This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.
Understanding and stratifying brain health through blood-based omics data

Danielle Alisha Gadd

THE UNIVERSITY
of EDINBURGH

Doctor of Philosophy
The University of Edinburgh
2024
Abstract

Brain health across the lifespan is dynamic and influenced by a complex interplay of genetics and the environment. Age-related neurological diseases are a growing burden on healthcare systems and society. Individuals that do not have overt diagnoses of neurological diseases will still experience declines in cognitive ability and reductions in grey matter volumes as they age. Understanding why some brains are healthier than others may provide insight into targets for the preservation of brain health. Early identification of individuals at high-risk of neurological diseases is a priority for preventative strategies.

Proteins are the effector molecules of disease in the body and are often the targets of therapeutic interventions. Blood samples can be used to derive measures of many thousands of blood proteins and are routinely collected in clinical and research settings. Another measure available from blood is DNA methylation (DNAm), which is an epigenetic mechanism that can regulate gene expression and protein levels. DNAm is thought to record the body’s response to a range of biological and environmental factors.

The first aim of this thesis is to perform methylome-wide association studies (MWAS) of circulating proteins, with a focus on those related to brain health. DNAm patterns can also be used to derive proxy scores for protein levels – an approach that is somewhat analogous to polygenic scores. These proxies are known as protein epigenetic scores (or EpiScores). In some cases, protein EpiScores outperform measured proteins in associations with brain imaging and lifestyle traits, possibly due to the relative lack of stability observed across some single time point protein measurements. Protein EpiScores could represent biomarkers for risk stratification. However, this has not been examined at scale. Consequently, the second aim of this thesis is to develop a comprehensive set of protein EpiScores and evaluate them as tools for risk stratification. For neurological diseases such as Alzheimer’s dementia, damage is thought to occur in the brain decades prior to symptom presentation. Dysfunction at the blood brain barrier can facilitate leakage of proteins into the bloodstream in the early stages of neurological disease. Similarly, peripherally-produced proteins may also serve as warning signatures. Therefore, the final aim of this thesis is to conduct an assessment of blood protein signatures of incident neurological diseases and associated morbidities.
In Chapters 1-3, I provide an overview of brain health and disease, blood-based molecular measures and key statistical approaches. In Chapter 4, I detail the population cohorts used in this thesis, before outlining my research aims and chosen methodologies in Chapter 5. In Chapter 6, I study serum measurements of S100 calcium-binding protein β (S100β) – a well-characterised marker with links to neuroinflammation and brain disease. I map the epigenetic and genetic signatures of this protein and test for evidence of a putative causal relationship between the protein and Alzheimer’s dementia. Chapter 7 extends this approach via a proteome-wide analysis. Instead of focusing on a single candidate biomarker, I conduct MWAS of 4,235 plasma proteins, identifying 2,928 associations (n ≥ 778 individuals). I also scan the proteome against fifteen brain health traits, identifying 405 associations involving 191 proteins. I integrate these signatures to highlight potential pathways between the methylome and proteome that may have relevance to brain health. In Chapter 8, I consider 953 possible plasma proteins for protein EpiScore development (N in the training set ranged from 706 to 944 individuals). I evaluate these EpiScores in independent populations, with 109 statistically significant EpiScores taken forward and modelled as biomarkers of incident diseases in 9,537 individuals. I build on this work in Chapter 9, where I generate protein EpiScores for GDF15 and NT-proBNP, which are two leading cardiovascular disease (CVD) markers implicated in brain health. I use a much-expanded sample size (n ≥ 16,963) to train these scores and show that they replicate protein-disease associations and associate with brain health outcomes. Finally, in Chapter 10, I test individual protein associations with 23 incident morbidities and death. I map whether proteins are markers for multiple neurological diseases, or specific to singular diseases. I then create ProteinScores for 10-year onset stratification of each incident outcome. ProteinScores for Alzheimer’s dementia and Parkinson’s disease are amongst the best-performing 10-year onset scores.

