dukart J, Eickhoff SB, Götz M, Hamdan S, *et al.* (2024): On leakage in machine learning pipelines. arXiv https://doi.org/10.48550/arXiv.2311.04179.
43. Nadeau C, Bengio Y (2003): Inference for the generalization error. Mach Learn 52:239–281.
44. Miller KL, Alfaro-Almagro F, Bangerter NK, Thomas DL, Yacoub E, Xu J, *et al.* (2016): Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nat Neurosci 19:1523–1536.
45. Carey CE, Shafee R, Wedow R, Elliott A, Palmer DS, Compitello J, *et al.* (2024): Principled distillation of UK Biobank phenotype data reveals underlying structure in human variation. Nat Hum Behav 8:1599–1615.
46. Wysocki AC, Lawson KM, Rhemtulla M (2022): Statistical control requires causal justification. Adv Methods Pract Psychol Sci 5: 25152459221095823.

47. Maxwell SE, Cole DA (2007): Bias in cross-sectional analyses of longitudinal mediation. Psychol Methods 12:23–44.
48. Pearl J (1995): Causal diagrams for empirical research. Biometrika 82:669–688.
49. Sprenger J, Weinberger N (2021): Simpson's paradox. Stanf Encycl Philos. Available at: https://plato.stanford.edu/entries/paradox-simpson/?ref=praxarchy.com. Accessed March 14, 2025.
50. Humphreys KL, King LS, Sacchet MD, Camacho MC, Colich NL, Ordaz SJ, *et al.* (2019): Evidence for a sensitive period in the effects of early life stress on hippocampal volume. Dev Sci 22:e12775.
51. Wang X, Cao Z, Yin S, Duan T, Sun T, Xu C (2025): Childhood maltreatment and depression: Mediating role of lifestyle factors, personality traits, adult traumas, and social connections among middle-aged and elderly participants. BMC Med 23:319.
52. Berkson J (1946): Limitations of the application of 4-fold table analysis to hospital data. Biometrics 2:47–53.
53. Elwert F, Winship C (2014): Endogenous selection bias: The problem of conditioning on a collider variable. Annu Rev Sociol 40:31–53.
54. Tönnies T, Kahl S, Kuss O (2022): Collider bias in observational studies. Dtsch Ärztebl Int 119:107–122.
55. Drevets WC, Thase ME, Moses-Kolko EL, Price J, Frank E, Kupfer DJ, Mathis C (2007): Serotonin-1A receptor imaging in recurrent depression: Replication and literature review. Nucl Med Biol 34:865–877.
56. Pariante CM, Lightman SL (2008): The HPA axis in major depression: Classical theories and new developments. Trends Neurosci 31:464–468.
57. Ulrich-Lai YM, Herman JP (2009): Neural regulation of endocrine and autonomic stress responses. Nat Rev Neurosci 10:397–409.
58. Pearl J (2009): Causal inference in statistics: An overview. Statist Surv 3:96–146.
59. Rohrer JM (2018): Thinking clearly about correlations and causation: Graphical causal models for observational data. Adv Methods Pract Psychol Sci 1:27–42.
60. VanderWeele TJ (2019): Principles of confounder selection. Eur J Epidemiol 34:211–219.
61. Pearl J, Mackenzie D (2018): The New Science of Cause and Effect. New York, NY: Basic Books.
62. Dinga R, Schmaal L, Penninx BWJH, Veltman DJ, Marquand AF (2020): Controlling for effects of confounding variables on machine learning predictions. bioRxiv https://doi.org/10.1101/2020.08.17.255034.
63. Snoek L, Miletić S, Scholte HS (2019): How to control for confounds in decoding analyses of neuroimaging data. Neuroimage 184:741–760.
64. Chyzhyk D, Varoquaux G, Milham M, Thirion B (2022): How to remove or control confounds in predictive models, with applications to brain biomarkers. GigaScience 11:giac014.
65. Hamdan S, Love BC, von Polier GG, Weis S, Schwender H, Eickhoff SB, Patil KR (2022): Confound-leakage: Confound removal in machine learning leads to leakage. arXiv https://doi.org/10.48550/arXiv.2210.09232.
66. Wiersch L, Hamdan S, Hoffstaedter F, Votinov M, Habel U, Clemens B, *et al.* (2023): Accurate sex prediction of cisgender and transgender individuals without brain size bias. Sci Rep 13:13868.
67. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al.* (2012): Scikit-learn: Machine learning in python. J Mach Learn Res 12:2825–2830.
68. Hernan MA, Robins JM (2024): Causal Inference: What If, 1st ed. Boca Raton: Taylor & Francis.
69. Spisak T (2022): Statistical quantification of confounding bias in machine learning models. GigaScience 11:giac082.
70. Marek S, Tervo-Clemmens B, Calabro FJ, Montez DF, Kay BP, Hatoum AS, *et al.* (2022): Reproducible brain-wide association studies require thousands of individuals. Nature 603:654–660.
71. Poldrack RA, Gorgolewski KJ (2014): Making big data open: Data sharing in neuroimaging. Nat Neurosci 17:1510–1517.
72. Ma Q, Zhang T, Zanetti MV, Shen H, Satterthwaite TD, Wolf DH, *et al.* (2018): Classification of multi-site MR images in the presence of heterogeneity using multi-task learning. Neuroimage Clin 19:476–486.
73. Bayer JMM, Thompson PM, Ching CRK, Liu M, Chen A, Panzenhagen AC, *et al.* (2022): Site effects how-to and when: An overview of retrospective techniques to accommodate site effects in multi-site neuroimaging analyses. Front Neurol 13:923988.
74. Solanes A, Palau P, Fortea L, Salvador R, González-Navarro L, Llach CD, *et al.* (2021): Biased accuracy in multisite machine-learning studies due to incomplete removal of the effects of the site. Psychiatry Res Neuroimaging 314:111313.
75. Li H, Smith SM, Gruber S, Lukas SE, Silveri MM, Hill KP, *et al.* (2020): Denoising scanner effects from multimodal MRI data using linked independent component analysis. Neuroimage 208:116388.
76. Fortin J-P, Cullen N, Sheline YI, Taylor WD, Aselcioglu I, Cook PA, *et al.* (2018): Harmonization of cortical thickness measurements across scanners and sites. Neuroimage 167:104–120.
77. Antonopoulos G, More S, Raimondo F, Eickhoff SB, Hoffstaedter F, Patil KR (2023): A systematic comparison of VBM pipelines and their application to age prediction. Neuroimage 279:120292.
78. Marcus DS, Wang TH, Parker J, Csernansky JG, Morris JC, Buckner RL (2007): Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J Cogn Neurosci 19:1498–1507.
79. Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack CR, Jagust W, *et al.* (2005): Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI). Alzheimers Dement 1:55–66.
80. Fowler C, Rainey-Smith SR, Bird S, Bomke J, Bourgeat P, Brown BM, *et al.* (2021): Fifteen years of the Australian imaging, biomarkers and lifestyle (AIBL) study: Progress and observations from 2,359 older adults spanning the spectrum from cognitive normality to Alzheimer's disease. J Alzheimers Dis Rep 5:443–468.
81. National Collaborating Centre for Mental Health (2010): The Classification of Depression and Depression Rating Scales/Questionnaires. Depression in Adults with a Chronic Physical Health Problem: Treatment and Management. Leicester: British Psychological Society. Available at: https://www.ncbi.nlm.nih.gov/books/NBK82926/. Accessed June 16, 2025.
82. Institute of Medicine (US) Committee on Assessing Interactions Among Social, Behavioral, and Genetic Factors in Health (2006): Genes, Behavior, and the Social Environment: Moving Beyond the Nature/Nurture Debate. Washington, DC: National Academies Press.
83. Andrade C (2013): Signal-to-noise ratio, variability, and their relevance in clinical trials. J Clin Psychiatry 74:479–481.
84. Hu F, Chen AA, Horng H, Bashyam V, Davatzikos C, Alexander-Bloch A, *et al.* (2023): Image harmonization: A review of statistical and deep learning methods for removing batch effects and evaluation metrics for effective harmonization. Neuroimage 274:120125.
85. Da-Ano R, Visvikis D, Hatt M (2020): Harmonization strategies for multicenter radiomics investigations. Phys Med Biol 65:24TR02.
86. Wang Y-W, Chen X, Yan C-G (2023): Comprehensive evaluation of harmonization on functional brain imaging for multisite data-fusion. Neuroimage 274:120089.
87. Nan Y, Ser JD, Walsh S, Schönlieb C, Roberts M, Selby I, *et al.* (2022): Data harmonisation for information fusion in digital healthcare: A state-of-the-art systematic review, meta-analysis and future research directions. Inf Fusion 82:99–122.
88. McElroy E, Wood T, Bond R, Mulvenna M, Shevlin M, Ploubidis GB, *et al.* (2024): Using natural language processing to facilitate the harmonisation of mental health questionnaires: A validation study using real-world data. BMC Psychiatry 24:530.
89. Yu Y, Cui HQ, Haas SS, New F, Sanford N, Yu K, *et al.* (2024): Brain-age prediction: Systematic evaluation of site effects, and sample age range and size. Hum Brain Mapp 45:e26768.
90. Marzi C, Giannelli M, Barucci A, Tessa C, Mascalchi M, Diciotti S (2024): Efficacy of MRI data harmonization in the age of machine learning: A multicenter study across 36 datasets. Sci Data 11:115.
91. Chen AA, Beer JC, Tustison NJ, Cook PA, Shinohara RT, Shou H, Alzheimer's Disease Neuroimaging Initiative (2022): Mitigating site

effects in covariance for machine learning in neuroimaging data. Hum Brain Mapp 43:1179–1195.
92. Nieto N, Eickhoff SB, Jung C, Reuter M, Diers K, Kelm M, *et al.* (2024): Impact of leakage on data harmonization in machine learning pipelines in class imbalance across sites. arXiv https://doi.org/10.48550/arXiv.2410.19643.
93. Abbasi S, Lan H, Choupan J, Sheikh-Bahaei N, Pandey G, Varghese B (2024): Deep learning for the harmonization of structural MRI scans: A survey. Biomed Eng OnLine 23:90.
94. Dewey BE, Zhao C, Reinhold JC, Carass A, Fitzgerald KC, Sotirchos ES, *et al.* (2019): DeepHarmony: A deep learning approach to contrast harmonization across scanner changes. Magn Reson Imaging 64:160–170.
95. Torbati ME, Tudorascu DL, Minhas DS, Maillard P, DeCarli CS, Hwang SJ (2021): Multi-scanner harmonization of paired neuroimaging data via structure preserving embedding learning. IEEE Int Conf Comput Vis Workshops 2021 3277–3286.
96. Cackowski S, Barbier EL, Dojat M, Christen T (2023): ImUnity: A generalizable VAE-GAN solution for multicenter MR image harmonization. Med Image Anal 88:102799.
97. Komandur D, Gupta U, Chattopadhyay T, Dhinagar NJ, Thomopoulos SI, Chen J-C, *et al.* (2023): Unsupervised harmonization of brain MRI using 3D CycleGANs and its effect on brain age prediction. In: 2023 19th International Symposium on Medical Information Processing and Analysis (SIPAIM), 1–5.
98. Tian D, Zeng Z, Sun X, Tong Q, Li H, He H, *et al.* (2022): A deep learning-based multisite neuroimage harmonization framework established with a traveling-subject dataset. Neuroimage 257:119297.
99. Maikusa N, Zhu Y, Uematsu A, Yamashita A, Saotome K, Okada N, *et al.* (2021): Comparison of traveling-subject and ComBat harmonization methods for assessing structural brain characteristics. Hum Brain Mapp 42:5278–5287.
100. Ibrahim A, Primakov S, Beuque M, Woodruff HC, Halilaj I, Wu G, *et al.* (2021): Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework. Methods 188:20–29.
101. Doshi-Velez F, Kim B (2017): Towards a rigorous science of interpretable machine learning. arXiv https://doi.org/10.48550/arXiv.1702.08608.
102. Chen K-Y, Rha S-W, Li Y-J, Jin Z, Minami Y, Park JY, *et al.* (2012): 'Smoker's paradox' in young patients with acute myocardial infarction. Clin Exp Pharmacol Physiol 39:630–635.
103. Molnar C (2018): Interpretable Machine Learning, 3rd ed. Leanpub. Available at: https://leanpub.next/interpretable-machine-learning. Accessed March 26, 2025.
104. Chen J, Ooi LQR, Tan TWK, Zhang S, Li J, Asplund CL, *et al.* (2023): Relationship between prediction accuracy and feature importance reliability: An empirical and theoretical study. Neuroimage 274:120115.
105. Haufe S, Meinecke F, Görgen K, Dähne S, Haynes J-D, Blankertz B, Bießmann F (2014): On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage 87:96–110.
106. Scholbeck CA, Molnar C, Heumann C, Bischl B, Casalicchio G (2020): Sampling, intervention, prediction, aggregation: A generalized framework for model-agnostic interpretations. In: Cellier P, Driessens K, editors. Machine Learning and Knowledge Discovery in Databases. Cham: Springer International Publishing, 205–216.
107. Breiman L (2001): Random forests. Mach Learn 45:5–32.
108. Lundberg SM, Lee S-I (2017): A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. Advances in Neural Information Processing Systems, vol. 30. Red Hook, NY: Curran Associates, Inc. Available at: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.
109. Shapley LS (1953): 17. A value for n-person games. In: Kuhn HW, Tucker AW, editors. (1953), Contributions to the Theory of Games, II: Princeton: Princeton University Press, 307–318.
110. López S, Saboya M (2009): On the relationship between Shapley and Owen values. Cent Eur J Oper Res 17:415–423.
111. Demšar J, Zupan B (2021): Hands-on training about overfitting. PLoS Comput Biol 17:e1008671.
112. Parekh P, Vivek Bhalerao G, ADBS consortium, John JP, Venkatasubramanian G (2022): Sample size requirement for achieving multisite harmonization using structural brain MRI features. Neuroimage 264:119768.
113. Garcia-Dias R, Scarpazza C, Baecker L, Vieira S, Pinaya WHL, Corvin A, *et al.* (2020): Neuroharmony: A new tool for harmonizing volumetric MRI data from unseen scanners. Neuroimage 220:117127.
114. Chen AA, Luo C, Chen Y, Shinohara RT, Shou H, Alzheimer's Disease Neuroimaging Initiative (2022): Privacy-preserving harmonization via distributed ComBat. Neuroimage 248:118822.
115. Beer JC, Tustison NJ, Cook PA, Davatzikos C, Sheline YI, Shinohara RT, *et al.* (2020): Longitudinal ComBat: A method for harmonizing longitudinal multi-scanner imaging data. Neuroimage 220:117129.
116. Chen H, Janizek JD, Lundberg S, Lee S-I (2020): True to the model or true to the data? arXiv https://doi.org/10.48550/arXiv.2006.16234.

5.2 Manuscript 2: How causal inference tools can support debiasing of machine learning models for meaningful brain-based predictions

Komeyer, V., Herrmann C., Eickhoff, S. B., Rathkopf C., Raimondo, F.[+], & Patil, K. R.[+] (2025). How causal inference tools can support debiasing of machine learning models for meaningful brain-based predictions. *MedRXiv*.

The manuscript is currently under consideration at *Artificial Intelligence Review*.

Own contributions according to CRediT

- Conceptualization (content and structure)
- Data curation (data management)
- Formal analysis
- Investigation
- Methodology
- Validation (integration/validation of framework with the literature)
- Visualization
- Writing – original draft
- Writing – review & editing

How causal inference tools can support debiasing of machine learning models for meaningful brain-based predictions

Vera Komeyer[1,2,3], Jun-Prof. Dr. Carolin Herrmann[4], Prof. Dr. Simon B. Eickhoff[1,2], Dr. Charles Rathkopf[1], Dr. Federico Raimondo[1,2 †] & Dr. Kaustubh R. Patil[1,2,†]*

[1]Institute of Neuroscience and Medicine, Brain and Behaviour (INM-7), Research Centre Juelich, Juelich, Germany
[2]Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Duesseldorf, Duesseldorf, Germany
[3]Department of Biology, Faculty of Mathematics and Natural Sciences, Heinrich Heine University Duesseldorf, Duesseldorf, Germany
[4]Mathematical Institute, Faculty of Mathematics and Natural Sciences, Heinrich Heine University Duesseldorf, Duesseldorf, Germany
[†]These authors contributed equally
* Correspondence to Vera Komeyer (v.komeyer@fz-juelich.de)

Abstract

Machine learning (ML) offers transformative opportunities for neurobiomedicine, yet predictive models often exploit confounding-driven associations rather than genuine biological mechanisms, undermining generalizability and neurobiomedical validity. Current practice commonly defines confounders heuristically (e.g., age, sex) or correlationally, risking confusion with colliders or mediators. To address this, we propose a pragmatically integratable, causally informed three-step framework for confounder selection and adjustment aimed to support debiased, meaningful neurobiomedical supervised ML (SML) models. Step 1 involves a domain-knowledge-driven causal analysis of a specific research question, formalized in a directed acyclic graph (DAG). Step 2 applies graph-theoretic rules to the DAG to identify valid deconfounding variables. Additionally, it provides strategies for unmeasured variables, including discussion of their theoretical and practical strengths and limitations. Step 3 integrates the causal justification with empirical associations, ensuring that only statistically relevant confounders are adjusted for. We illustrate the framework's practical application using a UK Biobank-based brain-behaviour prediction example and demonstrate the substantial impact of confounding on predictive models – underscoring the necessity of proper deconfounding. Despite the popularity of linear feature residualization, its reliance on linear assumptions and adjustment of only features (or target) limits its effectiveness. As a potential solution, we introduce double machine learning, originally developped for causal inference, and discuss its adaptability to associative SML. Importantly, causally informed deconfounded SML models should not be causally interpreted without further justifications. Nevertheless, they are essential for producing robust, generalizable, and neurobiomedically meaningful predictive insights.

Keywords: machine learning, (de-)confounding, causality, DAG

1. Introduction

1.1. Predictive analytics are a useful tool in neurobiomedicine if models are unbiased

Machine learning (ML) and artificial intelligence (AI) offer transformative potential in neurobiomedical research and deployment. Using large, high-dimensional and oftentimes observational datasets, ML enables the development of predictive models for identifying biomarkers and supporting diagnosis, prognosis and treatment decisions [1–3]. Predictive models thereby serve two main purposes: (1) supporting scientific discovery by uncovering neurobiological mechansims and (2) advancing precision medicine through clinical decision-making tools (e.g. [4]).

Both goals require reliable models that generalize across settings. Scientific models (case 1) aim to answer scientific questions such as identifying brain patterns linked to psychiatric conditions (e,g. depression or schizophrenia [5,6]). Achieving such a deeper understanding of neurobiomedical mechanisms requires reliable and generalisable insights. Clinical (tools) models (case 2) must provide generalizable predictions for consistent usability across hospitals and patient populations, akin to standardized diagnostic tests. However, in both cases, the conventional emphasis on merely maximizing model accuracy, can cause models to fail when applied to new conditions [7,8], e.g. due to data distribution shifts [9] or covariate shifts [10,11] if the high accuracy was based on models overfitting to specific datasets. Problematically, such failure of generalization is often accompanied by unreliability of predictions [12–15]. A sole focus on accuracy maximisation hence risks overlooking of biological and clinical meaningfulness of models.

One important but often underappreciated source of poor generalization and unreliable predictions – beyond issues such as small sample sizes, poor regularization or model misspecification-is bias. Biased models base their predictions on spurious associations rather than genuine biological relationships (**Figure 1**a), hindering generalizability to new data. A key contributor to biased supervised machine learning (SML) models is confounding, where a variable influences both the input features and target outcomes (**Box 1**; confounder bias, Simpson's Paradox). For example, in psychiatric research, comorbid conditions (e.g. anxiety, substance use disorder), age or medication use can influence both brain imaging features and psychiatric diagnoses, introducing misleading associations and making it difficult to disentangle disorder-specific neural signatures from overlapping, yet distinct, effects. For instance, a model might falsely attribute structural brain changes to schizophrenia when, these changes are actually (partially) driven by aging or long-term medication use. In essence, biased models incorrectly attribute effects such as a psychiatric condition to features (e.g. brain measures) while the effects are actually due to another variable (the confounder), which limits both generalizability and biological insights of a model.

Complex variable interdependencies in neurobiological research make it challenging to identify which third variables (those that are neither features nor targets) qualify as confounders (and which do not). In many biomedical disciplines it is common to correct for a conventionally established set of confounders (e.g. demographics, lifestyle factors etc.), without transparent and systematic justification [16–18]. When justifications are provided, they often rest solely on report of statistical associations between potential confounders and feature(s) (X) and/or target (Y) [18–20]. However, neither no justification nor justification through correlative patterns is enough. Instead, effective confounder identification requires understanding the causal roles of third variables in the context of a specific research question.

1.2. Causal justification is required for proper deconfounding

Correlation-based definitions alone are not enough because different types of third variables Z, namely confounders, colliders and mediators (**Box 1)** could produce the same correlation between Z

and both X and Y [18]. This is problematic because only confounders should be controlled for, while making sure to not correct for colliders as this would inadvertently introduce bias (collider bias (e.g. [21,22]), Berksons's paradox [23], **Box 1**). This implies that adjusting for a third variable is not always right or not adjusting for it is always wrong. Rather, the decision depends on the process that generated the data, which is what we are ultimately interested in to achieve both, better transportability and generalizability of models and to better understand biological mechanisms. Consequently, arbitrary or standard adjustments risk introducing rather than removing bias. While correlations do not differ between types of third variables, directionalities i.e. causalities do differ (**Box 1**).

Directed acyclic graphs (DAG, **Box 2**) offer a principled way to clarify variable roles and systematize relationships by encoding causal assumptions (directionalities) around the relationship of interest. Additionally, they enable transparent communication of assumptions and are hence critical for building unbiased SML models [24,25]. Constructing such DAGs requires domain expertise and literature-based causal reasoning. As first objective, here, we will illustrate, how building such a DAG can be achieved in the context of biological SML tasks, by leveraging a typically applied bottom-up strategy from classical statistical causal inference to foster unbiased meaningful model development.

1.3. Challenge of unmeasured confounders in neurobiomedical observational data

Even with a principled, DAG-informed approach to confounder selection, a critical challenge remains: relevant confounders may be unmeasured or entirely unobservable. This is especially common in neurobiomedical contexts, where important biological constructs (e.g. hormone levels, early-life adversity, genetic liabilities) may be latent, unrecorded, or infeasible to measure. In such cases, standard confounder selection strategies, such as the so-called backdoor adjustment (see chapter 2, step 2 and **Box 3**), fall short, and alternative strategies are necessary. To tackle this issue, here as second objective we will introduce and discuss how a range of established methods from the causal inference literature can be applied to the SML context to handle unmeasured confounding. We neither claim novelty nor completeness of presented methods, but aim to build bridges between disciplines by suggesting and discussing integrability and feasibility of tools for debiasing of neurobiomedical SML models.

In summary, unaddressed confounding or improper handling of other third variables can lead to biased SML models based on spurious associations, limiting their generalizability and interpretability – issues that undermine model utility for clinical applications and mechanistic insights. Correlation-based confounder selection, though commonly used, is inadequate. Because confounding is inherently causal, we here advocate for the integration of causal reasoning into the confounder selection process for SML workflows. Importantly, our goal is not to estimate treatment-outcome effects, generate counterfactual data or build causal discovery graphs (identifying the correct DAG from data) such as pursuit in traditional causal inference frameworks and causal ML approaches. Rather, we aim to make causal inference principles accessible and practically useful for researchers in neurobiomedical SML. Concretely, we propose an easy-to-follow stepwise framework for identification of a suitable set of third variables to adjust for, and offer practical guidance for unobserved or unmeasured confounders (chapter 2). Additionally, we discuss the limitations of post-hoc linear (feature) residualization and explore the potential of adopting Double Machine Learning (DML) as an alternative confounder adjustment strategy (chapter 3). While deconfounded ML can support unbiased models, it does not equate to causal inference (chapter 4). By making causal inference principles accessible to researchers working in neurobiomedical SML we aim to link statistical prediction with causal reasoning. This offers a structured way to remove bias so that confounder adjustment enhances rather than hinders model validity, reliability and generalizability.

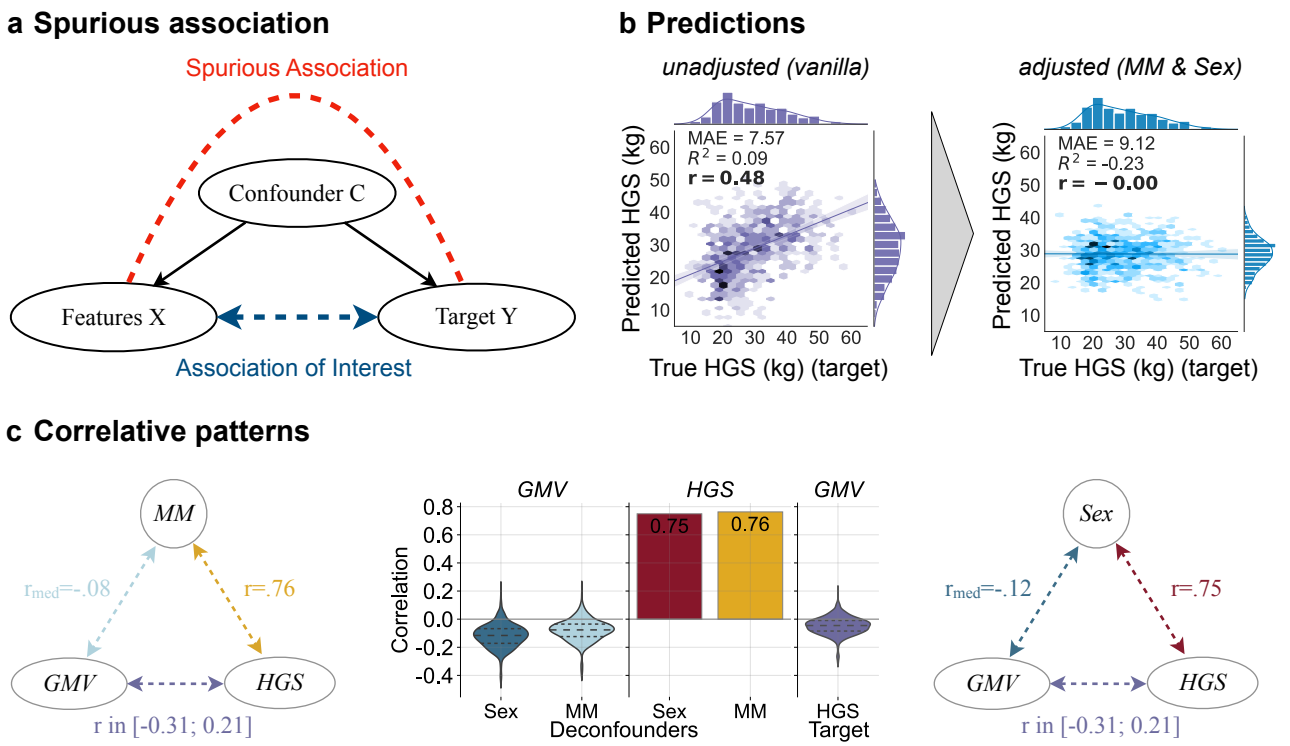


Figure 1. Illustration of concepts and statistical relationships of variables involved in the illustrative prediction of hand grip strength (HGS) from gray matter volume (GMV). **a**. When investigation the relationships between features X and a target Y, a confounding variable C can introduce a spurious association, biasing the actual association of interest as it influences both the features X and the target Y (inspired by [26]). **b.** Prediction of HGS from 1088 parcellated cortical, subcortical and cerebellar GMV features in N=3620 healthy subjects from the UK Biobank using a linear support vector regression (SVR) not adjusted for confounding (left) and adjusted for the identified deconfounders muscle mass (MM) and sex (linear feature regression) (right). **c.** Pairwise statistical association between features (GMV), target (HGS) and both deconfounders (MM, sex). GMV refers to a feature vector with 1088 features so that correlations with the deconfounders are indicated as median value (left and right plot, blue arrows) and correlation with the target as range (purple arrows). Violin plots (middle plot) show the respective distributions of parcel-wise feature associations with the deconfounders (blue) and the target (purple).

2. 3-step framework for confounder selection and adjustment

The core mechanism for unbiasing supervised ML models is through the identification of and adjustment for a correct set of deconfounders. While confounders are all variables that confound the X-Y relationship, deconfounders are a sufficient subset of confounders whose adjustment blocks non-causal (confounding) paths between input features (X) and the target (Y) and enables unbiased prediction (**Box 3**). Identification of a correct set of deconfounders requires a causal analysis around the relationship of interest (X-Y) to identify different possibilities for confounder adjustment (**Figure 2**, step 1). Additionally, unbiasing may require strategies to handle cases of unobserved confounders (once identified), an ubiquitous problem when using observational data as often the case in neurobiomedical supervised ML (**Figure 2**, step 2). Once all adjustment variables have been identified the statistical relevance of variables must be confirmed before adjusting the supervised ML model (**Figure 2**, step 3). In the following we will detail each step and discuss benefits and challenges in the context of deconfounding a supervised ML model. The fundament for all steps is built by a causal analysis which is summarised in a directed acyclic graph (DAG) (**Box 2**). As this analysis relies on domain knowledge about the process that generates the observational data all steps are

research-question dependent, i.e. there is no one-size-fits-all solution. We therefore exemplify each theoretical step with a real-world example of a supervised prediction.

Concretely, the SML example is the out of sample prediction of the target Hand Grip Strength (HGS) from T1w-MRI derived Grey Matter Volume (GMV) features in the UK Biobank (UKB) [27] as a large observational dataset (for methods see supplementary materials). In this example, a simple linear support vector regression (SVR) prediction model (see methods) without confounder consideration (*vanilla* model) can lead to a decent prediction as indicated by a correlation between true and predicted HGS of r=0.48 (**Figure 1**b, left). However, the following outlined stepwise approach for correct deconfounding will clarify that this model is biased, i.e. the decent prediction is biased by confounding signals.

2.1. Step 1 – The causal analysis

Conducting a causal analysis around the relationship of interest between input features (X) and the target variable (Y) is the cornerstone for all subsequent steps. This analysis is formalized using a DAG, which serves as a structured representation of the assumed causal relationships among all relevant variables. The DAG allows to differentiate between various types of third variables (e.g. confounders, mediators, colliders) and thereby helps to identify a suitable set of variables to adjust for to block confounding paths between X and Y.

We here suggest a bottom-up strategy to guide the causal analysis and determine influential factors on X and Y. This begins by asking about known and conceivable causes of the target Y. Starting with Y, additional variables are iteratively added based on their potential causal influence – either on Y or on other already included variables, until the network of relevant relationships is mapped. Constructing valid directed edges (arrows) in the DAG requires domain knowledge and literature justification to encode both empirically established and theoretically plausible cause-effect relationships [28]. While the process is inherently subjective, its strength lies in making modelling assumptions explicit and transparent, in contrast to arbitrary or purely correlation-based confounder selection. In the GMV-HGS example, the DAG might start with *lower arm/upper body muscle mass* as established physiological cause of HGS (muscle mass → HGS) and sex-specific influences on HGS independent of muscle-mass e.g. through sex differences in muscle function. Additionally, the GMV is added as conceivable cause of HGS as GMV-HGS is the hypothesized predictive relationship to be unbiased[1]. In the next iteration, known or conceivable causes of *muscle mass* could be *sex hormones*, *eating behaviour*, *strength training*, *age* etc.. The GMV features are influenced by *TIV*, age, *sex hormones* and further - potentially unmeasurable or unobserved - environmental and behavioural factors. Iterating this process builds a DAG (**Figure 3**), where the bottom-up approach allows to systematically disentangle the complex causal structure of intertwined biological, environmental, and behavioural factors typical for neurobiomedical data.

A key challenge in confounder selection for predictive modelling is knowing when and if all confounders were identified. This is where the concept of deconfounders becomes useful: Rather than attempting to include *all* confounders, the goal is to adjust for a sufficient subset that blocks all so-called backdoor paths between X and Y (see **Box 3** and step 2). The bottom-up construction process helps determine when adding further variables does not yield meaningful gains in bias reduction. For example, further specification of an unmeasured variable U_2 in **Figure 3** would not change the deconfounding status of the GMV-HGS relationship. In step 2 we will discuss existing options to identify a sufficient set of deconfounders.

[1] As we leverage tools from causal inference, we will use terminology such as "X effects/causes Y" and notation such as X→Y, but we do not directly imply that the deconfounded SML model allows for causal claims. We will discuss later in the paper where debiased SML can be positioned w.r.t. causal inference.

While the DAG relies on established causal relations, oftentimes not all links are well-known (yet). In cases where domain knowledge is incomplete or even contradictory, the DAG must rely on ambiguous cause-effect assumptions, so that multiple plausible DAGs may exist for the same research question. Importantly, causal assumptions encoded in a DAG cannot be empirically verified using observational data alone, and the bias from incorrect assumptions does not vanish with large sample sizes [29]. To address this well-known challenge, the field of causal graph discovery offers methods for learning DAGs from data (e.g. DAG-GNN, CAM, NOTEARS, LiNGAM, GES etc.), some of which have been applied to uncover causal structures in areas such as Alzheimer's disease biomarkers [30]. However, such approaches are often (computationally) comprehensive research projects and therefore not feasible as part of another project that targets the out-of-sample prediction of a feature-target relationship rather than modelling the causal assumptions around this relationship.

Despite its limitations, a hypothesis- and literature-based DAG remains a powerful tool. Through clear justification it enables formalization and transparent communication of assumptions, strengthening replicability of the causal reasoning underpinning a model. Additionally, the DAG provides a basis for interpreting model outputs contingent upon the used set of deconfounding variables, based on the assumptions the DAG advertises. Perhaps most importantly, the causal analysis forces researchers to precisely *think* about and critically engage with the question the model is about to answer - an aspect sometimes falling short in pure data-driven approaches.

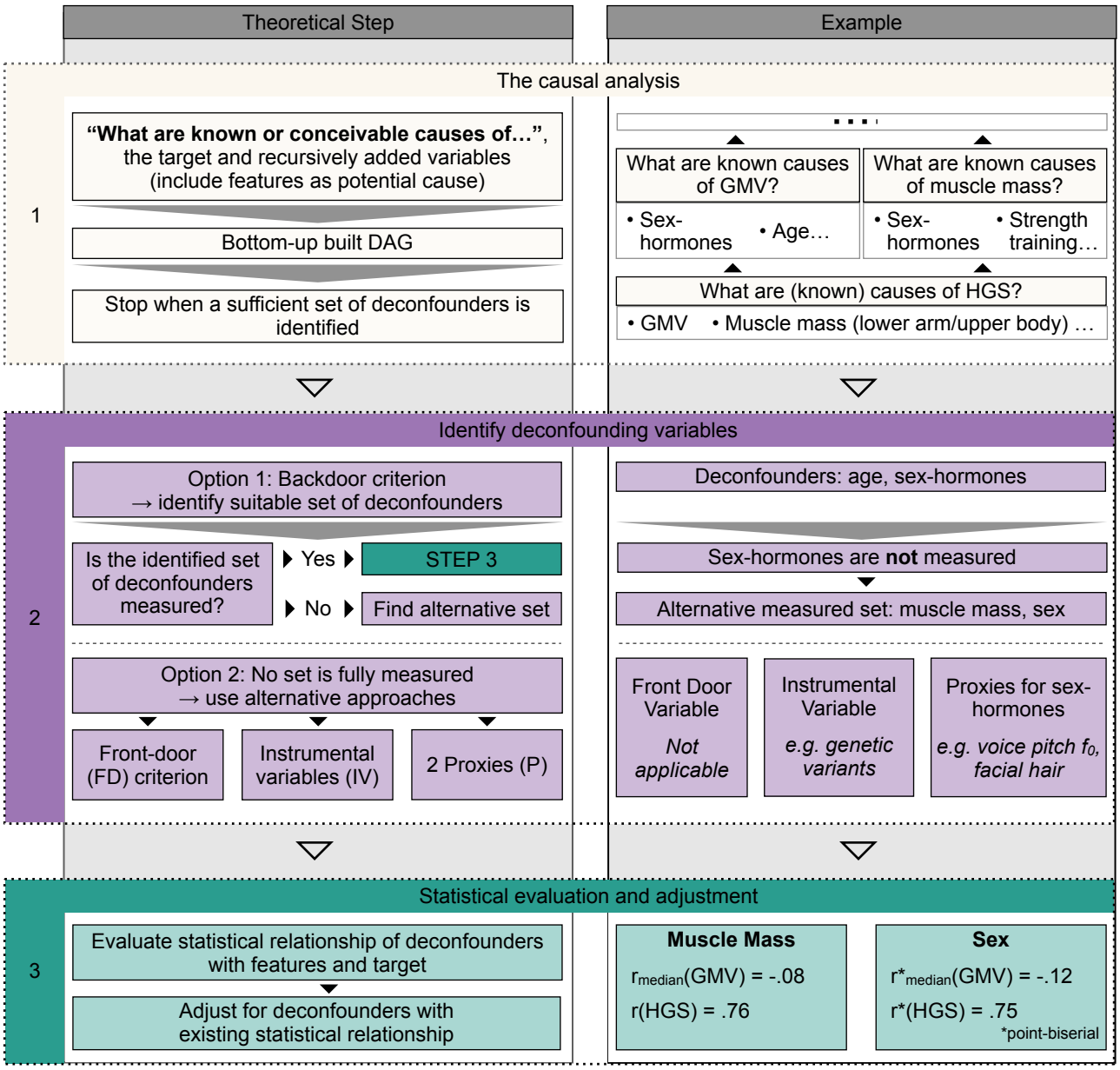


Figure 2. 3-step framework for confounder selection with left panel describing the theoretical step and the right panel illustrating it using the example of predicting hand grip strength (HGS) from parcellated gray matter volume (GMV). **Step 1** (beige) involves the causal analysis around the research question of interest, illustrating bottom-up building of the DAG. A sufficient set of deconfounders becomes clear from performing step 2 (chapter 2.2). **Step 2** (purple) involves different possibilities for identifying a proper set of deconfounding variables from the DAG built in step 1. The backdoor criterion is the easiest applicable approach (chapter 2.2.1), but in case of unmeasured deconfounders usage of the frontdoor criterium, instrumental variables or two proxies can serve as alternatives (chapter 2.2.2). **Step 3** covers the statistical evaluation of the identified deconfounders with the features and the target as well as the statistical adjustment of the model. In the example prediction, muscle mass was correlated with the 1088 GMV parcels in median by r_{median}=-0.08 and with HGS by r=.76 and sex was correlated (point-biserial correlation) with GMV by r_{median}=-0.12 and with HGS by r=.75 (**Figure 1**).

2.2. Step 2 – Identifying a suitable set of deconfounders and options in the case of unobserved deconfounders

2.2.1. Identifying a suitable set of deconfounders based on the backdoor criterion

Unlike purely correlative approaches, causal analysis as formalized in a DAG distinguishes confounding pathways from colliders and mediators (**Box 1**), enabling principled deconfounding strategies (**Box 3**). The most direct method for identifying a suitable set of deconfounders is based on the backdoor criterion [25]. A set of variables Z satisfies the backdoor criterion relative to the relationship X-Y if (1) no variable in Z is a descendant of X, and (2) Z (all variables in the set) blocks all backdoor paths from X to Y (**Box 3**). Graphically, this means that variables with arrows pointing into X in the DAG qualify as valid adjustment variables Z, which when conditioned on, block non-causal pathways (flow of information) between X and Y. Such deconfounder sets can be identified manually by following graph rules in the DAG or automatized using dedicated tools (e.g. DAGitty [28] or CausalFusion (https://causalfusion.net)).

In the GMV-HGS example DAG (**Figure 3**), several confounding (backdoor) paths exist between GMV and HGS. One valid deconfounder set includes *sex-hormone levels* and *age.* Adjusting for this set would block all backdoor paths and hence debias the prediction of HGS from GMV (**Figure 3**, green). In practice, however, identified deconfounders may not be measured or unobservable. For example, in the GMV-HGS prediction, although *sex-hormone levels* are included in the UKB, their measurement occurred on average 14.79 years prior to the brain imaging (GMV) and HGS assessment (session 0 vs. session 2). Given this substantial temporal gap, using *sex-hormone levels* would not be valid for adjustments.

The first, and easiest way to overcome the problem of an unmeasured (set of) deconfounder(s) is by still applying the backdoor criterion but trying to find alternative sets of deconfounders (alternative routes) that would block all non-causal pathways. In the GMV-HGS example such an alternative set could be *sex* and *muscle-mass* (**Figure 3**, purple), both of which are measured in the UKB for session 2. If such alternatives exist (as in our example) one can proceed to step 3 of the process. However, there is a multitude of cases which lack any alternative measured deconfounder sets that satisfy the backdoor criterion. The next sections therefore elaborate on three alternative strategies: front-door adjustment, instrumental variables and proxies.

2.2.2. Alternatives in the case of unobserved deconfounders

2.2.2.1. Front door criterion

One alternative if deconfounders identified through the backdoor criterion are unavailable is front-door adjustment [25]. It requires an intermediate variable F that meets three conditions: a) F intercepts all direct paths from X to Y (i.e., X → F → Y, b) there is no backdoor path from X to F and c) all backdoor paths from F to Y are blocked by X (**Box 3**). If these conditions are satisfied, the unbiased relationship between X and Y can be estimated by combining the estimate of effect X→F and of F→Y, circumventing the unobserved variable Z. In SML, this can be operationalised through two-stage models: the first model predicts F from X and the second predicts Y from predicted F.

In practice, however, finding such a variable F is especially challenging in neurobiomedical contexts. For example, in the GMV-HGS context, F would need to be causally affected exclusively by GMV and independently affect HGS, without being subject to the same confounding structure—this is typically hard to guarantee in neurobiomedical settings.

2.2.2.2. *Instrumental variables (IVs)*

Another frequently used alternative in the presence of unmeasured confounders are instrumental variables (IVs) [31]. A third variable V qualifies as IV if it satisfies three assumptions: (a) independence: The unmeasured confounder Z and V are independent (no arrow) (in short $V \perp Z$); (b) relevance: V causes X ($V \rightarrow X$); (c) exclusion restriction: V affects Y only through X ($V \rightarrow X \rightarrow Y$) (no direct causal connection $V \rightarrow Y$) (**Box 1**). Conceptually, IVs can be understood as mimicking randomization or simulating an experimental intervention. They thereby provide variation in X that is orthogonal (independent) to the confounding structure. As there are no confounders of the relation between V and Y, any observed association must be causal. Moreover, since V affects Y only via X, V allows to isolate the variation in X free of confounding, enabling causal estimation even when Z is unobserved. In SML practice, IVs are used in a two-stage process, where in the first stage V is used to predict X and this predicted X is used to predict Y. For example, in our neurobiological application, genetic variants (SNPs) identified in genome-wide association studies (GWAS) could serve as potential IV if there were a genetic variant known to affect GMV but assumed to not directly affect HGS.

While standard IV, as applied for causal inference, assumes linear relationships between V and X (e.g. modelling the first stage as $X=\pi V+\epsilon$), the particular strength of transporting this approach to the SML context is that using any kind of ML model for the first stage prediction can better capture complex, nonlinear and interactive relationships between V an X. This is especially suitable for the oftentimes high-dimensional setup of SML with potentially complex instruments and features. Therefore, ultimately, the usefulness of this approach relies on the strength of the IV V and on how well the features X are predicted from V: the better V and the better the prediction of X, the better and less biased the final estimation of Y (in which we are interested) is.

However, IV estimation comes with trade-offs. Most notably, it can exhibit higher variance compared to direct confounder adjustment methods, especially when the IV is weak (i.e. V only weakly predicts X). Weak instruments introduce noise in the first-stage prediction, leading to noisy and unstable second-stage estimates [29]. This resembles a classic bias-variance trade-off: while strong and valid IVs can eliminate systematic bias from unmeasured confounding, especially weak IVs can lead to high variance estimates. This issue can be exacerbated in neurobiomedical contexts, where identifying strong and valid IVs can be particularly challenging. Meeting both, the independence and the exclusion restriction assumption can be difficult given the biological complexity and interconnectedness of (neuro)biological systems, such as the brain and the presence of systemic and multiscale brain-body interactions. For example, an IV used to predict GMV might also directly affect HGS. Moreover, the multivariate nature of brain features can make it difficult to identify true paths of influence as the IV could affect different brain regions differently. Consequently, the effectiveness of the IV approach in the case of unmeasured confounders depends on the strength and validity of IVs which are difficult to find for multidimensional, multicollinear and causally intertwined data, such as neurobiological data.

It is important to highlight here once again that while in treatment effect estimation IVs are used to identify the causal effect of a treatment (e.g. drug assignment) on an outcome in a population, the goal in supervised ML deconfounding is to train a predictive model that avoids spurious associations due to confounding - improving generalizability and biological meaningfulness of models. Thus, while the statistical machinery is similar, the objectives diverge: unbiased function approximation in ML vs. estimating e.g. average treatment effects in causal inference.

2.2.2.3. *Two proxies*

Another strategy for debiasing SML models arising from unmeasured deconfounders - such as *sex-hormone levels* in the prediction of HGS from GMV - is the use of proxy variables. A proxy P is a third variable that is causally influenced by a variable Z, here e.g. the unmeasured deconfounder, but

does not itself directly affect the target or the feature(s) (**Box 1**). For instance, in the GMV-HGS prediction, potential proxies for *sex-hormone levels* include *voice pitch (F_0)*, *waist-to-hip-ratio (WHR)*, *facial hair* or *ratio of index finger to ring finger length (2D:4D)* - all of which are known to be biologically modulated by testosterone levels (**Figure 3**, orange). Miao et al. [32] formalize the conditions under which it is possible to use at least two proxies P1 and P2 to nonparametrically recover the influence of the unmeasured deconfounder Z. This method requires three key assumptions:

1) Conditional Independence

The proxies must be statistically independent of each other when the latent deconfounder Z is held constant.

- Formally: $P(P1, P2 \mid Z) = P(P1 \mid Z) \cdot P(P2 \mid Z)$
- Intuition: Any statistical association between, for example, *voice pitch* and *facial hair* should be explainable solely by shared dependence on *sex-hormone levels*. This assumption is biologically plausible since these traits arise from distinct physiological pathways—laryngeal development vs. follicular activation – with no evidence of a direct causal link between the two.

2) Relevance Condition

Each proxy must provide nontrivial information about the unobserved deconfounder Z.

- Formally: $P(P|Z=z_1) \neq P(P|Z=z_2)$ for some $z_1 \neq z_2$ (The conditional distribution P(P|Z) must change as Z changes)
- Intuition: Changes in Z must induce systematic and detectable variation in the probability distribution of each proxy. For example, testosterone levels during puberty lower voice pitch by growth and thickening of vocal folds (Harries et al., 1998) and stimulate facial hair growth via androgen receptor activation (Randall, 2008), i.e. both proxies respond systematically to changes in the deconfounder. In contrast, a variable like *2D:4D*, while causally downstream, it is only weakly associated with prenatal testosterone exposure [33]. This weak and noisy variation may thus make it fail to satisfy the relevance condition.

3) **Rank Condition** (generalization of earlier frameworks [34] that required functional links between proxies and Z)

The joint distribution of the proxies varies sufficiently across values of Z to allow for its reconstruction up to a transformation.

- Formally: The covariance matrix (conditional expectation operator) of the proxies given Z must be of full column rank. This guarantees that the mapping $Z \rightarrow (P1, P2)$ is injective, so that different values of Z produce distinguishable configurations of proxy values.
- Intuition: Both proxies are independent noisy reflections of Z, so that their combination uniquely pins down the value of Z. For example, *voice pitch* and *facial hair* reflect testosterone activity via partially distinct pathways, i.e. anatomical and functional independence of the larynx and facial hair follicles. This suggests that their joint variation provides non-redundant information about *sex-hormone levels*, making the rank condition plausibly satisfied.

A central challenge of proxy-based deconfounding is that the three conditions cannot be empirically verified in a data-driven way, since the deconfounder Z is unobserved. For example, conditional independence tests between *voice pitch* and *facial hair* (e.g., partial correlations, conditional mutual information) require conditioning on the unmeasured *sex-hormone levels*. Using a cause of the unobserved deconfounder as a surrogate variable (e.g. *hormone intake*) (**Figure 3**) might appear as a good strategy for data-driven testing of proxy quality, but is conceptually flawed. For example, *hormone intake* is a parent of *hormone levels*, not a noisy measure thereof, and thus violates the structure required of proxies. Even if *hormone intake* were a proxy, it would not be valid to

empirically assess the conditional independence of two proxies based on another proxy. Similarly, the relevance and rank conditions depend on the structure of the mapping from Z→P, which requires direct observation of Z. Consequently, the credibility of proxy-based deconfounding must rest on domain knowledge and theoretical justification. This includes extending the DAG from step 1 to incorporate candidate proxies by justifying each edge and the plausibility of the three conditions based on known biology. For instance, one can argue that *voice pitch* and *facial hair* fulfill all three conditions to serve as proxies for *sex-hormone level* given established physiological mechanisms, whereas *2D:4D* may fail the relevance criterion due to weak variation.

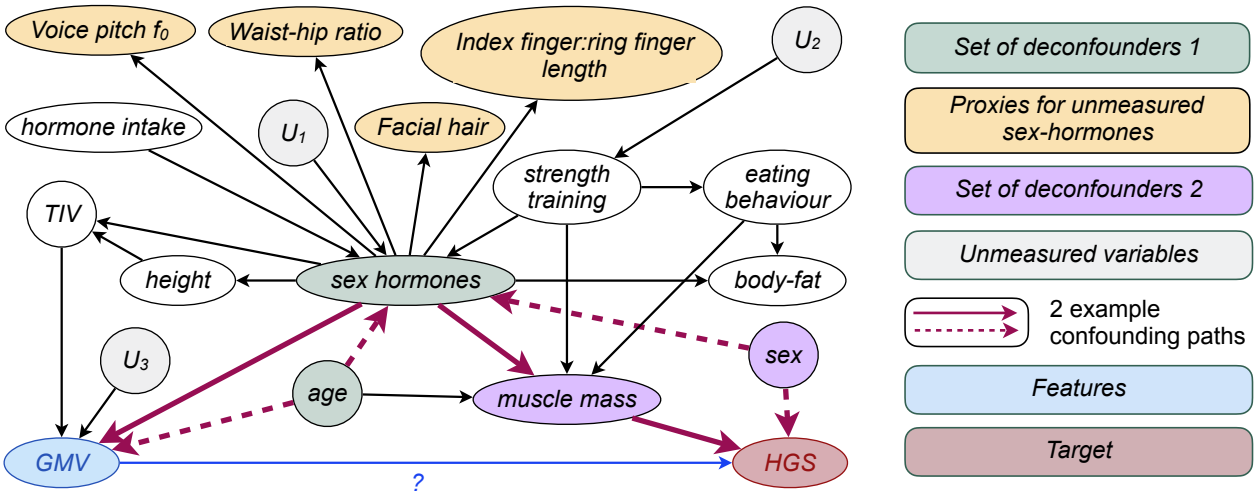


Figure 3. Example DAG for deconfounder identification for the GMV-HGS prediction example resulting from causal analysis following the framework outlined in Figure 2. Two example confounding paths are highlighted in red (not comprehensive). Considering all confounding pathways, one potential minimal set of adjustment variables (deconfounders) is *sex hormones* and *age* (set of deconfounders 1). As *sex hormones* are unmeasured in the exemplary data sample, an alternative sufficient set of deconfounders would be *muscle mass* and *sex* (purple). Both sets qualify to block all non-causal pathways following the backdoor criterion and are hence a sufficient subset of variables to debias the predictive GMV-HGS model.

2.2.2.4. Broader reflections on handling unmeasured confounders

Despite their appeal, the success of strategies such as front door-adjustment, IVs and proxies hinges on the availability of high-quality variables. For example the IV approach requires strong instruments V to avoid high variance and imprecise estimates. Proxy approaches in contrast risk bias if the proxies fail to adequately capture the unmeasured deconfounder's information.

Beyond variable quality, the validity of these strategies depends entirely on correct specification of the underlying causal structure. Mischaracterizing causal roles or dependencies can lead to biased estimates, undermining the very purpose of deconfounding. This emphasizes a central point: no deconfounding strategy can rescue missspecifications of causal relationships. If the DAG fails to reflect the data-generating process, all derived adjustments risk introducing rather than mitigating bias.

As a result, in practice, it is often advisable to favor the simplest deconfounding approach, i.e. where the least assumptions must be fulfilled, making the backdoor criterion an attractive choice, if the respective variables are available. The list of discussed alternative strategies is not exhaustive but reflects approaches that are (i) grounded in causal theory, (ii) transparent in their assumptions, and (iii) reasonably implementable in real-world datasets.

2.3. Step 3 – Statistical evaluation and adjustment

2.3.1. Statistical evaluation

In the GMV-HGS prediction example, the unmeasured deconfounder *sex-hormone levels* can be circumvented by using an alternative set of backdoor-justified deconfounders, namely *muscle-mass* and *sex* instead of *age* and *sex-hormone levels* (**Figure 3**, orange). Both alternatives are availably measured in the UKB at the same time point as HGS and GMV. and causally relevant as per the DAG (step 1).

In addition to causal relevance, deconfounders need to be statistically associated with both feature(s) and the target as causal relationships only become actionable when they manifest in the data [35]. We nonetheless put the causal analysis first because this allows for a bottom-up identification of deconfounders (as classically pursuit in statistics) in contrast to a top-down pre-definition of confounders which can create insecurities about what variables to include as confounders (see step 1).

The threshold for what qualifies as sufficient statistical association depends on context, comparable to no hard threshold in for example nul hypothesis testing (convention of $p<.05$ or $p<.01$). In our example, *sex* correlates[2] with HGS at $r=.75$ and with GMV at $r_{median}=-.12$; *muscle mass* correlates with HGS at $r=.76$ and with GMV at $r_{median}=-.08$ (**Figure 1c**). The relatively low median correlations between the deconfounders and GMV arise from the multi-dimensionality of the GMV features (1088 brain regions), many of which show heterogenous associations – positive in some regions, negative in others - resulting in a median near zero. This is important to note as statistically unrelated deconfounders should not be adjusted for. In the best case such adjustment would be irrelevant but in the worst case it can introduce bias by leaking information from the deconfounder into the feature or target in the adjustment process [36]. Lastly, there also must be a statistical association between GMV (X) and HGS (Y) (**Figure 1**c) to assure predictability (even though potentially biased) and thereby debiasing (deconfounding) meaningful.

2.3.2. Deconfounding using linear residualization

Once deconfounders are identified through both causal justification and empirical association (in our case: *sex* and *muscle mass*), models can be adjusted. In SML several (post-hoc) confounder mitigation strategies exist. One of the most established approaches is linear residualization, where confounder information is mass-univariately linearly regressed out of features or the target (e.g. [37]) (residualization). In our example, we residualized the GMV features for the identified deconfounders *muscle mass* (operationalised as lean mass) and *sex*. Using a linear SVR (L2) model trained on residualized features, we observed no correlation between true and predicted HGS ($r=0.00$, $R^2<0$[3]; **Figure 1**b, right) (for methods see supplementary materials).

This constrasts strongly with the unadjusted model, which yielded $r=.48$ ($R^2=.09$) between true and predicted HGS (section 2). Keeping in mind the combination of a simple linear SVR with linear confounder regression, the collapse in predictive performance after deconfounding implies that the earlier performance was largely driven by confounding bias. In other words, the unadjusted model did not learn meaningful biological relationships between GMV and HGS but rather exploited demographic and behavioural correlates. Such a model would likely fail in datasets with different distributions of *sex* or *muscle mass*, undermining both generalizability and scientific insights.

[2] Point-biserial correlation (see methods in supplementary materials for details).
[3] R^2 refers to the coefficient of determination as commonly used in the ML literature and not to squared correlation values.

On the other hand side, the poor performance of the unbiased model does not necessarily imply the absence of any meaningful GMV-HGS relationship. Instead, it signals that a simple linear model may be insufficient to capture more nuanced, biologically plausible patterns in the data. Further exploration using non-linear models or multimodal inputs may be required to recover valid signal under proper deconfounding constraints.

3. Limitations of linear (feature) residualization and alternative approaches

3.1. Limitations and strength of linear feature residualization

Linear residualization of features is commonly used in neurobiomedical predictive modelling for confounder adjustment. Despite its popularity, it has two main limitations relevant to this work, namely (1) the assumption of a parametric linear relationship between confounders and each feature and/or the target, and (2) the adjustment typically being applied to features or target, not both.

First, linear residualization effectively only removes linear confounding effects. While this may be sufficient with linear predictive models, it becomes problematic with non-linear prediction algorithms. These can leverage residual non-linear confounding information, resulting in biased predictions despite linear adjustment.

Second, standard practice in SML often involves residualizing either the features or the target, but rarely both. Typically, only features are adjusted to preserve interpretability of the target, which often represents the neurobiological entity of interest (e.g. diseases status, cognitive score). In the following, we elaborate why one-sided confounder adjustment can be of concern.

3.1.1. On the potential benefits of feature and target adjustment – signal contribution perspective

Linear residualization of only the features X implicitely assumes that all variation in the target Y that is associated with confounders Z is fully captured by X. In practice, this assumption rarely holds. While residualizing X removes the confounders' variance from X ("clean" features), it leaves confounder-related variance in Y. Thus, Y is still partially influenced by Z so that the statistical relationship between X and Y becomes misaligned: the model attempts to predict Y from features that no longer carry the Z-related information, while Y still contains it. This misalignment can have two main consequences. First, the one-sided adjusted model may remain partially biased. Since confounding variance remains in Y, the model may over- or under-estimate relationships between X and Y, depending on the structure of Z's effect. Second, predictive accuracy may be reduced because the remaining information in residualized X cannot explain the signal in Y linked to Z. In effect, while there is no formal guarantee that residualizing both X and Y is benefitial, it could improve statistical alignment between features and target, supporting potentially less biased and more accurate predictions.

3.1.2. Necessity of feature and target adjustment - causal perspective

As confounding is inherently a causal concept, causality-based reasoning offers a deeper lens to understand the limitations of only feature or only target residualization. Previously, we introduced the backdoor criterion as a method solely for confounder identification based on a DAG. But it does more than this: Being rooted in causal literature, it provides a formal basis to estimate the interventional causal effect of X on Y – expressed as $P(Y|do(X))$ (Box 2).

This interventional distribution represents the probability of observing Y=y when actively intervening to set X=x in the population. In contrast, the potentially confounded observational distribution $P(Y|X)$ reflects the probability of observing Y=y among individuals for whom X=x is observed, regardless

of other influencing factors. For instance, $P(HGS=29kg|GMV_{anterior\_globus\_pallidus}=300mm^3)$[4] might conflate effects of sex or body composition, whereas $P(HGS=29kg|do(GMV_{anterior\_globus\_pallidus}=300mm^3))$ isolates the interventional causal influence of GMV (**Box 2**).

According to the backdoor adjustment formula [38], for variables Z that satisfy the backdoor criterion, the interventional distribution can be computed as:

$$P(Y|do(X)) = \sum_z P(Y \mid X = x, Z = z)P(Z = z)$$

Here:

- $P(Y|X, Z=z)$ is the conditional observational distribution, i.e. how likely it is to observe a certain value Y, given that a certain value X=x and Z=z was observed (e.g., $P(HGS=29kg|GMV_{anterior\_globus\_pallidus}=300mm^3$, sex=female)), and
- $P(Z=z)$ is the marginal probability distribution of Z (e.g., P(sex=female) = 0.6 if 60% of the population are female)

This formula tells us that to estimate the interventional causal effect of X on Y, we must assess the conditional probabilities P(Y|X=x, Z=z) across all levels of Z (e.g. male, female), then aggregate them using the prevalence of each z in the population as weights. For example, the causal effect of a GMV of $300mm^3$ in the left anterior globus pallidus on HGS would be obtained by summation (discrete Z) or integration (continuous Z) of the weighted conditional distributions over all sexes: $P(\,Y \mid do(X = 300)\,) = \sum_{z\in\{male,female\}} P(Y \mid X = 300, Z = z)P(Z = z)$.

Evaluating each level of the confounding variable Z separately is essential. By holding Z constant, its influence on the X-Y relationship is effectively neutralized. This ensures that any observed variation in Y can be solely attributed to changes in X, comparable to the strategy of matching samples. Moreover, by aggregating across all levels of Z using a weighted sum (or integral), the approach estimates the effect of setting X=x in the entire population, rather than being limited to the subpopulation for which X=x is naturally observed. Together, by conditioning on Z, this strategy eliminates variation in the X-Y relationship from shared dependence on Z, effectively blocking information flow along the non-causal backdoor path from X to Y via $X \leftarrow Z \rightarrow Y$. As a result, any remaining association reflects the direct interventional causal influence of X on Y.

In predictive SML, we approximate this idea using residualization. Instead of adjusting additively as in the backdoor formula, residualization removes the influence of *Z* subtractively by regressing it out from both X and Y. This yields:

- $X_{resid} = \tilde{X} = X - \hat{X}(Z)$ (features orthogonal to (independent of) Z)
- $Y_{resid} = \tilde{Y} = Y - \hat{Y}(Z)$ (target orthogonal to (independent of) Z),

with $\hat{X}$ and $\hat{Y}$ being the respectively predicted values of X and Y from Z. A model is then trained to predict $\tilde{Y}$ from $\tilde{X}$. This procedure removes Z-related variation from both the features and the target, i.e. cutting off both $Z \rightarrow X \rightarrow Y$ and $Z \rightarrow Y$, so that the remaining signal in the model now reflects the component of X that explains variation in Y independent of the confounders Z. In this sense, dual residualization (adjusting both features and target) enables not just proper model debiasing, but could additionally provide the opportunity to examine causal effects of X on Y using predictive models. However, as discussed below, the conditions under which causal claims from deconfounded SML models are valid must be carefully examined.

[4] The probability of observing a HGS (Y) of 29kg (y) in people where the GMV of the left anterior globus pallidus (X) is $300mm^3$ (x).

3.1.3. Reasons linear (feature) residualization is used despite limitations

Despite well-founded arguments for residualizing both features and the target and for using non-linear models for confounder adjustment, linear feature residualization remains common, especially in fields such as brain-behaviour predictive modelling. One reason is its simplicity: assuming linear confounder-feature/target relationships limits overfitting risk, avoids hyperparameter optimisation, and makes the residualization easily integratable in SML cross validation (CV) pipelines without leakage. This makes it a practical choice for achieving debiased predictions.

Another key reason is that residualizing only features preserves interpretability of the target, which is typically the scientific or clinical interest (e.g. disease status, cognitive score, age). Residualizing the target can obscure its meaning. In such cases, quantifying confounder-target associations (see section 2, step 3) can support informed decision and transparent communication of the trade-off between confounding influence and target interpretability.

Additional reasons include established practice, limited tooling, and the sometimes implicit misconception that cofounders affect the target only via features ($Z \rightarrow X \rightarrow Y$), ignoring direct $Z \rightarrow Y$ pathways. More broadly, while linear feature residualization is an established convention, a causal approach to confounding is less widespread in certain SML communities. Finally, clear, accepted alternatives are scarce. While alternatives such as double/debiased machine learning (DML) [39] can offer more thorough bias removal, they are designed for causal inference and are not (yet) adopted for debiasing prediction models.

3.2. Double/Debiased machine learning to overcome the limitations of linear feature residualization for debiasing SML models?

Double/Debiased Machine Learning (DML) [39] is a method to estimate causal parameters, such as average treatment effects in the presence of high-dimensional confounding. It is both theoretically well established and practically implemented in various Python and R based toolboxes such as EconML (e.g. [40]) or DoubleML (e.g. [41]). While we leverage causal tools to unbias SML models, they leverage SML tools (correlative) to obtain unbiased causal parameter estimates. Even though originally developed for causal questions, DML provides valuable insights that may be repurposed to improve confounder adjustment in supervised machine learning (SML) under theoretical adaptation and practical considerations.

3.2.1. DML – the method

The DML framework targets the estimation of a causal parameter θ_0 from data $W = (Y, D, X_C)$, where Y is an outcome, D a treatment variable and X_C a potentially high-dimensional set of confounders[5]. The relationship between variables is described by two partially linear models:

(1) Treatment/Confounder-Outcome relationship:

$$Y = D \cdot \theta_0 + g_0(X_C) + U \quad (1)$$

with stochastic errors U, following $E[U|D,X_C]=0$

(2) Treatment-Confounder relationship:

$$D = m_0(X_C) + V \quad (2)$$

with stochastic error V, following $E[V|X_C]=0$

[5] We here stick with the variable naming as used by Chernozhukov et al. (2018), where they use X for confounders and not features like in the SML context. For distinction we label it here X_C. D would resemble a binary feature (0 or 1) in the SML context.

While the treatment-outcome relationship is assumed to be linearly described by the causal parameter θ_0, the nuisance functions g_0 and m_0 that model the respective relationship with the confounders can be high-dimensional, non-linear and complex, hence the suggested usage of regularised SML tools, such as LASSO or l_1-penalized neural networks, to get an estimate of g_0, i.e. $\hat{g}_0$. Regularised algorithms are useful to resolve bias-variance trade-offs in a prediction context but would lead to a biased estimate of the regression coefficient θ_0 if the prediction $\hat{g}_0$ would be directly plugged into (1). To counter this, DML introduces a cross-fitting strategy, splitting the data into an auxiliary and main subset: The nuisance functions m_0 and g_0 are learned on the auxiliary data and treatment D and outcome Y are residualized ($\tilde{D}$, $\tilde{Y}$) with their predicted values $\hat{D}$ and $\hat{Y}$ obtained using (2) and (1), respectively. $\tilde{D}$ and $\tilde{Y}$ are orthogonal (independent) to X_C and can therefore be used to get a debiased estimate θ_0 on the separate main data, ensuring so-called Neyman orthogonality. This orthogonality condition is key to robustness, as it guarantees that small errors in nuisance estimation $\hat{g}_0$ and $\hat{m}_0$ do not substantially bias the estimate θ_0.

3.2.2. Transferring the DML approach to debiasing SML models – challenges and opportunities

Directly translating DML framework into the SML context poses nontrivial methodological and practical challenges[6]. Fundamentally, the objectives differ: while DML aims to recover a treatment effect θ_0, SML seeks to build predictive models of Y that are robust to confounding.

Methodologically, an essential aspect of DML's theoretical foundation lies in its use of sample splitting into auxiliary and main datasets (or generalized to k-fold), which allows the scaled estimation error of the target parameter θ_0 (i.e.,$\sqrt{n}(\widehat{\theta_0} - \theta_0)$) to be decomposed into three terms which under certain constraints vanish asymptotically. Specifically, for the third term in this decomposition (equation 1.6 in Chernozhukov et al (2018): $\frac{1}{\sqrt{n}}\sum_{i\epsilon I} V_i(\hat{g}_0(X_i) - g_0(X_i)$, I: main sample split of size n) its probability vanishes due to the use of the sample-splitting strategy. However, this result has only been rigorously proven for the partially linear model. In contrast, SML typically seeks to estimate complex, potentially non-linear relationships between the target Y and the features X. Therefore, before DML can be validly applied in SML contexts, it would be necessary to establish that the required assumptions still hold under such general, non-linear settings.

While the features X in SML are seen as the conceptual anologous of the treatment D in DML, dimensionality roles are reversed. DML is developed under the assumption of a binary treatment D and high-dimensional confounders X_C, whereas in SML the roles are reversed: features X are high-dimensional, and confounders C are few, which can fundamentally alter the estimation landscape. Although the original DML framework allows some flexibility in dimensionality of the treatment variable (footnote 1 in Chernozhukov et al (2018)), using a high-dimensional feature matrix X as a "treatment" analogue can introduce complications. With high-dimensional features X, such as 1088 brain regions in the GMV-HGS example, modelling the confounder-feature relationship involves predicting multivariate targets because the features of the actual prediction aim become the targets of the confounder-feature modelling[7]. This presents two modelling options: (1) fitting a multi-output model with e.g. 1088 targets and only few inputs (e.g. the confounders *sex* and *muscle mass*), or (2) training e.g. 1088 separate models in a mass-univariate fashion. While the latter is common in typical linear feature residualization, more complex models require different ML strategies such as nested cross-validation and are more prone to overfitting. Crucially, if the orthogonalization step that relies on residualizing X and Y is not implemented with sufficient precision due to model misspecification

[6] In the following we will adopt the following notation: Y: outcome or target, Z: confounders, X: features in the SML sense, D: treatment in the DML sense. The features X in SML are the conceptual anologous to the treatment D in DML.
[7] Comparable to equation (2) for the treatment estimate D, but while D is binary in DML, X here is high-dimensional.

or overfitting, then the cross-fitted bias term (equation 1.6. in Chernozhukov et al (2018)) might not vanish as expected, leaving residual bias even after cross-fitting.

Adopting DML to the SML context practically introduces challenges to model training. The DML procedure only requires cross fitting for confounder modelling but uses a linear regression for treatment effect estimation. In contrast, SML relies on complex models for predicting Y from X that already require nested cross-validation. When now feature and target residualization additionally require complex ML models which each need nested cross-validation, the resulting data partitioning becomes intricate. Concretely, it requires a first split into auxiliary and main folds (as per the DML framework), and then internally within the auxiliary and main fold further splits for nested cross-validation for hyper parameter optimisation (inner loop) and estimation of generalization error (outer loop) of all nuisance models (in the auxiliary fold) as well the actual SML model (in the main fold) (**Figure 4**). This hierarchical structure not only sharply reduces the effective sample size available for training and adds substantial computational burden, it also increases susceptibility to data leakage, especially when additional preprocessing or feature selection steps are needed.

Finally, the conceptual concerns remain regarding target residualization and interpretability. In causal inference, residualizing the outcome is unproblematic because the causal estimand θ_0 is the parameter of interest. In SML, however, the original target is the quantity of interest, such as HGS in the above example, not some derivative thereof. Hence, it is problematic that residualized targets are not directly neurobiologically interpretable.

Despite these challenges, adapting DML principles to the SML setting opens promising avenues for improving confounder-adjusted predictions. First, it strengthens the rationale for residualizing both features and the target – a deliberate trade-off between complete removal of confounding information and preservation of target interpretability. Second, DML provides a mathematically well-theorised framework for employing flexible ML estimators in residualization, extending beyond simple linear regression. Third, the cross-fitting scheme central to DML highlights the need for separate data partitions to prevent overfitting of complex nuisance estimators. SML pipelines do rely on data-splitting through nested cross-validation for hyperparameter optimisation (inner loop) and out-of-sample performance generalization estimation (outer loop). However, confounder adjustment is conducted within each fold of the innermost data split rather than on independent partitions. Incorporating an additional cross-fitting step (**Figure 4**), inspired by DML, could therefore enhance SML practice by yielding out-of-sample confounder-adjusted feature residuals, ensuring robustness of confounder adjustment models to overfitting even when complex models are employed.

In summary, the DML framework provides a powerful tool for debiased causal estimation, with potential to inspire more thorough confounder adjustment in SML. However, its direct application to building debiased SML models requires further theoretical and practical adaptations.

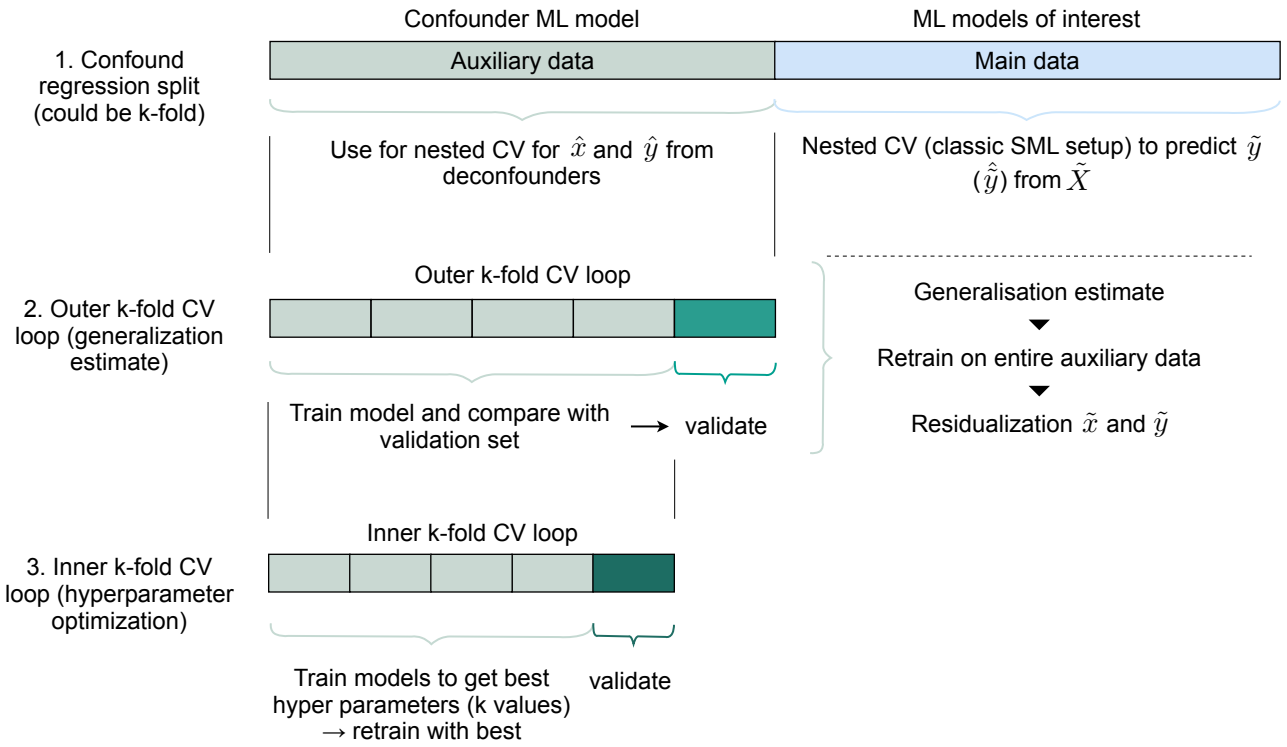


Figure 4. Theoretical cross validation scheme in case of the application of core DML principles to SML. Illustration of the hierarchical data partitioning required when combining Double/Debiased Machine Learning (DML) sample splitting with nested cross-validation (CV) for supervised machine learning (SML). First, the dataset is split into auxiliary and main folds according to the DML framework, enabling cross-fitting of nuisance models on the auxiliary data to prevent overfitting. Within each fold, further nested CV is required: an inner loop for hyperparameter optimization and an outer loop for out-of-sample performance estimation of both the nuisance models (in the auxiliary fold) and the main SML model (in the main fold). This multi-level split could support proper residualization of features (and targets) by enabling more complex models for confounder adjustment through limitating the risk of overfitting. However, it substantially reduces effective sample size, increases computational cost, especially when the auxiliary-main split is also extended to k-fold splits. Additionally, it raises the risk of data leakage if additional preprocessing or feature selection is performed.

4. Positioning of deconfounded SML between predictive modelling and causal inference

Deconfounding supervised machine learning (SML) is necessary to mitigate bias in predictive models but additionally can bring SML closer to causal inference. SML is fundamentally associative, capturing statistical patterns based on conditional, marginal, or joint distributions. However, sufficient deconfounding, e.g. by drawing inspiration from double machine learning (DML), could enable causal interpretation in SML models. Therefore, the question arises: how far can SML go beyond prediction toward interpretable causal claims?

To contextualize the question, a recap of basic causal inference concepts is helpful. Accordings to Pearl's ladder of causation, SML operates at the lowest, associational rung (e.g. P(Y|X)), whereas causal inference aims for higher rungs, namely interventions (P(Y|do(X))) or counterfactuals, which require additional structural assumptions [26,42]. Central to causal inference is treatment effect estimation, either at the individual level (ITE) or averaged across a population (ATE). Randomized control trials (RCTs) enable causal identification via randomization, ensuring exchangeability/ignorability (no unmeasured confounders). When RCTs are unfeasible frameworks such as Structural Causal Modeling (SCM) [43], Rubin's potential outcomes framework [44,45] and

Structural Equation Modeling (SEM; [46] offer alternatives. SCM, in particular, fromalizes causal assumptions about the data-generating processes via DAGs, enabling identification of confounders, mediators, and colliders.

4.1. Distinction of causal Machine Learning from debiased SML

Causal machine learning (causal ML) aims to bridge ML and causal inference, shifting focus from prediction to estimating interventional effects. For example, instead of predicting diabetes risk, causal ML might estimate how a lifestyle intervention changes diabetes risk.

As outlined by Feuerriegel et al. [47], causal ML typically involves:
(1) Problem setup: Defining the causal question quantity of interest (e.g., ITE/ATE, Conditional ATE) and relevant variables (treatment, confounders, outcome) and assessing the plausibility of assumptions;
(2) Choosing and fitting the causal ML methods (e.g. causal forest, meta-learners, Bayesian networks).
Even though the problem setup evaluates plausibility of assumptions, integrating an explicit step to model causal variable relations would further strengthen the selection of relevant variables and ensure appropriate confounder adjustment (here: chapter 2). Beyond causal ML for treatment effect estimation, newer directions include counterfactual data generation e.g. through deep learning and causal graph discovery to learn DAGs from data to complement expert-driven causal modelling.

Despite these causal ML applications, the question remains, if and how debiased SML could allow for causal insights not in the sense of estimating treatment effects but in the sense of understanding causal feature-target mechanisms. For instance, rather than associative prediction of HGS from parcellated GMV, causal SML would ask if the GMV in the left anterior globus pallidus is a causal driver of HGS, aiming for biological mechanisms rather than statistical associations.

4.2. Assumptions required for causal interpretability and positioning of debiased SML

To evaluate the potential of SML for causal insights, key assumptions from causal inference must be considered and evaluated whether they can be met in debiased SML models, particularly in neurobiological contexts.

1. **Ignorability**: All confounders must be observed and accounted for. Our DAG-guided framework and DML discussion aims for this, though high-dimensional, intertwined neurobiomedical data and methodologcial limitations make full coverage challenging.

2. **Causal Markov Assumption**: Variable must be conditionally independent of their non-descendants given their direct causes. For instance, HGS is conditionally independent of *sex hormones* given *muscle mass* (**Figure 3**). This assumption must be assumed in expert-defined DAGs, but biological systems may require or inadvertently contain (hidden) cycles, undermining causal interpretation.

3. **Positivity**: All levels of the treatment (features, e.g. low/high GMV) must occur across all confounder strata (e.g. young/old). For instance, usually low GMV is particularly associated with older age, but not observed in young adults, so it cannot be identified what would happen if a young person had low GMV, violating positivity. In ML terms, this corresponds to poor extrapolation, covariate shift or domain mismatch challenges. Even a deconfounded model cannot learn a causal relationship in regions of the feature space that are unobserved or underrepresented.

4. **Faithfulness (Absence of Coincidental Independencies)**: All observed conditional independencies (**Box 2**) in the data correspond to those implied by the structure of the causal DAG

and are not the result of numerical coincidences or cancellation effects. Faithfulness links statistical patterns to causal structure. Faithfulness would for example be violated in a setup with GMV-> HGS, physical activity-> HGS and physical activity-> GMV, if the effects of GMV-> HGS and physical activity-> HGS are equal in magnitude but opposite in direction. The net observed association between GMV and HGS could be close to zero, despite a true underlying causal pathway. This creates a conditional independence in the data that does not match the actual DAG. Sometimes unfaithfulness can be prevented by confounder correction, but generally is a bigger concept, ruling out all statistical independencies that result from exact numerical cancellations, not just dues to confounidng. Although strong and ultimately untestable, faithfulness is often regarded as a reasonable default assumption—unless there is domain-specific reason to suspect causal cancellations.

5. **Consistency:** An individual's observed outcome must equal their potential outcome under a certain treatment. In our context, if a person has a specific GMV level, their observed HGS is assumed to reflect what would occur under that same GMV level in a counterfactual scenario. Additionally, it assumes no interference between individuals, akin to the independent and identically distributed (i.i.d.) assumption in ML. Violations occur when identical treatment/feature values (e.g., GMV) arise from different biological processes – such as lifelong physical training in one person versus genetic predisposition in another – potentially leading to different outcomes (HGS) despite the same observed GMV. This illustrates a key limitation of applying debiased SML for causal insight in complex biological systems with latent heterogeneity.

4.3. How much causal claims do properly debiased SML models allow for?

Given the discussed constraints, how far can a well-deconfounded SML model support causal interpretation of the relationship between X and Y? Consider a well debiased SML model: Relevant confounders are correctly identified, features and the target are appropriately residualized (linearly and non-linearly), and predictive performance is high. Can we then conclude that X causes Y?

Not necessarily. While such a model may hint towards a causal relationship, it does not guarantee one. Violations of assumptions, such as unobserved confounding (violating ignorability), unaccounted DAG cycles (violating the causal Markov condition), or features reflecting proxies for latent biological processes (violating consistency), can still hinder a causal interpretation.

Even if all assumptions hold, interpreting the learned association P(Y|X) as the interventional distribution P(Y|do(X)) remains problematic due to ambiguity of causal direction. SML models estimate P(Y|X), which is direction-agnostic[8]. For instance, predicting blood pressure (Y) from drug dosage (X), reflects learning in causal direction (X → Y), while predicting disease status (Y) from brain structure (X) may reflect learning in anti-causal direction (Y → X) - yet in both cases, the model is trained to predict Y from X [48].

In the GMV-HGS example, predictions by a SML model may reflect GMV affecting HGS (e.g. motor cortex GMV determines strength) or HGS influencing GMV (e.g. strength training induced neuroplasticity). Both pathways are biologically plausible. Moreover, high-dimensional models, such as the exemplary one using 1088 GMV parcels, likely capture mixtures of causal, anti-causal, and non-causal patterns, reflecting the dynamic, bidirectional nature of brain-behaviour relationships. Disentangling these directions requires additional information: domain knowledge, experimental interventions, or further modelling assumptions such as additive noise models, invariance, or algorithmic asymmetries (see e.g. work from Schölkopf et al. for details).

[8] Direction-agnostic in the causal sense that it doesn't tell whether X→Y or Y→X, not in the symmetric sense, as P(Y|X) ≠ P(X|Y).