The work done in this thesis provides information to help us identify those at the highest risk of developing neurological diseases (and associated morbidities), up to a decade prior to onset. My findings also tell us about the individual protein and DNAm patterns that associate with brain health and disease. Taken together, these results indicate that profiling epigenetic and proteomic information from our blood may improve our understanding of brain ageing. This work sits within the ethos of early detection and prevention, which should be at the heart of healthcare as we age.
Lay Summary

Blood tests are an accessible way to measure the biology occurring in the human body. Biological markers from the blood can provide insight into the impact that our lifestyles and environments may have on our brain health. Biological markers from the brain can also be released into the blood, which can warn us that damage may be occurring in the brain before we even experience symptoms.

This PhD asks whether information from a single blood test can identify markers of brain health and predict whether individuals will develop neurological diseases. To do this, I combine biological measurements from the blood with individuals' GP and hospital records, their brain scans and their performance in cognitive tests. I work with large-scale datasets such as Generation Scotland, which has over 20,000 participants. I focus on two main markers from the blood in this work. The first are proteins, which are the building blocks of our bodies that keep us healthy, but are often also the drivers of disease. The second is DNA methylation, which represents small chemical changes to the DNA that makes up our genome. These chemical changes can act like dimmer switches to turn genes up or down, thereby changing the levels of proteins we produce. DNA methylation can be influenced by lifestyle and environmental factors. I study links between DNA methylation and proteins in the blood to understand if these patterns can be used to identify individuals at high-risk of developing diseases.

In my first two studies, I mapped the DNA methylation patterns for over 4000 proteins from the blood. I also pinpointed proteins in the blood that were linked to how well people performed on a cognitive test, or what their brain imaging measures were. In my third and fourth studies, I derived DNA methylation patterns that are predictive of over 900 blood proteins. I narrowed this down to a set of 109 predictors, which were associated with the onset of diseases, over 16 years of follow-up. These predictors helped identify people at the highest risk for diseases like ischaemic stroke and type 2 diabetes. Even though type 2 diabetes is considered a metabolic disease, it can have a huge impact on our brain health. Identifying people at risk of diseases that often predispose us to poorer brain health as we age may therefore be important to keep our brains as healthy as possible, for as long as possible.

In my fifth study, I scanned 1,468 proteins from the blood to identify those that were early markers of disease over 16 years. I identified which proteins were associated with the
onset of a range of diseases including Parkinson’s disease, Alzheimer’s dementia, amyotrophic lateral sclerosis and ischaemic stroke. I integrated the results from this study into a visualisation tool that allows users to browse my results.

These findings are likely to be useful in identifying those at high risk of neurological diseases, sometimes decades before a diagnosis. The information also builds on our understanding of how our lifestyle, genetics and environments shape our brain health. My results may be used to identify biological targets for drug development to prevent neurological disease. The predictors I built show that from a single blood test we can identify individuals at high risk of developing Alzheimer’s dementia, Parkinson’s disease and other diseases like type 2 diabetes. There is potential for these types of predictors to be included in clinical or commercial settings eventually. For example, in the way that you can pay to learn about your genetic ancestry and predisposition to certain diseases, you might be able to have a DNA methylation or protein analysis of your health in the future.
Declaration of Originality

I declare that this thesis is my own composition and that it has not been submitted for any other degree or professional qualification at this university or any other institution. Parts of the work comprising this thesis have been previously published. The included publications are my own work, except where indicated otherwise.

The work presented in Chapter 6 has been published in *Wellcome Open Research*. Key author contributions are as follows: D.A.G and R.E.M conceived and designed the research; D.A.G and R.F.H. co-supervised R.I.M. in initial analyses. D.A.G. then re-ran and expanded on the original statistical analyses; D.A.G and R.E.M drafted the article, D.L.McC., S.E.H., S.R.C, R.A.S. and N.J.A. contributed to data preparation and processing; all authors reviewed the manuscript.

The work presented in Chapter 7 has been published in *Nature Communications*. Key author contributions are as follows: D.A.G and R.E.M conceived and designed the research; D.A.G, led the statistical analyses, supported by R.E.M and N.A.R.; D.A.G. and R.E.M drafted the article; R.F.H., D.L.Mc.C., L.S., M.C.B., T.C. and A.M.M. consulted on methodology; all authors reviewed the manuscript.