In summary, debiased SML models can offer insights consistent with causality, but not proof of it. In addition to proper deconfounding, causal assumptions must be satisfied and the model's predictive direction must align with the true data-generating process. Because SML inherently learns P(Y|X), interpreting this as P(Y|do(X)) requires a strong justification without which any causal interpretation remains suggestive rather than definitive.

5. Conclusion

Confounding is a major source of bias in neurobiomedical supervised predictive modeling, and principled confunder selection - beyond correlation- or heuristic-based justification - is essential for meaningful predictions. We propose a three-step framework for confounder selection - causal analysis via DAGs, guidance for appropriate deconfounder selection strategy, and statistical validation - that integrates tools from causal inference into associative supervised machine learning (SML). While linear (feature) residualization is commonly applied for confounder adjustment, it is limited in handling complex, non-linear confounding and may leave residual confounding signal. Approaches such as Double/Debiased Machine Learning can inspire improved adjustments in SML workflows but theoretical and practical challenges remain. Importantly, even with appropriate deconfounding, SML remains fundamentally associative, and causal interpretations would require additional assumptions. Nonetheless, properly deconfounded models are critical for generalizable and neurobiomedically informative supervised predictive models, forming an indispensable foundation for neurobiomedical ML research.

Box 1 – Types of third variables and associated biases

When investigating the relationship between a predictor (feature) X and an outcome (target) Y, third variables can be related to X and Y in different ways. The different natures can be best visualized by using directed acyclic graphs (DAGs) (**Box 2**).

Types of third variables

A **confounder Z** is a (direct or indirect) common cause of the feature X and the target Y (Fig. B1.1A). Formally, a confounder can be defined as a variable Z that leads to a discrepancy between the conditional probability of Y given X (*seeing*) and the probability when intervening on X (*doing*): $P(Y \mid X) \neq P(Y \mid do(X))$. Not controlling for a confounder will obscure the causal effect of X on Y. One can either control for the confounder itself or any variable that lies on the path X ← Z → Y.

A **collider C** is the common effect of a feature X and a target Y. Conditioning on a collider induces a spurious (i.e. non-causal) association between X and Y (Berkson's paradox) [21,23,49] (Fig. B1.1B). In other words, if X and Y were independent to begin with, conditioning on Z will make them dependent. (see Berkson's paradox for an example).

A **mediator M** is caused by X and is a cause of Y [50–52] (Fig. B1.1C). For example blood pressure might mediate the relationship between a drug and the risk for a heart attack such that the drug decreases the risk for a heart attack via lowering blood pressure. When interested in the total effect of the predictor on the outcome (X → Y and X → Z → Y), conditioning on M blocks the causal path X → M → Y and will hence only reveal a partial effect. When only interested in the direct effect X → Y conditioning on M can nonetheless lead to biased estimates, if the mediator and the outcome share a common cause because then the mediator is a collider for the predictor and this common cause.

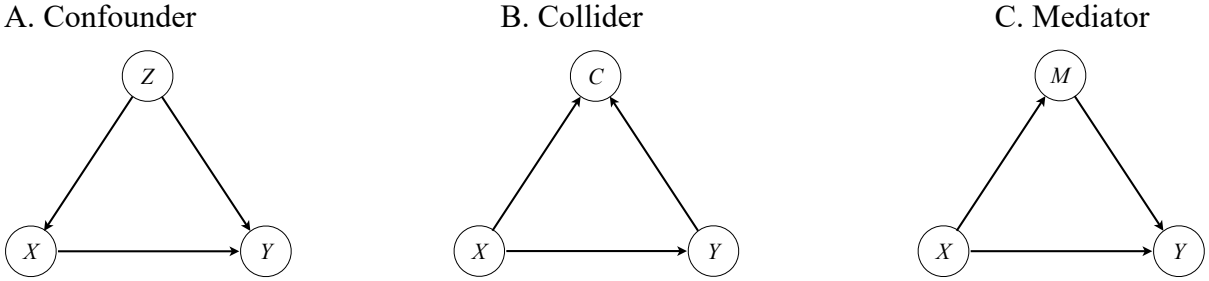


Fig. B1.1 DAG of a confounder (A), a collider (B) and a mediator (C).

Note: Defining confounding via correlations and not as a causal note is not sufficient because each of the causal structures A.-C. produces a correlation between the third variable and both X and Y, which could all produce the same correlation matrix. Consequently, correlations cannot help to distinguish between a confounder, a collider or a mediator [24,25].

A **proxy P** is caused by X but has no causal relation to Y [38] (Fig. B1.2A, B). If the feature X is a perfectly reliable measure of the construct of interest, then controlling for a proxy will not affect the path X → Y. However, in many disciplines X is an unreliable measure of the true causal variable, e.g. a MRI scan for the underlying morphology. In this case, the proxy is a second unreliable measure of the same true predictor (e.g. morphology) and conditioning on this proxy will partition the true predictive effect between the two unreliable proxies so that neither of the unreliable measures will capture the full causal effect [18]. The same logic applies for proxies of confounders.

An **instrumental variable I** for the confounded relationship between X and Y needs to fulfill the following criteria (Fig. B1.2.C):

a) Z and I are independent, i.e. there is no arrow between Z and I.
b) There is an arrow between I and X.
c) There is no direct arrow between I and Y, i.e. no direct causal connection.

There are no confounders of the relation between I and Y, so that any observed association must be causal. Likewise, since the effect of I on Y goes through X, one can conclude that the observed association between X and Y must also be causal. An instrumental variable is therefore similar to a coin flip, which simulates a variable with no incoming arrow.

A. Proxy of the feature B. Proxy of the confounder C. Instrumental Variable

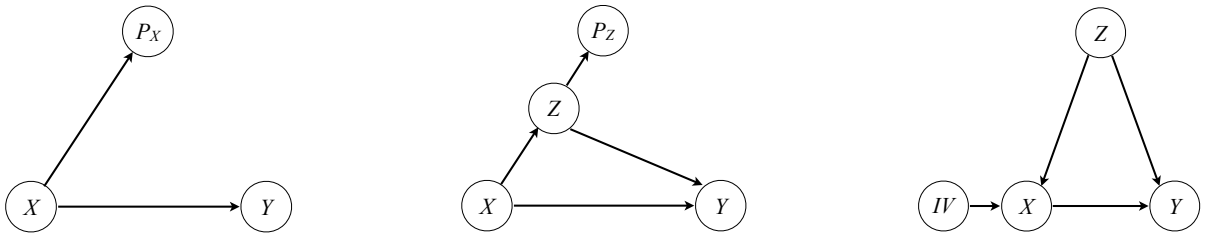


Fig. B1.2 DAG of a proxy of the feature (A) or a confounder (B) and an instrumental variable (C).

Types of biases (paradoxes) associated with third variables

Simpson's paradox (confounder bias)

Simpson's paradox is a statistical phenomenon in which the statistical relationship between two variables in a population can appear, disappear, or reverse when splitting the population in subgroups or when aggregating two heterogenous subgroups into a population. For example, two variables might be positively associated in the overall population but either not or negatively associated within the subgroups[53]. More generally, it is characterized by the statistical results of the subgroups differing from the aggregated population. It alerts to cases where at least one of the statistical trends (either in the aggregated data, the partitioned data, or both) cannot represent the causal effects[42].

Berkson's paradox (collider bias)

Berkson's paradox is the opposite of the Simpson's paradox, i.e. it occurs when falsely conditioning on a variable that is the effect of both the feature(s) and the target (collider). Conditioning on such a collider creates a spurious association between the feature(s) and the target. For example, performing a study on patients who are hospitalized, one controls for/conditions on hospitalization. However, if only a disease 1 and a disease 2 together could lead to hospitalization in the first place (with no causal relation between the diseases), conditioning on hospitalization (by performing the study only on hospitalized patients) would introduce non-existing relation between disease 1 and disease 2.

Box 2 – Basic Causal Concepts and Terminology

Directed Acyclic Graph (DAG)

Formally, a graph *G* is a collection of nodes and edges that connect (some of) the nodes. In a directed graph the edges are directed, i.e. pointing from one node to another. Visually, arrows indicate the directions. Nodes connected by one edge are called adjacent. A path in a graph is

any sequence of adjacent nodes, regardless of direction of the edges that join them, e.g. X ← Z → Y is a path, but not a directed path but X → Z → Y is a directed path. A directed cycle is a directed path that starts from a node X and ends in X. A directed acyclic graph (DAG) is a directed graph with no directed cycles. DAGs obey the local Markov assumption that given its parents in the DAG a node X is independent of all its non-descendants.

Practically, a DAG supports formalization of causal relations or assumptions between variables (nodes), where the arrows (edges) represent directions of known or suspected causal relationships between two variables. For example, X → Y means that X is a direct cause of Y, i.e. the arrow implicitly says that some probability rule or function specifies how Y would change if X were to change. The rule according to which this change happens might either be known (e.g. previous research) or has to be estimated from data. The practicality of DAGs is that they allow to express how a joint distribution over a set of random variables factorizes because they allow to encode (conditional) independence relationships.

Probabilistic language

To formally express and intuitively exemplify different types of probabilities we use the neuroimaging example of gray matter volume (GMV) in the left anterior globus pallidus (denoted as G) and hand grip strength (HGS, denoted as H).

1. **Marginal probability**: P(G=g)
   The probability that a randomly selected individual has GMV G=g in the left anterior globus pallidus.
   Example: What is the probability that a person's GMV in this region is 400 mm³?
2. **Joint probability**: P(G=g, H=h)
   The probability that an individual simultaneously has a GMV of g and HGS of h.
   Example: What is the probability that someone has a GMV of 400 mm³ and a HGS of 35 kg?
3. **Conditional probability**: P(H=h|G=g)
   The probability distribution HGS among individuals with a given GMV value.
   Example: Among individuals with a GMV of 400 mm³, what is the probability distribution of their HGS? ⇒ associational reasoning based on observation.
4. **Interventional probability (causal)**: P(H=h|do(G=g))
   The probability of HGS if we were to hypothetically intervene and set the GMV to g for all individuals.
   Example: What would the distribution of HGS look like if we could biologically set GMV in this brain region to 400 mm³ for everyone?
   ⇒ causal reasoning under intervention. Corresponds to what is estimated in experimental or simulated interventions.
5. **Conditional independence**: X⊥Y|Z
   Two variables X and Y are conditionally independent given a third variable Z if, once Z is known, knowing X provides no further information about Y, and vice versa.
   Example: Suppose physical activity (P) affects both GMV (G) and HGS (H). Then GMV and HGS may appear correlated. However, once physical activity is accounted for, GMV and HGS may become conditionally independent: G⊥H|P.

Box 3 – Two classic ways to account for confounding influences based on DAGs.

There are several ways to identify and account for confounding influences. Two of the most known ones are the backdoor and frontdoor criterion.

Backdoor criterion

A backdoor path is any path from X to Y that starts with an arrow pointing into X, for example X ← Z → Y in Fig. B3.1A. Backdoor paths are *non-causal* paths. To deconfound X and Y one needs to block every *non-causal* path between X and Y without blocking or perturbing causal paths. This can be achieved by adjusting for variables with an incoming arrow into X based on the respective DAG (Fig. B3.1A). These variables are called deconfounders and can differ from the set of confounders, for example when a confounder does not need to be controlled for because the backdoor path is already blocked by a collider (Fig. B3.1B) or when the actual confounder is unmeasured, but the backdoor path can be blocked by controlling for a measured deconfounder (Fig. B3.1C). Randomized control trials (RCTs) avoid confounding as non-causal pathways are blocked through randomization.

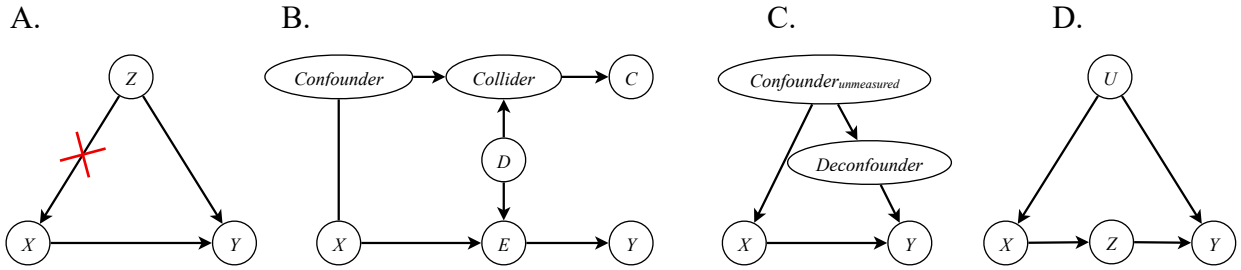


Fig. B3.1 Different positions of a confounder in the DAG structure require different confounder adjustment strategies.

Frontdoor criterion

The backdoor criterion is not feasible when one (or all) deconfounders cannot be measured or are not available. In this case the frontdoor criterion can be applied. The frontdoor criterion (Fig. B3.1D) requires a variable Z that

a) intercepts all direct paths from X → Y
b) there is no backdoor path from X to Z
c) all backdoor paths from Z to Y are blocked by X.

As the relationship between X and Y is confounded by the unobserved variable U, in the frontdoor criterion the effect of X on Y is estimated indirectly by combining the estimate of effect X → Z and of Z → Y.

6. Acknowledgments

This research has been conducted using the UK Biobank Resource under application number 41655. This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 431549029 - Collaborative Research Centre CRC1451 on motor performance project B05.

7. Competing Interests

The authors declare no competing interests.

8. Author Contributions

VK: conceptualization, formal analysis, methodology, visualization, writing – original draft, writing – review and editing; **CH**: writing – review and editing; **SBE**: funding acquisition, resources; **CR**: writing – review and editing; **FR**: supervision; **KRP**: supervision, writing – review and editing

9. References

1. Berisha V, Krantsevich C, Hahn PR, et al. Digital medicine and the curse of dimensionality. *Npj Digit Med*. 2021;4(1):153. doi:10.1038/s41746-021-00521-5

2. Darcy AM, Louie AK, Roberts LW. Machine Learning and the Profession of Medicine. *JAMA*. 2016;315(6):551. doi:10.1001/jama.2015.18421

3. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol*. 2017;2(4):230-243. doi:10.1136/svn-2017-000101

4. Schölkopf B, Locatello F, Bauer S, et al. Towards Causal Representation Learning. Published online February 22, 2021. doi:10.48550/arXiv.2102.11107

5. Chen J, Patil KR, Weis S, et al. Neurobiological divergence of the positive and negative schizophrenia subtypes identified on a new factor structure of psychopathology using non-negative factorization: an international machine learning study. *Biol Psychiatry*. 2020;87(3):282-293.

6. Winter NR, Blanke J, Leenings R, et al. A systematic evaluation of machine learning–based biomarkers for major depressive disorder. *JAMA Psychiatry*. 2024;81(4):386-395.

7. Chekroud AM, Hawrilenko M, Loho H, et al. Illusory generalizability of clinical prediction models. *Science*. 2024;383(6679):164-167. doi:10.1126/science.adg8538

8. Kapoor S, Narayanan A. Leakage and the Reproducibility Crisis in ML-based Science. Published online July 14, 2022. Accessed January 31, 2023. http://arxiv.org/abs/2207.07048

9. Quinonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND. *Dataset Shift in Machine Learning*. Mit Press; 2008.

10. Huyen C. *Designing Machine Learning Systems: An Iterative Process for Production-Ready Applications*. First edition. O'Reilly Media, Inc; 2022.

11. Sugiyama M, Kawanabe M. *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. MIT press; 2012.

12. Arbabshirani MR, Plis S, Sui J, Calhoun VD. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*. 2017;145:137-165. doi:10.1016/j.neuroimage.2016.02.079

13. Benkarim O, Paquola C, Park B yong, et al. The Cost of Untracked Diversity in Brain-Imaging Prediction. *bioRxiv*. Published online June 2021:34. doi:https://doi.org/10.1101/2021.06.16.448764

14. Pulini AA, Kerr WT, Loo SK, Lenartowicz A. Classification Accuracy of Neuroimaging Biomarkers in Attention-Deficit/Hyperactivity Disorder: Effects of Sample Size and Circular Analysis. *Biol Psychiatry Cogn Neurosci Neuroimaging*. 2019;4(2):108-120. doi:10.1016/j.bpsc.2018.06.003

15. Woo CW, Chang LJ, Lindquist MA, Wager TD. Building better biomarkers: brain models in translational neuroimaging. *Nat Neurosci*. 2017;20(3):365-377. doi:10.1038/nn.4478

16. Becker TE. Potential Problems in the Statistical Control of Variables in Organizational Research: A Qualitative Analysis With Recommendations. *Organ Res Methods*. 2005;8(3):274-289. doi:10.1177/1094428105278021

17. Bernerth JB, Aguinis H. A Critical Review and Best-Practice Recommendations for Control Variable Usage. *Pers Psychol*. 2016;69(1):229-283. doi:10.1111/peps.12103

18. Wysocki AC, Lawson KM, Rhemtulla M. Statistical Control Requires Causal Justification. *Advances in Methods and Practices in Psychological Science*. 2022;5(2).

19. Atinc G, Simmering MJ, Kroll MJ. Control Variable Use and Reporting in Macro and Micro Management Research. *Organ Res Methods*. 2012;15(1):57-74. doi:10.1177/1094428110397773

20. Carlson KD, Wu J. The Illusion of Statistical Control: Control Variable Practice in Management Research. *Organ Res Methods*. 2012;15(3):413-435. doi:10.1177/1094428111428817

21. Elwert F, Winship C. Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. *Annu Rev Sociol*. 2014;40(1):31-53. doi:10.1146/annurev-soc-071913-043455

22. Tönnies T, Kahl S, Kuss O. Collider bias in observational studies: consequences for medical research part 30 of a series on evaluation of scientific publications. *Dtsch Ärztebl Int*. 2022;119(7):107.

23. Berkson J. Limitations of the Application of Fourfold Table Analysis to Hospital Data. *Biom Bull*. 1946;2(3):47. doi:10.2307/3002000

24. Maxwell SE, Cole DA. Bias in cross-sectional analyses of longitudinal mediation. *Psychol Methods*. 2007;12(1):23-44. doi:10.1037/1082-989X.12.1.23

25. Pearl J. Causal diagrams for empirical research. *Biometrika*. 1995;82(4):669-710.

26. Kaddour J, Lynch A, Liu Q, Kusner MJ, Silva R. Causal machine learning: A survey and open problems. *ArXiv Prepr ArXiv220615475*. Published online 2022.

27. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med*. 2015;12(3):e1001779. doi:10.1371/journal.pmed.1001779

28. Textor J, Hardt J, Knüppel S. DAGitty: A Graphical Tool for Analyzing Causal Diagrams. *Epidemiology*. 2011;22(5):745. doi:10.1097/EDE.0b013e318225c2be

29. Sharma A, Syrgkanis V, Zhang C, Kıcıman E. Dowhy: Addressing challenges in expressing and validating causal assumptions. *ArXiv Prepr ArXiv210813518*. Published online 2021.

30. Abdulaal A, Hadjivasiliou A, Montaña-Brown N, et al. Causal modelling agents: Causal graph discovery through synergising metadata-and data-driven reasoning. In: *12th International Conference on Learning Representations, ICLR 2024*. Vol 2024. International Conference on Learning Representations (ICLR); 2024.

31. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *J Am Stat Assoc*. 1996;91(434):444-455.

32. Miao W, Geng Z, Tchetgen Tchetgen EJ. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*. 2018;105(4):987-993.

33. Ventura T, Gomes M, Pita A, Neto M, Taylor A. Digit ratio (2D: 4D) in newborns: influences of prenatal testosterone and maternal environment. *Early Hum Dev*. 2013;89(2):107-112.

34. Kuroki M, Pearl J. Measurement bias and effect restoration in causal inference. *Biometrika*. 2014;101(2):423-437.

35. Rohrer JM. Thinking clearly about correlations and causation: Graphical causal models for observational data. *Adv Methods Pract Psychol Sci*. 2018;1(1):27-42.

36. Hamdan S, Love BC, von Polier GG, et al. Confound-leakage: confound removal in machine learning leads to leakage. *GigaScience*. 2023;12.

37. Chyzhyk D, Varoquaux G, Milham M, Thirion B. How to remove or control confounds in predictive models, with applications to brain biomarkers. *GigaScience*. 2022;11:giac014. doi:10.1093/gigascience/giac014

38. Pearl J. Causal inference in statistics: An overview. *Stat Surv*. 2009;3(none). doi:10.1214/09-SS057

39. Chernozhukov V, Chetverikov D, Demirer M, et al. Double/debiased machine learning for treatment and structural parameters. Published online 2018.

40. Oprescu M, Syrgkanis V, Battocchi K, Hei M, Lewis G. EconML: A machine learning library for estimating heterogeneous treatment effects. In: *33rd Conference on Neural Information Processing Systems*. ; 2019:6.

41. Bach P, Chernozhukov V, Kurz MS, Spindler M. DoubleML-an object-oriented implementation of double machine learning in python. *J Mach Learn Res*. 2022;23(53):1-6.

42. Pearl J, Mackenzie D. *The Book of Why: The New Science of Cause and Effect*. Basic Books; 2018.

43. Pearl J. Causality: models, reasoning, and inference. Published online 2000.

44. Imbens GW, Rubin DB. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge university press; 2015.

45. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66(5):688.

46. Bollen KA. *Structural Equations with Latent Variables*. John Wiley & Sons; 1989.

47. Feuerriegel S, Frauen D, Melnychuk V, et al. Causal machine learning for predicting treatment outcomes. *Nat Med*. 2024;30(4):958-968.

48. Schölkopf B, Janzing D, Peters J, Sgouritsa E, Zhang K, Mooij J. On causal and anticausal learning. *ArXiv Prepr ArXiv12066471*. Published online 2012.

49. Rohrer JM. Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data.

50. Baron RM, Kenny DA. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J Pers Soc Psychol*. 1986;51(6):1173.

51. Hayes AF. Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Commun Monogr*. 2009;76(4):408-420.

52. Judd CM, Kenny DA. Process analysis: Estimating mediation in treatment evaluations. *Eval Rev*. 1981;5(5):602-619.

53. Sprenger J, Weinberger N. Simpson's paradox. In: Zalta EN, ed. *The Stanford Encyclopedia of Philosophy*. Summer 2021. Metaphysics Research Lab, Stanford University; 2021. https://plato.stanford.edu/archives/sum2021/entries/paradox-simpson/

5.3 Manuscript 3: Hand grip strength as a behavioral read-out of distributed but specific system-level brain integrity: A large-scale multi-modal machine learning study

Komeyer, V., Eickhoff, S. B., Kasper, J., Patil, K. R. [+], & Raimondo, F.[+] (2025). Hand grip strength as a behavioural read-out of distributed but specific system-level brain integrity: A large-scale multi-modal machine learning study.

Own contributions according to CRediT

- Conceptualization (design of the study)
- Data curation (data management and feature extraction)
- Formal analysis
- Investigation
- Methodology
- Software (implementation of analytical procedure)
- Visualization
- Validation
- Writing – original draft
- Writing – review & editing

Hand Grip Strength as a Behavioural Read-out of Distributed but Specific System-level brain integrity: A Large-scale Multi-modal Machine Learning study

Vera Komeyer[1,2,3], Simon B. Eickhoff[1,2], Jan Kasper[1,2], Christian Grefkes[4,5], Federico Raimondo[1,2,*] & Kaustubh R. Patil[1,2,*]

[1]Institute of Systems Neuroscience, Heinrich Heine University Düsseldorf, Düsseldorf, Germany;
[2]Institute of Neuroscience and Medicine (INM-7: Brain and Behaviour), Research Centre Jülich, Jülich, Germany;
[3]Department of Biology, Faculty of Mathematics and Natural Sciences, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany
[4]Department of Neurology, Goethe University Frankfurt, University Hospital, Frankfurt am Main, Germany;
[5]Cooperative Brain Imaging Center CoBIC, Goethe University Frankfurt, Frankfurt am Main, Germany

* Those authors contributed equally

Abstract

Background

Hand grip strength (HGS) is a simple yet powerful and highly reliable clinical assessment and robust predictor of a versatile set of outcomes including frailty, cognitive decline, and mortality. As such, we here operationalize it as a composite marker of non-fragility. Yet, its broad prognostic value cannot be explained by peripheral motor function or muscular-skeletal integrity alone. Instead, explaining its extensive informativeness requires a deeper but currently lacking understanding of the system-level neural architecture supporting HGS.

Methods

We used interpretable predictive modelling on large-scale multi-modal neuroimaging data from the UK Biobank to identify robust and generalizable out-of-sample brain predictors of HGS while enabling insights into individual-level differences. Structural, diffusion, and functional MRI-derived measures served as input features in both uni- and multi-modal setups. Additionally, our comprehensive machine learning workflow included various linear and non-linear learning algorithms as well as an in depth confounder justification protocol and multicollinearity-aware feature importance analysis. The confounder selection procedure converged on training separate models for women and men with additional linear adjustment for age. While a presumably standard setup, the underlying justification procedure confirmed this as sufficient setup for maximizing interpretability of neuronal signal contributions while minimizing non-neuronal contributions to predictions. Feature clustering preceded post-hoc feature importance analysis of best generalizing out-of-sample models. This allowed to account for ubiquitous multicollinearities in the neuroimaging features in the subsequent SHAP-based feature importance analysis. Unraveling the functioning of succesful models, this setup allowed to identify the most relevant neuroimaging-derived features and their interactions underlying succesful prediction of HGS in previously unseen subjects. To identify consistently most informative neural predictors across best out-of-sample models, in each sex a weighted feature-success ratio was derived.

Results

Multi-modal models outperformed uni-modal approaches but neurobiological signatures remained stable across modeling strategies. Across sexes, HGS was primarily constrained by distributed but specific systems including microstructural integrity of various white matter tracts and morphology of subcortical nuclei rather than

focal or isolated (cortical) motor regions. Primarily, microstructural integrity of the medial lemniscus and gray matter volume of the anterior globus pallidus emerged as dominant, modality-independent predictors. Long-range associative fibres, such as the inferior longitudinal fasciculus, and thalamocortical tracts further contributed, whereas markers of diffuse brain aging such as white matter hyperintensities were comparatively uninformative. Beyond these sex-indifferent findings, we saw higher contribution of thalamus morphology and integrity of descending motor pathway and cerebellar afferents and in women. In men, HGS prediction relied more on widespread cortical morphology including higher-order motor planning systems. Feature importances redistributed with increasing modality integration, reinforcing the relevance of complementary multi-modal information for HGS prediction. Inter-individual heterogeneity in brain feature importance suggested flexible neurobiological strategies converging on similar behavioral output.

Discussion

HGS reflects the integrity of large-scale neural communication spanning sensorimotor, associative, and cognitive-motivational systems. Rather than indexing isolated motor output, HGS functions as a read-out of specific but distributed brain-wide system-level integrity, providing a mechanistic explanation for its versatile predictive value within but also beyond motor outcomes (e.g. cogntive decline). These findings position HGS as a scalable behavioural assay of incipient large-scale neural (dys-)function preceding overt motor or cognitive decline and offer a target for early interventional strategies.

Introduction

Hand grip strength (HGS), the maximal voluntary force generated by the hand [1], is a highly reliable and scalable biomarker traditionally viewed as measure of peripheral muscle properties, muscular-skeletal integrity or physical fitness [2–4]. However, its predictive value extends far beyond musculoskeletal health, encompassing mobility limitations, frailty, falls, fracture risk cognitive performance and cognitive decline, dementia, cardiovascular events, and all-cause mortality [5–13]. Remarkably, HGS outperforms traditional risk indicators such as systolic blood pressure or obesity in predicting mortality [14,15]. Critically, HGS measurements are not purely mechanical, but are influenced by motivation, attention, and affective drive [6,7,16–20], which may contribute substantially to performance in elderly and clinical populations. This aligns with associations between HGS and cognition across multiple domains – including processing speed, memory and verbal and spatial abilities [5,7,8] – where decreased HGS can precede and track overt cognitive decline [6,21]. This broad informativeness establishes HGS as an integrative health biomarker across the lifespan [6,7], so that we propose that it is best understood as a proxy and composite marker for general non-fragility. As an easily acquirable and ubiquitous clinical measure, HGS offers the opportunity to study the biological mechanisms linking physical capability, cognitive alterations and systemic health at scale.

HGS relies on fast and precise coordination of sensory, motor, and associative processes [22,23] across distributed neural circuits, involving cortical, subcortical, cerebellar, and spinal systems [23,24]. It thereby indexes fundamental brain-body communication and the integrity of nervous system function [22,23]. Neuroimaging studies have linked HGS to global brain measures, including total brain volume and white matter hyperintensities, as well as regional indices of gray and white matter integrity, such as hippocampal volume and microstructural properties of white matter tracts, such as the corticospinal tract (CST) and cortico-cerebellar circuits [25–28]. Functional neuroimaging further implicates coordinated activity within motor, premotor and parietal regions, and cerebellar networks during force generation [29–31]. Crucially, while peripheral musculoskeletal factors influence baseline HGS, longitudinal changes in HGS across adulthood are increasingly governed by neural mechanisms [23]. This dual relevance of HGS at both the between-individual (population) level and the within-individual (longitudinal) level suggests the presence of shared, global neural mechanisms alongside substantial inter-individual heterogeneity in the activation, modulation, and interaction of underlying neural circuits and executive strategies. This positions HGS as a window into large-scale brain organization and brain health beyond pure motor control.

Consequently, addressing the multi-facettedness of HGS requires whole-brain characterizations of neural structure and function, which is offered by multi-modal neuroimaging. Structural MRI (sMRI) measures – including gray matter volume (GMV), cortical thickness (CT), surface area, and gray-white matter contrast (GWC) – capture complementary aspects of brain architecture and neurodegenerative processes. Diffusion-weighted imaging (DWI) enables in vivo characterization of white matter (WM) microstructure [32], with tensor-based metrics providing sensitivity to overall tract integrity [33] and more advanced models, such as neurite orientation dispersion and density imaging (NODDI), offering greater biological specificity regarding dendritic density, coherence, and myelination [34–37]. Resting state functional MRI (rsfMRI) quantifies intrinsic neural activity and local and global connectivity patterns supporting sensorimotor integration.

Capturing the system-level nature of HGS and its potentially heterogeneous neural mechanisms across individuals, requires moving beyond uni-variate group-level associations and the perspective of HGS as a downstream correlate of isolated brain regions and peripheral function. Prior work largely relied on single modalities, predefined regions [28] or univariate, hypothesis-driven within-group analyses. This may not capture the multi-variate, interdependent nature of neuroimaging data. Additionally, this limits insights into how distributed brain systems jointly support HGS, which patterns generalize across samples, and how they vary across individuals. Multivariate machine learning approaches based on multi-modal neuroimaging features provide a means to these limitations. They leverage features and their interactions to arrive at generalizable out-of-sample predictions, while allowing for individual-level insights. However, for unbiased predictions and reliable interpretations, confounding factors and feature multicollinearities must be accounted for, respectively. The former is needed to isolate neuronal from non-neuronal sources of information, while the latter is required for unambigous neurobiological interpretation. Fullfilling those methodological constraints allows understanding the neurobiological predictors of HGS which can help to provide a mechanistic explanation for its broad prognostic utility [6,23]. Such insights can build a foundation for brain-based markers whichs' prognostic value may precede those of the behavioural readout HGS, supporting early interventions before overt clinical declines.

In this study, we used large-scale multi-modal neuroimaging from the UK Biobank [38] to identify the system-level neural architecture underlying HGS as a composite marker of non-fragility. Integrating structural, diffusion, and functional MRI with interpretable, confounder-controlled machine learning, we show that HGS reflects structure and integrity of distributed but specific sensorimotor, cognitive and affective architectures. Those include microstructural integrity of medial lemniscus, thalamocortical radiations, associative fibres, pontine-cerebellar afferents, as well as GMV in basal ganglia (mainly anterior globus pallidus) and thalamic nuclei, rather than focal cortical structure and function or diffuse global white matter integrity. These neural signatures generalize at the population level while revealing inter-individual and sex-specific heterogeneity. Our findings establish HGS as a system-level neural phenotype and composite marker for non-fragility by providing mechanistic insights linking HGS to brain health and early vulnerability preceding physical and cognitive decline in the general population.

Results

1. Large-scale machine learning framework for identifying neuroimaging predictors of HGS (Fig. 1)

To identify robust neurobiological predictors of hand grip strength (HGS), we implemented a large-scale, confounder-controlled machine learning (ML) workflow (Fig. 1). While the workflow is described in large depth in method sections 2 to 6, in brief we trained 1078 models in a nested CV scheme (section 2 methods) for generalization estimation and hyperparameter optimisation. They resulted from combining 11 neuroimaging-derived feature groups across four modalities (sMRI, rsfMRI, DWI, multimodal) with seven learning algorithms and 14 confounder-adjustment strategies. Each feature group contained p features, which resembles for example the parcels of the used brain atlases or the number of white matter tracts and their microstructural measures (Fig. 1a). Based on a theoretical and empirical selection approach we justified the best deconfounding setup, i.e. separately trained models per sex with additional linear adjustment for age (section 3 methods). In a model selection process (Fig. 1b, section 4 methods, section 2 results), we identified significantly best performing models and ranked surviving models based on outer CV performances. Among those, we selected the top performing model per feature group to be used for out-of-sample (OOS) prediction on previously held-out 20% of the data, resulting in five unique feature group (e.g. GMV) best performing models per sex (*winning* models). OOS prediction models underwent post-hoc feature importance analysis (Fig. 1c, section 5&6 methods, section 3&4 results), based on SHAP values, including feature clustering to account for feature multi-collinearities, which is essential for reliable model interpretation. This resulted in Owen values reflecting cluster- and feature importances per winning OOS model. To identify consistently most informative neural predictors across winning models, in each sex a weighted feature-success ratio (FSR) was derived.

A thorough confounder selection procedure is key to enable detection of neuronal underpinnings of HGS, while minimizing non-neuronal influences as much as possible. While the detailed process is described in method section 3, in brief it was built on both a theory-driven (DAG-based) and prediction-based procedure. Theoretically, we applied a causally informed selection framework [39] to derive a directed acyclic graph (DAG) of presumed biological connections, which identified age and sex as a sufficient minimal adjustment set (deconfounders) based on graph rules (S-Fig. 1). Although alternative sufficient sets were admissible under the DAG (e.g. sex hormones and muscle mass), these variables are less reliably and consistently measurable at scale. Given the apparent minimality of sex and age as deconfounders, we validated it empirically by adjusting for a wide set of confounding setups, including sex, age, TIV, waist circumference, BMI, body fat percentage, and whole-body fat free mass. This empirical approach confirmed two things: First, without confounder adjustment one can reach considerably high predictive performance (e.g. R^2 = 0.4, S-Fig. 2; see also Komeyer et al. 2024). Second, there is a saturation effect of drop in predictive performance when adjusting for more variables than sex and age (S-Fig. 2). This confirms the previous result that sex and age adjustment is sufficient. Nonetheless, an additional analysis confirmed that under the use of non-linear learning algorithms, models must be separately trained on males and females to avoid non-linear sex-related confounding influences (S-Fig. 3, methods section 3), converging on sex-

separate models with additional linear age adjustment as best adjustment setup. This comprehensive selection and justification procedure is necessary to allow transparent communication of the selection process and respective confounder influences to foster reproducibility of results. Additionally, it is indispensable for the neurobiological interpretations of this work as it assures that interpretations are based on neurobiological sources of signal while minimizing non-neuronal influences.

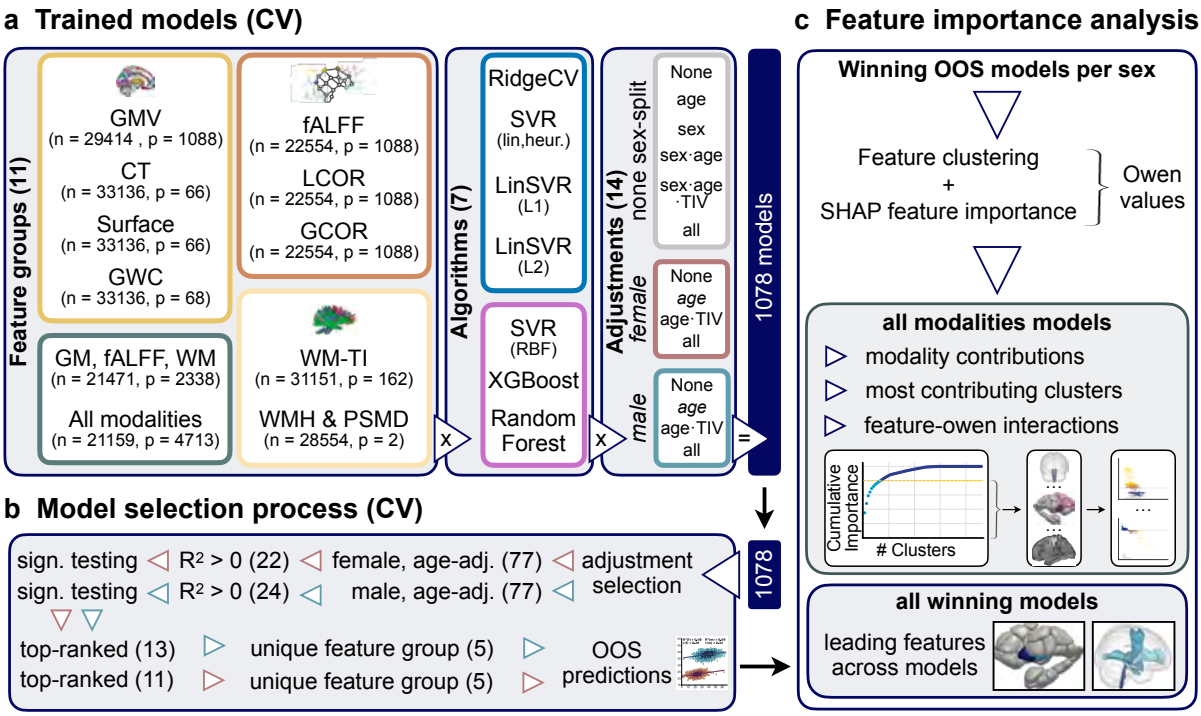


Fig. 1 | Computational framework for multi-modal HGS predicion and neurobiological machine learning model interpretation. a) Overview of trained models and confounder selection. A nested cross-validation (CV) scheme was employed for hyperparameter optimization and generalization estimate of 1,078 machine learning model configurations, combining 11 feature groups (spanning sMRI, rsfMRI, DTI, and multi-modal modality) with 7 learning algorithms and 14 confounder-adjustment settings. Evaluation of confounder-adjustment settings served empirical verification of theoretically identified optimal confounders (see methods, S-Fig. 1, S-Fig. 2). Sex-split models with additional linear age adjustment identified as robust setup for downstream analysis. **b) Model selection process and out-of-sample (OOS) predictions.** From the 154 fully confounder-adjusted models (female+age-adj., male+age-adj.) we selected models with positive R^2 values in all outer CV folds, which survived significance testing ($p<.05$) based on five error metrics (R^2, pearson r, spearman r, mean absolute error, root mean squared error) and which did not perform significantly worse than the top model in a post-hoc pairwise comparison after omnibus testing (male 13, female 11). From the top ranked models we kept one model per feature group and tested its generalizability though out-of-sample (OOS) predictions on previously held-out 20% of the data (winning OOS models). **c) Multicollinearity aware model interpreation and feature importance.** Winning OOS models were interpreted uisng SHAP feature importance analysis and XGBoost-based feature clustering to address feature multicollinearities. Predictive relevance was quantified via Owen values at both feature and cluster levels. Owen values and clusters were used for further inspections of clusters explaining 80% cumulative importance per

winning model and for calculation of a weighted feature success ratio (FSR) across all winning models per sex, identifying the 20 globally most important brain features that form the neurobiological backbone of HGS. *n: sample size, p: number of features per feature group (e.p. parcels of an atlas), GMV: gray matter volume, CT: cortical thickness, GWC: gray-white contrast, fALFF: fractional amplitude of low freuqnecy fluctuations, LCOR: local correlations, GCOR: global correlations, WMM-TI: white matter tract integrity, WMH: white matter hyperintensities, PSMD: peak skeletonized mean diffusivity, CV: cross-validation, OOS: out-of-sample*.

2. Model selection: Multi-modal feature groups with non-linear tree-based models best predict HGS (Fig. 2)

In the model selection process (Fig. 1b, Fig. 2), we evaluated the CV-results of the 154 fully-adjusted models. This aimed to identify the best setups for OOS prediction and feature importance analysis. In a first selection step, we kept only those models with positive R^2 across 10-fold CV. This resulted in 22 models for women and 24 for men (Fig. 1b).

2.1 CV-based model evaluation

Among these 46 models, all-modalities, three-modalities and White Matter Tract Integrity (WM-TI) feature groups exhibited strongest variability of CV validation-set R^2 across folds, algorithms and sexes (see Fig. 2a for performances). Based on average CV R^2, non-linear tree-based algorithms consistently performed best, particularly with multi-modal feature groups (Fig. 2b). This suggests that exploitation of multi-modal feature interactions is critical for HGS prediction. While WM-TI, followed by fALFF and GMV in average emerged as strong uni-modal predictors, White Matter Hyperintensities&Peak Skeletonized Mean Diffusivity (WMH&PSMD) and white surface carried little predictive value (Fig. 2a, Fig. 2b).

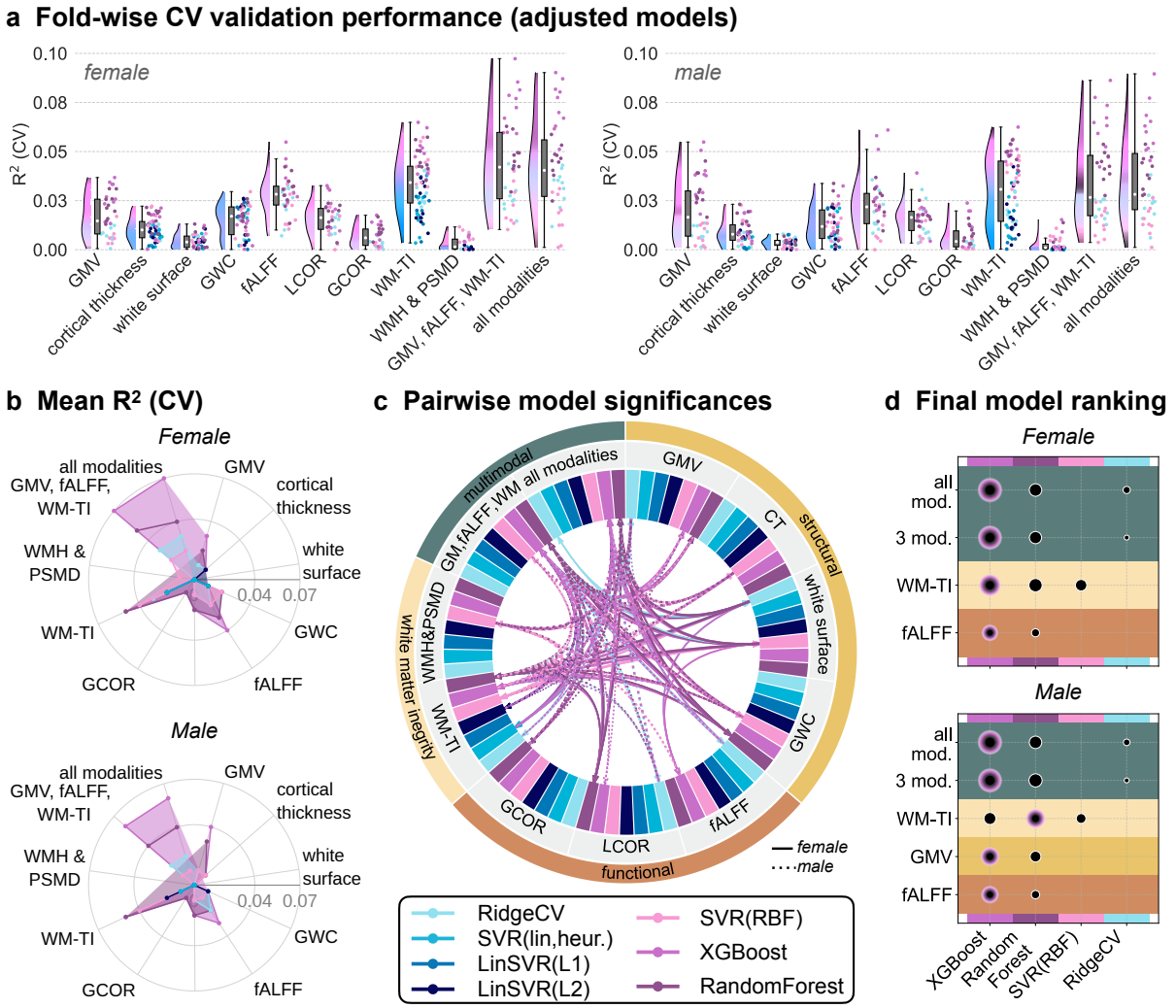


Fig. 2 | Model selection process: statistical ranking and out-of-sample validation. a) CV performance variation across algorithms and folds. Distribution of R^2 values over 10-fold cross validation (CV) performance of sex-split models additionally linearly adjusted for age with $R^2 > 0$ in all folds. **b) Average positive R^2 CV performances.** Mean CV R2 perfromances per feature group illustrate advantage of non-linear, tree-based algorithms especially for multi-modal models. **c) Statistical model comparisons.** Pairwise statistical model performance comparisons over outer CV folds resulting from Nemenyi post-hoc test after significant Friedmann omnibus effect confirmation. Arrow thickness indicates -log10(p-value) of the cross-metric harmonic mean of p-values from the Nemenyi test per outer CV error measure. Arrow direction and colour indicate the significantly better performing model (p<.05), showing an overall better performance of non-linear algorithms (purple) and higher

informativeness of certain feature groups. **d) Final model selection and ranking.** Ranking of significantly best models based on average performance over all error metrics and folds, keeping the best algorithm per brain feature group (halo). Best ranked, unique brain feature models were selected for OOS prediction, with exception of additionally using GMV-XGBoost model in women and XGBoost instead of random forest in men. **e) OOS generalization of selected models.** Predictive performance on the 20% hold-out dataset, separately trained on male and female population but shown in the same scatter plot per feature group for comparability. All-modalities models performed with XGBoost consistently yielded highest prediction accuracy across both sexes. *OOS: Out-of sample, CV: Cross-validation, GMV: Gray matter volume, 3 mod.: multi-modal model with GMV, fALFF, WM-TI feature groups*.

2.2 Model ranking and selection

Omnibus Friedman test was based on all metrics (R2, Pearson r, Spearman rho, MAE, RMSE) and identified significant performance differences between models ($p_{harm\_mean\_women}$ = 0.036, $p_{harm\_mean\_men}$ < 0.001; supplementary Table 1). Post-hoc Nemenyi pairwise comparisons (Fig. 2c, S-Fig. 4, S-Fig. 5) confirmed superior performance of non-linear, especially tree-based algorithms. While multi-modal feature groups were mostly informative with non-linear (especially tree-based) algorithms, the WM-TI feature group also led to significantly better predictions with linear algorithms (e.g. ridge regression, linear SVR).

Ranking models based on their average performance across metrics and folds (Fig. 2d), confirmed all modalities with XGBoost as top-performing model in both sexes. Keeping models that did not perform significantly worse than this top model per sex confirmed that XGBoost and random forest were consistently best in informative feature groups. Those included all-modalitites, three-modalities, WM-TI and, to a lesser extent, fALFF and GMV (Fig. 1a). Ridge regression as only linear algorithm achieved statistically comparable performance to the top model for the two multi-modal feature groups. Even though the GMV feature group did not survive significance thresholding in women, the GMV-XGBoost model ranked above the multi-modal ridge regression model surviving significance thresholding (methods section 4.2, supplementary Table 2), prompting its inclusion in subsequent analyses. In men, random forest ranked one place higher than XGBoost for the WM-TI feature group, but not significantly (p_{harm_mean} < 0.90, Fig. S2) so that we continued with the XGBoost-WM-TI in males for consistency and computational efficiency.

2.3 OOS generalization

The five selected models per sex generalized well on hold-out data (Fig. 2e), with all-modalities models achieving the highest OOS performance (r(m) = 0.32, r(f) = 0.29). Performance rankings remained largely stable across sexes, following all-modalities > three-modalities > uni-modal models (WM-TI, GMV, fALFF) (Fig. 2e). Although moderate, these performances align with the typical range of accuracies reported in population-level neuroimaging studies (Schulz et al., 2024), especially in large cohorts that can converge on lower but more realistic effect sizes (Sui et al., 2020). Additionally, they reflect the conservative nature of our confounder-adjustment. In line with Leenings et al (2022), we argue that this level of performance, coupled with high explainability and robustness as aimed for in our study, is more scientifically insightful than inflated accuracies, potentially built on spurious associations. Overall, these results confirm that integrating diverse neural modalities enhances HGS prediction,

providing the basis for exploring the specific brain systems underlying the all-modalities OOS model per sex as best performers.

3. Neural systems underlying HGS prediction in all-modalities models (Fig. 3)

3.1 White Matter Tract Integrity dominates HGS prediction in both sexes

Feature-group decomposition of the all-modalities model revealed that WM-TI features accounted for the largest share of predictive importance in both sexes (Fig. 3a). In women, WM-TI was followed by GMV and in men by CT and GMV, contributing comparibly. WMH & PSMD in both sexes and surface area in men provided moderate contributions, while functional modalities and GWC contributed minimally.

3.2 Key predictive feature clusters

To account for feature multi-collinearities, features were clustered based on their shared information with respect to HGS before undergoing feature importance analysis. While clusters inform about the hierarchical dependency structure among neuroimaging features, the resulting feature importances (Owen values) reflect both aggregated cluster and fine-grained feature contributions to predictions (methods sections 5&6). A small number of feature clusters accounted for 80% of cumulative model importance (nine clusters in women, 23 in men, Fig. 3b). Cluster importance varied substantially across individuals, exhibiting both continuous and discrete subgroup distributions (Fig. 3c/S-Fig. 6, scatter). This indicates heterogeneous neurobiological strategies for achieving high HGS.

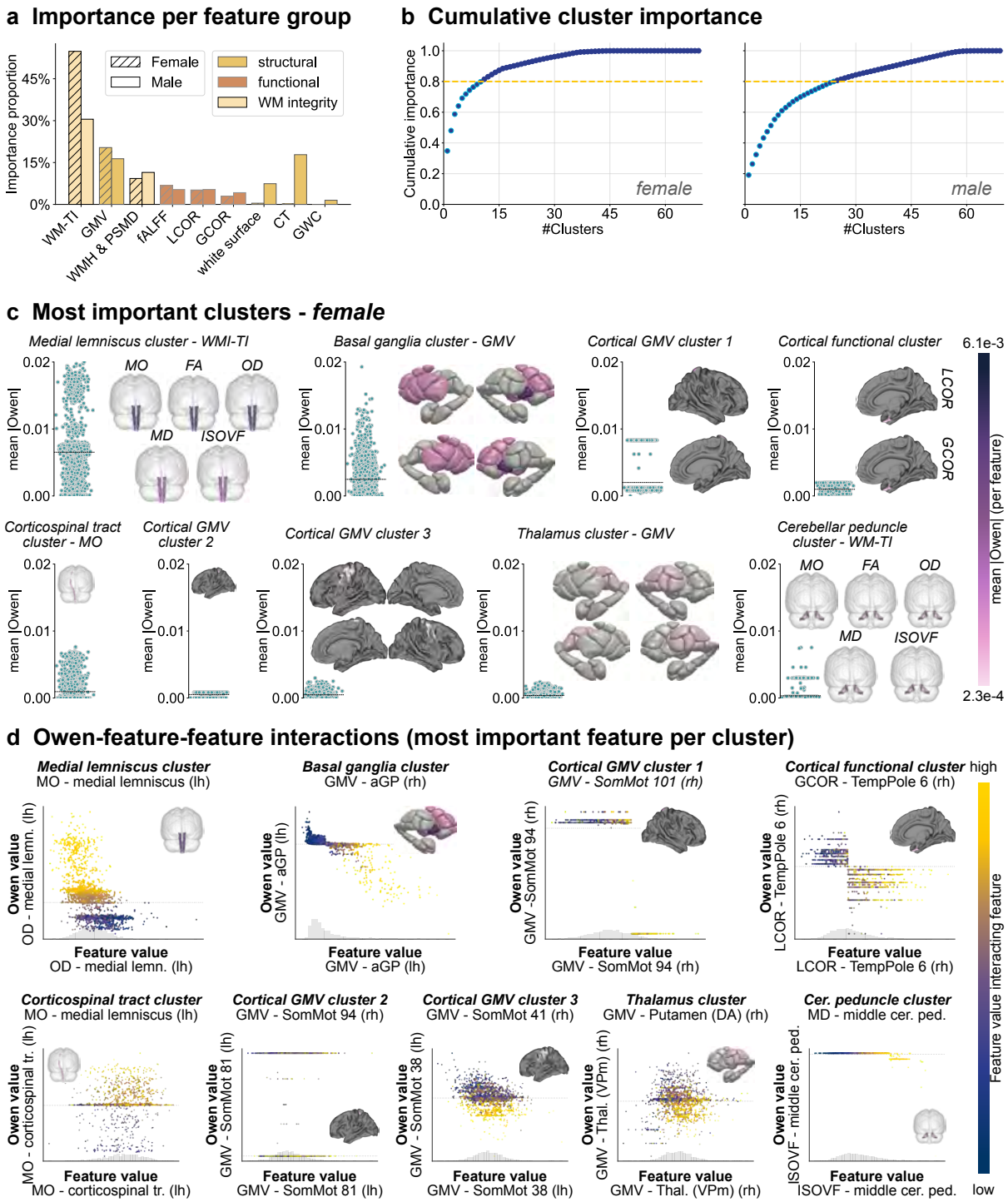


Fig. 3 | Owen feature importances and predictive directionalities and interactions of the all-modalities OOS model. a) Relative modality and feature group contributions. The WM-TI feature group contributed most to predictions in both sexes followed by GMV in females and CT in males. Functional feature groups and GWC were of minor importance in both sexes, whereas white surface area and CT were notably more relevant in men than in women. **b) Cumulative cluster importance.** Cumulative importance curves identifying the minimal set of feature clusters explaining 80% of total model importance. Predictive signals were more concentrated in women (9 clusters) than men (23 clusters), defining the scope for downstream neurobiological inspection. **c) Composition of dominant feature clusters in women.** (Left/scatter) Distribution of mean absolute subject-speciic Owen

values across cluster features. (Right/brain plots) Spatial mapping of feature-level mean absolute Owen values across subjects (global explanation; range: 2.3e-4 to 6.1e-3). For WM-TI features, all microstructural metrics are visualized concurrently to show tract-specific metric variability. **d) Feature interactions and directionality of predictions.** Dependence plot visualizing the relationship between a cluster's leading feature value (x-axis) and its predictive contribution (Owen value; y-axis) to derive directionality of predictive importance. Colors represent the value of the most relevantly interacting feature (feature name beneath cluster name), uncovering essential feature interplay exploited by XGBoost and providing explanations for the inferior performance of linear learning algorithms. Inset brain plots show the top feature within its cluster as shown in panel c. *WM-TI: White Matter Tract Integrity, GMV: Gray Matter Volume, MO: Diffusion Tensor Mode, OD: Orientation Dispersion, MD: Mean Diffusivity, ISOVF: Isotropic Volume Fraction, FA: Fractional Anisotropy, ICVF: Intra-Cellular Volume Fraction, lh: left hemisphere, rh: right hemisphere*.

3.3 Core neurobiological predictors supporting HGS in women

In women, among the nine most important clusters **microstructural WM-TI metrics** of the bilateral medial lemniscus formed the most dominant cluster, with considerable inter-individual variability (Fig. 3c). Lower left OD, especially in interaction with higher MO, predicted higher HGS (Fig. 3d), suggesting that coherent microstructural organization in this sensory ascending pathway supports greater HGS. MO of the left corticospinal tract (CST) - another WM-TI cluster - exhibited comparable subject-specific variability (Fig. 3c), with predictive directionality mainly evolving through interaction: Low MO in the left medial lemniscus at any level of left CST MO predicted lower HGS and vice versa (Fig. 3d). This hints that intact ascending sensory integrity may be necessary for descending motor pathways to contribute effectively to HGS. A third WM-TI cluster involved multiple microstructural measures of the middle cerebellar peduncle (MCP), forming discrete importance sub-groups (Fig. 3c). High ISOVF interacting with high MD showed a non-linear tendency of predicting lower HGS (Fig. 3d).

A cluster of **basal ganglia GMV**, especially in the anterior globus pallidus (aGP) but also putamen, caudate, and nucleus accumbens ranked second in importance (Fig. 3c). Low aGP GMV was non-linearly linked to higher HGS and middle-to-high aGP GMV to lower HGS prediction (Fig. 3d). A cluster encompassing all thalamic nuclei contributed modestly to predictions (Fig. 3c), nonetheless in interaction with high dorsal anterior putamen GMV it was predictive for lower HGS (Fig. 3d).

Cortical contributions were limited and heterogeneous. While in cortical GMV cluster 1 (right S1 hip/knee regions) higher GMV was predictive for lower HGS (Fig. 3d) in a subset of subjects (Fig. 3c), cortical GMV cluster 2 and 3 (bilateral S1/M1) were of overall low importance (Fig. 3c). Functionally, interaction of higher LCOR and GCOR in TempPole 6 (right entorhinal cortex) predicted below-average HGS (Fig. 3c, d).

3.4 Core neurobiological predictors supporting HGS in men

In men, **WM-TI clusters** again highlighted bilateral medial lemniscus microstructural measures as dominant predictor with left lower OD/higher MO predicting higher HGS (S-Fig. 6, S-Fig. 7). Additionally, MO in bilateral inferior fronto-occipital (IFOF) and inferior longitudinal fasciculi (ILF) was mainly informative in a sub-group of men (S-Fig. 6), with particularly high MO of left ILF leading to an abrupt decrease in predicted HGS (S-Fig. 7). A notable cross-modal interaction linked a cluster of bilateral

microstrucutral measures of the parahippocampal part of the cingulum (PHpC) with Crus I GMV: Low-to-mid-level MD in left PHpC with low GMV in left cerebellar Crus I predicted either low or high HGS, whereas high GMV in Crus I at any PHcP MD level yielded average HGS predictions (S-Fig. 6, S-Fig. 7).

A **basal ganglia GMV** cluster including aGP and posterior globus pallidus (pGP) ranked second in overall importance (S-Fig. 6). GMV in aGP again showed non-linear informativeness for HGS prediction, with middle-to-high bilateral GMV predicting lower HGS (S-Fig. 7).

Cortical GMV clusters formed three macro-systems:

(i) primary sensorimotor execution,

(ii) parietal proprioceptive and visuomotor integration, and

(iii) higher-order motor planning.

The **primary sensorimotor execution system** involved eight M1/S1 clusters (S-Fig. 6 cortical GMV clusters 4, 5, 7, 8, 11, 15, 16, 17), with strongest cluster-features originating from M1 hand/forearm, S1 hand/trunk, medial M1/S1 leg representations, and dorsal premotor cortex areas (S-Fig. 6). High GMV in a right M1 “elbow/forearm” parcel (SomMot 84) in non-linear interaction with middle-to-high left S1/M1 GMV predicted lower HGS (S-Fig. 7).

The **parietal proprioceptive-visuomotor integration system** summarized six clusters involving regions of the superior parietal lobule (SPL), intraparietal sulcus (IPS), supramarginal gyrus, and sensorimotor precuneus (S-Fig. 6 cortical GMV clusters 1, 2, 3, 6, 9, 18). Higher GMV in SPL and IPS parcels tended to predict greater HGS. Key interactions with a left supramarginal/IPS transition zone feature (DorsAttn Post 33) and posterior/visual precuneus (Cont pCun 3) occasionally modulated HGS prediction-directionalities non-linearly (S-Fig. 7).

The higher-order **motor planning system** included three clusters comprising dorsal premotor cortex (PMd) and supplementary motor areas (SMA, pre-SMA) (S-Fig. 6 cortical GMV clusters 12, 13, 14). A left SMA/PMd parcel (SomMot 75) emerged as key feature. Its higher GMV, especially when co-ocurring with high GMV in adjacent PMd/SMA parcels, led to discontinuously increased HGS, suggesting modulation of force recruitment efficiency through morphometric variation in higher-order planning regions.

Beyond GMV, a distributed CT cluster spanning most cortical regions showed moderate importance (S-Fig. 6). Higher CT in the left banks of the superior temporal sulcus (bankssts) tended to increase HGS prediction, especially in interaction with left pericalcarine (V1) CT and vice versa (S-Fig. 7).

Across sexes, HGS prediction of the all-modalities model relied on WM microstructural integrity, particular of the ascending sensory medial lemniscus, and of basal ganglia, especially aGP GMV. In women, predictions generally relied primarily on WM microstructural integrity and subcortical GMV with minimal cortical morphometry and functional connectivity involvement. In men, cortical GMV and CT contributed more strongly, suggesting a broader reliance on cortical macro-systems and higher-order motor planning regions in contrast to the higher relevance of intact signal transfer across systems in women. In both sexes, non-linear interactions within and across modalities were essential for accurate HGS prediction.

4. Feature stability across models and overall most successful predictors of HGS

4.1 Cross-model stability and modality-dependent redistribution

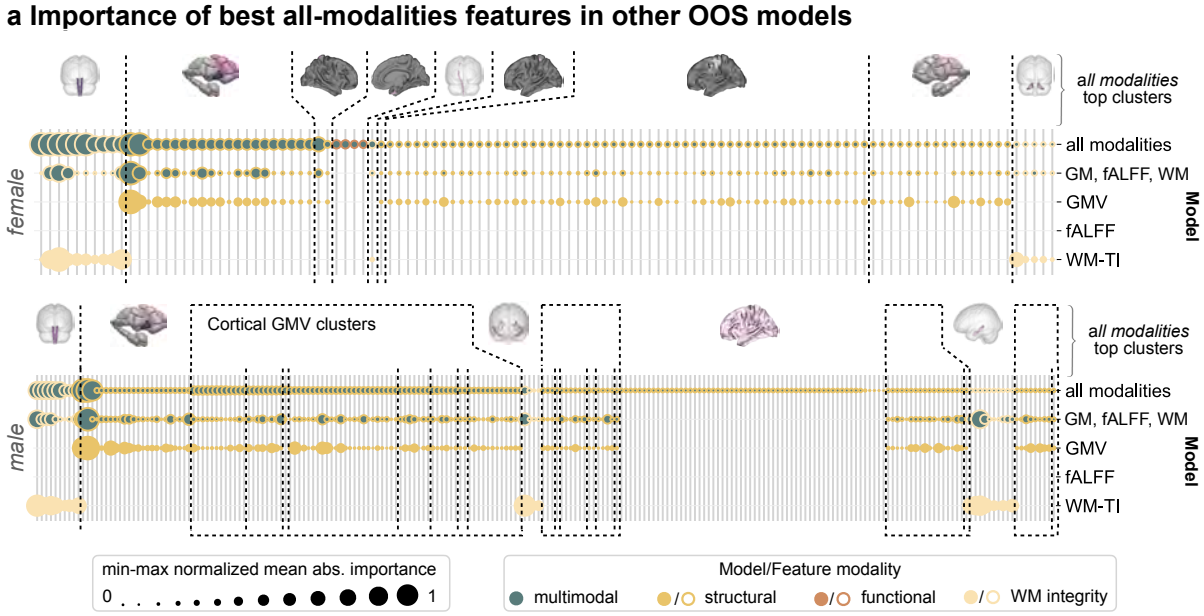
Feature importance was remarkably stable across winning OOS models, indicating that predictive signals represent robust neurobiological signatures rather than model-specific artifacts (Fig. 4a). However, increasing modality integration induced a "homogenizing" effect: all-modalities models distributed importance evenly across features within clusters, whereas uni-modal models exaggerated the dominance of individual "leading" predictors. This was most evident in cortical GMV, where diffuse, moderate importance in multi-modal models contrasted with localized dominance in uni-modal models, e.g in somatomotor and parietal regions in men. Similarly, individual WM-TI microstructural metrics were less dominant in all-modalities settings, suggesting that multi-modal integration captures complementary information, reducing reliance on within-modality variance.

Sex-specific nuances mirrored this redistribution. In women, importance of the MCP (WM-TI) and thalamic (GMV) clusters was higher and more differentiated in uni-modal than multi-modal models In men, the dominance of MO in left ILF (fasciculus cluster) was heightened in the WM-TI-only model and of ISOVF in the PHpC in WM-TI-only and three-modalities models.

Despite these shifts, core neurobiological contributors to HGS remained stable. Microstructural integrity of the medial lemniscus was primary informative predictor across models, dominated by MO, ISOVF, and OD in women and right OD in men. Likewise, basal ganglia GMV showed robust cross-model predictive power, anchored by bilateral aGP. This confirms the modality-independent biological relevance of both structures rather than model-specific effects.

4.2 Globally most informative predictors

The globally 20 most important features across models based on the weighted feature success ratio (FSR) identified a core neurobiological signature of HGS that is largely sex-invariant (Fig. 4b). In both sexes, the bilateral aGP (GMV) and various microstructural metrics of the medial lemniscus emerged as the top-ranking predictors. In women, this core was supplemented by mainly right microstructural measures of anterior and superior thalamic radiations (ATR, STR), MCP and ILF. In men, bilateral PHpC, left ILF, right STR and ATR, left posterior thalamic radiation (PTR), and left cingulate gyrus part of cingulum (CGpC) were additionally important.



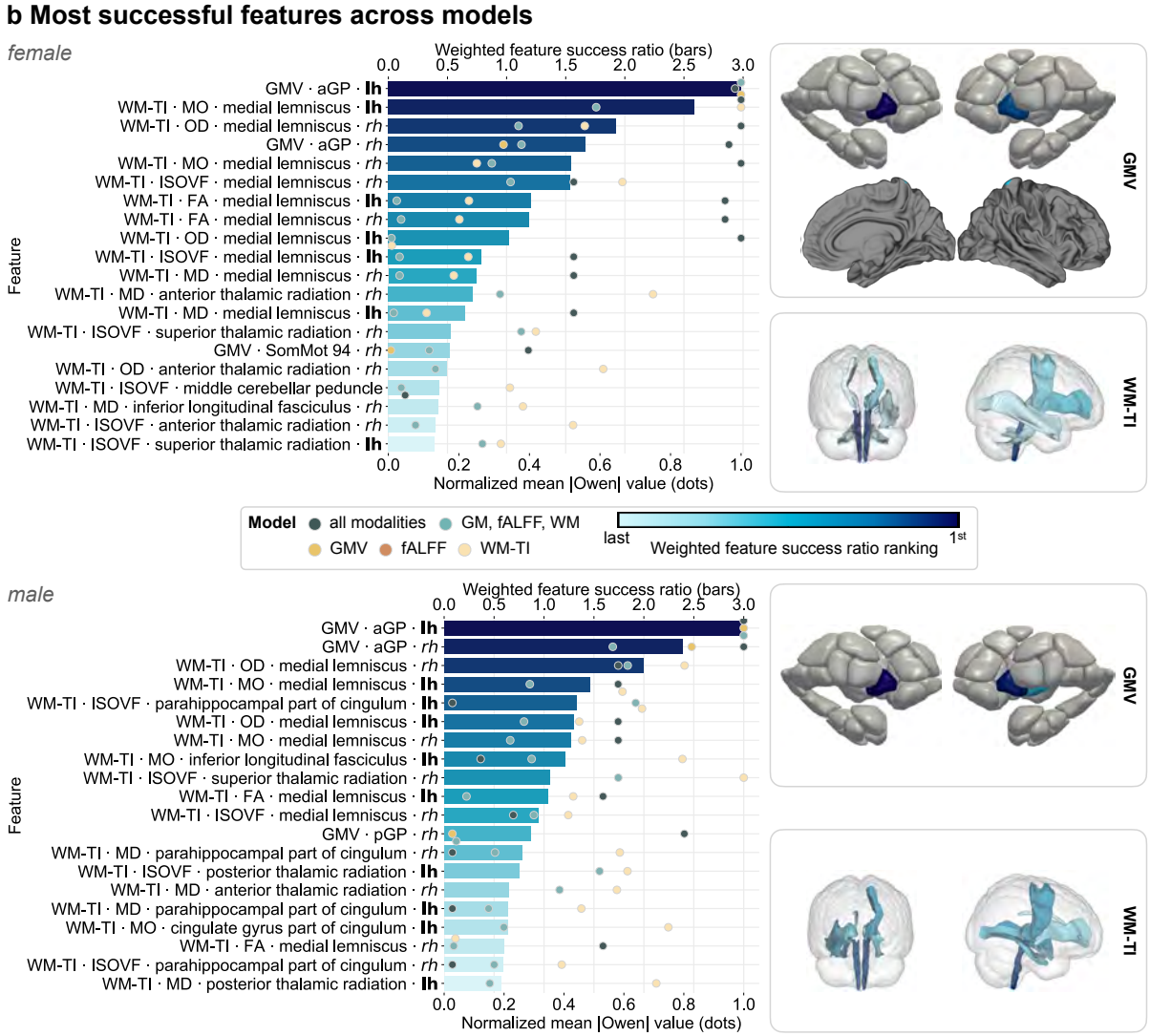


Fig. 4 | Cross-model feature stability and global ranking of HGS predictors. a) Consistency of predictive importances across winning OOS models. Comparative analysis of leading features from the all-modalities model (top row)

across the remaining winning OOS models fro females (top) and males (bottom). Dot sizes represent within model min-max normalized mean absolute Owen values, facilitating comparison of importances between models. Dot and outline colour denote modalities of features contained in the model modality of the feature itself, respectively. Persistence of findings across models highlights robustness of our modeling approach and a stable neurbiological core for HGS prediction. **b) Identification of the globally most informative neuroimaging features.** (Left) Ranking of the 20 most sucessful features across winning OOS models for each sex. Individual dots represent feature-specific min-max normalized mean absolute Owen values colour-coded by model type. Bars display the weighted feature success ratio (FSR), a composite metric integrating feature occurrence frequency across top-performing models with its importance when occuring, thereby effectively indexing how often a feature was within top 80% cumulative importance clusters and how important it was when being used. (Right) Spatial visualization of top 20 features in the brain (once per measure), sorted by modality. Colors correspond to the global FSR rank, highlighting the bilateral aGP and medial lemniscus integrity as primary, sex-invariant anchors of HGS prediction. *aGP: anterior globus pallidus.*

Discussion

Using multimodal neuroimaging in the UK Biobank, we identified neuronal predictors of hand-grip strength (HGS), beyond confounding, using explainable ML.

1. Feature importance stability and multi-modal synergy

We observed high feature stability across winning OOS models. Core neural signatures identified in the all-modalities model remained important as top predictors in the winning uni- and three-modalities models. This cross-model convergence suggests that our results capture a robust, generalizable neurobiological signal rather than model-specific artifacts.

While uni-modal models exploit fine-grained within-modality signals, superior predictive performance is achieved through multi-modal integration of diverse and complementary neuronal resources. For instance, fALFF features were informative in the uni-modal setting but absent from the all-modalities model and the top cross-model features (FSR top 20), while previously insufficient feature groups (CT, LCOR/GCOR) became informative. This suggests superiority of complementary signal integration over fine-grained exploitation for HGS prediction.

2. Sex-invariant neuronal predictors of HGS

2.1 Medial lemniscus integrity and basal ganglia structure as fundamental bottleneck

Microstructural integrity of the medial lemniscus and bilateral aGP GMV consistently dominated as strongest predictors of HGS across all models. The medial lemniscus forms a core conduit for proprioceptive and tactile signals from dorsal column nuclei to thalamocortical circuits, essential for precision grip and force scaling [40,41]. We found that reduced microstructural coherence and increased dispersion in this pathway predicted lower HGS, aligning with previous evidence [42]. It suggests that afferent sensory precision represents a fundamental bottleneck for motor performance, even in healthy subjects with preserved corticospinal output capacity.

Complementing this sensory conduit, the aGP – a basal ganglia territory linked to limbic and associative functionality [43–45] – emerged as critical cognitive-motivational hub. Intriguingly, higher aGP GMV predicted lower HGS. As aGP clustered with putamen, caudate, and nucleus accumbens regions – entry points for motor, associative, and limbic loops, respectively [46,47] – this inverse relationship may suggest that larger aGP GMV reflects less efficient action selection or increased cognitive-motivational “noise” regarding goal decision, reward seeking or aversive avoidance that interferes with core motor commands [48–52]. Regardless, the persistent cross-model importance of the aGP reinforces that HGS is gated by non-motor, cognitive-motivational processes, explaining its versatility as a marker for global neural health.

2.2. Structural integrity of long-ranging white-matter tracts as information flow scaffold

HGS performance further relies on the integrity of long-range associative and thalamocortical tracts (FSR top 20). The inferior longitudinal fasciculus (ILF) - typically associated with visual recognition and semantic processing – emerged as a stable predictor, alone and when clustering with inferior fronto-occipital fasciculus – a relevant tract for connecting visual information with higher cognitive functioning. Higher ILF microstructural coherence paradoxically predicted lower HGS in a subgroup of men. This pattern suggests a sub-group of individuals, where simplified

fibre geometry may reflect a loss of cross-fibre complexity or early pathological reorganization rather than preserved microstructure. Such structural simplification may impair the visuomotor integration required for maximal force production before overt functional decline occurs.

This reliance on large-scale integrative systems is underscored by the importance of tissue integrity and structural organization (MD, ISOVF) of the anterior and superior thalamic radiations (ATR/STR) (FSR top 20). The ATR provides a scaffold for spatial and contextual memory and attention allocation [53,54]. Its predictive power together with the unimportance of anterior thalamic nuclei or frontal lobe and cingulate gyrus – regions connected by the ATR [55], suggests that signal transfer fidelity is a more sensitive marker of HGS than regional GMV or neuronal activation. Complementing this cognitive scaffold, the STR facilitates reciprocal sensory feedback and motor commands between thalamic ventral nuclei and cortical S1/M1 [55–57]. Aligning with our medial lemniscus findings, its relevance reinforces the importance of afferent sensory precision for HGS. Together, these white-matter signatures demonstrate that successful HGS is a functional readout of the brain's ability to coordinate high-fidelity information transfer across both motor and non-motor architectures.

3. Sex-specific neural architectures supporting HGS

There emerged modest sex-specific differences, with varying stability across models.

3.1 HGS in women: sensorimotor transmission, cerebellar coordination and subcortical modulation

In women, HGS is primarily constrained by the integrity of an integrated sensorimotor transmission system. The middle cerebellar peduncle (MCP) emerged as robust predictor across all models**.** Thereby reduced tract compactness and increased structural disorganization of this major pontine-cerebellar afferent pathway predicted lower HGS in a subset of individuals. This highlights the relevance of intact corticopontocerebellar information transfer, aligning with the previous importance of afferent signal quality. Given the cerebellum's broader roles, it additionally suggests that HGS is a refined motor output that requires integration of bilateral cortical error correction signals and vestibular information [58,59]. Notably, the discontinuous importance distribution of the MCP (Fig. 3c, scatter) suggests that cerebellar contributions may only become salient in individuals with emerging microstructural alterations, potentially marking a pre-clinical vulnerability in sensorimotor coordination.

The all-modalities model (but not the FSR top 20) further revealed that efferent motor output in women is fundamentally gated by afferent quality. Specifically, CST (motor efferent) coherence was predictive for higher HGS only when medial lemniscus (sensory afferent) integrity was preserved. This interaction aligns with evidence that proprioceptive input modulates corticospinal excitability and enables precise force calibration [60–62], reinforcing the theory that sensory fidelity acts as a rate-limiting factor for maximal grip output even when descending motor pathways are structurally intact.

This gated architecture is supported by a subcortical interaction between the ventral posteromedial (VPm) thalamus and the dorsal anterior putamen. While the VPm relays orofacial sensory input, it is part of the broader somatosensory thalamus [63]. Its coupling with the associative putamen territory – a key hub in corticostriatal loops – suggests that female HGS is not merely a sensorimotor process but depends on cognitive modulation, including working memory, action selection and decision making [64]. This aligns with the sex-invariant aGP findings. Finally, cortical GMV contributions

were modest and localized outside canonical hand representations (e.g. hip/knee S1), potentially reflecting motor overflow or stabilizing co-activation patterns [65]. Collectively, these findings indicate that HGS in women is an index of high-fidelity information transfer and subcortical cognitive-motivational gating rather than regional cortical morphology.

3.2 HGS in men: distributed associative scaffolds and executive buffering

Contrasting female-specific HGS predictors, the male neural architecture of HGS integrates more core cortical and non-cortical motor structures with higher-order associative hubs. Men uniquely relied on the pGP – a basal ganglia sensorimotor territory [43] – and exhibited a stable dependency on integrity of long-range integration tracts such as the posterior thalamic radiation (PTR), ILF, and cingulate gyrus part of cingulum (CGpC). These pathways facilitate visuomotor, spatial, and contextual incorporation into motor planning, as well as necessary motivational drive, error detection, and executive control [55]. This suggests that male HGS relies on multi-modal systems, including core sensorimotor territory but also structural integrity of visuomotor, motivational, and cognitive-executive information transfer.

A key finding was the non-linear, cross-modality interaction between microstructural organization of the parahippocampal part of the cingulum (PHpC) and cerebellar Crus I GMV. Thereby, efficient cognitive and spatial information transmission [66,67] through high-fidelity PHpC seems to not translate to successful HGS when higher cognitive involvement from Crus I is low (low GMV) [68]. However, low Crus I GMV and disorganized PHpC is paradoxically a predictor for high HGS, possibly through reduction of noisy inference, the recruitment of alternative networks or further interactions. These multi-modal dynamics reinforce that in men, HGS performance is gated by the brain's ability to orchestrate cognitive-contextual loops.

This integrative strategy is further supported by cortical importances and interactions in the all-modalities model. Here, GMV in primary hand/forearm sensorimotor regions, parietal hubs (SPL, IPS, supramarginal gyrus), and motor planning regions (SMA, PMd) emerged as predictive. The parietal hubs provide the substrate for visuomotor transformation and proprioceptive integration, required for internal models of limb position; whereas planning regions support action selection and preparation, and sequencing of complex movements. Additionally, importance of global CT reflects the role of widespread cortical preservation. Notably, the observed interaction between sensory and associative cortical regions indicates that intact low-level sensory processing is a prerequisite for higher-order cortical structure to confer functional advantages. Collectively, these cortical patterns indicate that in men, HGS increasingly depends on widespread cortical preservation and on the structure of a broad cortical integration architecture – spanning sensorimotor execution, visuomotor transformation, and motor planning – suggesting that cortical morphology acts as a system-level coordinator of information flow rather than as a localized determinant of force output.

4. HGS as an integrative readout of large-scale neural communication

Our results support the conceptualization of HGS as a behavioural composite marker of non-fragility, contrasting a simple test of peripheral strength or isolated motor output. Consistent with this view, HGS was not primarily constrained by focal motor regions or diffuse global brain measures. Instead, we observed predominance of distributed systems linking integrity of sensory afferents, thalamocortical relays, cerebellar

pathways, and associative connections with morphology of subcortical nuclei, especially in non-core motor areas. While we observed a focus on cortical morphology, including core-motor areas, especially in men, this retained valid only in specific models. These findings position HGS as a selective readout of brain-wide coordination.

Critically, this integrative signature is anatomically selective. While HGS reflects distributed integrity, it does not act as a generic proxy for diffuse brain aging or global white-matter damage (e.g. WMH). Instead, it captures the vulnerability of functionally meaningful architectures essential for sensorimotor precision, volitional effort regulation, and broader cognitive functioning – processes that are particularly vulnerable in ageing and clinical populations. This specificity aligns with the notion that HGS performance depends on more than muscle capacity alone, incorporating motivational, attentional and cognitive components and suggests that HGS captures selective system vulnerability rather than undifferentiated decline.

Additionally, the found architecture provides an explanation for the link between HGS and cognitive health. We propose that motor execution and higher-order cognition, including the capacity to mobilize voluntary effort, converge on a shared integrative substrate. The thalamocortical and associative tracts required for maximal HGS are overlapping with circuits supporting attention, memory, and action selection. Consequently, HGS serves as a sensitive behavioural assay of incipient network dysfunction. Subtle microstructural degradation my impair the rapid integration required for force production well before it triggers detectable failures or overt deficits in standard cognitive assessments. This positioning transforms HGS from a narrowly defined measure of strength into a scalable, high-fidelity window into the efficiency of large-scale neural health and non-fragile functioning.

5. Neurobiological heterogeneity hints towards strategic flexibility

The marked inter-individual variation in feature importance suggests that HGS is not governed by single canonical architecture, but is a final common outcome achieved through flexible neurobiological strategies. The observed continuous importance gradients and discrete subgroup-forming distributions point to biological redundancy, where individuals may differentially weight sensory, subcortical, or associative systems to maintain performance. Such many-to-one mapping between neural architecture and behaviour reinforces the conceptualization of HGS as a composite marker of systemic non-fragility rather than a localized readout of motor capacity.

Sex-specific differences in predictive patterns can be understood within this heterogeneity framework. Despite a shared and highly standardized behavioural endpoint, women and men exhibited partially distinct feature importances, potentially reflecting variation in how distributed systems contribute to the same physiological output. Rather than implying fundamentally different mechanisms of grip force generation, observed differences suggest flexibel but structured variability in the reliance on sensory, sensorimotor, and cognitive-motivtional-affective processes mirrored in the respective involvement of neuronal substrates. Such variability would be consistent with the idea that non-fragile performance can be maintained through multiple neural configurations of sensory integration, sensorimotor precision, subcortical modulation, cortical planning and commanding, and generally large-scale coordination. Thereby, the exact neural implementation may be shaped by lifelong

differences in body composition, physical activity patterns and broader biological context, such as hormonal milieu, but also psychoaffective conditions.