The work presented in Chapter 8 has been published in *eLife*. Author contributions are as follows: D.A.G, R.F.H., D.L.Mc.C., S.B.Z., K.S. and R.E.M conceived and designed the research; D.A.G, R.F.H., S.B.Z., D.L.Mc.C. and R.E.M conducted the statistical analyses; D.A.G. and R.E.M drafted the article; all authors reviewed the manuscript.

The work presented in Chapter 9 has been submitted for publication. Author contributions are as follows: D.A.G, R.E.M, P.W., and N.S., conceived and designed the research; D.A.G conducted the statistical analyses; D.A.G and R.E.M drafted the article; all authors reviewed the manuscript.

The work presented in Chapter 10 has been submitted for publication. Author contributions are as follows: D.A.G, R.E.M, C.N.F., R.F.H. and B.B.S. conceived and designed the research; D.A.G conducted the statistical analyses; D.A.G and R.E.M drafted the article; all authors reviewed the manuscript.

Signed: Danni Gadd Date: 02/03/24
Acknowledgements

Firstly, I would like to thank my supervisor, Prof Riccardo Marioni for his unwavering support. It is rare to find a mentor with such integrity, academic excellence and dedication and I am truly privileged to have worked with you. I would also like to thank my co-supervisors Prof Craig Ritchie, Prof Graciela Muniz Terrera and Dr Diego Oyarzún, and thesis chair Prof Veronique Vitart for their advice during my studies.

Thank you to the volunteers of the Generation Scotland, Cooperative Health Research in the Region of Augsburg (KORA), UK Biobank and Lothian Birth cohorts. Your contributions to the datasets available in these cohorts have made my PhD possible. I also acknowledge Wellcome for funding my PhD studies. Thank you to Prof David Porteous, Dr Simon Cox, Dr Kathy Evans and Prof Andrew McIntosh for their assistance with a number of my projects. Thank you also to Prof Karsten Suhre, Dr Shaza Zaghlool, Dr Benjamin Sun and Dr Chris Foley for their collaborations on empirical work included in this thesis. Thank you also to Dr Allan McRae and Prof Naomi Wray for their collaboration as part of my research visit to Brisbane in the autumn of 2022.

A huge thank you to Dr Robert Hillary and Dr Daniel McCartney for their mentorship, patience and counsel. Thank you to all lab members past and present, Anna Stevenson, Yipeng Cheng, Ella Davyson, Ola Chybowska and Elena Bernabeu for their friendship and expertise. Thank you to all my fellow members of the Translational Neuroscience PhD programme, Kevin, Eleanor, Amelia, Tyler, to name a few, for their support and friendship over the past four years. Thank you to the directors of the programme and to Jane Haley and Marja Main for an enjoyable and enriching learning experience.

Thank you to Ian and Zoe, for welcoming me to Glasgow, extending your friendship, and bringing joy to my weekends. Thank you to Emma for all the voice notes. Thank you to my parents, Jane and Simon, for a lifetime of support. Thank you to Oliver, for encouraging me to pursue this PhD, and for all the laughs, silliness and love along the way. Finally, thank you to my Nan, Susan – though I did not know you for as long as I’d have hoped, your brain tumour was the reason I became interested in science. I hope I’ve done and continue to do you proud.
## Contents

1 BRAIN HEALTH .................................................................................................................. 24

1.1 CENTRAL THEMES IN BRAIN HEALTH ........................................................................... 24
  1.1.1 The ageing brain ......................................................................................................... 24
  1.1.2 The burden of neurological diseases ......................................................................... 27
  1.1.3 Human complexity ................................................................................................... 29
  1.1.4 Characterising neurodegenerative diseases ............................................................ 30

1.2 MEASUREMENT OF BRAIN HEALTH ............................................................................. 31
  1.2.1 Neuroimaging ........................................................................................................... 31
  1.2.2 Cognitive tests .......................................................................................................... 34

1.3 MOLECULAR INDICATORS OF BRAIN HEALTH ............................................................. 35
  1.3.1 Neuroinflammation .................................................................................................. 36
  1.3.2 Proteomic profiling of CSF ....................................................................................... 36
  1.3.3 Proteomic profiling of Blood .................................................................................... 37
  1.3.4 Genetic measures ..................................................................................................... 38
  1.3.5 DNA methylation measures ..................................................................................... 39
  1.3.6 Variability in DNAm measures ................................................................................ 41
  1.3.7 DNAm and brain health ........................................................................................... 42
  1.3.8 Molecular architecture of proteins .......................................................................... 44
  1.3.9 Instability of protein measures ................................................................................ 44

1.4 MOTIVATION FOR STUDYING OMICS MARKERS OF BRAIN HEALTH .................... 48

1.5 SUMMARY ....................................................................................................................... 51

2 LINKING THE PROTEOME AND METHYLOME ............................................................. 52

2.1 THE CIRCULATING PROTEOME ..................................................................................... 52
  2.1.1 Proteomic profile of blood ........................................................................................ 52
  2.1.2 Proteomic technologies ............................................................................................ 54

2.2 MAPPING ASSOCIATIONS BETWEEN THE PROTEOME AND METHYLOME ......... 57
  2.2.1 MWAS ....................................................................................................................... 57
  2.2.2 Rationale for MWAS of proteins .......................................................................... 59
  2.2.3 MWAS of proteins .................................................................................................. 59
  2.2.4 Challenges around causality in pQTM .................................................................. 68

2.3 EPIGENETIC SCORING ..................................................................................................... 68
2.3.1  EpiScores ................................................................. 69
2.3.2  Methods for EpiScore generation .................................. 69
2.3.3  EpiScore use case .................................................... 74
2.3.4  Protein EpiScores ..................................................... 74
2.3.5  Rationale for generating Protein EpiScores ....................... 83
2.4  SUMMARY .................................................................. 87

3  PROTEIN-INFORMED EVALUATION OF BRAIN HEALTH ........... 88

3.1  EVALUATING BIOMARKERS ............................................ 88
  3.1.1  Rationale for early blood biomarker detection .................. 88
  3.1.2  Cox PH analyses ...................................................... 89
  3.1.3  Cox PH assumptions .................................................. 90
  3.1.4  Non-methylation omics composite scores ......................... 90

3.2  METHODOLOGICAL CONSIDERATIONS ................................ 91
  3.2.1  Causal inference ...................................................... 91
  3.2.2  Approaches to causal inference .................................... 92
  3.2.3  Colocalisation ......................................................... 94
  3.2.4  Multiple testing considerations ..................................... 96

3.3  SUMMARY .................................................................. 96

4  STUDY COHORTS AND MOLECULAR SAMPLES ....................... 99

4.1  GENERATION SCOTLAND .................................................. 99
  4.1.1  Ethics and funding ..................................................... 100
  4.1.2  Genetic data ........................................................... 100
  4.1.3  DNA methylation data .............................................. 101
  4.1.4  Proteomics data ....................................................... 101

4.2  THE LOTHIAN BIRTH COHORTS 1936 AND 1921 .................... 102
  4.2.1  Ethics and funding ..................................................... 102
  4.2.2  Genetic data ........................................................... 103
  4.2.3  DNA methylation data .............................................. 103
  4.2.4  Proteomic data ....................................................... 104

4.3  THE KORA F4 COHORT .................................................... 105
  4.3.1  Ethics and funding ..................................................... 105
  4.3.2  Genetic data ........................................................... 105
9 CREATION OF GDF15 AND NT-PROBNP EPISCORES AND ASSESSMENT OF THEIR RELEVANCE TO BRAIN HEALTH .............................................................. 176

9.1 INTRODUCTION ........................................................................................................... 176

9.2 METHYLATION-BASED PREDICTORS OF SERUM GDF15 AND NT-PROBNP TRACK ONSET OF CLINICAL MORBIDITIES AFFECTING THE BODY AND BRAIN ............................... 177

9.3 BACKGROUND .............................................................................................................. 179

9.4 RESULTS ...................................................................................................................... 182

9.4.1 Sample populations .................................................................................................. 182

9.4.2 GDF15 and NT-proBNP disease associations ......................................................... 182

9.4.3 GDF15 and NT-proBNP epigenetic associations .................................................. 183

9.4.4 EpiScores for GDF15 and NT-proBNP in Generation Scotland .............................. 185

9.4.5 EpiScore replication of protein biomarker associations ......................................... 186

9.4.6 EpiScore application to the LBC1936 external cohort ........................................... 187

9.4.7 EpiScore assessment in LBC1936 .......................................................................... 188

9.5 DISCUSSION ............................................................................................................... 190