This heterogeneity has important implications for vulnerability and decline. If individuals depend on different neuronal patterns to achieve comparable HGS, deteroriation in HGS may reflect distinct circuit-level vulnerabilities and failures across individuals (and sexes). Accordingly, reductions in HGS may signal heterogenous neurobiological liabilities rather than a uniform pathway of degeneration, underscoring the need for personalized subtyping in longitudinal risk profiling.

6. Methodological rigor and scope

While our predictive accuracy was modest – a common benchmark in large-scale neurimaging – it must be interpreted within the context of our stringent confounder control. By rigorous deconfounding, we prioritized biological specifictiy over raw variance explained. This trade-off ensures that the identified signatures reflect actual brain contributions to behaviour rather than demographic proxies. While pragmatic thresholds were necessary for interpretability and differing dimensionalities across modalities may have influenced performances, the robustness of results across five performance metrics per model and various models reinforces the validity of the captures signals.

7. Translational implications and conclusions

Our findings provide a mechanistic explanation for HGS as a scalable, integrative biomarker of brain health. By identifying the specific neural signatures that constrain HGS, we shift its utility from a simple frailty screen to a window into the integrity of the brain's global communication infrastructure. Its system-level dependence explains why HGS captures early neural vulnerability in circuits that precede overt motor or cognitive impairment, offering a physiological rationale for its predictive value across diverse clinical outcomes.

Furthermore, the observed inter-individual variability suggests opportunities for neurobiological subtyping, allowing further investigations on roles and vulnerabilities of engaged sub-systems. Future longitudinal work should evaluate whether these HGS-linked neural signatures serve as leading indicators with higher prognostic value than the behavioural measure itself. Ultimately, HGS emerges as a low-cost, high-fidelity behavioural assay of incipient large-scale network (dys-)function, providing a target for early interventional strategies to preserve functional outcomes across the lifespan.

Methods

1. Data and pre-processing

We used data of the 1st scanning session (ses-2) of the UK Biobank [69]), recorded at three different sites in the UK (Cheadle, Reading, Newcastle). The exact acquisition protocol and parameters for the MRI scans can be found in Miller et al. [38].

1.1 Neuroimaging derived features

1.1.1 Processing of structural imaging data

Structural MRI (sMRI) derived features included Gray matter volume (GMV), cortical thickness (CT), surface area, grey white contrast (GWC) and white matter hyperintensities (WMH). For the GMV feature group raw T1w images were segmented into grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF) and normalized by using the default preprocessing of CAT 12.7 toolbox (MNI152 space; 1.5mm isotropic) [70] to arrive at computations of voxel-based morphometry (VBM) [71]. For non-linear registraion, the Shooting method [72] with modulation was used. Concretely, this means we used the m0wp1 output from the CAT 12.7 default-settings preprocessing (only non-linearly modulated (m0), warped/spatially normalized using Dartel/Shooting (w), partial volume segmentation (p), GM (1)). We extracted the parcel-wise GMV as the winsorized mean (limits 10%) of the voxel-wise values per parcel using cortical Schaefer et al. [73] atlas (1000 ROIs, 7 networks), subcortical Tian et al. [74] (S4 3T, 54 ROIs) and cerebellar Diedrichsen et al. [75] (34 ROIs, SUIT space) atlas to arrive at p=1088 parcels.

CT, surface area and GWC feature groups were derived from raw T1 images by the UKB [76] uisng FreeSurfer version (6.0) [77–79]. We obtained parcellated mean thickness, surface area and GWC based on Desikan-Killinay atlas [80] as provided by the UKB (CT: p=66, surface area: p=66, GWC: p=68).

The total volume of WMH was also derived from the UKB. They derived WMH primarily from T2 Flair data but also T1w images with automatic lesion segmentation using the BIANCA tool [81]. We calculated WMH loads from this data by dividng the WMH provided by the UKB with the total intracranial volume (TIV) of a subject to adjust for brain size. The adjusted WMH loadings are in the following referred to as WMH.

1.1.2 Processing of resting-state functional imaging data

rsfMRI derived feature groups included fractional amplitude of low frequency flutuations (fALFF), local correlations (LCOR) and global correlations (GCOR). rsfMRI data has undergone initial preprocessing by the UKB [82]. Subsequently, we normalized these images to MNI space, smoothed, and bandpass-filtered to improve the signal-to-noise ratio of neuronal activtiy in the BOLD signal. To account for motion, white matter, and cerebrospinal fluid (CSF) and thereby reduce potential confounding influences, control measures were applied. To improve data quality further, images that exhibited distortions and artifacts, including those attributed to within-scanner motion were discarded. Voxel-wise fALFF [83] was computed as the power ratio of neuronal activity-related oscillations to the total detectable frequency range in the BOLD signal. Voxel-wise LCOR [84] and GCOR [85] were used to quantify neuronal synchrony by assessing similarity in BOLD signal fluctuations. Specifically, they capture correlations between each voxel and either with its local neighborhood (LCOR) or with all other voxels (GCOR) [86]. Voxel-wise measures were aggregated

following the same parcellation scheme as for the GMV, resulting in p=1088 parcels per feature group.

1.1.3 Processing of Diffusion tensor imaging data

Diffusion-tensor-imaging (DTI) derived features included 6 white matter microstructural characteristics in 27 major white matter tracts as processed by the UK Biobank . The characteristics included fractional anisotropy (FA), a marker of axonal damage [87], the diffusion tensor mode (MO), a probabilistic measure in crossing-fibre tracography [87] and mean diffusivity (MD), an inverse measure of membrane density [87]. They were derived by feeding the b=1000 shell (50 directions) into the DTI fitting tool DTFIT [82]. In addition to the DTI fitting, the dMIR data was fed into NODDI (Neurite Orientation Dispersion and Sensity Ianging) using the AMICO tool to generate three voxelwise microstrucutral parmeters [82], namely isotropic volume fraction (ISOVF), an index of free water [88], intra-cellular volume fraction (ICVF), an index of white matter neurite density [88], and the orientation dispersion index (OD), a measure of within-voxel tract disorganisation providing a more specific metric of axonal damage than FA. Voxelwise information was combined as the weighted-mean value of the respective DTI/NODDI parameter within a tract, with 27 major tracts being defined by standard-space start/stop ROI masks defined by AutoPtx [89]. Those features were derived from the UKB directly (for details in the processing see [82]). Those features are combined in the feature group white matter tract integrity (WM-TI) (n=31151, p=162).

We computed peak width of skeletonized mean diffusivtiy (PSMD), an established global (whole brain) imaging marker of small vessel disease, sensitive to microstructural white matter changes [90] by following containerized standard procedure (https://github.com/miac-research/psmd/tree/main) based on raw DTI data from the UKB, projecting it on the white matter skeleton in standard (MNI) space and calculating PSMD as the difference between the 95th and 5th percentile of MD values.

1.1.4 Combination of feature groups based on modality

We combined the described feature groups from the three MRI modalities in the following way to arrive at 11 feature groups categorized in four modality groups (**Fig. 1**a). GMV (n=29414, p=1088), CT (n=33136, p=66), surface area (n=33136, p=66) and GWC (n=33136, p=68) formed each a feature group belonging to the structural modality. fALFF (n=22554, p=1088), LCOR (n=22554, p=1088) and GCOR (n=22554, p=1088) formed each a feature group, belonging to the functional modality. The white matter integrity modality contained the WM-TI feature group as well as the combination of structural WMH and DTI derived PSMD (feature group WMH & PSMD, n=28554, p=2), as they are both global markers of cerebral small vessel disease and white matter integrity. In the multimodal modality group we combined the GMV, the WM-TI and fALFF feature groups as a multi-modal condition containing the major feature group of each modality (feature group GM, fALFF, WM; n=21471, p=2338). Additionally, we combined all 9 feature groups into the feature group all modalities (n=21159, p=4713).

1.2 Behavioural variables

Non-imaging variables were also obtained from first imaging session (ses-2) of the UKB but behavioural assessments were measured outside of the scanner. Healthy subjects were defined by excluding the ICD-10 criteria chapters F, G, and I60 to I69, which excludes subjects with a history of mental and behavioural disorders, diseases of the nervous system or with a cerebrovascular disease.

For the target variable hand grip strength (HGS), outliers were defined as values exceeding 4th standard deviaton and were excluded from the data. HGS was averaged over left and right hand to include left- and right-handed subjects and minimize lateralization effects.

2. Modelling setup

2.1 Data splits and nested cross validation

For each feature group the data were split into a training (0.8) and test (0.2) set. The training data were used to perform a stratified nested cross validation (CV) with one repetition. The inner CV served hyperparameter optimization and the outer CV provided a generalization estimate of each model's performance, allowing to identify potential overfitting and served to select final models to be used on the hold out test set. A final estimator for the "winning" models (see model selection process) was retrained on the entire 80% training data and used for predictions on the 20% held out test data to get out-of-sample (OOS) predictions. All applied splits were stratified for binned age, binned HGS (2 bins) and sex (if models were trained on the not sex-split samples).

The nested scheme implies that the inner CV is performed on each outer fold's training data. Concretely, we used a 5-fold inner CV with successive halving grid search with a halving factor of 2, coefficient of determination (R^2) as optimization metric and an algorithm dependent grid. Outer CV performance was evaluated using root mean squared error (RMSE), mean absolute error (MAE), R^2, Pearson r and Spearman r, to get a transparent estimate of each model's performance.

Confounder adjustments were computed within the nested CV scheme to avoid data leakage. Therein, for each feature, a linear regression was fit using the confounders as predictors and each feature as dependent variable. Residuals of this fitted linear regression (original features minus predicted/fitted features) served as deconfounded features. Within the CV scheme, continuous features and confounders were z-scored (zero mean, unite variance) and categorical confounders (if applicable) were one-hot-encoded before confounder adjustment. Residualized features were again z-scored after adjustments as some algorithms are sensitive to feature distributions.

Table 1. Overview of sample sizes.

Feature group	$n_{train,female}$ (nested CV)	$n_{test,\ female}$ (OOS)	$n_{train,male}$ (nested CV)	$n_{test,\ male}$ (OOS)	#Features p
GMV	11076	2769	10101	2526	1088
Cortical Thickness	12490	3123	11367	2842	66
Surface	12490	3123	11367	2842	66
GWC	12490	3123	11367	2842	68
fALFF	8559	2140	7679	1920	1088
LCOR	8559	2140	7679	1920	1088
GCOR	8559	2140	7679	1920	1088

WM-TI	11813	2954	10614	2654	162
WMH & PSMD	10876	2719	9682	2421	2
GM, fALFF, WM	8099	2025	7359	1840	2338
All modalities	8008	2003	7224	1807	4713

2.2 Algorithms

Following this scheme we trained seven algorithms, namely a ridge regression, three version of a linear support vector regression (SVR) (linear), SVR with radial basis function (RBF) kernel, XGBoost and Random Forest (non-linear) (**Fig. 1**a). Implementations were either based on scikit-learn [91] or custom-written code.

3. Confounder selection process

As our core interest was in the neuronal underpinnings of HGS, we applied a thourough confounder identification process, to rule out as many non-brain related influences as possible. Confounding is an inherently causal concept [92,93], so that for proper confounder selection and justification causal reasoning must be applied to distinguish confounders from colliders and mediators, which is essential to not introduce additional bias [94]. To this end we used a DAG approach to identify relevant sets of deconfounders, following the approach suggested in Komeyer et al. [39], allowing us to theoretically identify sex and age as appropriate adjustment variables (S-Fig. 1).

To back up the theoretical justification, we additionally followed a more established approach in the field and adjusted the models for a variety of different adjustment scenarios to then evaluate the related performance drop under those different adjustment scenarios. Namely we evaluated no adjustment, age (alone), sex (alone), sex and age together, sex, age and TIV together as well as sex, age, TIV, waist circumference, BMI, body fat percentage and whole-body fat free mass together (the latter in the following is referred to as adjustment set “all”). Using this empirical approach, we could identify that the drop in predictability saturated when removing more confounding influences than sex and age (S-Fig. 2).

Sex (biological sex either based on NHS records or as self-reported) was the confounding variable with the strongest influece. As we used linear and non-linear prediction algorithms but a linear confounder adjustment strategy, we suspected that non-linear algorithms might pick up non-linear sex information. This could be confirmed (to different degrees in the different feature groups) by an additional investigation setup, where we used the Xgboost algorithm (non-linear) to classify sex from the respective brain feature group once with and without linearly residualizing the features for sex influences. While we observed a drop in predictability when adjusting for sex we had a remaining predictability from the residuals (should be at chance level) for all feature groups but WMH & PSMD, confirming that non-linear algorithms can potentially pick up non-linear sex information (S-Fig. 3).

Consequently, as a yet additional adjustment option, we trained models separately on the female and male subpopulation, to rule out any sex-related influences (sex-split predictions). All evaluations, the theoretical justification and the empirical approaches suggest that using sex-split samples for prediction with additional linear feature residualization for age is a robust adjustment setup, to control for as much non-linear

confounding influences as possible, to investigate neuronal influences on HGS, so that we continued with this setup.

In total this led to 14 adjustment scenarios of which 12 could be discarded in the confounder selection process and only the two final adjustment setups (female and age-adjusted, male and age-adjusted) were considered in the model selection process (**Fig. 1**a, italic).

4. Model selection process

The combination of 11 feature groups, 7 algorithms and 14 adjustment options resulted in a total of 1078 models. The confounder selection process allowed to identify the two valid adjustment options, namely separating female and male subpopulations and additionally adjusting for age per subpopulation (male & age-adj., female & age-adj.). This led to 77 models per sex being considered in the model selection process to identify based on the 10-fold outer CV results which feature group – algorithm combination to use for OOS predictions, separately for women and men (**Fig. 1**b).

4.1 Model comparison (significance testing)

Models were compared using the performance measure distributions of the 10 folds of the outer CV to identify if models performed significantly different. We considered all 5 error measures from the outer CV (R^2, pearson r, spearman r, mean absolute error (MAE), root mean squared error (RMSE)) for comparisons to get a broad perspective on model performance as different error metrics are sensitive to different aspects.

In a first filtering step, we discarded all models with a negative R2 value in any of the outer CV folds. Negative R2 values are not meaningful and an indicator of poor model performance. To keep the same number of outer folds for model comparisons, we discarded models even if they had only one fold with a negative R2 value. This led to 22 remaining models for females and 24 for males.

4.1.1 Friedmann chi-square test for general (omnibus) effect testing

We used a Friedmann chi-square test to test for a general effect of performance differences between models. It is a non-parametric alternative to a repeated measurement ANOVA for three or more dependent samples that does not require fold-performances to be normally distributed as analyses are based on rankings. We calculated the Friedmann test for all 5 error metrics (single metric results see supplementary Table 1) and aggregated results across metrics by building the harmonic mean of metric-wise p-values to arrive a one overall decision criteria. Results revealed significant model performance differences (alpha = 0.05) for male and female subsample (female & age-adj: $p_{harm,f} = 0.04$, male & age-adj: $p_{harm,m} = 2.8e\text{-}05$).

4.1.2 Nemenyi post-hoc pairwise model comparison

We performed Nemenyi post-hoc pairwise comparisons between model performances for all 5 metrics (females: 22 comparisons per metric, males: 24 comparisons per metric) and calculated the harmonic mean of p-values across metrics for each pair of models (all: S-Fig. 4, S-Fig. 5, significant: **Fig. 2**c). Nemenyi post-hoc test is based on the studentized range statistic (q distribution) and controls the family-wise error rate across pairwise model comparisons so that no additional correction for multiple comparisons was required. Additionally, there was no correction for multiple

comparisons for testing of the female and male models required, because we did not compare those setups with each other but performed them independently.

4.2 Model ranking

To determine an overall ranking of model performances, we first ranked models per outer CV fold and per metric (e.g. model x performed best in fold 0 according to MAE). We then took for each model per fold the average rank over metrics (e.g. model x in average performed best in fold 0) and then took the average performance rank of each model over folds, arriving at one average rank (over metrics and folds) per model (supplementary Table 2, 3).

We used the average best ranked model per sex as top performing reference model (supplementary Table 2, 3, orange) in combination with the previous Nemenyi pairwise model comparison results, to identify all models that performed significantly worse than this top performing model. Surviving models are hence statistically not distinguishable in performance from the top performing model. This led to 11 surviving models in women (supplementary Table 2) and 13 in men (supplementary Table 3). Of note, the GMV-XGBoost model in the female subpopulation was not among the surviving equally performing models. However, it's average performance ranking was better than the ranking of two other models that performed not distinguishably worse from the top model. Even though the significance testing was also rank based, this can happen because the average ranking combines all metrics directly for the ranking, while the test results were only aggregated over metrics at the level of the p-value, i.e. they were performed separately for each metric. We hence decided to include the GMV-XGBoost model additionally to the surviving models.

From the surviving top performing models we only selected the top performing model per feature group to be used for the OOS predictions, as we are interested in the neuronal aspects underlying each model's prediction and therefore do not need duplicate feature group models (supplementary Table, 2, 3, blue). This resulted in five unique feature group best performing models per sex (**Fig. 2**d). In the female subpopulation those models used the feature groups all-modalities, GM & fALFF & WM, WM-TI, fALFF and GMV (additionally included) always with XGBoost as algorithm. In the male subpopulation selected models used the same feature groups, also with using XGBoost as algorithm but for the WM-TI best ranked model Random Forest performed marginally better. As the difference between the WM-TI Random Forest and XGBoost models was not significant (p_{harm} = 0.90), we here also selected the XGBoost based model for better comparability with the other models and because XGBoost has a more efficient computing behaviour. Those five models per sex were then used for OOS predictions.

4.3 OOS predictions of winning models

For OOS predictions, the algorithms of the winning models from the CV per sex were retrained (including 5-fold inner CV for HPO) on the entire training set from the outer CV (80% of the data) and OOS predictions were performed on the previously held out 20% of the data (see *Data splits and nested cross validation* and Table 1 for sample sizes).

All OOS used XGBoost as algorithm. From XGBoost's hyperparameters max_depth, learning_rate, gamma, min_child_weigh, reg_lambda and subsample were tuned in the inner CV (grid and final hyperparameters: supplementary Table 4), whereas the following were specified but not tuned, n_estimators=1000, max_delta_step=0,

colsampple_bytree=1, eval_metric="rmsle", and for all others default settings were used.

OOS performances were evaluated with the same five error metrics as the outer CV, but for visualization purposes only R^2 and pearson r are displayed. All OOS predictions were performed separately for the female and male subpopulation, as before for the nested CV, but are visualized in the same plot (**Fig. 2**e).

5. Post-hoc feature importance analysis

To interpret the final out-of-sample prediction model, we performed a post-hoc feature importance analysis designed to identify the brain features most relevant for predicting hand grip strength (HGS), while explicitly accounting for inter-feature dependencies. Our interpretability workflow builds upon the SHapley Additive Explanations (SHAP) framework but extends it to handle associated predictors through feature clustering and Owen value decomposition.

5.1 SHAP

The SHAP (SHapley Additive exPlanations) framework [95] quantifies the contribution of individual features to a model's predictions. SHAP values are derived from cooperative game theory and represent each feature's marginal contribution to the model output, averaged over all possible feature combinations. In essence, SHAP decomposes each prediction into additive contributions from features. On a subject-level (local explanation), positive SHAP values indicate that a given feature increases the predicted target (HGS), while negative values indicate a decreased prediction. Averaging absolute SHAP values across individuals (global explanation) provides direction- and subject-independent measures of feature relevance.

SHAP assumes statistically independent features, an assumption that is often violated in high-dimensional neuroimaging data, where structural and functional brain measures are inherently correlated. In such cases, SHAP values may overestimate the importance of correlated features or split shared importance arbitrarily among them, resulting in misleading interpretations.

To overcome this limitation, we adopted a dependency-aware attribution approach based on feature clustering and Owen values [96,97], which extend the SHAP framework to account for feature inter-dependencies. The clustering step identifies groups of features with shared information with respect to the target, and the Owen value computation distributes importance both between and within the feature groups as identified in the clustering.

5.2 Feature clustering

Hierarchical clustering was implemented using shap.utils.hclust with a XGBoost-based distance metric (metric="xgboost_distances_r2"), quantifying the degree to which features share predictive information about the target (HGS) and clustering them accordingly. The resulting hierarchical dendrogram (average linkage) represents pairwise feature similarities.

5.3 Owen values

Owen values were computed to obtain feature attributions that explicitely incorporate the hierarchical dependency structure among the neuroimaging features. Following the hierarchical feature clustering, the resulting cluster-tree was passed to a SHAP

partition masker (shap.maskers.Partition), which enforces the clustering as a coalition structure in the Owen game [98]. When combined with this cluster-based masker, the SHAP partition explainer (shap.explainers.Partition) recursively evaluates coalition splits at every level of the clustering-tree, yielding Owen values for all cluster partitions from the root (full feature set) to the leaves (individual features). This procedure ensures that attribution reflects both aggregated cluster contributions and fine-grained feature contributions within clusters. The explainer was initialized on the training data and applied to the held-out test set, with the number of model evaluations set to 2p + 1 (p: number of features) to approximate the exact solution [95]. In practice, the derived Owen values indicate per feature how much the feature contributed to the model's predictions and in sum per cluster indicate how much the cluster contributed to the model's predictions, both locally (per-subject) and globally (averaged over subjects).

6. Cluster selection and interpretation

As Owen values were derived at every level of the hierarchical clustering, for further investigations, the tree must be cut at a certain level to interpret clusters. To this end, we set the number of clusters to sqrt(p) (p: number of features), to have a fair cluster number criterion given the different number of features in the different models (Table 2). Of those, we selected the first n clusters that explained 80% of cumulative importance (**Fig. 3**b, among others).

Table 2. Number of clusters per OOS model.

		All modalities	GM, fALFF, WM	WM-TI	GMV	fALFF
Overall number of clusters (female/male)		69	48	13	33	33
n clusters selected explaining 80% of cumulative importance	Female	9	20	8	17	21
	Male	23	31	8	20	20

6.1 Cluster investigations of the all-modalities model (Fig. 3)

6.1.1 Importance per feature group

The all-modalities model performed best in both sexes in the OOS predictions, and contained all feature groups, which makes it suitable for neuroscientific broad and accurate information gain. To get a general overview of feature contributions to the all-modalities models' predictions, we first calculated the proportion of contribution of importances per feature group and modality within the all-modalities model per sex. To this end, we divided for each feature group the sum of mean absolute Owen values by the number of features per group (e.g. GMV: p=1088), to correct for feature group size. This group-size corrected value was then divided by the sum of all corrected values (over all feature groups) to get the proportion of the importance of one feature group among all feature groups. For further summary, we colour coded each feature group by the modality (structural, functional, WM integrity) to which the feature group belonged (**Fig. 3**a).

6.1.2 Inspection of neuroscientific meaning of most important clusters

For the n most important clusters per sex of the all-modalities model we first investigated the mean absolute Owen value per subject over cluster features (female:

Fig. 3c, male: S-Fig. 6, scatter). Additionally, we evaluated the mean absolute Owen value per feature over subjects to arrive at a global explanation of how important each feature in the cluster was to arrive the HGs predictions (female: **Fig. 3**c, male: S-Fig. 6, brain plots).

6.1.3 Interactions of most important feature per cluster

While Owen values provide general information on the importance of a feature, they do not allow on their own to determine the directionality of a feature's influence. Therefore, for each of the n most important clusters we picked the feature with the highest mean absolute Owen value, i.e. the highest global importance (top feature) and investigated the interaction between its Owen and its feature value to determine the directionality of its influence. In contrast to linear models, tree-based models often rely on complex feature-feature interactions to arrive at predictions. To determine the most important interacting feature with the top feature of each cluster we trained a shallow random forest regression (n_estimators=100, max_depth=4) to predict the top feature's Owen value based on all features of the test set (excluding the top feature itself). From the sorted ranking of this random forest's feature importances we either picked the most import feature, or if importances were close, the one better explaining the interactions based on visual inspection among the top 3 ranked features. This allowed to inspect for each top feature per cluster the interaction between its Owen value, feature value and the most relevantly interacting feature's feature value (female: **Fig. 3**d, male: S-Fig. 7).

6.2 Identifying most relevant features across models (Fig. 4)

6.2.1 Importance of best all modalities model features in the other OOS models

To evaluate the importance of the most important features of the all-modalities model in the other OOS models, we first min-max normalized the mean absolute Owen value of features within each model, to arrive at importance values that are comparable in magnitude between models. We then extracted for each OOS model per sex the n most important clusters as described above. Per OOS model we picked those features that also appeared in the top n clusters of the all-modalities model and compared their min-max normalized mean absolute Owen value between models (note: not all features can theoretically appear in all models) (**Fig. 4**a). In the corresponding visualization, the mean absolute Owen value is indicated by circle size, whereas circle area colour codes the modality of the model (multi-modal, functional, structural, WM integrity) and edge colour indicates the modality of the feature in the model (structural, functional, WM integrity). Therefore, for uni-modal models area and edge colour always match, while for the multimodal models they can differ.

6.2.2 Most successful features across best models

To identify the most successful features independent of cluster belonging across all OOS models per sex, we calculated a weighted feature success ratio. To this end, we took the n most important clusters per model as describe before and combined the importance of those top features across models. Concretely, we counted in how many of the five models per sex each feature appeared in a top cluster. Additionally, we defined how often a model can theoretically appear, i.e. features of the uni-modal feature group models (GMV, fALFF, WM-TI) can appear in the respective uni-modal model, in the GM, fALFF, WM model and in the all-modalities model, i.e. three times, whereas all other features can maximally appear once (in the all-modalities model). We calculated the feature success ratio as the number of models the feature appears

in the top clusters divided by the number of models in can theoretically appear in. As this does not account for the respective importance of the feature in the model but only identifies if and how often the model appeared, we additionally weighted this ratio by a feature's importance. We calculated the weights by first min-max (0-1) normalizing the mean absolute Owen value over subjects per feature within a model to ensure value comparability between models. For each feature appearing in any of the models' supra-threshold clusters we summed this feature's min-max normalized mean absolute Owen value across models in which the feature appeared. This weight was multiplied with the previous feature success ratio to arrive at the weighted feature success ratio. We ranked models according to this score and took for females and males the 20 top features for neuroscientific interpretation. The weighted feature success ratio is visualized as bar plot (numerical values (0-3), with bar colour encoding the ranking of the model according to the score. Additionally, we highlighted the min-max normalized mean absolute Owen value (0-1) per model as data points as it is the driving factor for the weight calculation. For brain visualizations, if a feature appeared multiple times in the ranking with a different measurement (e.g. OD and MO of medial lemniscus rh), we display the higher ranked measure (**Fig. 4**b).

References

1. Surgent O, Guerrero-Gonzalez J, Dean DC, et al. How we get a grip: Microstructural neural correlates of manual grip strength in children. *NeuroImage*. 2023;273:120117. doi:10.1016/j.neuroimage.2023.120117

2. Gale CR, Martyn CN, Cooper C, Sayer AA. Grip strength, body composition, and mortality. *Int J Epidemiol*. 2007;36(1):228-235.

3. Gell M, Eickhoff SB, Omidvarnia A, et al. How measurement noise limits the accuracy of brain-behaviour predictions. *Nat Commun*. 2024;15(1):10678.

4. Wind AE, Takken T, Helders PJM, Engelbert RHH. Is grip strength a predictor for total muscle strength in healthy children, adolescents, and young adults? *Eur J Pediatr*. 2010;169(3):281-287. doi:10.1007/s00431-009-1010-4

5. Adamo DE, Anderson T, Koochaki M, Fritz NE. Declines in grip strength may indicate early changes in cognition in healthy middle-aged adults. Wylie GR, ed. *PLOS ONE*. 2020;15(4):e0232021. doi:10.1371/journal.pone.0232021

6. Firth J, Stubbs B, Vancampfort D, et al. Grip Strength Is Associated With Cognitive Performance in Schizophrenia and the General Population: A UK Biobank Study of 476559 Participants. *Schizophr Bull*. 2018;44(4):728-736. doi:10.1093/schbul/sby034

7. Firth JA, Smith L, Sarris J, et al. Handgrip Strength Is Associated With Hippocampal Volume and White Matter Hyperintensities in Major Depression and Healthy Controls: A UK Biobank Study. *Psychosom Med*. 2020;82(1):39-46. doi:10.1097/PSY.0000000000000753

8. Jiang R, Westwater ML, Noble S, et al. Associations between grip strength, brain structure, and mental health in > 40,000 participants from the UK Biobank. *BMC Med*. 2022;20(1):286. doi:10.1186/s12916-022-02490-2

9. Legrand D, Vaes B, Matheï C, Adriaensen W, Van Pottelbergh G, Degryse J. Muscle Strength and Physical Performance as Predictors of Mortality, Hospitalization, and Disability in the Oldest Old. *J Am Geriatr Soc*. 2014;62(6):1030-1038. doi:10.1111/jgs.12840

10. Ling CH, Taekema D, De Craen AJ, Gussekloo J, Westendorp RG, Maier AB. Handgrip strength and mortality in the oldest old population: the Leiden 85-plus study. *CMAJ Can Med Assoc J J Assoc Medicale Can*. 2010;182(5):429-435.

11. Metter EJ, Talbot LA, Schrager M, Conwit R. Skeletal muscle strength as a predictor of all-cause mortality in healthy men. *J Gerontol A Biol Sci Med Sci*. 2002;57(10):B359-B365.

12. Newman AB, Kupelian V, Visser M, et al. Strength, but not muscle mass, is associated with mortality in the health, aging and body composition study cohort. *J Gerontol A Biol Sci Med Sci*. 2006;61(1):72-77.

13. Sasaki H, Kasagi F, Yamada M, Fujita S. Grip Strength Predicts Cause-Specific Mortality in Middle-Aged and Elderly Persons. *Am J Med*. 2007;120(4):337-342. doi:10.1016/j.amjmed.2006.04.018

14. Kim Y, Wijndaele K, Lee D chul, Sharp SJ, Wareham N, Brage S. Independent and joint associations of grip strength and adiposity with all-cause and cardiovascular disease mortality in 403,199 adults: the UK Biobank study. *Am J Clin Nutr*. 2017;106(3):773-782.

15. Leong DP, Teo KK, Rangarajan S, et al. Prognostic value of grip strength: findings from the Prospective Urban Rural Epidemiology (PURE) study. *The Lancet*. 2015;386(9990):266-273.

16. Cass K, Blalock DV, Manwaring J, Wesselink D, Lundberg C, Mehler PS. Changes in grip strength, depression, and cognitive functioning during medical stabilization for anorexia nervosa: Exploring the utility of grip strength as a marker of severity. *J Rehabil Res Pract*. 2024;5(1):14-22.

17. Ganipineni VDP, Idavalapati ASKK, Tamalapakula SS, et al. Depression and hand-grip: unraveling the association. *Curēus*. 2023;15(5).

18. Laredo-Aguilera JA, Carmona-Torres JM, Cobo-Cuenca AI, García-Pinillos F, Latorre-Román PÁ. Handgrip strength is associated with psychological functioning, mood and sleep in women over 65 years. *Int J Environ Res Public Health*. 2019;16(5):873.

19. Rinne P, Hassan M, Fernandes C, et al. Motor dexterity and strength depend upon integrity of the attention-control system. *Proc Natl Acad Sci*. 2018;115(3):E536-E545.

20. Watson J, Ring D. Influence of psychological factors on grip strength. *J Hand Surg*. 2008;33(10):1791-1795.

21. Duchowny KA, Ackley SF, Brenowitz WD, et al. Associations Between Handgrip Strength and Dementia Risk, Cognition, and Neuroimaging Outcomes in the UK Biobank Cohort Study. *JAMA Netw Open*. 2022;5(6):e2218314. doi:10.1001/jamanetworkopen.2022.18314

22. Alfaro-Acha A, Snih SA, Raji MA, Kuo YF, Markides KS, Ottenbacher KJ. Handgrip strength and cognitive decline in older Mexican Americans. *J Gerontol A Biol Sci Med Sci*. 2006;61(8):859-865.

23. Carson RG. Get a grip: individual variations in grip strength are a marker of brain health. *Neurobiol Aging*. 2018;71:189-222. doi:10.1016/j.neurobiolaging.2018.07.023

24. Borra E, Gerbella M, Rozzi S, Luppino G. The macaque lateral grasping network: a neural substrate for generating purposeful hand actions. *Neurosci Biobehav Rev*. 2017;75:65-90.

25. Schulz R, Park CH, Boudrias MH, Gerloff C, Hummel FC, Ward NS. Assessing the integrity of corticospinal pathways from primary and secondary cortical motor areas after stroke. *Stroke J Cereb Circ*. 2012;43(8):2248-2251.

26. Schulz R, Frey BM, Koch P, et al. Cortico-cerebellar structural connectivity is related to residual motor output in chronic stroke. *Cereb Cortex*. 2017;27(1):635-645.

27. Schulz R, Park E, Lee J, et al. Interactions between the corticospinal tract and premotor–motor pathways for residual motor output after stroke. *Stroke J Cereb Circ*. 2017;48(10):2805-2811.

28. Weitnauer L, Frisch S, Melie-Garcia L, et al. Mapping grip force to motor networks. *NeuroImage*. 2021;229:117735. doi:10.1016/j.neuroimage.2021.117735

29. Castiello U. The neuroscience of grasping. *Nat Rev Neurosci*. 2005;6(9):726-736.

30. Errante A, Ziccarelli S, Mingolla GP, Fogassi L. Decoding grip type and action goal during the observation of reaching-grasping actions: A multivariate fMRI study. *NeuroImage*. 2021;243:118511.

31. King M, Rauch H, Stein DJ, Brooks SJ. The handyman's brain: a neuroimaging meta-analysis describing the similarities and differences between grip type and pattern in humans. *Neuroimage*. 2014;102:923-937.

32. Nemanich ST, Mueller BA, Gillick BT. Neurite orientation dispersion and density imaging quantifies corticospinal tract microstructural organization in children with unilateral cerebral palsy. *Hum Brain Mapp*. 2019;40(17):4888-4900. doi:10.1002/hbm.24744

33. Oschwald J, Mérillat S, Jäncke L, Seidler RD. Fractional Anisotropy in Selected, Motor-Related White Matter Tracts and Its Cross-Sectional and Longitudinal Associations With Motor Function in Healthy Older Adults. *Front Hum Neurosci*. 2021;15:621263. doi:10.3389/fnhum.2021.621263

34. Dick F, Tierney AT, Lutti A, Josephs O, Sereno MI, Weiskopf N. In vivo functional and myeloarchitectonic mapping of human primary auditory areas. *J Neurosci*. 2012;32(46):16095-16105.

35. Hoy AR. *Diffusion Tensor Imaging with Free Water Elimination*. phd. The University of Wisconsin-Madison; 2015.

36. Stüber C, Morawski M, Schäfer A, et al. Myelin and iron concentration in the human brain: a quantitative study of MRI contrast. *Neuroimage*. 2014;93:95-106.

37. Zhang H, Schneider T, Wheeler-Kingshott CA, Alexander DC. NODDI: practical in vivo neurite orientation dispersion and density imaging of the human brain. *Neuroimage*. 2012;61(4):1000-1016.

38. Miller KL, Alfaro-Almagro F, Bangerter NK, et al. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat Neurosci*. 2016;19(11):1523-1536. doi:10.1038/nn.4393

39. Komeyer V, Herrmann C, Eickhoff SB, Rathkopf C, Raimondo F, Patil KR. How causal inference tools can support debiasing of machine learning models for meaningful brain-based predictions. *MedRxiv Prepr Serv Health Sci*. Published online 2025:2024-09.

40. Al-Chalabi M, Reddy V, Alsalman I. Neuroanatomy, posterior column (dorsal column). Published online 2018.

41. Navarro-Orozco D, Bollu PC. Neuroanatomy, medial lemniscus (reils band, reils ribbon). Published online 2018.

42. Proske U, Gandevia SC. The proprioceptive senses: their roles in signaling body shape, body position and movement, and muscle force. *Physiol Rev*. Published online 2012.

43. Bertino S, Basile GA, Anastasi G, et al. Anatomical characterization of the human structural connectivity between the pedunculopontine nucleus and globus pallidus via multi-shell multi-tissue tractography. *Medicina (Mex)*. 2020;56(9):452.

44. Singh-Bains MK, Waldvogel HJ, Faull RL. The role of the human globus pallidus in H untington's disease. *Brain Pathol*. 2016;26(6):741-751.

45. Tian Y, Margulies DS, Breakspear M, Zalesky A. Topographic organization of the human subcortex unveiled with functional connectivity gradients. *Nat Neurosci*. 2020;23(11):1421-1432.

46. Alexander GE, DeLong MR, Strick PL. Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annu Rev Neurosci*. 1986;9(1):357-381.

47. Parent A, Hazrati LN. Functional anatomy of the basal ganglia. I. The cortico-basal ganglia-thalamo-cortical loop. *Brain Res Rev*. 1995;20(1):91-127.

48. Pessiglione M, Schmidt L, Draganski B, et al. How the brain translates money into force: a neuroimaging study of subliminal motivation. *science*. 2007;316(5826):904-906.

49. Prodoehl J, Corcos DM, Vaillancourt DE. Basal ganglia mechanisms underlying precision grip force control. *Neurosci Biobehav Rev*. 2009;33(6):900-908. doi:10.1016/j.neubiorev.2009.03.004

50. Saad P, Shendrik KS, Karroum PJ, Azizi H, Jolayemi A. The Anterior Globus Pallidus Externus of Basal Ganglia as Primarily a Limbic and Associative Territory. *Cureus*. Published online December 2, 2020. doi:10.7759/cureus.11846

51. Saga Y, Hoshi E, Tremblay L. Roles of multiple globus pallidus territories of monkeys and humans in motivation, cognition and action: an anatomical, physiological and pathophysiological review. *Front Neuroanat*. 2017;11:30.

52. Xia J, Lin X, Yu T, et al. Aberrant functional connectivity of the globus pallidus in the modulation of the relationship between childhood trauma and major depressive disorder. *J Psychiatry Neurosci*. 2024;49(4):E218-E232. doi:10.1503/jpn.240019

53. Nelson AJD. The anterior thalamic nuclei and cognition: A role beyond space? *Neurosci Biobehav Rev*. 2021;126:1-11. doi:10.1016/j.neubiorev.2021.02.047

54. Wright NF, Vann SD, Aggleton JP, Nelson AJD. A Critical Role for the Anterior Thalamus in Directing Attention to Task-Relevant Stimuli. *J Neurosci*. 2015;35(14):5480-5488. doi:10.1523/JNEUROSCI.4945-14.2015

55. George K, Das JM. Neuroanatomy, thalamocortical radiations. Published online 2019.

56. Kamali A, Milosavljevic S, Gandhi A, et al. The cortico-limbo-thalamo-cortical circuits: an update to the original papez circuit of the human limbic system. *Brain Topogr*. 2023;36(3):371-389.

57. Snell RS. *Clinical Neuroanatomy*. Lippincott Williams & Wilkins; 2010.

58. De Benedictis A, Rossi-Espagnet MC, De Palma L, Carai A, Marras CE. Networking of the Human Cerebellum: From Anatomo-Functional Development to Neurosurgical Implications. *Front Neurol*. 2022;13:806298. doi:10.3389/fneur.2022.806298

59. Purves D, Augustine GJ, Fitzpatrick D, et al. Projections to the cerebellum. In: *Neuroscience. 2nd Edition*. Sinauer Associates; 2001.

60. Asmussen MJ, Jacobs MF, Lee KGH, Zapallow CM, Nelson AJ. Short-Latency Afferent Inhibition Modulation during Finger Movement. J. Chacron M, ed. *PLoS ONE*. 2013;8(4):e60496. doi:10.1371/journal.pone.0060496

61. Davare M, Parikh PJ, Santello M. Sensorimotor uncertainty modulates corticospinal excitability during skilled object manipulation. *J Neurophysiol*. 2019;121(4):1162-1170. doi:10.1152/jn.00800.2018

62. Tokimura H, Di Lazzaro V, Tokimura Y, et al. Short latency inhibition of human hand motor cortex by somatosensory input from the hand. *J Physiol*. 2000;523(Pt 2):503.

63. Du Y, Lin SD, Wu XQ, et al. Ventral posteromedial nucleus of the thalamus gates the spread of trigeminal neuropathic pain. *J Headache Pain*. 2024;25(1):140.

64. Geiger LS, Moessnang C, Schäfer A, et al. Novelty modulates human striatal activation and prefrontal–striatal effective connectivity during working memory encoding. *Brain Struct Funct*. 2018;223(7):3121-3132. doi:10.1007/s00429-018-1679-0

65. Rosa GH de M, Moretto GH, Zhang K, Chagas T de J, Araujo JE de. Modulation of handgrip strength by contralateral hip flexor motor overflow: randomized controlled trial. *Acta Fisiátrica*. Published online 2025:88-94.

66. Aminoff EM, Kveraga K, Bar M. The role of the parahippocampal cortex in cognition. *Trends Cogn Sci*. 2013;17(8):379-390. doi:10.1016/j.tics.2013.06.009

67. Catani M. The connectional anatomy of the temporal lobe. *Handb Clin Neurol*. 2022;187:3-16.

68. Urbini N, McNabb CB, Jones DK, et al. The cognitive cerebellum: linking microstructure to cognitive functions in a healthy population. *NeuroImage*. 2025;317:121356. doi:10.1016/j.neuroimage.2025.121356

69. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med*. 2015;12(3):e1001779. doi:10.1371/journal.pmed.1001779

70. Gaser C, Dahnke R, Thompson PM, et al. CAT: a computational anatomy toolbox for the analysis of structural MRI data. *Gigascience*. 2024;13:giae049.