9.6 METHODS .................................................................................................................. 194

9.6.1 Generation Scotland ............................................................................................... 194

9.6.2 Lothian Birth Cohort 1936 .................................................................................... 194

9.6.3 Epigenome-wide association studies in Generation Scotland .............................. 195

9.6.4 EpiScore development ............................................................................................ 196

9.6.5 EpiScore testing ...................................................................................................... 197

9.6.6 Cox proportional hazards analyses in Generation Scotland ................................. 197

9.6.7 COVID-19 analyses in Generation Scotland ........................................................ 198

9.6.8 EpiScore associations with brain health traits in LBC1936 .................................... 199

9.7 CONCLUSION ............................................................................................................. 201

10 PLASMA PROTEIN PROFILES ASSOCIATED WITH THE ONSET OF NEUROLOGICAL DISEASES AND ASSOCIATED MORBIDITIES ......................................................... 202

10.1 INTRODUCTION ......................................................................................................... 202

10.2 BLOOD PROTEIN ASSESSMENT OF LEADING INCIDENT DISEASES AND MORTALITY IN UK BIOBANK ........................................................................................................... 203

10.3 BACKGROUND .......................................................................................................... 204

10.4 RESULTS ................................................................................................................... 205

10.4.1 The UKB-PPP sample ............................................................................................ 205
**Abbreviations**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>α-syn</td>
<td>Alpha-synuclein</td>
</tr>
<tr>
<td>A2M</td>
<td>Alpha-2-macroglobulin</td>
</tr>
<tr>
<td>Aβ</td>
<td>Amyloid beta</td>
</tr>
<tr>
<td>AD</td>
<td>Alzheimer’s dementia</td>
</tr>
<tr>
<td>ADHD</td>
<td>Attention deficit hyperactivity disorder</td>
</tr>
<tr>
<td>AHRR</td>
<td>Aryl-hydrocarbon receptor repressor</td>
</tr>
<tr>
<td>ALB</td>
<td>Albumin</td>
</tr>
<tr>
<td>ALS</td>
<td>Amyotrophic lateral sclerosis</td>
</tr>
<tr>
<td>ANX</td>
<td>Anxiety or phobic disorders</td>
</tr>
<tr>
<td>APOA</td>
<td>Apolipoprotein A</td>
</tr>
<tr>
<td>APOE</td>
<td>Apolipoprotein E</td>
</tr>
<tr>
<td>ASD</td>
<td>Autism spectrum disorder</td>
</tr>
<tr>
<td>AUC</td>
<td>Area under the curve</td>
</tr>
<tr>
<td>B2M</td>
<td>Beta 2-microglobulin protein</td>
</tr>
<tr>
<td>BBB</td>
<td>Blood brain barrier</td>
</tr>
<tr>
<td>BIN1</td>
<td>Bridging integrator 1</td>
</tr>
<tr>
<td>BMAA</td>
<td>Neurotoxin β-methylamino-l-alanine</td>
</tr>
<tr>
<td>BMI</td>
<td>Body mass index</td>
</tr>
<tr>
<td>BP</td>
<td>Bipolar disorder</td>
</tr>
<tr>
<td>CCL11</td>
<td>C-C motif chemokine 11</td>
</tr>
<tr>
<td>CD</td>
<td>Cluster of differentiation</td>
</tr>
<tr>
<td>CHI3L1</td>
<td>Chitinase 3 like 1</td>
</tr>
<tr>
<td>CI</td>
<td>Confidence interval</td>
</tr>
<tr>
<td>CNS</td>
<td>Central nervous system</td>
</tr>
<tr>
<td>COPD</td>
<td>Chronic obstructive pulmonary disease</td>
</tr>
<tr>
<td>COJO</td>
<td>Conditional and joint analyses</td>
</tr>
<tr>