71. Wagner AS, Waite LK, Wierzba M, et al. FAIRly big: A framework for computationally reproducible processing of large-scale data. *Sci Data*. 2022;9(1):80.

72. Ashburner J, Friston KJ. Diffeomorphic registration using geodesic shooting and Gauss–Newton optimisation. *Neuroimage*. 2011;55(3):954-967.

73. Schaefer A, Kong R, Gordon EM, et al. Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cereb Cortex*. 2018;28(9):3095-3114. doi:10.1093/cercor/bhx179

74. Tian Y, Margulies DS, Breakspear M, Zalesky A. Topographic organization of the human subcortex unveiled with functional connectivity gradients. *Nat Neurosci*. 2020;23(11):1421-1432. doi:10.1038/s41593-020-00711-6

75. Diedrichsen J, Balsters JH, Flavell J, Cussans E, Ramnani N. A probabilistic MR atlas of the human cerebellum. *NeuroImage*. 2009;46(1):39-46. doi:10.1016/j.neuroimage.2009.01.045

76. Alfaro-Almagro F, Jenkinson M, Bangerter NK, et al. Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *NeuroImage*. 2018;166:400-424. doi:10.1016/j.neuroimage.2017.10.034

77. Dale AM, Fischl B, Sereno MI. Cortical surface-based analysis: I. Segmentation and surface reconstruction. *Neuroimage*. 1999;9(2):179-194.

78. Fischl B, Van Der Kouwe A, Destrieux C, et al. Automatically parcellating the human cerebral cortex. *Cereb Cortex*. 2004;14(1):11-22.

79. Fischl B, Sereno MI, Dale AM. Cortical surface-based analysis: II: inflation, flattening, and a surface-based coordinate system. *Neuroimage*. 1999;9(2):195-207.

80. Desikan RS, Ségonne F, Fischl B, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*. 2006;31(3):968-980.

81. Griffanti L, Zamboni G, Khan A, et al. BIANCA (Brain Intensity AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities. *Neuroimage*. 2016;141:191-205.

82. Smith S, Alfaro-Almagro F, Miller KL. UK Biobank brain imaging documentation. *Publ Dec*. Published online 2020.

83. Zou QH, Zhu CZ, Yang Y, et al. An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: fractional ALFF. *J Neurosci Methods*. 2008;172(1):137-141.

84. Deshpande G, LaConte S, Peltier S, Hu X. Integrated local correlation: a new measure of local coherence in fMRI data. *Hum Brain Mapp*. 2009;30(1):13-23.

85. Saad ZS, Reynolds RC, Jo HJ, et al. Correcting brain-wide correlation differences in resting-state FMRI. *Brain Connect*. 2013;3(4):339-352.

86. Kasper J, Eickhoff SB, Caspers S, et al. Local synchronicity in dopamine-rich caudate nucleus influences Huntington's disease motor phenotype. *Brain J Neurol*. 2023;146(8):3319-3330.

87. Tae WS, Ham BJ, Pyun SB, Kang SH, Kim BJ. Current clinical applications of diffusion-tensor imaging in neurological disorders. *J Clin Neurol*. 2018;14(2):129-140.

88. Kitzbichler MG, Martins D, Bethlehem RA, et al. Two human brain systems micro-structurally associated with obesity. *Elife*. 2023;12:e85175.

89. De Groot M, Vernooij MW, Klein S, et al. Improving alignment in tract-based spatial statistics: evaluation and optimization of image registration. *Neuroimage*. 2013;76:400-411.

90. Baykara E, Gesierich B, Adam R, et al. A novel imaging marker for small vessel disease based on skeletonization of white matter tracts and diffusion histograms. *Ann Neurol*. 2016;80(4):581-592.

91. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in python. *J Mach Learn Res*. 2011;12:2825-2830.

92. Pearl J. Causal diagrams for empirical research. *Biometrika*. 1995;82(4):669-710.

93. Pearl J. Causal inference in statistics: An overview. *Stat Surv*. 2009;3(none). doi:10.1214/09-SS057

94. Wysocki AC, Lawson KM, Rhemtulla M. Statistical Control Requires Causal Justification. *Advances in Methods and Practices in Psychological Science*. 2022;5(2).

95. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;30.

96. Aas K, Jullum M, Løland A. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artif Intell*. 2021;298:103502.

97. Owen G. Multilinear extensions of games. *Manag Sci*. 1972;18(5-part-2):64-79.

98. Owen G. Values of games with a priori unions. In: *Mathematical Economics and Game Theory: Essays in Honor of Oskar Morgenstern*. Springer; 1977:76-88.

6 Discussion

6.1 Integrative summary of findings across studies

6.1.1 From methodological challenge to biological insight

The overarching goal of this dissertation was to elucidate system-level neural architecture that underpins HGS, in order to better understand and explain its robust predictive value for diverse health outcomes. This was achieved using machine learning on large-scale, multi-modal neuroimaging data. Addressing this question required a critical examination of the methodological foundations upon which such investigations rest. Consequently, the thesis was structured as a sequential research program progressing from the identification of methodological obstacles (Study 1), development of theoretically grounded solutions (Study 2), and application of these solutions to a concrete neurobiological question (Study 3). This rigorous methodological consideration of issues such as confounding and feature multicollinearity allowed for empirical neurobiological interpretations in an unbiased and reliable manner.

Across studies, a common methodological insight emerged that brain-behaviour prediction in observational neuroimaging is feasible but fragile. Neuroimaging-based predictive models operate under conditions of low signal-to-noise ratio, multicollinearity among neural features, and substantial influence of third variables (neither features nor target) that covary with both brain measures and behavioural outcomes. This thesis demonstrates that these limitations do not invalidate predictive modelling as a neuroscientific tool per se, but that these conditions require a deliberate synthesis of data-driven methodology incorporating domain-specific theoretical and biological knowledge. This leads directly to the distinction of the intended utility of a model between outcome prediction or neurobiological explanation. While general methodological concerns apply to both scenarios, this thesis focuses methodologically and empirically on the later. In this realm, study 1 reviews challenges in cases of pure domain-content-focus without methodological rigor, an imbalance that risks inflated performance estimates, limited generalizability, or overconfident neurobiological conclusions. Taking the opposite perspective, study 2 demonstrates that purely data-driven approaches are particularly inadequate when addressing confounding, which requires the inclusion of causal reasoning based on domain-knowledge. In the absence of such reasoning, models may exploit spurious associations and again, overpromise biological insight. In convergence, study 3 illustrates what becomes neurobiologically visible – despite limited signals – once these issues are addressed and complemented by interpretation strategies that account for feature multicollinearity. Concretely, study 3 elucidated the neural signatures most used by models to predict HGS, providing insights into the system-level brain-architecture supporting HGS. Together, the thesis demonstrates that robust neuroscientific insight emerges not from methodological or theoretical considerations alone, but from their systematic integration.

From a neurobiological perspective, the empirical findigns of study 3 converge on a coherent system-level interpretation of HGS. Predictive signal was not diffusely distributed across the cortex nor dominated by functional connectivtiy measures, but is concentrated in a distributed yet specific set of white matter pathways and subcortical structures. In particular, the integrity of long-range tracts, supporting sensorimotor and associative transmission and integration, as well as structural properties of basal ganglia and thalamic nuclei, emerged as central contributors. These findings were remarkably consistent across modelling strategies and sexes, underscoring their robustness. Rather than reflecting isolated or focal motor representations, individual differences in HGS appear to index the efficiency and integrity of distribtued neural systems that synchronize motor execution with sensory feedback, subcortical modulation, and higher-order control processes. This pattern supports the interpretation of HGS as a maker of system-level neural integrity and coordination rather than localized motor strength alone.

6.1.2 Conceptual synthesis: Why is HGS such a versatile marker?

At first glance, and particularly in light of the strong influence of body composition and other non-neuronal factors, HGS may appear to be a predominantly peripheral musculoskeletal measure. Indeed, muscle mass, biomechanics, and general somatic health contribute substantially to inter-individual variance in HGS and constitute important sources of confounding in neuroimaging analyses. However, the neuronal findings synthesized in this thesis demonstrate that HGS is neither a pure muscle measure nor a simple cortical motor output. Instead, they suggest that HGS constitutes a global read-out of the brain's capacity to coordinate, transmit, and integrate signals across hierarchical levels of neural organization and across task-specialized motor and non-motor systems. The execution of maximal grip force does not rely solely on primary motor output, but requires the integrity of ascending and descending pathways, subcortical gating and scaling mechanisms, cerebellar timing, and extra-motor processes related to attention, motivation, and affective drive. In this sense, HGS can be conceptualized as a global stress test of distributed neural integrity.

Despite its apparent simplicity, this system-level dependence provides a parsimonious explanation for the remarkable breadth of epidemiological associations observed for HGS. Because HGS reflects the integrity and processes of distributed neural systems, it is sensitive to a wide range of pathological and age-related aspects that compromise global, but specific brain health, including cerebrovascular burden, neurodegeneration, and dysregulation of neuromodulatory systems. As a result, HGS exhibits robust links to mortality, cognitive decline, frailty, and psychiatric vulnerability. The central conceptual takeaway emerging from the work done in this thesis is therefore that HGS is a uniquely informative phenotype because it is effortful and systematically demanding as it requires the coordinated mobilization of motor, sensory, cognitive, and motivational resources toward a unified physiological goal. As such, HGS occupies a distinctive position among behavioural phenotypes, offering an accessible yet biologically rich window into global brain and organismal health across the adult lifespan.

6.2 Methodological contributions to brain–behaviour predictive modelling

The methodological studies included in this thesis are not auxiliary to the neuroscientific investigation but constitute a necessary foundation for it. Observational large-scale neuroimaging data together with machine learning offer wide opportunities for the study of brain-behaviour relationships. Yet, under the accompanying challenges, methodological choices can determine whether predictive models yield neuroscientifically meaningful insights or merely reflect spurious or indirect associations. The first two studies therefore address core methodological challenges that must be resolved to build brain-based predictions of HGS that allow for neurobiologically interpretation of brain features informative for the model.

6.2.1 From prediction performance to neuroscientific meaning

A centrally stressed conceptual perspective in this thesis is that prediction is not explanation. From this directly follows the distinction between predictive success of a model in the sense of high predictive performance and explanatory success as in providing informative neurobiological explanations. While those two are not necessarily mutually exclusive, high performance can arise from non-neural correlates of feaures or target, in the case of HGS for example age, sex, body size, or peripheral physiology. However, predictive performance is determined by prediction metrics which can only depict accuracy but not its neurobiological informativeness. Consequently, uncritical reliance on prediction metrics risks misinterpretation of effective model functioning on neuronal constructs, when actual model success is based on demographic predictors or proxies of peripheral biological factors. This thesis therefore shifts the focus from the question of whether a phenotype can be predicted to the question of what the prediction can tell us about neurobiology.

6.2.2 Confounding as a reflection of biological entanglement

Confounding represents a particularly critical challenge in brain-behaviour predictive modelling because variables commonly labeled as confounders in neuroimaging are often biologically meaningful and embedded in shared causal pathways linking brain, body, behaviour, and environmental factors. This entanglement renders determination of causal influence-directionalities difficult, whereas correlative structures between variables remain the same. Problematically, correlation-based confounder selection is inherently unreliable as confounding is a causal concept (Wysocki et al., 2022). The complexity of cause-effect structures in neurobiology therefore makes it particularly difficult to distinguish confounders from mediators or colliders. To overcome limitations of current practices of correlation-based confounder handling in neuroimaging research, study 2 therefore introduces causal inference concepts, including directed acyclic graphs and graph-theoretical adjustment rules, into the context of associative neuroimaging prediction, with the aim to offer a pragmatic and adoptable solution for confounder handling in neuroimaging. The core idea relies on a theory-data hybrid approach by integration of theoretical cause-effect domain knowledge into data-driven predictive modelling. It thereby improves transparency of confounder treatment decisions, makes assumptions explicit, and

allows for principled differentiation between types of third variables. Study 2 thereby briges two gaps: First, between disciplines, namely through the combination of causal inference tools and associative machine learning, and second between paradigms, concretely between theory- and data-driven approaches. Importantly, despite causally informed, deconfounded models remain associative in nature: Causal reaosning reduces spurious associations and improve interpretatbility but it does not convert prediction into causal inference.

Beyond confounder identification, the biological entanglement of variables also challenges standard confounder adjustment strategies. Linear feature residualization, while commonly applied in neuroimaging-based predictive modelling (Chyzhyk et al., 2022; Snoek et al., 2019), can be insufficient in high-dimensional and non-linear machine learning settings and cannot account for complex confounding structures (Hamdan et al., 2023a). Study 2 of this thesis therefore discussed potential alternatives and the implementation of sex-split models in study 3 is a direct consequence with the aim to reduce non-linear sex-related confounding that cannot be captured by linear feature residualization. Nonetheless, future work is needed to develop theoretically grounded but pragmatically feasible improvements to linear feature residulaization for confounder adjustment in neuroimaging-based multivariate and non-linear machine learning.

6.2.3 Multicollinearity-aware interpretation as a prerequisite for system neuroscience

A further methodological concern in this thesis is the interpretation of predictive models in the presence of strong multicollinearties, which is inherent to neuroimaging features. Correlated features can be used interchangeably by predictive models, rendering feature-wise importance estimates unstable and potentially misleading (Jiang, Woo, et al., 2022; Kraha et al., 2012). To address this issue, study 2 adopts a hierarchical-cluster-based interpretation strategy that aggregates features into higher-order structures based on their informativeness about the target. Interpretations are performed at every hierarchy of the aggregation process, so that they recover system-level structures and offer insights into cooperative meaning (López & Saboya, 2009). This approach yields more robust and biologically meaningful insights by shifting the focus from isolated brain features to systems and shared signale-carriers, aligning model interpretation with system-level views of brain organization.

Collectively, the methodological contributions of this thesis extend beyond specific application to HGS. Confounding, multicollinearity, and methodological pitfalls such as data leakage or biased model evaluations represent pervasive challenges in neuroimaging-based predictive modelling (Alfaro-Almagro et al., 2021; Benkarim et al., 2021; Chyzhyk et al., 2022; Hamdan et al., 2023b; Rao et al., 2017), with large datasets exacerbating the especially the confounding issue through high sensitivity to artifactual associations (Smith & Nichols, 2018b). By combining a systematic overview of common pitfalls (Study 1), a causally informed framework for confounder handling (Study 2), and a concrete empirical implementation that accounts for the previously identified challenges as well as feature multicollinearity (Study 3), this thesis establishes a concise yet comprehensive methodological

foundation for extracting neurobiologically meaningful information from large-scale brain–behaviour predictive models.

6.3 Neural systems supporting the prediction of HGS

6.3.1 A system-level perspective on HGS

The central neurobiological insight emerging from study 3 is that HGS primarily reflects the efficiency and integrity of neural signal transmission across distributed systems rather than the isolate capacity for force generation within focal motor regions. This interpretation is strongly supported by the dominance of white matter tract integrity measures among the most informative predictors. The predictive relevance of multiple complementary microstructural indices (e.g. diffusion tensor mode (MO), orientation dispersion index (OD), isotropic volume fraction (ISOVF), fractional anisotropy (FA), mean diffusivity (MD)) indicates that HGS depends on diverse properties of white matter organization, including axonal density, orientation coherence, and extracellular diffusion. These measures do not capture the magnitude of the initial motor command, but rather the quality, reliability, and efficiency with which signals are transmitted and integrated across hierarchical levels of the motor system. Consistent with this view, long-range projection and association fibers emerged as central contributors to the prediction of HGS. Ascending somatosensory pathways, particularly the medial lemniscus, and STR, a motor-specific thalamocortical connector, formed a coherent sensorimotor integration axis. Importantly, CST integrity was predictive primarily in conjunction with medial lemniscus integrity, underscoring the critical role of afferent feedback in maximal force production. Efficient sensory feedback is required to stabilize force output, calibrate scaling, and prevent maladaptive central inhibition (Bolognini et al., 2016; Purves, Augustine, Fitzpatrick, Katz, et al., 2001b). Consequently, functional maximal force generation can be limited not through weakness of efferent drive, but through inefficient sensorimotor integration and disrupted internal state estimation (Asan et al., 2022; Nowak & Hermsdörfer, 2006). This aligns with findings that increased somatosensory input from the paretic hand after stroke, for instance by using somatosensory stimulation, may improve motor function (Conforto et al., 2002). Beyond classic sensorimotor pathways, non-motor thalamic radiations (most prominently the anterior thalamic radiation (ATR)) contributed to HGS prediction. The ATR provides structural connectivity between prefrontal regions and thalamic nuclei and supports attention allocation, contextual processing, and executive modulation (Nelson, 2021; Wright et al., 2015). Empirical evidence links ATR microstructure to attention direction under conflicting cues (Mamiya et al., 2018) as well as to greater adherence to physical activity in older adults (Gujral et al., 2018). Its involvement in HGS suggests that maximal grip strength is not a purely motor output, but depends on executive engagement, sustained attention, and cognitive drive required to initiate and maintain voluntary effort. This interpretation is further reinforced by the contribution of long-range association fibres, including the ILF and cingulum bundle (gyrus and hippocampal part), which integrate sensory, attentional, and cognitive information. Recent work indicates that such

associative tracts contribute not only to cognition but also to proprioception and state estimation (Chilvers et al., 2022), providing a structural substrate for the well-established association between HGS and cognitive performance. The prominence of white matter integrity as a predictor of HGS also aligns with evidence from aging-related changes, vascular pathology, and neurodegeneration, where diffuse white matter pathology such as small-vessel disease, as reflected by WMH, was seen to precede cognitive decline and neurodegenerative processes (Brickman et al., 2015; Debette et al., 2019; Prins & Scheltens, 2015; Wardlaw et al., 2015). Even mild white matter damage has been associated with reduced motor speed, coordination, and muscular strength (Sachdev, 2005), and stronger grip was associated with reduced WMH people with major depressive disorder (J. A. Firth et al., 2020). Interestingly though, in our study WMH did not emerge as a strong predictor of HGS but exhibited lower performance than more specific measures of white matter tract integrity. In combination, this suggests that HGS can be understood as a behavioural readout of global, yet specific structure and integrity of defined sub-systems rather than diffuse decline.

Opposing the prominent role of white matter, cortical features contributed comparatively little to HGS prediction. This may appear surprising given the functional imaging evidence demonstrating linear increases in motor cortical activation with increasing force output (Dettmers et al., 1995; Thickbroom et al., 1999; N. S. Ward & Frackowiak, 2003). One explanation however can be that HGS reflects power grip rather than precision grip and places minimal demands on fine motor planning and dexterity. Cortical specialization is particularly critical for skilled, fractioned movements and motor learning. In contrast, power grip relies on robust, redundant, and evolutionarily conserved circuitry, as it can be interpreted as the cortical appropriation of foundational grasping predominantly mediated by spinal and brainstem circuits (Capute & Accardo, 1996; Futagi et al., 2012; Marques De Moraes et al., 2017; Stephens-Sarlós et al., 2025; Zafeiriou, 2004). This hints towards a neural architecture of HGS optimized for efficiency and redundancy, rendering HGS relatively robust to focal cortical variation while remaining sensitive to distributed system integrity.

In contrast, subcortical gray matter emerged as core contributor to HGS across modelling setups and sexes with basal ganglia structures forming one of the most informative clusters. Basal ganglia act in a gate-keeping function, by selecting or inhibiting motor programs or preparing upper motor neuron circuits for initiation of movement (Purves, Augustine, Fitzpatrick, Hall, et al., 2001). This gate-keeping function was particularly attributed to sub-structures such as the STN and GPi (more posterior basal ganglia), exhibiting higher activity with higher force amplitude (Prodoehl et al., 2009; Spraker et al., 2007; Vaillancourt et al., 2004), whereas regions such as the caudate nucelus and anterior putamen (more anterior) previously were more associated with the preparatory phase, showing higher activity based on the predictability of the required force (Wasson et al., 2010). Such a anterior-posterior distinction also underlies the functional connectivity gradient based subcortical parcellation of the globus pallidus as established by (2020) and used in study 3. The globus pallidus along this anterior-posterior gadient is topographically organized in limbic, associative (anterior globus pallidus (aGP)), and sensorimotor

(posterior globus pallidus (pGP)) functional zones (Bertino et al., 2020). Study 3 identified the aGP as one of the dominating predictors, with larger aGP GMV being associated with lower HGS. This is interesting, both w.r.t to the involvement of the aGP in motivational control (reward seeking, aversive avoidance) and processing (goal decision, action selection), cognitive processing and effort-based decision-making, rather than direct motor execution (Pessiglione et al., 2007; Prodoehl et al., 2009; Saad et al., 2020; Saga et al., 2017; Xia et al., 2024), as well as w.r.t. this inverse relationship of higher volume-lower HGS. This combination could reflect less efficient action selection through overactivity concerning goal decision or reward/aversive processing and reinforces that HGS is influenced by cognitive-motivational-affective, i.e. non-motor processes. While such interpretations of volume-based features can be ambigous (hypertrophy, gliosis, developmental variance), the predictive importance of aGP and the respective motivational-affective component is consistent with findings for example of reduced motivation, attention, or affective drive in depressive states being associated with diminished grip strength independent of muscle mass or corticospinal integrity (Cass et al., 2024; J. Firth et al., 2018; J. A. Firth et al., 2020; Rinne et al., 2018). Beyond the basal ganglia, additional involvment of thalamus nuclei in women, highlights the relevance of the thalamus as an integration hub coordinating basal ganglia–cerebellar–cortical loops to constrain force output according to predictive internal models rather than maximal muscle recruitment (Bosch-Bouju et al., 2013; Opri et al., 2019; Takahashi et al., 2021; Wasson et al., 2010).

Together, findings from study 3 position HGS as an emergent property of integrated motor, cognitive, motivational, and neuromodulatory systems rather than as a readout of isolated motor capacity.

6.3.2 Many-to-one mapping and strategic flexibility in HGS

The inter-individual variability in feature importance observed in Study 3 suggests that HGS is not supported by a single canonical architecture, but reflects a fial common behavioural outcome achievable through flexible, rdundant neurobiological strategies. This many-to-one mapping between neural systems and HGS aligns with the conceptualization of HGS as a phenotypic readout into the current state of a complex system. Observed sex-specific patterns are naturally embedded within this framework of strategic flexibility. While HGS represents a standardized behavioural endpoint, neurobiological predictors in women and men exhibited great overlap but also distinct patterns. Rather than implying divergent mechanisms of force generation, these patterns point to variability in how systems spanning sensory integration, subcortical modulation, and cognitive-motivational control are recruited. In an older population such as the present UKB cohort, sex potentially could mirror lifelong differences in occupational exposure, physical activity, or psychosocial stressors. Such exposomic factors may play a role in the recruitment of varying neural strategies for HGS execution. Within this context, sex-specific predictive patterns likely reflect differences in the distribution of reliance across redundant neural systems, rather than residual confounding or categorical biological divergence. That these patterns in sex-stratified, age-adjusted models hints that they represent meaningful vriation in

system-level organization rather than artefacts of scaling or peripheral confounding. HGS thus appears to be supported by multiple viable neural configurations, whose relative weighting is shaped by an individual's lifelong biological and environmental context.

This heterogeneity has implications for vulnerability and decline. If HGS can be maintained through diverse neural configurations, its decline may reflect failures in distinct circuit-level components across individuals. HGS thus serves as a sensitive behavioral readout of individual differences in neural integrity, indexing heterogeneous neurobiological liabilities rather than a uniform pathway of degeneration.

6.3.3 Hand grip strength, aging, and the lifespan–healthspan gap

Understanding HGS as a marker of distributed neural scaffolding provides an explanation for its remarkable predictive power across diverse health outcomes. Strong grip strength may reflect preserved structural integrity on the one hand and on the other hand mirror the capacity to compensate for emerging neural degradation through efficient signal and process rerouting. For example, subcortical and cerebellar involvement, including sex-specific contributions of pontocerebellar afferents (middle cerebellar peduncle (MCP) in study 3), aligns with theories of neural compensation in aging, where increased reliance on subcortical circuits supports stable motor output in the presence of degradation of CST integrity and M1 excitability (Noble et al., 2011). A high ability for rerouting and compensation is also benefitial in impaired motor system conditions such as post-stroke. There, for example both, up-regulation of subcortical systems after cortical stroke (Ejaz et al., 2018) as well as higher recruitment of secondary motor networks after subcortical stroke (N. S. Ward et al., 2006) could be observed. Although such strategies do not necessarily lead to successful compensation and restoration of function (e.g. unwished mirror movements), they do highlight the flexibility of the motor system for reorganization. The reason for HGS serving as such a verstile marker in health and disease may therfore be that it reflects the system's ability for compensation and rerouting. By relying on wide-ranging structural (tract) integrity that likely supports signal integration across systems, HGS becomes sensitive to vulnerability across functional neuronal domains. Recruitment of extra-motor systems as found in study 3 aligns with broad evidence that motor behaviour in general and HGS in particular includes the recruitment of extra-motor structures and networks both in healthy subjects and in motor recovery, for example after stroke (Guo et al., 2017; Johnson et al., 2017; Lam et al., 2018; Mattos et al., 2023; Park et al., 2011; Rezaei et al., 2025). More specifically, previous research has found that motor and cognitive processes develop, interact and decline bidirectionally across the lifespan (Basile & Sardella, 2021) and that physical fitness interventions possibly improve cognition and delay dementia onset (Ahlskog et al., 2011; Berryman et al., 2013; Guiney & Machado, 2013; Guo et al., 2017; Voelcker-Rehage & Niemann, 2013). Together this provides plausible explanations why HGS, albeit a simple motor measure, is predictive not only of physical decline and frailty but also cognitive decline and general morbidity. It supports the conceptualization of HGS as a system-level readout.

Age-related alterations of HGS further highlight its potential for the lifespan-healthspan challenge. Neuromodulatory alterations, vascular burden, and the vulnerability of long-range white matter tracts can contribute to declining HGS with aging, often beginning in midlife before overt motor or cognitive impairments. Evidence from longitudinal studies for example shows that midlife HGS predicts functional limitations decades later (Rantanen et al., 1999), suggesting that HGS captures early neural vulnerability, not only late-stage disability. By elucidating the distributed neural predictors of HGS, study 3 provides explanatory insights for this preditive capacity and suggests that those neural markers may allow even earlier identification of individual risk. The specific neuronal contributors to HGS thereby have the opportunitiy to be more informative than diffuse global markers such as WMHs, thanks to their higher specificity and sub-group informativeness. In the context of increasing lifespans without proportional extensions in healthspan, HGS already functions as a versatile marker at the behavioural level. Extending this utility, the identified specific and integrative underlying neuronal structures and systems, sensitive to early biological aging, can offer potential for individually and systemically targeted and earlier diagnoses, interventions and lifestyle adaptations.

6.4 Limitations, implications, and future directions

The findings in this thesis should be interpreted within the constraints imposed by the underlying data and design. All analyses were conducted in the UKB. While offering unprecedented access to large sample sizes, restrictions apply due to volunteer bias, survivor effects, and our focus of analyses of a healthy, geographically restricted sub-population, effecting variance in health-related traits and limit generalizability to clinical, more vulnerable and generally more diverse populations. After validation and evaluation in a CV scheme, predictive models were used on previously held-out and hence unseen data. In future work it will be interesting to evaluate predictions and interpretaitons on an external non-UKB cohort. Additionally, moderate predictive performances require caution about neurobiological interpretations as all interpretations only explain model behaviour in the realm of low explained variance (H. Chen et al., 2020; J. Chen, Ooi, et al., 2023). While causal inference tools were used for confounder selection and justification for study 3 as promoted in study 2, the observational nature of the data and associative nature of the used machine learning setup preclude any causal interpretation of results in study 3.

Despite contraints, the insights gained in this thesis carry several implications and motivate future directions. Methodologcially, they underscore the opportunities for neurobiological insights when combining considerate data-driven methodology with theoretic domain knowledge, for confounder handling but also beyond, especially given the highly confounded and low neuronal signal-to-noise ratios in neuroimaging based predictive modelling. Coneptually, the findings support the conceptualization of HGS not as a peripheral motor measure, but as a behavioural index of distributed but specific neural integrity, integrating motor, cognitive, motivational, and neuromodulatory systems. This persepctive helps explain its robust association with aging- and impariment related outcomes and

positions HGS as a low-cost, scalable marker of biological resilience. By identifying specific neural systems that constrain HGS, this thesis further suggests that grip strength may serve not only as a screening tool but also as a window into the brain's global communication architecture, with potential relevance for early risk stratification and longitudinal monitoring. Future research using longitudinal and multi-measure designs (e.g. integration of vascular, metabolic, hormonal and inflammatory markers) should evaluate whether these HGS-linked neural signatures serve as leading indicators with higher prognostic and interventional value than the behavioural measure itself.

6.5 Conclusion

This thesis characterizes the system-level neural architecture underlying HGS and embeds it within a framework for brain-behaviour predictive modelling in large-scale observational neuroimaging. By integrating critical evaluations of methodological challenges, practical solutions for confounder selection and justification, and multi-modal neuroimaging application, the work demonstrates that robust neuroscientific insights require a combination of data- and theory-driven paradigms, underscoring that concerns such as low predictive performance, confounding, and feature multicollinearities must be handled as inferential constraints rather than as technical afterthoughts. Findings indicate that (1) inappropriately designed predictive models can obscure neural meaning by privileging demographic or peripheral correlates, whereas causally-informed deconfounding and multicollinearity-aware model explanation enable interpretable extraction of neural signal; and (2) when these conditions are met, HGS is best understood as an emergent readout of distributed but specific signal transmission and integration across white matter pathways and subcortical control systems, rather than as a localized cortical motor output. HGS thus reflects the integrity and compensatory capacity of large-scale neural communication architectures, providing parsimonious explanation for its robust associations with aging. Related decline, cognition, and mortality. Yet, open questions remain regarding the temporal evolution, modifiability, and cross-cohort generalization of these neural signatures. As population neuroscience continues to advance through larger, more diverse samples and multimodal integration, the methodological and conceptual synthesis presented here offers a scalable framework for extracting biologically meaningful insight from complex brain-behaviour relationships and for situating simple behavioural phenotypes, such as HGS, as windows into organismal resilience across the lifespan.

8 References

Abalay, A., Cemel, Y., Varhan, B., & Yavuzer, M. G. (2024). The Relationships Between Wrist Joint Position Sense, Anthropometric Characteristics and Grip Strength of the Hand in Healthy Individuals. *Bakirkoy Tip Dergisi / Medical Journal of Bakirkoy*, 182–188. https://doi.org/10.4274/BMJ.galenos.2023.2023.4-4

Abe, T., & Loenneke, J. P. (2015). Handgrip strength dominance is associated with difference in forearm muscle size. *Journal of Physical Therapy Science*, *27*(7), 2147–2149. https://doi.org/10.1589/jpts.27.2147

Ahlskog, J. E., Geda, Y. E., Graff-Radford, N. R., & Petersen, R. C. (2011). Physical exercise as a preventive or disease-modifying treatment of dementia and brain aging. *Mayo clinic proceedings*, *86*, Article 9.

Alfaro-Almagro, F., McCarthy, P., Afyouni, S., Andersson, J. L., Bastiani, M., Miller, K. L., Nichols, T. E., & Smith, S. M. (2021). Confound modelling in UK Biobank brain imaging. *NeuroImage*, *224*, 117002.

Amante-da-Rosa-Cardoso, A., Cunha-Rodrigues, M., Guerra, R., Mendes, J., Sousa, A., Sousa-Santos, A., Silva, C., Santos, A., Borges, N., Amaral, T., & Valdiviesso, R. (2025). Association between handgrip strength and whole-body derived bone mineral density in adults: A cross-sectional study. *Athena Health & Research Journal*, *2*(Suppl.). https://doi.org/10.62741/ahrj.v2iSuppl..89

Andrade, C. (2013). Signal-to-noise ratio, variability, and their relevance in clinical trials. *The Journal of clinical psychiatry*, *74*(5), 19584.

Arbabshirani, M. R., Plis, S., Sui, J., & Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*, *145*, 137–165. https://doi.org/10.1016/j.neuroimage.2016.02.079

Asan, A. S., McIntosh, J. R., & Carmel, J. B. (2022). Targeting Sensory and Motor Integration for Recovery of Movement After CNS Injury. *Frontiers in Neuroscience*, *15*, 791824. https://doi.org/10.3389/fnins.2021.791824

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience, Vol 34*, *28*(1), 403–450.

Baker, S. N., & Perez, M. A. (2017). Reticulospinal Contributions to Gross Hand Function after Human Spinal Cord Injury. *The Journal of Neuroscience*, *37*(40), 9778–9784. https://doi.org/10.1523/JNEUROSCI.3368-16.2017

Barnes, L., Bauer, P., Bordes Edgar, V., Gershon, R., Giordani, B., Guralnik, J., Hendrie, H., Hook, J., & Lin, F. (2025). Motor. *NIHToolbox*. https://nihtoolbox.org/domain/motor/

Basile, G., & Sardella, A. (2021). From cognitive to motor impairment and from sarcopenia to cognitive impairment: A bidirectional pathway towards frailty and disability. *Aging Clinical and Experimental Research*, *33*(2), 469–478. https://doi.org/10.1007/s40520-020-01550-y

Beauchet, O., Allali, G., Montero-Odasso, M., Sejdić, E., Fantino, B., & Annweiler, C. (2014). Motor phenotype of decline in cognitive performance among community-dwellers without dementia: Population-based study and meta-analysis. *PloS one*, *9*(6), e99318.

Bekena, S., Singh, R. K., Zhu, Y., Carr, D. B., & Babulal, G. M. (2025). Sensorimotor function as an early marker of cognitive decline and alzheimer's biomarker burden. *GeroScience*. https://doi.org/10.1007/s11357-025-02055-0

Bello, L., Freyschlag, C. F., & Rech, F. (2021). Motor control. In *Intraoperative mapping of cognitive networks: Which tasks for which locations* (S. 3–19). Springer.

Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. *Journal of machine learning research*, *5*(Sep), 1089–1105.

Benkarim, O., Paquola, C., Park, B., Kebets, V., Hong, S.-J., de Wael, V., Zhang, S., Yeo, B. T. T., Eickenberg, M., Ge, T., Poline, J.-B., Bernhardt, B., & Bzdok, D. (2021). The Cost of Untracked Diversity in Brain-Imaging Prediction. *bioRxiv*, 34. https://doi.org/10.1101/2021.06.16.448764

Berisha, V., Krantsevich, C., Hahn, P. R., Hahn, S., Dasarathy, G., Turaga, P., & Liss, J. (2021). Digital medicine and the curse of dimensionality. *Npj Digital Medicine*, *4*(1), 153. https://doi.org/10.1038/s41746-021-00521-5

Berryman, N., Bherer, L., Nadeau, S., Lauzière, S., Lehr, L., Bobeuf, F., Kergoat, M. J., Vu, T. T. M., & Bosquet, L. (2013). Executive functions, physical fitness and mobility in well-functioning older adults. *Experimental gerontology*, *48*(12), 1402–1409.

Bertino, S., Basile, G. A., Anastasi, G., Bramanti, A., Fonti, B., Cavallaro, F., Bruschetta, D., Milardi, D., & Cacciola, A. (2020). Anatomical Characterization of the Human Structural Connectivity between the Pedunculopontine Nucleus and Globus Pallidus via Multi-Shell Multi-Tissue Tractography. *Medicina*, *56*(9), 452. https://doi.org/10.3390/medicina56090452

Bohannon, R. W. (2019). Grip strength: An indispensable biomarker for older adults. *Clinical interventions in aging*, 1681–1691.

Bolognini, N., Pascual-Leone, A., & Fregni, F. (2009). Using non-invasive brain stimulation to augment motor training-induced plasticity. *Journal of NeuroEngineering and Rehabilitation*, *6*(1), 8. https://doi.org/10.1186/1743-0003-6-8

Bolognini, N., Russo, C., & Edwards, D. J. (2016). The sensory side of post-stroke motor rehabilitation. *Restorative Neurology and Neuroscience*, *34*(4), 571–586. https://doi.org/10.3233/RNN-150606

Borich, M. R., Brodie, S. M., Gray, W. A., Ionta, S., & Boyd, L. A. (2015). Understanding the role of the primary somatosensory cortex: Opportunities for rehabilitation. *Neuropsychologia*, *79*, 246–255. https://doi.org/10.1016/j.neuropsychologia.2015.07.007

Borra, E., Gerbella, M., Rozzi, S., & Luppino, G. (2017). The macaque lateral grasping network: A neural substrate for generating purposeful hand actions. *Neuroscience & Biobehavioral Reviews*, *75*, 65–90.

Bosch-Bouju, C., Hyland, B. I., & Parr-Brownlie, L. C. (2013). Motor thalamus integration of cortical, cerebellar and basal ganglia information: Implications for normal and parkinsonian conditions. *Frontiers in Computational Neuroscience*, *7*. https://doi.org/10.3389/fncom.2013.00163

Brickman, A. M., Zahodne, L. B., Guzman, V. A., Narkhede, A., Meier, I. B., Griffith, E. Y., Provenzano, F. A., Schupf, N., Manly, J. J., Stern, Y., Luchsinger, J. A., & Mayeux, R. (2015). Reconsidering harbingers of dementia: Progression of parietal lobe white matter hyperintensities predicts Alzheimer's disease incidence. *Neurobiology of Aging*, *36*(1), 27–32. https://doi.org/10.1016/j.neurobiolaging.2014.07.019

Buchman, A. S., & Bennett, D. A. (2011). Loss of motor function in preclinical Alzheimer's disease. *Expert review of neurotherapeutics*, *11*(5), 665–676.

Buchman, A. S., Boyle, P. A., Wilson, R. S., Bienias, J. L., & Bennett, D. A. (2007). Physical activity and motor decline in older persons. *Muscle & Nerve*, *35*(3), 354–362. https://doi.org/10.1002/mus.20702

Buracchio, T., Dodge, H. H., Howieson, D., Wasserman, D., & Kaye, J. (2010). The trajectory of gait speed preceding mild cognitive impairment. *Archives of neurology*, *67*(8), 980–986.

Camicioli, R., Howieson, D., Oken, B., Sexton, G., & Kaye, J. (1998). Motor slowing precedes cognitive impairment in the oldest old. *Neurology*, *50*(5), 1496–1498.

Capute, A. J., & Accardo, P. J. (1996). The infant neurodevelopmental assessement: A clinical interpretive manual for CAT-CLAMS in the first Two years of life, part 2. *Current problems in pediatrics*, *26*(8), 265–306.

Carson, R. G. (2018). Get a grip: Individual variations in grip strength are a marker of brain health. *Neurobiology of Aging*, *71*, 189–222. https://doi.org/10.1016/j.neurobiolaging.2018.07.023

Cass, K., Blalock, D. V., Manwaring, J., Wesselink, D., Lundberg, C., & Mehler, P. S. (2024). Changes in grip strength, depression, and cognitive functioning during medical stabilization for anorexia

nervosa: Exploring the utility of grip strength as a marker of severity. *Journal of Rehabilitation Research and Practice*, *5*(1), 14–22.

Chai, L., Zhang, D., & Fan, J. (2024). Comparison of grip strength measurements for predicting all-cause mortality among adults aged 20+ years from the NHANES 2011–2014. *Scientific Reports*, *14*(1), 29245. https://doi.org/10.1038/s41598-024-80487-y

Chakrabarty, S., & Martin, J. H. (2011). Co-development of proprioceptive afferents and the corticospinal tract within the cervical spinal cord. *European Journal of Neuroscience*, *34*(5), 682–694.

Chekroud, A. M., Hawrilenko, M., Loho, H., Bondar, J., Gueorguieva, R., Hasan, A., Kambeitz, J., Corlett, P. R., Koutsouleris, N., Krumholz, H. M., Krystal, J. H., & Paulus, M. (2024). Illusory generalizability of clinical prediction models. *Science*, *383*(6679), 164–167. https://doi.org/10.1126/science.adg8538

Chen, H., Janizek, J. D., Lundberg, S., & Lee, S.-I. (2020). *True to the Model or True to the Data?* (arXiv:2006.16234). arXiv. http://arxiv.org/abs/2006.16234

Chen, J., Ooi, L. Q. R., Tan, T. W. K., Zhang, S., Li, J., Asplund, C. L., Eickhoff, S. B., Bzdok, D., Holmes, A. J., & Yeo, B. T. T. (2023). Relationship between prediction accuracy and feature importance reliability: An empirical and theoretical study. *NeuroImage*, *274*, 120115. https://doi.org/10.1016/j.neuroimage.2023.120115

Chen, J., Patil, K. R., Yeo, B. T. T., & Eickhoff, S. B. (2023). Leveraging Machine Learning for Gaining Neurobiological and Nosological Insights in Psychiatric Research. *Biological Psychiatry*, *93*(1), 18–28. https://doi.org/10.1016/j.biopsych.2022.07.025

Chilvers, M. J., Low, T. A., & Dukelow, S. P. (2022). Beyond the dorsal column medial lemniscus in proprioception and stroke: A white matter investigation. *Brain Sciences*, *12*(12), 1651.

Chong, J. S. X., Chua, K. Y., Ng, K. K., Chong, S. W., Leong, R. L. F., Chee, M. W. L., Koh, W. P., & Zhou, J. H. (2024). Higher handgrip strength is linked to higher salience ventral attention functional network segregation in older adults. *Communications Biology*, *7*(1), 214. https://doi.org/10.1038/s42003-024-05862-x

Chyzhyk, D., Varoquaux, G., Milham, M., & Thirion, B. (2022). How to remove or control confounds in predictive models, with applications to brain biomarkers. *GigaScience*, *11*, giac014. https://doi.org/10.1093/gigascience/giac014

Clark, B. C. (2019). Neuromuscular changes with aging and sarcopenia. *The Journal of frailty & aging*, *8*(1), 7–9.

Conforto, A. B., Kaelin-Lang, A., & Cohen, L. G. (2002). Increase in hand muscle strength of stroke patients after somatosensory stimulation. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, *51*(1), 122–125.

Craig, A. D. (2009). How do you feel — now? The anterior insula and human awareness. *Nature Reviews Neuroscience*, *10*(1), 59–70. https://doi.org/10.1038/nrn2555

Cruz-Jentoft, A. J., Bahat, G., Bauer, J., Boirie, Y., Bruyère, O., Cederholm, T., Cooper, C., Landi, F., Rolland, Y., Sayer, A. A., Schneider, S. M., Sieber, C. C., Topinkova, E., Vandewoude, M., Visser, M., Zamboni, M., Writing Group for the European Working Group on Sarcopenia in Older People 2 (EWGSOP2), and the Extended Group for EWGSOP2, Bautmans, I., Baeyens, J.-P., … Schols, J. (2019). Sarcopenia: Revised European consensus on definition and diagnosis. *Age and Ageing*, *48*(1), 16–31. https://doi.org/10.1093/ageing/afy169

Dacre, J., Colligan, M., Clarke, T., Ammer, J. J., Schiemann, J., Chamosa-Pino, V., Claudi, F., Harston, J. A., Eleftheriou, C., Pakan, J. M. P., Huang, C.-C., Hantman, A. W., Rochefort, N. L., & Duguid, I. (2021). A cerebellar-thalamocortical pathway drives behavioral context-dependent movement initiation. *Neuron*, *109*(14), 2326-2338.e8. https://doi.org/10.1016/j.neuron.2021.05.016

Danielson, T. L., Gould, L. A., DeFreitas, J. M., MacLennan, R. J., Ekstrand, C., Borowsky, R., Farthing, J. P., & Andrushko, J. W. (2024). Activity in the pontine reticular nuclei scales with handgrip force in humans. *Journal of Neurophysiology*, *131*(5), 807–814.

Davis, M., Wang, Y., Bao, S., Buchanan, J. J., Wright, D. L., & Lei, Y. (2022). The Interactions Between Primary Somatosensory and Motor Cortex during Human Grasping Behaviors. *Neuroscience*, *485*, 1–11. https://doi.org/10.1016/j.neuroscience.2021.11.039

Debette, S., Schilling, S., Duperron, M.-G., Larsson, S. C., & Markus, H. S. (2019). Clinical significance of magnetic resonance imaging markers of vascular brain injury: A systematic review and meta-analysis. *JAMA neurology*, *76*(1), 81–94.

Dedeoğlu, M., Gafuroğlu, Ü., Yilmaz, Ö., & Bodur, H. (2013). The Relationship Between Hand Grip and Pinch Strengths and Disease Activity, Articular Damage, Pain, and Disability in Patients with Rheumatoid Arthritis. *Archives of Rheumatology*, *28*(2), 069–077. https://doi.org/10.5606/tjr.2013.2742

Demšar, J., & Zupan, B. (2021). Hands-on training about overfitting. *PLOS Computational Biology*, *17*(3), e1008671. https://doi.org/10.1371/journal.pcbi.1008671

Dettmers, C., Fink, G. R., Lemon, R. N., Stephan, K. M., Passingham, R. E., Silbersweig, D., Holmes, A., Ridding, M. C., Brooks, D. J., & Frackowiak, R. S. (1995). Relation between cerebral activity and force in the motor areas of the human brain. *Journal of Neurophysiology*, *74*(2), 802–815. https://doi.org/10.1152/jn.1995.74.2.802

Ding, Q., Triggs, W. J., Kamath, S. M., & Patten, C. (2019). Short Intracortical Inhibition During Voluntary Movement Reveals Persistent Impairment Post-stroke. *Frontiers in Neurology*, *9*, 1105. https://doi.org/10.3389/fneur.2018.01105

Doshi-Velez, F., & Kim, B. (2017). *Towards A Rigorous Science of Interpretable Machine Learning* (arXiv:1702.08608). arXiv. https://doi.org/10.48550/arXiv.1702.08608

Duchowny, K. A., Ackley, S. F., Brenowitz, W. D., Wang, J., Zimmerman, S. C., Caunca, M. R., & Glymour, M. M. (2022). Associations Between Handgrip Strength and Dementia Risk, Cognition, and Neuroimaging Outcomes in the UK Biobank Cohort Study. *JAMA Network Open*, *5*(6), e2218314. https://doi.org/10.1001/jamanetworkopen.2022.18314

Dudzińska-Griszek, J., Szuster, K., & Szewieczek, J. (2017). Grip strength as a frailty diagnostic component in geriatric inpatients. *Clinical Interventions in Aging*, *Volume 12*, 1151–1157. https://doi.org/10.2147/CIA.S140192

Duval, L., Stinear, C. M., & Byblow, W. D. (2024). Modulation of motor cortex inhibition during manual dexterity tasks: An adaptive threshold hunting study. *Journal of Neurophysiology*, *132*(4), 1223–1230. https://doi.org/10.1152/jn.00262.2024

Ejaz, N., Xu, J., Branscheidt, M., Hertler, B., Schambra, H., Widmer, M., Faria, A. V., Harran, M. D., Cortes, J. C., Kim, N., Celnik, P. A., Kitago, T., Luft, A. R., Krakauer, J. W., & Diedrichsen, J. (2018). Evidence for a subcortical origin of mirror movements after stroke: A longitudinal study. *Brain*, *141*(3), 837–847. https://doi.org/10.1093/brain/awx384

Elwert, F., & Winship, C. (2014). Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. *Annual Review of Sociology*, *40*(1), 31–53. https://doi.org/10.1146/annurev-soc-071913-043455

Federico, P., & Perez, M. A. (2017). Distinct Corticocortical Contributions to Human Precision and Power Grip. *Cerebral Cortex*, *27*(11), 5070–5082. https://doi.org/10.1093/cercor/bhw291

Ferreiro De Andrade, K. N., & Conforto, A. B. (2018). Decreased short-interval intracortical inhibition correlates with better pinch strength in patients with stroke and good motor recovery. *Brain Stimulation*, *11*(4), 772–774. https://doi.org/10.1016/j.brs.2018.01.030

Fielding, R. A., Vellas, B., Evans, W. J., Bhasin, S., Morley, J. E., Newman, A. B., Abellan Van Kan, G., Andrieu, S., Bauer, J., Breuille, D., Cederholm, T., Chandler, J., De Meynard, C., Donini, L., Harris, T., Kannt, A., Keime Guibert, F., Onder, G., Papanicolaou, D., … Zamboni, M. (2011). Sarcopenia: An Undiagnosed Condition in Older Adults. Current Consensus Definition: Prevalence, Etiology, and Consequences. International Working Group on Sarcopenia. *Journal of the American Medical Directors Association*, *12*(4), 249–256. https://doi.org/10.1016/j.jamda.2011.01.003

Firth, J. A., Smith, L., Sarris, J., Vancampfort, D., Schuch, F., Carvalho, A. F., Solmi, M., Yung, A. R., Stubbs, B., & Firth, J. (2020). Handgrip Strength Is Associated With Hippocampal Volume and White Matter Hyperintensities in Major Depression and Healthy Controls: A UK Biobank Study. *Psychosomatic Medicine*, *82*(1), 39–46. https://doi.org/10.1097/PSY.0000000000000753

Firth, J., Stubbs, B., Vancampfort, D., Firth, J. A., Large, M., Rosenbaum, S., Hallgren, M., Ward, P. B., Sarris, J., & Yung, A. R. (2018). Grip Strength Is Associated With Cognitive Performance in Schizophrenia and the General Population: A UK Biobank Study of 476559 Participants. *Schizophrenia Bulletin*, *44*(4), 728–736. https://doi.org/10.1093/schbul/sby034

Fritz, N. E., McCarthy, C. J., & Adamo, D. E. (2017). Handgrip strength as a means of monitoring progression of cognitive decline – A scoping review. *Ageing Research Reviews*, *35*, 112–123. https://doi.org/10.1016/j.arr.2017.01.004

Fugiel, J., Czyż, S. H., Rohan, A., Lindner, K., Winkel, I., & Sobieszczańska, M. (2025). Motor function tests as early indicators of cognitive and functional decline in older adults: A correlational study. *PloS one*, *20*(12), e0338646.

Futagi, Y., Toribe, Y., & Suzuki, Y. (2012). The Grasp Reflex and Moro Reflex in Infants: Hierarchy of Primitive Reflex Responses. *International Journal of Pediatrics*, *2012*, 1–10. https://doi.org/10.1155/2012/191562

Gale, C. R., Martyn, C. N., Cooper, C., & Sayer, A. A. (2007). Grip strength, body composition, and mortality. *International journal of epidemiology*, *36*(1), 228–235.

Gallagher, S., Moore, J. S., Stobbe, T. J., McGlothlin, J. D., & Bhattacharya, A. (2000). Physical strength assessment in ergonomics. In *Handbook of industrial automation* (S. 797–827). CRC Press.

Ganipineni, V. D. P., Idavalapati, A. S. K. K., Tamalapakula, S. S., Moparthi, V., Potru, M., Owolabi, O. J., KUMAR, I. A. S. K., Sowrab, T. S., & Vagdevi, M. (2023). Depression and hand-grip: Unraveling the association. *Curēus*, *15*(5).

Garmany, A., & Terzic, A. (2025). Healthspan-lifespan gap differs in magnitude and disease contribution across world regions. *Communications Medicine*, *5*(1), 381. https://doi.org/10.1038/s43856-025-01111-2

Gell, M., Eickhoff, S. B., Omidvarnia, A., Küppers, V., Patil, K. R., Satterthwaite, T. D., Müller, V. I., & Langner, R. (2024). How measurement noise limits the accuracy of brain-behaviour predictions. *Nature Communications*, *15*(1), 10678.

Gianfredi, V., Nucci, D., Pennisi, F., Maggi, S., Veronese, N., & Soysal, P. (2025). Aging, longevity, and healthy aging: The public health approach. *Aging Clinical and Experimental Research*, *37*(1), 125. https://doi.org/10.1007/s40520-025-03021-8

Gibson, A. S. C., & Noakes, T. (2004). Evidence for complex system integration and dynamic neural regulation of skeletal muscle recruitment during exercise in humans. *British journal of sports medicine*, *38*(6), 797–806.

Glover, I. S., & Baker, S. N. (2022). Both Corticospinal and Reticulospinal Tracts Control Force of Contraction. *The Journal of Neuroscience*, *42*(15), 3150–3164. https://doi.org/10.1523/JNEUROSCI.0627-21.2022

Guiney, H., & Machado, L. (2013). Benefits of regular aerobic exercise for executive functioning in healthy populations. *Psychonomic bulletin & review*, *20*(1), 73–86.

Gujral, S., McAuley, E., Oberlin, L. E., Kramer, A. F., & Erickson, K. I. (2018). Role of Brain Structure in Predicting Adherence to a Physical Activity Regimen. *Psychosomatic Medicine*, *80*(1), 69–77. https://doi.org/10.1097/PSY.0000000000000526

Guo, Y., Wang, Z., Prathap, S., & Holschneider, D. P. (2017). Recruitment of prefrontal-striatal circuit in response to skilled motor challenge. *NeuroReport*, *28*(18), 1187–1194. https://doi.org/10.1097/WNR.0000000000000881

Halsband, U., Matsuzaka, Y., & Tanji, J. (1994). Neuronal activity in the primate supplementary, pre-supplementary and premotor cortex during externally and internally instructed sequential movements. *Neuroscience research*, *20*(2), 149–155.

Hamdan, S., Love, B. C., von Polier, G. G., Weis, S., Schwender, H., Eickhoff, S. B., & Patil, K. R. (2023a). Confound-leakage: Confound removal in machine learning leads to leakage. *GigaScience*, *12*.

Hamdan, S., Love, B. C., von Polier, G. G., Weis, S., Schwender, H., Eickhoff, S. B., & Patil, K. R. (2023b). Confound-leakage: Confound removal in machine learning leads to leakage. *GigaScience*, *12*, giad071.

Haugen, I. K., Aaserud, J., & Kvien, T. K. (2021). Get a Grip on Factors Related to Grip Strength in Persons With Hand Osteoarthritis: Results From an Observational Cohort Study. *Arthritis Care & Research*, *73*(6), 794–800. https://doi.org/10.1002/acr.24385

He, S.-Q., Dum, R. P., & Strick, P. L. (1993). Topographic organization of corticospinal projections from the frontal lobe: Motor areas on the lateral surface of the hemisphere. *Journal of Neuroscience*, *13*(3), 952–980.

He, S.-Q., Dum, R. P., & Strick, P. L. (1995). Topographic organization of corticospinal projections from the frontal lobe: Motor areas on the medial surface of the hemisphere. *Journal of Neuroscience*, *15*(5), 3284–3306.

Hunter, S. K. (2025). Motor performance and aging in males and females. *Journal of Electromyography and Kinesiology*, *85*, 103066. https://doi.org/10.1016/j.jelekin.2025.103066

Hunter, S. K., Pereira, H. M., & Keenan, K. G. (2016). The aging neuromuscular system and motor performance. *Journal of applied physiology*.

Inoue, R., & Nishimune, H. (2023). Neuronal Plasticity and Age-Related Functional Decline in the Motor Cortex. *Cells*, *12*(17), 2142. https://doi.org/10.3390/cells12172142

Javed, K., Reddy, V., & Lui, F. (2018). *Neuroanatomy, lateral corticospinal tract*.

Jiang, R., Westwater, M. L., Noble, S., Rosenblatt, M., Dai, W., Qi, S., Sui, J., Calhoun, V. D., & Scheinost, D. (2022). Associations between grip strength, brain structure, and mental health in > 40,000 participants from the UK Biobank. *BMC Medicine*, *20*(1), 286. https://doi.org/10.1186/s12916-022-02490-2

Jiang, R., Woo, C.-W., Qi, S., Wu, J., & Sui, J. (2022). Interpreting Brain Biomarkers: Challenges and solutions in interpreting machine learning-based predictive neuroimaging. *IEEE Signal Processing Magazine*, *39*(4), 107–118. https://doi.org/10.1109/MSP.2022.3155951

Johnson, J. J., Breault, M. S., Sacre, P., Kerr, M. S. D., Johnson, M., Bulacio, J., Gonzalez-Martinez, J., Sarma, S. V., & Gale, J. T. (2017). The role of nonmotor brain regions during human motor control. *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2498–2501. https://doi.org/10.1109/EMBC.2017.8037364

Kapoor, S., & Narayanan, A. (2022). *Leakage and the Reproducibility Crisis in ML-based Science* (arXiv:2207.07048). arXiv. http://arxiv.org/abs/2207.07048

Klotzbier, T. J., & Schott, N. (2025). Scaffolding theory of maturation, cognition, motor performance, and motor skill acquisition (SMART COMPASS): A revised and comprehensive framework for understanding motor-cognitive interactions across the lifespan. *Frontiers in Human Neuroscience*, *19*, 1631958.

Kraha, A., Turner, H., Nimon, K., Zientek, L. R., & Henson, R. K. (2012). Tools to support interpreting multiple regression in the face of multicollinearity. *Frontiers in psychology*, *3*, 44.

Kulwattho, N., Chieh, H.-F., Lin, C.-J., Chen, W.-J., Pintavirooj, C., Ma, C. C., & Su, F.-C. (2025). The brain activation on upper extremity motor control tasks in different forces levels. *Scientific Reports*, *15*(1), 42842. https://doi.org/10.1038/s41598-025-20727-x

Lam, T. K., Dawson, D. R., Honjo, K., Ross, B., Binns, M. A., Stuss, D. T., Black, S. E., Chen, J. J., Levine, B. T., Fujioka, T., & others. (2018). Neural coupling between contralesional motor and frontoparietal networks correlates with motor ability in individuals with chronic stroke. *Journal of the neurological sciences*, *384*, 21–29.

Landsmeer, J. (1962). Power grip and precision handling. *Annals of the rheumatic diseases*, *21*(2), 164.

Lawman, H. G., Troiano, R. P., Perna, F. M., Wang, C.-Y., Fryar, C. D., & Ogden, C. L. (2016). Associations of Relative Handgrip Strength and Cardiovascular Disease Biomarkers in U.S. Adults, 2011–2012. *American Journal of Preventive Medicine*, *50*(6), 677–683. https://doi.org/10.1016/j.amepre.2015.10.022

Leisman, G., Moustafa, A., & Shafir, T. (2016). Thinking, Walking, Talking: Integratory Motor and Cognitive Brain Function. *Frontiers in Public Health*, *4*. https://doi.org/10.3389/fpubh.2016.00094

Lemon, R. N. (2008). Descending Pathways in Motor Control. *Annual Review of Neuroscience*, *31*(1), 195–218. https://doi.org/10.1146/annurev.neuro.31.060407.125547

Leong, D. P., Teo, K. K., Rangarajan, S., Lopez-Jaramillo, P., Avezum, A., Orlandini, A., Seron, P., Ahmed, S. H., Rosengren, A., Kelishadi, R., & others. (2015). Prognostic value of grip strength: Findings from the Prospective Urban Rural Epidemiology (PURE) study. *The Lancet*, *386*(9990), 266–273.

Li, J., Bzdok, D., Chen, J., Tam, A., Ooi, L. Q. R., Holmes, A. J., Ge, T., Patil, K. R., Jabbi, M., Eickhoff, S. B., Yeo, B. T. T., & Genon, S. (2022). Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Science Advances*, *8*(11), eabj1812. https://doi.org/10.1126/sciadv.abj1812

Ling, C. H., Taekema, D., De Craen, A. J., Gussekloo, J., Westendorp, R. G., & Maier, A. B. (2010). Handgrip strength and mortality in the oldest old population: The Leiden 85-plus study. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*, *182*(5), 429–435.

Long, C. I., Conrad, P., Hall, E., & Furler, S. (1970). Intrinsic-extrinsic muscle control of the hand in power grip and precision handling: An electromyographic study. *JBJS*, *52*(5), 853–867.

López, S., & Saboya, M. (2009). On the relationship between Shapley and Owen values. *Central European Journal of Operations Research*, *17*(4), 415–423.

Mamiya, P. C., Richards, T. L., & Kuhl, P. K. (2018). Right forceps minor and anterior thalamic radiation predict executive function skills in young bilingual adults. *Frontiers in psychology*, *9*, 118.

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., … Dosenbach, N. U. F. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, *603*(7902), 654–660. https://doi.org/10.1038/s41586-022-04492-9

Marín-Jiménez, N., Cruz-León, C., Perez-Bey, A., Conde-Caveda, J., Grao-Cruces, A., Aparicio, V. A., Castro-Piñero, J., & Cuenca-García, M. (2022). Predictive Validity of Motor Fitness and Flexibility Tests in Adults and Older Adults: A Systematic Review. *Journal of Clinical Medicine*, *11*(2), 328. https://doi.org/10.3390/jcm11020328

Marques De Moraes, M. V., Dionisio, J., Tan, U., & Tudella, E. (2017). Palmar Grasp Reflex in Human Newborns. *Pediatrics & Therapeutics*, *07*(01). https://doi.org/10.4172/2161-0665.1000309

Mattos, D. J., Rutlin, J., Hong, X., Zinn, K., Shimony, J. S., & Carter, A. R. (2023). The role of extra-motor networks in upper limb motor performance post-stroke. *Neuroscience*, *514*, 1–13.

McGrath, R., Johnson, N., Klawitter, L., Mahoney, S., Trautman, K., Carlson, C., Rockstad, E., & Hackney, K. J. (2020). What are the association patterns between handgrip strength and adverse health conditions? A topical review. *SAGE Open Medicine*, *8*, 2050312120910358. https://doi.org/10.1177/2050312120910358

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., Bartsch, A. J., Jbabdi, S., Sotiropoulos, S. N., Andersson, J. L. R., Griffanti, L., Douaud, G., Okell, T. W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., … Smith, S. M. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*, *19*(11), 1523–1536. https://doi.org/10.1038/nn.4393

Molnar, C. (2020). *Interpreting Machine Learning Models With SHAP*.

Monzée, J., Lamarre, Y., & Smith, A. M. (2003). The effects of digital anesthesia on force control using a precision grip. *Journal of neurophysiology*, *89*(2), 672–683.

Moreno-López, Y., Olivares-Moreno, R., Cordero-Erausquin, M., & Rojas-Piloni, G. (2016). Sensorimotor Integration by Corticospinal System. *Frontiers in Neuroanatomy*, *10*. https://doi.org/10.3389/fnana.2016.00024

Navarro-Orozco, D., & Bollu, P. C. (2018). *Neuroanatomy, medial lemniscus (reils band, reils ribbon)*.

Nelson, A. J. D. (2021). The anterior thalamic nuclei and cognition: A role beyond space? *Neuroscience & Biobehavioral Reviews*, *126*, 1–11. https://doi.org/10.1016/j.neubiorev.2021.02.047

Ni, Z., Gunraj, C., & Chen, R. (2007). Short interval intracortical inhibition and facilitation during the silent period in human. *The Journal of Physiology*, *583*(3), 971–982. https://doi.org/10.1113/jphysiol.2007.135749

Noble, J. W., Eng, J. J., Kokotilo, K. J., & Boyd, L. A. (2011). Aging effects on the control of grip force magnitude: An fMRI study. *Experimental Gerontology*, *46*(6), 453–461. https://doi.org/10.1016/j.exger.2011.01.004

Nowak, D. A., & Hermsdörfer, J. (2006). Predictive and reactive control of grasping forces: On the role of the basal ganglia and sensory feedback. *Experimental Brain Research*, *173*(4), 650–660. https://doi.org/10.1007/s00221-006-0409-7

Opri, E., Cernera, S., Okun, M. S., Foote, K. D., & Gunduz, A. (2019). The Functional Role of Thalamocortical Coupling in the Human Motor Network. *The Journal of Neuroscience*, *39*(41), 8124–8134. https://doi.org/10.1523/JNEUROSCI.1153-19.2019

Oschwald, J., Mérillat, S., Jäncke, L., & Seidler, R. D. (2021). Fractional Anisotropy in Selected, Motor-Related White Matter Tracts and Its Cross-Sectional and Longitudinal Associations With Motor Function in Healthy Older Adults. *Frontiers in Human Neuroscience*, *15*, 621263. https://doi.org/10.3389/fnhum.2021.621263

Padoa-Schioppa, C., & Conen, K. E. (2017). Orbitofrontal Cortex: A Neural Circuit for Economic Decisions. *Neuron*, *96*(4), 736–754. https://doi.org/10.1016/j.neuron.2017.09.031

Park, C., Chang, W. H., Ohn, S. H., Kim, S. T., Bang, O. Y., Pascual-Leone, A., & Kim, Y.-H. (2011). Longitudinal changes of resting-state functional connectivity during motor recovery after stroke. *Stroke; a journal of cerebral circulation*, *42*(5), 1357–1362.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, *3*(none). https://doi.org/10.1214/09-SS057

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.

Perrier, J.-F., Rasmussen, H. B., Christensen, R. K., & Petersen, A. V. (2013). Modulation of the intrinsic properties of motoneurons by serotonin. *Current pharmaceutical design*, *19*(24), 4371–4384.

Pessiglione, M., Schmidt, L., Draganski, B., Kalisch, R., Lau, H., Dolan, R. J., & Frith, C. D. (2007). How the brain translates money into force: A neuroimaging study of subliminal motivation. *science*, *316*(5826), 904–906.

Peterson, M. D., Collins, S., Meier, H. C., Brahmsteadt, A., & Faul, J. D. (2023). Grip strength is inversely associated with DNA methylation age acceleration. *Journal of Cachexia, Sarcopenia and Muscle*, *14*(1), 108–115.

Prins, N. D., & Scheltens, P. (2015). White matter hyperintensities, cognitive impairment and dementia: An update. *Nature Reviews Neurology*, *11*(3), 157–165. https://doi.org/10.1038/nrneurol.2015.10

Prodoehl, J., Corcos, D. M., & Vaillancourt, D. E. (2009). Basal ganglia mechanisms underlying precision grip force control. *Neuroscience & Biobehavioral Reviews*, *33*(6), 900–908. https://doi.org/10.1016/j.neubiorev.2009.03.004

Purves, D., Augustine, G., Fitzpatrick, D., Hall, W., LaMantia, A., McNamara, J., & White, L. (2001). The regulation of muscle force. *Neuroscience*.

Purves, D., Augustine, G. J., & Fitzpatrick, D. (Hrsg.). (2001). Neural Centers Responsible for Movement. In *Neuroscience. 2nd edition*. Sinauer Associates.

Purves, D., Augustine, G. J., Fitzpatrick, D., Katz, L. C., LaMantia, A.-S., McNamara, J. O., & Williams, S. M. (2001a). Modulation of Movement by the Cerebellum. In *Neuroscience. 2nd edition*. Sinauer Associates.

Purves, D., Augustine, G. J., Fitzpatrick, D., Katz, L. C., LaMantia, A.-S., McNamara, J. O., & Williams, S. M. (2001b). The major afferent pathway for mechanosensory information: The dorsal column-medial lemniscus system. *Neuroscience*, 189–207.

Quattrocchi, A., Garufi, G., Gugliandolo, G., De Marchis, C., Collufio, D., Cardali, S. M., & Donato, N. (2024). Handgrip Strength in Health Applications: A Review of the Measurement Methodologies and Influencing Factors. *Sensors*, *24*(16), 5100. https://doi.org/10.3390/s24165100

Rantanen, T., Guralnik, J. M., Foley, D., Masaki, K., Leveille, S., Curb, J. D., & White, L. (1999). Midlife hand grip strength as a predictor of old age disability. *JAMA : the journal of the American Medical Association*, *281*(6), 558–560.

Rao, A., Monteiro, J. M., & Mourao-Miranda, J. (2017). Predictive modelling using neuroimaging data in the presence of confounds. *NeuroImage*, *150*, 23–49. https://doi.org/10.1016/j.neuroimage.2017.01.066

Reeve IV, T. E., Ur, R., Craven, T. E., Kaan, J. H., Goldman, M. P., Edwards, M. S., Hurie, J. B., Velazquez-Ramirez, G., & Corriere, M. A. (2018). Grip strength measurement for frailty assessment in patients with vascular disease and associations with comorbidity, cardiac risk, and sarcopenia. *Journal of vascular surgery*, *67*(5), 1512–1520.

Rezaei, A., Areshenkoff, C. N., Gale, D. J., De Brouwer, A. J., Nashed, J. Y., Flanagan, J. R., & Gallivan, J. P. (2025). Transfer of motor learning is associated with patterns of activity in the default mode network. *Plos Biology*, *23*(8), e3003268.

Richardson, A. G., Attiah, M. A., Berman, J. I., Chen, H. I., Liu, X., Zhang, M., Van Der Spiegel, J., & Lucas, T. H. (2016a). The effects of acute cortical somatosensory deafferentation on grip force control. *Cortex*, *74*, 1–8. https://doi.org/10.1016/j.cortex.2015.10.007

Richardson, A. G., Attiah, M. A., Berman, J. I., Chen, H. I., Liu, X., Zhang, M., Van Der Spiegel, J., & Lucas, T. H. (2016b). The effects of acute cortical somatosensory deafferentation on grip force control. *Cortex*, *74*, 1–8. https://doi.org/10.1016/j.cortex.2015.10.007

Riemann, B. L., & Lephart, S. M. (o. J.). *The Sensorimotor System, Part I: The Physiologic Basis of Functional Joint Stability*.

Rijk, J. M., Roos, P. R., Deckx, L., Van Den Akker, M., & Buntinx, F. (2016). Prognostic value of handgrip strength in people aged 60 years and older: A systematic review and meta-analysis. *Geriatrics & Gerontology International*, *16*(1), 5–20. https://doi.org/10.1111/ggi.12508

Rinne, P., Hassan, M., Fernandes, C., Han, E., Hennessy, E., Waldman, A., Sharma, P., Soto, D., Leech, R., Malhotra, P. A., & others. (2018). Motor dexterity and strength depend upon integrity of the attention-control system. *Proceedings of the National Academy of Sciences*, *115*(3), E536–E545.

Roberts, H. C., Denison, H. J., Martin, H. J., Patel, H. P., Syddall, H., Cooper, C., & Sayer, A. A. (2011). A review of the measurement of grip strength in clinical and epidemiological studies: Towards a standardised approach. *Age and ageing*, *40*(4), 423–429.

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in methods and practices in psychological science*, *1*(1), 27–42.

Roland, P. E., Larsen, B., Lassen, N. A., & Skinhoj, E. (1980). Supplementary motor area and other cortical areas in organization of voluntary movements in man. *Journal of neurophysiology*, *43*(1), 118–136.

Roth, R. H., & Ding, J. B. (2024). Cortico-basal ganglia plasticity in motor learning. *Neuron*, *112*(15), 2486–2502. https://doi.org/10.1016/j.neuron.2024.06.014

Saad, P., Shendrik, K. S., Karroum, P. J., Azizi, H., & Jolayemi, A. (2020). The Anterior Globus Pallidus Externus of Basal Ganglia as Primarily a Limbic and Associative Territory. *Cureus*. https://doi.org/10.7759/cureus.11846

Sachdev, P. S. (2005). White matter hyperintensities are related to physical disability and poor motor function. *Journal of Neurology, Neurosurgery & Psychiatry*, *76*(3), 362–367. https://doi.org/10.1136/jnnp.2004.042945

Sadowski, B. (2008). Plasticity of the Cortical Motor System. *Journal of Human Kinetics*, *20*(2008), 5–22. https://doi.org/10.2478/v10078-008-0014-x

Saga, Y., Hoshi, E., & Tremblay, L. (2017). Roles of multiple globus pallidus territories of monkeys and humans in motivation, cognition and action: An anatomical, physiological and pathophysiological review. *Frontiers in neuroanatomy*, *11*, 30.

Salamone, J. D., Correa, M., Yohn, S., Cruz, L. L., San Miguel, N., & Alatorre, L. (2016). The pharmacology of effort-related choice behavior: Dopamine, depression, and individual differences. *Behavioural Processes*, *127*, 3–17.

Sanes, J. N., & Donoghue, J. P. (2000). Plasticity and Primary Motor Cortex. *Annual Review of Neuroscience*, *23*(1), 393–415. https://doi.org/10.1146/annurev.neuro.23.1.393

Sasaki, H., Kasagi, F., Yamada, M., & Fujita, S. (2007). Grip Strength Predicts Cause-Specific Mortality in Middle-Aged and Elderly Persons. *The American Journal of Medicine*, *120*(4), 337–342. https://doi.org/10.1016/j.amjmed.2006.04.018

Sasse, L., Nicolaisen-Sobesky, E., Dukart, J., Eickhoff, S. B., Götz, M., Hamdan, S., Komeyer, V., Kulkarni, A., Lahnakoski, J. M., Love, B. C., Raimondo, F., & Patil, K. R. (2025). Overview of leakage scenarios in supervised machine learning. *Journal of Big Data*, *12*(1), 135. https://doi.org/10.1186/s40537-025-01193-8

Sayyid, Z. N., Wang, H., Cai, Y., Gross, A. L., Swenor, B. K., Deal, J. A., Lin, F. R., Wanigatunga, A. A., Dougherty, R. J., Tian, Q., & others. (2024). Sensory and motor deficits as contributors to early cognitive impairment. *Alzheimer's & Dementia*, *20*(4), 2653–2661.

Shang, X., Meng, X., Xiao, X., Xie, Z., & Yuan, X. (2021). Grip training improves handgrip strength, cognition, and brain white matter in minor acute ischemic stroke patients. *Clinical Neurology and Neurosurgery*, *209*, 106886. https://doi.org/10.1016/j.clineuro.2021.106886

Shi, P., & Feng, X. (2022). Motor skills and cognitive benefits in children and adolescents: Relationship, mechanism and perspectives. *Frontiers in Psychology*, *13*, 1017825. https://doi.org/10.3389/fpsyg.2022.1017825

Shinohara, M., Li, S., Kang, N., Zatsiorsky, V. M., & Latash, M. L. (2003). Effects of age and gender on finger coordination in MVC and submaximal force-matching tasks. *Journal of Applied Physiology*, *94*(1), 259–270. https://doi.org/10.1152/japplphysiol.00643.2002

Singh, N. M., Harrod, J. B., Subramanian, S., Robinson, M., Chang, K., Cetin-Karayumak, S., Dalca, A. V., Eickhoff, S., Fox, M., Franke, L., Golland, P., Haehn, D., Iglesias, J. E., O'Donnell, L. J., Ou, Y., Rathi, Y., Siddiqi, S. H., Sun, H., Westover, M. B., … Gollub, R. L. (2022). How Machine Learning is Powering Neuroimaging to Improve Brain Health. *Neuroinformatics*, *20*(4), 943–964. https://doi.org/10.1007/s12021-022-09572-9

Smith, S. M., & Nichols, T. E. (2018a). Statistical Challenges in "Big Data" Human Neuroimaging. *Neuron*, *97*(2), 263–268. https://doi.org/10.1016/j.neuron.2017.12.018

Smith, S. M., & Nichols, T. E. (2018b). Statistical Challenges in "Big Data" Human Neuroimaging. *Neuron*, *97*(2), 263–268. https://doi.org/10.1016/j.neuron.2017.12.018

Snoek, L., Miletić, S., & Scholte, H. S. (2019). How to control for confounds in decoding analyses of neuroimaging data. *NeuroImage*, *184*, 741–760. https://doi.org/10.1016/j.neuroimage.2018.09.074

Song, J., Liu, T., Zhao, J., Wang, S., Dang, X., & Wang, W. (2022). Causal associations of hand grip strength with bone mineral density and fracture risk: A mendelian randomization study. *Frontiers in Endocrinology*, *13*, 1020750. https://doi.org/10.3389/fendo.2022.1020750

Spampinato, D., & Celnik, P. (2021). Multiple Motor Learning Processes in Humans: Defining Their Neurophysiological Bases. *The Neuroscientist*, *27*(3), 246–267. https://doi.org/10.1177/1073858420939552

Spirduso, W. W., & MacRae, P. G. (1990). Motor Performance and Aging. In *Handbook of the Psychology of Aging* (S. 183–200). Elsevier. https://doi.org/10.1016/B978-0-12-101280-9.50017-6

Spraker, M. B., Yu, H., Corcos, D. M., & Vaillancourt, D. E. (2007). Role of Individual Basal Ganglia Nuclei in Force Amplitude Generation. *Journal of Neurophysiology*, *98*(2), 821–834. https://doi.org/10.1152/jn.00239.2007

Stephens-Sarlós, E., Horváth-Pápai, A., Tóth, E. E., Ihász, F., Somogyi, A., & Szabo, A. (2025). Relationship between primitive reflexes, functional fitness, handgrip strength, and physical activity in older adults aged 65 and over. *Physiological Reports*, *13*(7), e70229.

Stinear, C. M., Coxon, J. P., & Byblow, W. D. (2009). Primary motor cortex and movement prevention: Where Stop meets Go. *Neuroscience & Biobehavioral Reviews*, *33*(5), 662–673. https://doi.org/10.1016/j.neubiorev.2008.08.013

Surgent, O., Guerrero-Gonzalez, J., Dean, D. C., Kirk, G. R., Adluru, N., Kecskemeti, S. R., Alexander, A. L., & Travers, B. G. (2023). How we get a grip: Microstructural neural correlates of manual grip strength in children. *NeuroImage*, *273*, 120117. https://doi.org/10.1016/j.neuroimage.2023.120117

Suzuki, M., & Nishimura, Y. (2022). The ventral striatum contributes to the activity of the motor cortex and motor outputs in monkeys. *Frontiers in Systems Neuroscience*, *16*, 979272. https://doi.org/10.3389/fnsys.2022.979272

Syddall, H. E., Westbury, L. D., Dodds, R., Dennison, E., Cooper, C., & Sayer, A. A. (2017). Mortality in the Hertfordshire Ageing Study: Association with level and loss of hand grip strength in later life. *Age and Ageing*, *46*(3), 407–412.

Takahashi, N., Moberg, S., Zolnik, T. A., Catanese, J., Sachdev, R. N., Larkum, M. E., & Jaeger, D. (2021). Thalamic input to motor cortex facilitates goal-directed action initiation. *Current Biology*, *31*(18), 4148–4155.

Tazoe, T., & Perez, M. A. (2017). Cortical and reticular contributions to human precision and power grip. *The Journal of Physiology*, *595*(8), 2715–2730. https://doi.org/10.1113/JP273679

Thickbroom, G. W., Phillips, B. A., Morris, I., Byrnes, M. L., Sacco, P., & Mastaglia, F. L. (1999). Differences in functional magnetic resonance imaging of sensorimotor cortex during static and dynamic finger flexion. *Experimental Brain Research*, *126*(3), 431–438. https://doi.org/10.1007/s002210050749

Tian, Q., Bair, W.-N., Resnick, S. M., Bilgel, M., Wong, D. F., & Studenski, S. A. (2018). β-amyloid deposition is associated with gait variability in usual aging. *Gait & posture*, *61*, 346–352.

Tian, Q., Resnick, S. M., Bilgel, M., Wong, D. F., Ferrucci, L., & Studenski, S. A. (2017). β-Amyloid burden predicts lower extremity performance decline in cognitively unimpaired older adults. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, *72*(5), 716–723.

Tian, Y., Margulies, D. S., Breakspear, M., & Zalesky, A. (2020). Topographic organization of the human subcortex unveiled with functional connectivity gradients. *Nature neuroscience*, *23*(11), 1421–1432.

Tönnies, T., Kahl, S., & Kuss, O. (2022). Collider bias in observational studies: Consequences for medical research part 30 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International*, *119*(7), 107.

Vaillancourt, D. E., Mayka, M. A., Thulborn, K. R., & Corcos, D. M. (2004). Subthalamic nucleus and internal globus pallidus scale with the rate of change of force production in humans. *NeuroImage*, *23*(1), 175–186. https://doi.org/10.1016/j.neuroimage.2004.04.040

Vaishya, R., Misra, A., Vaish, A., Ursino, N., & D'Ambrosi, R. (2024). Hand grip strength as a proposed new vital sign of health: A narrative review of evidences. *Journal of Health, Population and Nutrition*, *43*(1), 7. https://doi.org/10.1186/s41043-024-00500-y

VanderWeele, T. J. (2019). Principles of confounder selection. *European Journal of Epidemiology*, *34*(3), 211–219. https://doi.org/10.1007/s10654-019-00494-6

Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, *180*, 68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061

Verghese, J., Lipton, R. B., Hall, C. B., Kuslansky, G., Katz, M. J., & Buschke, H. (2002). Abnormality of gait as a predictor of non-Alzheimer's dementia. *New England Journal of Medicine*, *347*(22), 1761–1768.

Voelcker-Rehage, C., & Niemann, C. (2013). Structural and functional brain changes related to different types of physical activity across the life span. *Neuroscience & Biobehavioral Reviews*, *37*(9), 2268–2295.

Ward, H. B., Beermann, A., Manzanarez Felix, K., Narisetti, L., Lewandowski, K. E., Coleman, M., Bouix, S., Holt, D., Öngür, D., Breier, A., & others. (2025). Grip strength as a marker of resting-state network integrity and well-being in early psychosis. *American Journal of Psychiatry*, appi-ajp.

Ward, N. S., & Frackowiak, R. S. J. (2003). Age-related changes in the neural correlates of motor performance. *Brain*, *126*(4), 873–888. https://doi.org/10.1093/brain/awg071

Ward, N. S., Newton, J. M., Swayne, O. B. C., Lee, L., Frackowiak, R. S. J., Thompson, A. J., Greenwood, R. J., & Rothwell, J. C. (2007). The relationship between brain activity and peak grip force is modulated by corticospinal system integrity after subcortical stroke. *European Journal of Neuroscience*, *25*(6), 1865–1873. https://doi.org/10.1111/j.1460-9568.2007.05434.x

Ward, N. S., Newton, J. M., Swayne, O. B. C., Lee, L., Thompson, A. J., Greenwood, R. J., Rothwell, J. C., & Frackowiak, R. S. J. (2006). Motor system activation after subcortical stroke depends on corticospinal system integrity. *Brain*, *129*(3), 809–819. https://doi.org/10.1093/brain/awl002

Wardlaw, J. M., Valdés Hernández, M. C., & Muñoz-Maniega, S. (2015). What are white matter hyperintensities made of? Relevance to vascular cognitive impairment. *Journal of the American Heart Association*, *4*(6), e001140.