<td>COVID-19</td>
<td>Coronavirus disease 2019</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>CpG</td>
<td>Cytosine-Guanine dinucleotide</td>
</tr>
<tr>
<td>CRP</td>
<td>C-reactive protein</td>
</tr>
<tr>
<td>CSF</td>
<td>Cerebrospinal fluid</td>
</tr>
<tr>
<td>CVD</td>
<td>Cardiavascular disease</td>
</tr>
<tr>
<td>CXCL</td>
<td>Motif chemokine ligand</td>
</tr>
<tr>
<td>DALYs</td>
<td>Disability-adjusted life years</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>DNAm</td>
<td>DNA methylation</td>
</tr>
<tr>
<td>EA</td>
<td>Educational attainment</td>
</tr>
<tr>
<td>eQTL</td>
<td>Expression quantitative trait locus</td>
</tr>
<tr>
<td>EpiScore</td>
<td>Epigenetic score</td>
</tr>
<tr>
<td>EpiSmokEr</td>
<td>Epigenetic Smoking status Estimator</td>
</tr>
<tr>
<td>EU</td>
<td>European Union</td>
</tr>
<tr>
<td>EWAS</td>
<td>Epigenome-wide association studies</td>
</tr>
<tr>
<td>F2RL3</td>
<td>F2R-like thrombin or trypsin receptor 3</td>
</tr>
<tr>
<td>FCGR3B</td>
<td>Fc Gamma Receptor IIIb</td>
</tr>
<tr>
<td>FDR</td>
<td>False discovery rate</td>
</tr>
<tr>
<td>FGA</td>
<td>Fibrinogen A</td>
</tr>
<tr>
<td>GCTA</td>
<td>Genome-wide complex trait analysis</td>
</tr>
<tr>
<td>GDF15</td>
<td>Growth differentiation factor 15</td>
</tr>
<tr>
<td>GP</td>
<td>General Practitioner</td>
</tr>
<tr>
<td>GRM</td>
<td>Genetic relationship matrix</td>
</tr>
<tr>
<td>GS</td>
<td>Generation Scotland</td>
</tr>
<tr>
<td>GWAS</td>
<td>Genome-wide association studies</td>
</tr>
<tr>
<td>HAGH</td>
<td>Hydroxyacylglutathione hydrolase</td>
</tr>
<tr>
<td>HbA1c</td>
<td>Glycated haemoglobin</td>
</tr>
<tr>
<td>hQTL</td>
<td>Histone quantitative trait locus</td>
</tr>
<tr>
<td>HR</td>
<td>Hazard ratio</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>HRC</td>
<td>Haplotype Reference Consortium</td>
</tr>
<tr>
<td>HPPA</td>
<td>Human Plasma PeptideAtlas</td>
</tr>
<tr>
<td>HPPP</td>
<td>Human Plasma Proteome Project</td>
</tr>
<tr>
<td>HUPO</td>
<td>The Human Proteome Organisation</td>
</tr>
<tr>
<td>HWE</td>
<td>Hardy-Weinberg equilibrium</td>
</tr>
<tr>
<td>ICD</td>
<td>International classification of disease codes</td>
</tr>
<tr>
<td>IHD</td>
<td>Ischemic heart disease</td>
</tr>
<tr>
<td>ICD</td>
<td>International classification of diseases</td>
</tr>
<tr>
<td>Ig</td>
<td>Immunoglobulin</td>
</tr>
<tr>
<td>IL</td>
<td>Interleukin</td>
</tr>
<tr>
<td>IL18R1</td>
<td>Interleukin 18 receptor 1</td>
</tr>
<tr>
<td>ITIH1/3/4</td>
<td>Inter-alpha-trypsin inhibitor heavy chain H1/H3/H4</td>
</tr>
<tr>
<td>IV</td>
<td>Instrumental variable</td>
</tr>
<tr>
<td>Kb</td>
<td>Kilobase</td>
</tr>
<tr>
<td>KORA</td>
<td>Cooperative Health Research in the Region of Augsburg</td>
</tr>
<tr>
<td>LAG3</td>
<td>Lymphocyte Activation Gene 3</td>
</tr>
<tr>
<td>LASSO</td>
<td>Least absolute shrinkage and selection operator</td>
</tr>
<tr>
<td>LBC1921</td>
<td>Lothian Birth Cohort 1921</td>
</tr>
<tr>
<td>LBC1936</td>
<td>Lothian Birth Cohort 1936</td>
</tr>
<tr>
<td>LD</td>
<td>Linkage disequilibrium</td>
</tr>
<tr>
<td>MAF</td>
<td>Minor allele frequency</td>
</tr>
<tr>
<td>MAPT</td>
<td>Microtubule associated protein tau</td>
</tr>
<tr>
<td>MATN3</td>
<td>Matrilin-3</td>
</tr>
<tr>
<td>Mb</td>
<td>Megabase</td>
</tr>
<tr>
<td>MDD</td>
<td>Major depressive disorder</td>
</tr>
<tr>
<td>MDGA1</td>
<td>MAM Domain Containing Glycosylphosphatidylinositol Anchor 1</td>
</tr>
<tr>
<td>MMP</td>
<td>Matrix metalloproteinase</td>
</tr>
<tr>
<td>MMSE</td>
<td>Mini-mental state examination</td>
</tr>
</tbody>
</table>
MOMENT: multi-component MLM-based omic association excluding the target

mQTL: Methylation quantitative trait locus

MR: Mendelian randomisation

MRI: Magnetic resonance imaging

mRNA: Messenger RNA

MS4A4A: Membrane-spanning 4-domains A4A

Mspec: Mass spectrometry

MWAS: Methylome-wide association study

NEFL: Neurofilament light chain

NEP: Neprilysin

NG: Neurogranin

NLRC5: NOD-like receptor family CARD domain containing 5

NT-proBNP: N-terminal-pro B-type natriuretic peptide

OR: Odds ratio

OSCA: OmicS-data-based Complex trait Analysis

PAPPA: Pregnancy-associated plasma protein-A

PCA: Principal component analysis

PCR: Polymerase chain reaction

PEA: Proximity extension assay

PET: Positron emission tomography

PH: Proportional hazards

PIP: Posterior inclusion probability

pQTL: Protein quantitative trait locus

pQTM: Protein quantitative trait methylation locus

PRS: Polygenic risk score

p-tau: Hyperphosphorylated tau

QMDiab: Qatar Metabolomics Study on Diabetes

S100β: S100 calcium-binding protein β
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCZ</td>
<td>Schizophrenia</td>
</tr>
<tr>
<td>SD</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>SE</td>
<td>Standard error</td>
</tr>
<tr>
<td>SEM</td>
<td>Structural equation model</td>
</tr>
<tr>
<td>SELEX</td>
<td>systematic evolution of ligands by exponential enrichment</td>
</tr>
<tr>
<td>SERPINA1</td>
<td>Alpha-1-antitrypsin</td>
</tr>
<tr>
<td>SIMD</td>
<td>Scottish index of multiple deprivation</td>
</tr>
<tr>
<td>SNP</td>
<td>Single-nucleotide polymorphism</td>
</tr>
<tr>
<td>SOMAmer</td>
<td>Slow Off-rate Modified Aptamer</td>
</tr>
<tr>
<td>sQTL</td>
<td>Splicing quantitative trait locus</td>
</tr>
<tr>
<td>STRADL</td>
<td>Stratifying Resilience and Depression Longitudinally</td>
</tr>
<tr>
<td>TNF</td>
<td>Tumour necrosis factor</td>
</tr>
<tr>
<td>TREM2</td>
<td>Triggering Receptor Expressed On Myeloid Cells 2</td>
</tr>
<tr>
<td>t-tau</td>
<td>total-tau</td>
</tr>
<tr>
<td>TTR</td>
<td>Thyroxine binding prealbumin</td>
</tr>
<tr>
<td>UK</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>UKB</td>
<td>UK Biobank</td>
</tr>
<tr>
<td>UKB-PPP</td>
<td>UK Biobank Pharma Proteomics Project</td>
</tr>
<tr>
<td>USA</td>
<td>United States of America</td>
</tr>
<tr>
<td>VGF</td>
<td>VGF Neurosecretory protein VGF (non-acronymic)</td>
</tr>
<tr>
<td>WFDC2</td>
<td>Whey acidic protein four-disulfide core domain protein 2</td>
</tr>
</tbody>
</table>
Figures