Wasson, P., Prodoehl, J., Coombes, S. A., Corcos, D. M., & Vaillancourt, D. E. (2010). Predicting grip force amplitude involves circuits in the anterior basal ganglia. *NeuroImage*, *49*(4), 3230–3238. https://doi.org/10.1016/j.neuroimage.2009.11.047

Welniarz, Q., Gallea, C., Lamy, J., Méneret, A., Popa, T., Valabregue, R., Béranger, B., Brochard, V., Flamand-Roze, C., Trouillard, O., Bonnet, C., Brüggemann, N., Bitoun, P., Degos, B., Hubsch, C., Hainque, E., Golmard, J., Vidailhet, M., Lehéricy, S., … Roze, E. (2019). The supplementary motor area modulates interhemispheric interactions during movement preparation. *Human Brain Mapping*, *40*(7), 2125–2142. https://doi.org/10.1002/hbm.24512

Wen, Z., Gu, J., Chen, R., Wang, Q., Ding, N., Meng, L., Wang, X., Liu, H., Sheng, Z., & Zheng, H. (2023). Handgrip Strength and Muscle Quality: Results from the National Health and Nutrition Examination Survey Database. *Journal of Clinical Medicine*, *12*(9), 3184. https://doi.org/10.3390/jcm12093184

Wilkinson, J., Arnold, K. F., Murray, E. J., Van Smeden, M., Carr, K., Sippy, R., De Kamps, M., Beam, A., Konigorski, S., Lippert, C., Gilthorpe, M. S., & Tennant, P. W. G. (2020). Time to reality check the promises of machine learning-powered precision medicine. *The Lancet Digital Health*, *2*(12), e677–e680. https://doi.org/10.1016/S2589-7500(20)30200-4

Winter, L., Huang, Q., Sertic, J. V. L., & Konczak, J. (2022). The Effectiveness of Proprioceptive Training for Improving Motor Performance and Motor Dysfunction: A Systematic Review. *Frontiers in Rehabilitation Sciences*, *3*, 830166. https://doi.org/10.3389/fresc.2022.830166

Wolpert, D. M., & Flanagan, J. R. (2001). Motor prediction. *Current Biology*, *11*(18), R729–R732. https://doi.org/10.1016/S0960-9822(01)00432-8

World Health Organization. (2025, Februar 21). *Ageing: Global population*. https://www.who.int/news-room/questions-and-answers/item/population-ageing#:~:text=Globally%2C%20life%20expectancy%20at%20birth,significant%20implications%20for%20public%20health.

Wright, N. F., Vann, S. D., Aggleton, J. P., & Nelson, A. J. D. (2015). A Critical Role for the Anterior Thalamus in Directing Attention to Task-Relevant Stimuli. *The Journal of Neuroscience*, *35*(14), 5480–5488. https://doi.org/10.1523/JNEUROSCI.4945-14.2015

Wysocki, A. C., Lawson, K. M., & Rhemtulla, M. (2022). *Statistical Control Requires Causal Justification*. *5*(2).

Xia, J., Lin, X., Yu, T., Yu, H., Zou, Y., Luo, Q., & Peng, H. (2024). Aberrant functional connectivity of the globus pallidus in the modulation of the relationship between childhood trauma and major depressive disorder. *Journal of Psychiatry and Neuroscience*, *49*(4), E218–E232. https://doi.org/10.1503/jpn.240019

Xie, X., Li, D., Zhou, M., Wang, Z., & Zhang, X. (2025). Effects of hand strength and walking speed combined and in isolation on the prediction of cognitive decline and dementia in middle-aged and older adults: A systematic review and meta-analysis. *Journal of the American Medical Directors Association*, *26*(6), 105576.

Zaaimi, B., Edgley, S. A., Soteropoulos, D. S., & Baker, S. N. (2012). Changes in descending motor pathway connectivity after corticospinal tract lesion in macaque monkey. *Brain*, *135*(7), 2277–2289. https://doi.org/10.1093/brain/aws115

Zafeiriou, D. I. (2004). Primitive reflexes and postural reactions in the neurodevelopmental examination. *Pediatric neurology*, *31*(1), 1–8.

Zapparoli, L., Mariano, M., & Paulesu, E. (2022). How the motor system copes with aging: A quantitative meta-analysis of the effect of aging on motor function control. *Communications Biology*, *5*(1), 79. https://doi.org/10.1038/s42003-022-03027-2

Zhang, X., Huang, M., Yuan, X., Zhong, X., Dai, S., Wang, Y., Zhang, Q., Wongwitwichote, K., & Jiang, C. (2025). Lifespan trajectories of motor control and neural oscillations: A systematic review of magnetoencephalography insights. *Developmental Cognitive Neuroscience*, *72*, 101529. https://doi.org/10.1016/j.dcn.2025.101529

9 Publications

Publications, preprints and manuscripts included in this thesis

Komeyer, V., Nieto, N., Eickhoff, S. B., Raimondo, F., & Patil, K. R. (2025). Overview of Challenges in Brain-based Predictive Modeling: Towards meaningful predictive insights. *Biological Psychiatry*.

Komeyer, V., Herrmann, C., Eickhoff, S. B., Rathkopf, C., Raimondo, F., & Patil, K. R. (2025). How causal inference tools can support debiasing of machine learning models for meaningful brain-based predictions. *medRxiv*.

Komeyer, V., Eickhoff, S. B., Kasper, J., Patil, K. R. $^{+}$, & Raimondo, F.$^{+}$ (2025). Hand grip strength as a behavioural read-out of distributed but specific system-level brain integrity: A large-scale multi-modal machine learning study.

Peer-reviewed publications not subject to this thesis

Wiersch, L., Friedrich, P., Hamdan, S., **Komeyer, V.,** Hoffstaedter, F., Patil, K. R., ... & Weis, S. (2024). Sex classification from functional brain connectivity: Generalization to multiple datasets. *Human brain mapping*, *45*(6), e26683.

Hamdan, S., More, S., Sasse, L., **Komeyer, V.,** Patil, K. R., Raimondo, F., & Alzheimer's Disease Neuroimaging Initiative. (2024). Julearn: an easy-to-use library for leakage-free evaluation and inspection of ML models. *Gigabyte*, *2024*, 1-16.

Sasse, L., Nicolaisen-Sobesky, E., Dukart, J., Eickhoff, S. B., Götz, M., Hamdan, S., **Komeyer, V.**, ... & Patil, K. R. (2025). Overview of leakage scenarios in supervised machine learning. *Journal of Big Data*, *12*(1), 135.

Further preprints

Komeyer, V., Eickhoff, S. B., Grefkes, C., Patil, K. R., & Raimondo, F. (2024). Confounder control in biomedicine necessitates conceptual considerations beyond statistical evaluations. *medRxiv*, 2024-02.

Raimondo, F., Bi, H., **Komeyer, V.**, Kasper, J., Primus, S., Hoffstaedter, F., ... & Patil, K. R. (2024). Can we predict sleep health based on brain features? A large-scale machine learning study using UK Biobank. *bioRxiv*. *Accepted in Brain Communications*

Nazarzadeh, K., Eickhoff, S. B., Antonopoulos, G., Hensel, L., Tscherpel, C., **Komeyer, V.**, ... & Patil, K. R. (2024). Machine Learning-Driven Correction of Handgrip Strength: A Novel Biomarker for Neurological and Health Outcomes in the UK Biobank. *medRxiv*.

Nazarzadeh, K., Eickhoff, S. B., Antonopoulos, G., Hensel, L., Tscherpel, C., **Komeyer, V.**, ... & Patil, K. R. (2025). Predicting Brain Volumes from Anthropometric and Demographic Features: Insights from UK Biobank Neuroimaging Data. *bioRxiv*. *Accepted in Brain Structure and Function*

10 Author Declaration

Ich versichere an Eides Statt, dass die Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf“ erstellt worden ist.

Date: _____ Signature: _____

11 Appendix

Contains:

Supplementary information for Study 2

Supplementary information for Study 3

Supplementary Materials

How causal inference tools can support debiasing of machine learning models for meaningful brain-based predictions

Contents:

Methods
Supplementary references

1. Methods

1.1. Data and pre-processing

For our example predictions we used data of the 1st scanning session (ses-2) of the UK Biobank[1], recorded at three different sites in the UK (Cheadle, Reading, Newcastle). The exact protocol and acquisition parameters of the structural imaging can be found in Miller et al. (2016)[2]. The structural pre-processing was carried out by pipelines developed and run by the UKB[3].

For the Grey Matter Volume (GMV) features 41,180 T1-weighted pre-processed images were retrieved from UKB and converted into a DataLad[4] dataset for provenance tracking with subsequent computations of voxel-based morphometry (CAT 12.7 (default settings); MNI152 space; 1.5mm isotropic)[5]. We extracted the parcel-wise GMV as the winsorized mean (limits 10%) of the voxel-wise values per parcel using the cortical Schaefer et al. (2018)[6] atlas (1000 ROIs), subcortical Tian et al. (2020)[7] (S4 3T) and cerebellar Diedrichsen et al. (2009)[8] (SUIT space) atlas.

All non-imaging variables, including the exemplarily target Hand Grip Strength (HGS) and the investigated example confounders were obtained directly from the UK Biobank[9]. We chose HGS as a robust, objective and reliable target[10–13] to avoid further conceptual problems oftentimes coming along with more latent variables as targets, such as intelligence or executive functioning measures[14]. Healthy subjects were (rather conservatively) defined by excluding the ICD-10 criteria chapters F, G and I60 to I69, which excludes subjects with a history of mental and behavioural disorders, diseases of the nervous system or with a cerebrovascular disease. All NaN values and outliers larger than the 4th standard deviation were removed from the non-imaging data. Additionally, the HGS was averaged over left and right hand and there was a check for balance of sex distributions in the HGS.

1.2. Modelling

10% of the data were set apart to be used as a locked test set for a related project and left untouched for this project. The remaining 90% of the data were split into a training (0.8) and test (0.2) set. Learning algorithms were fitted on the training set by using a cross-validation (CV) scheme. The CV on the training set served to control for the fitting behaviour (e.g. overfitting) of the model and to get an impression of the generalization error. A final estimator, retrained on the entire training set (using root mean squared error; RMSE) was eventually used to make the predictions on the initially held-out test set. These predictions were used to report and visualize the predictive performances. All applied splits were stratified for binned *age*, binned HGS (2 bins) and *sex* (as either defined in the NHS central registry or self-reported). Within the CV scheme, continuous features were z-scored (mean of zero and unit variance). We used a (stratified) 5-fold strategy with one repetition for the CV. We scored the CV using RMSE, mean absolute error (MAE), coefficient of determination (R^2), Pearson's r and Spearman's r. The confound removal was applied within the CV to avoid data leakage. Therein, for each feature, a linear regression was fit using the confounds as independent variables and the features as dependent variables. The new, confound-free features were calculated as the residuals of the fitted linear regression (original features minus predicted/fitted features).

1.3. Algorithms, sample sizes and statistical evaluation

The example predictions of HGS from GMV with no confounder adjustment (*vanilla* model) or adjustment for *muscle mass* and *sex* as confounders as illustrated in Fig. 1b (main manuscript) was performed using scikit-learn's[15] linear support vector regression (SVR) with a squared epsilon insensitive loss (L2) and a heuristically calculated hyperparameter C ($C = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\sqrt{features^2}}$, [16]). This heuristic C value was calculated in a CV consistent manner, i.e. it was calculated only on the training data within the respective fold of the CV. Due to using a heuristic estimate of the hyperparameter C, no nested CV setup was necessary for hyper parameter optimization.

To have comparable models between the unadjusted *vanilla* model and the confounder adjusted model, sample sizes were matched to the variable with the least amount of subjects measured at ses-2, which was

muscle mass (operationalized through UKB's variable *total lean mass*). This resulted in the shown out-of-sample predictions being performed on N=3620 (N_{train} = 2606, N_{test} = 652) subjects. The 5-fold CV was previously performed on the N_{train} = 2606 subjects.

All correlations were calculated on these same N=3620 subjects. Parcel-wise correlations between parcellated GMV and HGS, parcellated GMV and *muscle mass* as well as HGS and *muscle mass* (all continuous variables) were calculated using Pearson's r (Fig. 1c, Fig. 2 bottom, main manuscript). All correlations including *sex* were calculated using point-biserial correlation coefficient to account for the discrete nature of this variable.

1.4. Code availability

Custom code generated for this project was made publicly available in a GitHub repository. The repository contains further detailed information on used python packages (and versions), code execution and necessary steps for replication of computations.

2. References

1. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med*. 2015;12(3):e1001779. doi:10.1371/journal.pmed.1001779

2. Miller KL, Alfaro-Almagro F, Bangerter NK, et al. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat Neurosci*. 2016;19(11):1523-1536. doi:10.1038/nn.4393

3. Alfaro-Almagro F, Jenkinson M, Bangerter NK, et al. Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *NeuroImage*. 2018;166:400-424. doi:10.1016/j.neuroimage.2017.10.034

4. Halchenko YO, Meyer K, Poldrack B, et al. DataLad: distributed system for joint management of code, data, and their relationship. *J Open Source Softw*. 2021;6(63):3262. doi:10.21105/joss.03262

5. Wagner AS, Waite LK, Wierzba M, et al. FAIRly big: A framework for computationally reproducible processing of large-scale data. :25.

6. Schaefer A, Kong R, Gordon EM, et al. Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cereb Cortex*. 2018;28(9):3095-3114. doi:10.1093/cercor/bhx179

7. Tian Y, Margulies DS, Breakspear M, Zalesky A. Topographic organization of the human subcortex unveiled with functional connectivity gradients. *Nat Neurosci*. 2020;23(11):1421-1432. doi:10.1038/s41593-020-00711-6

8. Diedrichsen J, Balsters JH, Flavell J, Cussans E, Ramnani N. A probabilistic MR atlas of the human cerebellum. *NeuroImage*. 2009;46(1):39-46. doi:10.1016/j.neuroimage.2009.01.045

9. Brandes N, Linial N, Linial M. PWAS: proteome-wide association study—linking genes and phenotypes by functional variation in proteins. *Genome Biol*. 2020;21(1):173. doi:10.1186/s13059-020-02089-x

10. Alonso AC, Ribeiro SM, Luna NMS, et al. Association between handgrip strength, balance, and knee flexion/extension strength in older adults. Sergi G, ed. *PLOS ONE*. 2018;13(6):e0198185. doi:10.1371/journal.pone.0198185

11. Bobos P, Nazari G, Lu Z, MacDermid JC. Measurement Properties of the Hand Grip Strength Assessment: A Systematic Review With Meta-analysis. *Arch Phys Med Rehabil*. 2020;101(3):553-565. doi:10.1016/j.apmr.2019.10.183

12. Bohannon RW, Schaubert KL. Test–retest reliability of grip-strength measures obtained over a 12-week interval from community-dwelling elders. *J Hand Ther*. 2005;18(4):426-428.

13. Reuter SE, Massy-Westropp N, Evans AM. Reliability and validity of indices of hand-grip strength and endurance: EVALUATION OF GRIP STRENGTH AND ENDURANCE. *Aust Occup Ther J*. 2011;58(2):82-87. doi:10.1111/j.1440-1630.2010.00888.x

14. Gell M, Eickhoff SB, Omidvarnia A, et al. How measurement noise limits the accuracy of brain-behaviour predictions. *Nat Commun*. 2024;15(1):10678.

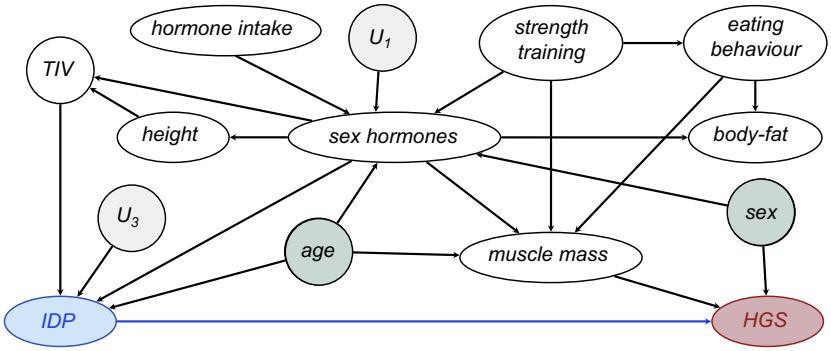
15. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in python. *J Mach Learn Res*. 2011;12:2825-2830.

16. R: Fast Heuristics For The Estimation Of the C Constant Of A... Accessed December 9, 2022. https://search.r-project.org/CRAN/refmans/LiblineaR/html/heuristicC.html

Supplementary materials

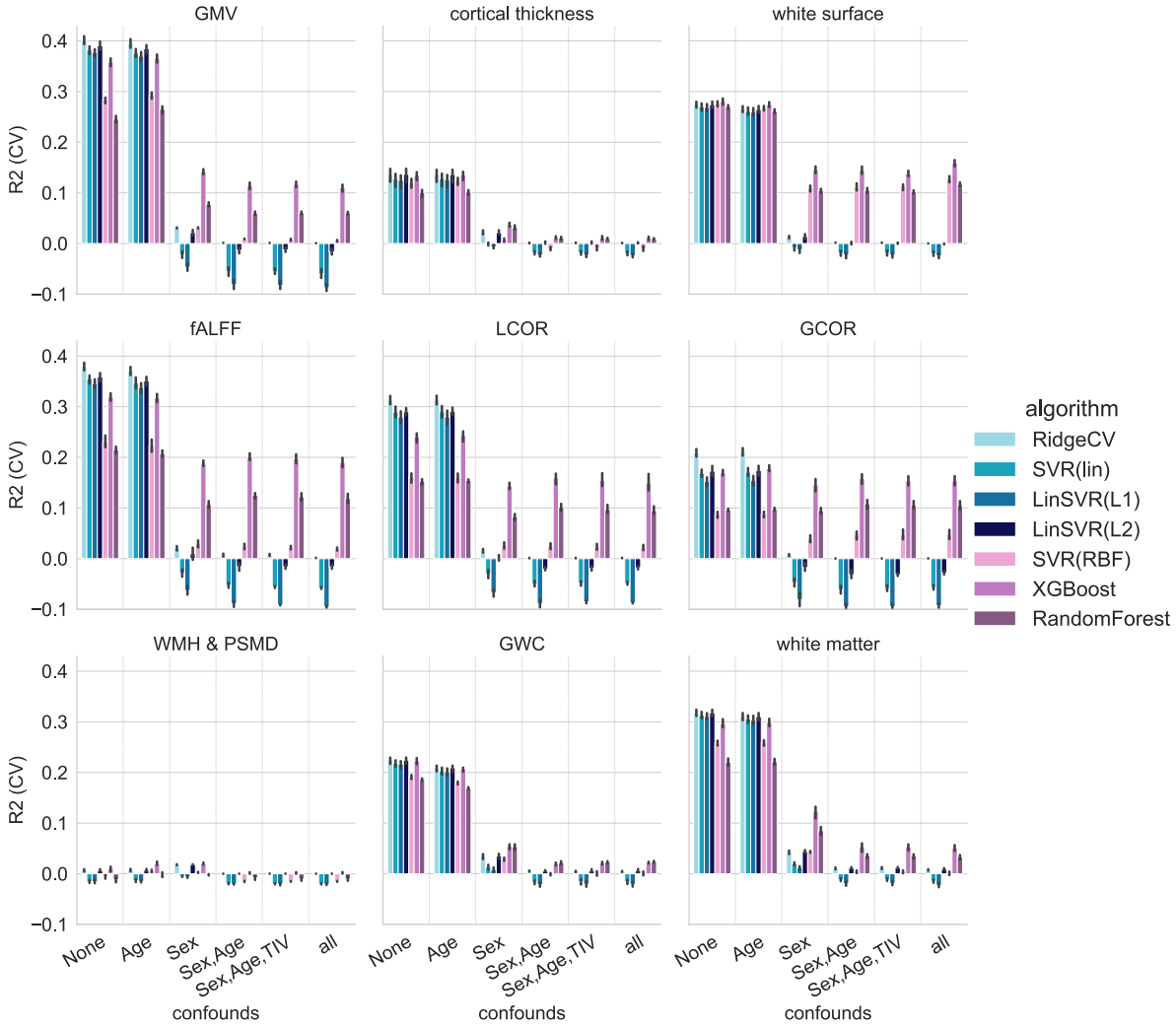
Supplementary methods

1. Theoretical identification of confounding pathways and deconfounders



S-Fig. 1 | DAG for deconfounder selection.

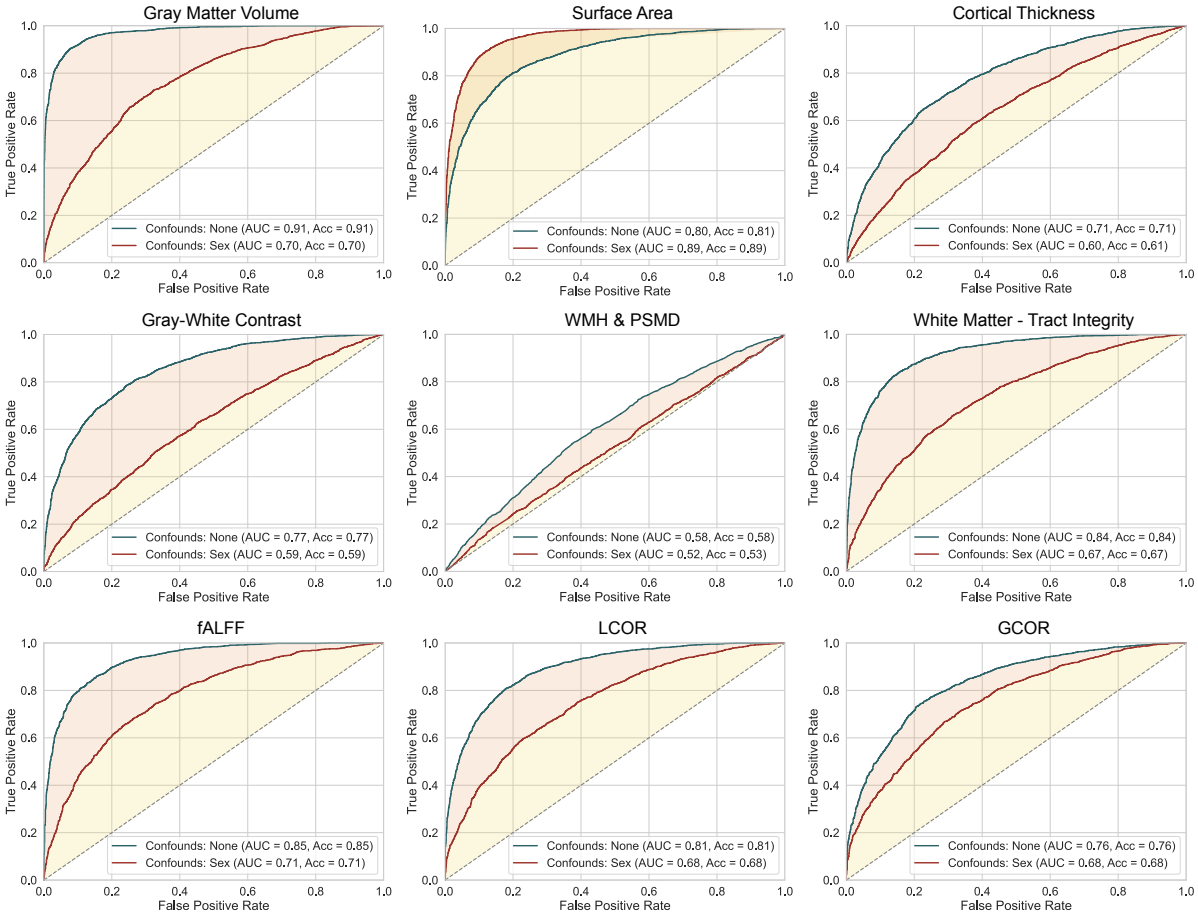
2. Empirical justification of confounding variables and CV results for discarded adjustment scenarios



S-Fig. 2 | Saturation effect of drop in predictive performance when adjusting for more variables than sex and age. all: age, sex, TIV, waist circumference, BMI, body fat percentage, whole-body fat free mass.

3. Empirical justification for choice of sex-split subsamples

a residual predictability with XGBoost after linear sex-adjustment



S-Fig. 3 | Residual predictability of sex using XGBoost as a non-linear algorithm after linear sex-adjustment. The residual predictability motivates to train models on sex-split subsamples to exlcude any sex-related information to be used for predictions.

Supplementary results

1. Statistical model comparisons

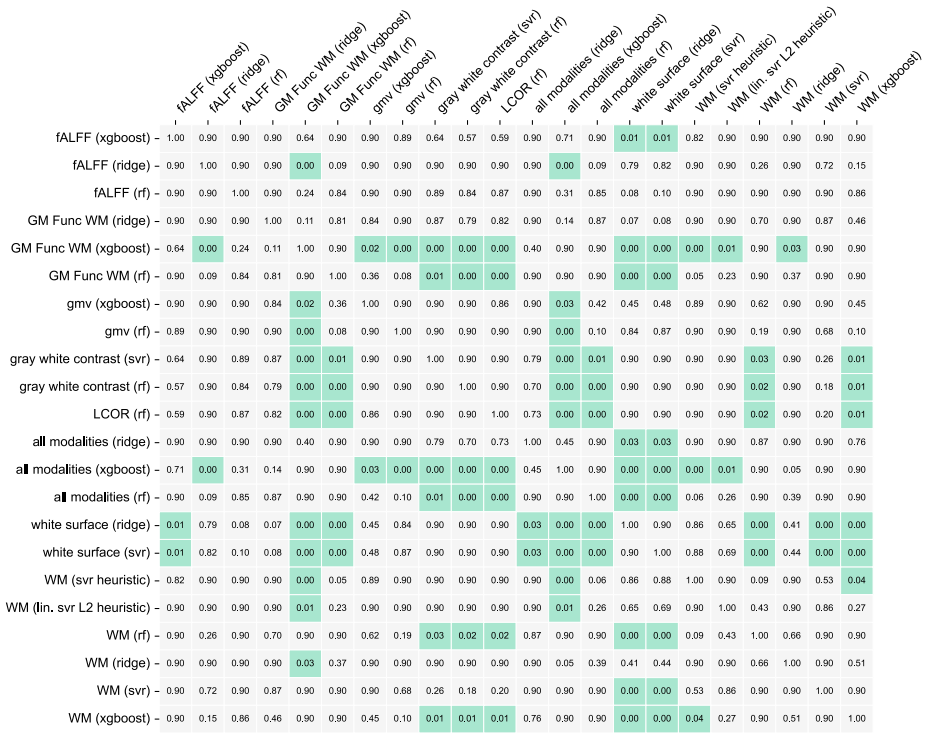
1.1 Friedmann omnisbus-test results

Table 1. Test statistics from Friedmann omnibus test per sex, and metric. Degrees of freedom originate from the number of comapred models with positive R^2.

sex	metric	chisquare	p-value
female	R2	chisquare(21, N=10)=18.76	0.02
	Pearson r	chisquare(21, N=10)=14.41	0.10
	Spearman r	chisquare(21, N=10)=17.10	0.04
	MAE	chisquare(21, N=10)=18.87	0.02
	RMSE	chisquare(21, N=10)=9.27	0.41
male	R2	chisquare(23, N=10)=15.14	0.08
	Pearson r	chisquare(23, N=10)=15.41	0.08
	Spearman r	chisquare(23, N=10)=21.68	0.01
	MAE	chisquare(23, N=10)=38.94	1.18e-05
	RMSE	chisquare(23, N=10)=39.18	1.07e-05

1.2 Nemenyi post-hoc pairwise comparisons

Female



S-Fig. 4 | Nemenyi post-hoc pairwise comparisons of model performances (women). Cell-wise harmonic mean over p-values of the five model evaluation error

metrics (R2, pearson r, spearman r, mean absolute error, root mean squared error) on each of which the Nemenyi post-hoc comparison was performed. Green indicates $p < 0.05$.

Male

S-Fig. 5 | Nemenyi post-hoc pairwise comparisons of model performances (men). Cell-wise harmonic mean over p-values of the five model evaluation error metrics (R2, pearson r, spearman r, mean absolute error, root mean squared error) on each of which the Nemenyi post-hoc comparison was performed. Green indicates $p < 0.05$.

2. Model ranking

2.1 Average ranking per model over metrics and folds

Female

Table 2. Average ranking per model over metrics and folds (women) and cross-metric harmonic mean of p-values of statistical performance difference with top (reference) model (Note: not significant means that the model performed statistically not distinguishably well from the top model, i.e. the model is as good as the top model). Green: not-significantly worse performing than top model. Orange: top (reference) model. Blue: Best performing algorithm per brain feature group.

Model	Average ranking	p_{harm}-value with best
All modalities – xgboost	4.14	1.00
GM, fALFF, WM – xgboost	4.46	0.90

WM-TI – xgboost	6.04	0.90
WM-TI – rf	6.78	0.90
All modalities – rf	7.58	0.90
GM, fALFF, WM – rf	7.84	0.90
WM-TI – svr	8.58	0.90
fALFF - xgboost	9.72	0.71
fALFF - rf	11.18	0.31
GMV - xgboost	11.38	0.03
All modalities - ridge	11.68	0.45
GM, fALFF, WM - ridge	12.80	0.14
WM-TI - ridge	13.04	0.05
GMV – rf	13.18	< 0.001
WM-TI – lin. svr (L2)	13.92	0.01
fALFF – ridge	13.98	< 0.001
GWC – rf	14.34	< 0.001
GWC - svr	14.48	< 0.001
WM-TI – svr	15.18	< 0.001
LCOR – rf	16.16	< 0.001
Surface - svr	18.24	< 0.001
Surface - ridge	18.30	< 0.001

Male

Table 3. Average ranking per model over metrics and folds (men) and cross-metric harmonic mean of p-values of statistical performance difference with top (reference) model (Note: not significant means that the model performed statistically not distinguishably well from the top model, i.e. the model is as good as the top model). Green: not-significantly worse performing than top model. Orange: top (reference) model. Blue: Best performing algorithm per brain feature group.

Model	Average ranking	p_{harm}-value with best
All modalities – xgboost	3.86	1.00
GM, fALFF, WM – xgboost	4.18	0.90
GM, fALFF, WM – rf	6.94	0.90
All modalities – rf	7.76	0.90
WM-TI – rf	8.12	0.81
WM-TI – xgboost	*8.36*	*0.80*

GMV -xgboost	8.44	0.90
fALFF -xgboost	9.36	0.87
GMV - rf	9.54	0.80
WM-TI – svr	10.50	0.81
fALFF – rf	11.66	0.30
All modalities - ridge	12.80	0.27
GM, fALFF, WM -ridge	13.66	0.06
LCOR - xgboost	13.96	0.02
fALFF - ridge	14.08	0.01
GM, fALFF, WM - svr	15.54	< 0.001
WM-TI - ridge	16.10	0.02
LCOR -ridge	16.34	< 0.001
GWC - svr	16.80	< 0.001
WM-TI – lin. SVR (L2)	17.20	< 0.001
GMV – ridge	17.50	< 0.001
CT – svr	18.20	< 0.001
GCOR - rf	18.78	< 0.001
WMH & PSMD - svr	20.32	< 0.001

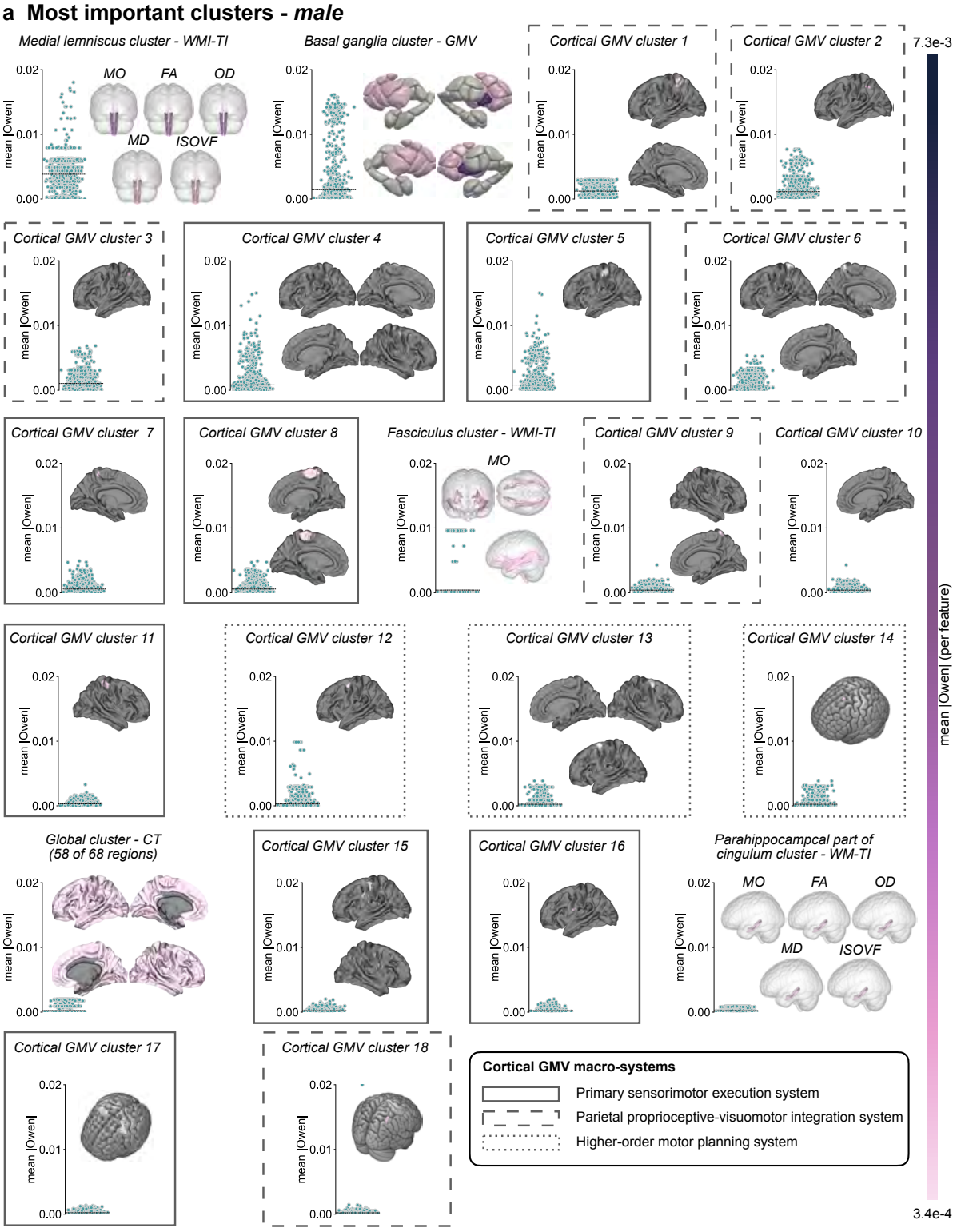
3. OOS prediction of winning models

Table 4. Hyperparameters used for final OOS winning models (all XGBoost).

Model	sex	Used HPOs					
		γ	learning rate	max depth	min child weight	ƛ	sub-sample
All modalities - xgboost	female	0	0.01	6	100	2	0.5
	male	0	0.01	6	100	1.5	0.5
GM, fALFF, WM - xgboost	female	0	0.01	6	150	2	0.5
	male	0	0.01	6	100	1	0.5
WM-TI - xgboost	female	0	0.005	6	100	2	1
	male	0	0.005	10	200	1.5	1
fALFF - xgboost	Female	0	0.01	6	100	1	0.5
	male	0	0.01	6	150	1.5	0.5
	female	0	0.01	6	100	1	0.5

GMV - XGBoost	male	0	0.005	10	200	1.5	1
		Grid					
		0, 0.1	0.005, 0.01, 0.02	6, 10	100, 150, 200, 300	1, 1.5, 2	0.5, 1

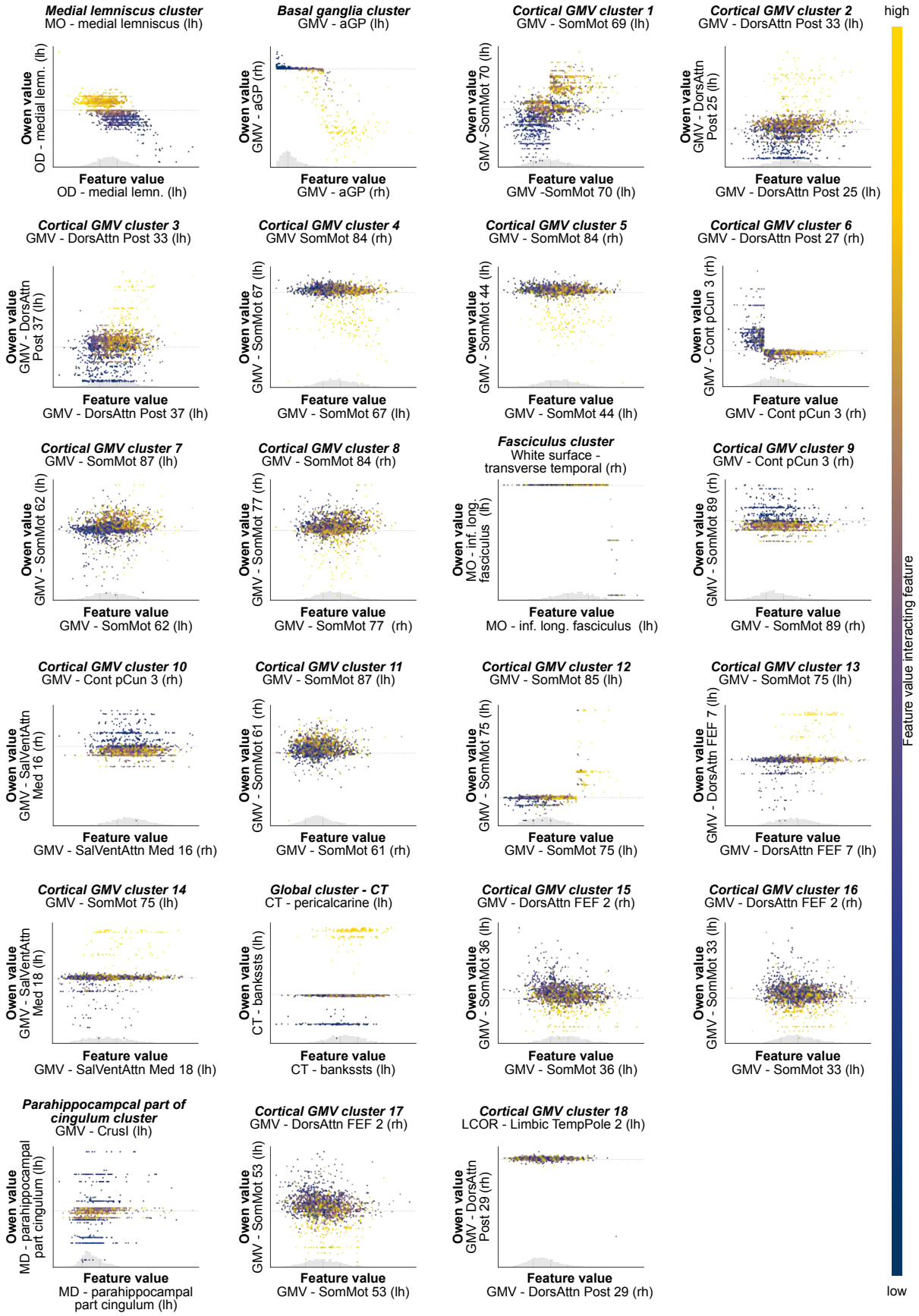
4. Most important clusters – male (corresponding to Fig. 3c females)



S-Fig. 6 | Top 23 most important feature clusters of the male subpopulation model. The scatter plot shows the mean absolute Owen value across all features of the clusetr per subject. The brain plots show the mean absolute Owen value per feature of each cluster across subjects (global explanation), ranging from 3.4e-4 to 7.3e-3 across all clusters. For the WM-TI feature group, the same tract could appear multiple times per cluster with a different measure, so that all measures are displayed with their respective value. Cortical GMV clusters were sorted into 3 major macro-systems. WM-TI: White Matter Tract Integrity, GMV: Gray Matter Volume,

MO: Diffusion Tensor Mode, OD: Orientation Dispersion, MD: Mean Diffusivity, ISOVF: Isotropic Volume Fraction, FA: Fractional Anisotropy

5. Owen-feature-feature interactions – male (corresponding to Fig. 3d females)



S-Fig. 7 | Interaction of each cluster's top feature's Owen value with its feature value and the feature value of the most relevantly interacting feature to derive

directionality of feature-importance. y-axis: Top feature's Owen value, x-axis: Top feature's feature value, feature beneath cluster name: most relevant interacting feature, colour: feature value of the interacting feature. Indent brain plots show the top feature within its cluster as shown in **S-Fig. 6**. WM-TI: White Matter Tract Integrity, GMV: Gray Matter Volume, MO: Diffusion Tensor Mode, OD: Orientation Dispersion, MD: Mean Diffusivity, ISOVF: Isotropic Volume Fraction, FA: Fractional Anisotropy, ICVF: Intra-Cellular Volume Fraction, lh: left hemisphere, rh: right hemisphere.