Figure 1-1. Dynamic changes in brain morphology and the onset of neurological diseases across the life course. .......................................................... 26
Figure 1-2. Disability-adjusted life-years resulting from neurological diseases. ...... 28
Figure 1-3. Individual differences in ageing hallmarks of the brain.................... 33
Figure 1-4. Crystallised and fluid intelligence over the human life course. ........... 35
Figure 1-5. Infinium DNA methylation detection example................................ 40
Figure 1-6. Sources of variability in DNAm.................................................. 43
Figure 1-7. Test-retest reliability for four inflammatory blood markers............. 46
Figure 1-8. Test-retest variability of plasma protein markers of Alzheimer's dementia. ................................................................. 47
Figure 1-9. Motivation for integrating omics markers of brain health. .............. 50
Figure 2-1. Proteins detected in human blood across three platforms. .............. 53
Figure 2-2. Olink Proximity Extension Assay................................................. 55
Figure 2-3. Multiplexed SOMAmer affinity assay. ....................................... 56
Figure 2-4. The rise in MWAS from 2009 to 2022. .................................... 57
Figure 2-5. 98 pQTM that replicated across the KORA and QMDiab populations . 61
Figure 2-6. CRP EpiScore in the LBC1936 population by age and inter-wave stability. .............................................................................. 78
Figure 2-7. Motivation for generating additional protein EpiScores............... 86
Figure 3-1. Value of metabolomic risk scores is endpoint dependent. ............ 91
Figure 3-2. Mendelian randomisation principles and assumptions.................. 93
Figure 3-3. Causal variant situations that are tested in colocalisation approaches. ... 95
Figure 3-4. Motivation for biomarker approaches in brain health. .................. 98
Figure 5-1. Cohort study populations used in Chapters 6-10 of this thesis. ....... 116
Figure 9-1. Study design for this assessment of GDF15 and NT-proBNP EpiScores as biomarkers................................................................. 181
Figure 9.2. Disease associations for GDF15 and NT-proBNP in Generation Scotland (N ≥ 16,963).

Figure 9.3. Comparison of EpiScores versus measured protein equivalents in fully-adjusted associations with incident diseases in the Generation Scotland test sample (N ≥ 2,808).

Figure 9.4. External assessment of the GDF15 and NT-proBNP EpiScores in LBC1936.

Figure 10.1. Individual protein associations with incident outcomes in the UK Biobank (N=47,600).

Figure 10.2. Value offered by ProteinScores for incident outcomes in the UK Biobank.

Figure 10.3. Exploration of the type 2 diabetes ProteinScore.
Tables

Table 2-1. Summary of key MWAS/EWAS studies profiling blood protein levels. ..... 65
Table 2-2. Examples of common DNA-methylation based EpiScore approaches.... 73
Table 2-3. Comparison of protein EpiScores versus measured proteins in associations with time-to-mortality in the GrimAge...................... 76
Table 2-4. Summary of studies generating EpiScores for blood protein levels identified by the literature search. .......................................................... 80
Table 9-1. EWAS of GDF15 and NT-proBNP levels in Generation Scotland (N ≥ 16,963). ....................................................................................... 185
Table 10-1. The 24 incident outcomes profiled over a maximum of 15 years of follow-up in the UK Biobank (N=47,600). ......................................................... 208
Introduction to Thesis

As we age, our risk of developing neurological diseases increases. Detecting individuals that are at high risk of developing a disease can provide opportunities for intervention and prevention. Identifying markers that associate with better or poorer brain ageing may also provide insight on how we can keep our brain as health as possible, for as long as possible. Blood-based ‘omics’ data refers to measures of a range of molecular phenotypes from blood samples. These measures can include genetic, epigenetic, proteomic and metabolomic markers. Each of these molecular layers may contribute information that reflects the health state of an individual. This information may be used to identify 1) potential mechanistic pathways leading to disease, or 2) those at the greatest risk of developing a disease. As proteins are often causal mediators of disease and are typically the targets of therapeutic interventions, they are central in achieving these aims. Blood proteins can be markers of peripheral processes such as inflammation, or can enter the bloodstream through breakdown of blood-brain barrier integrity. Lifestyle and environmental exposures can elicit changes in DNA methylation (DNAm) – an epigenetic modification that is known to associate with circulating protein levels. Unlike genetic information that is typically set from birth, DNAm and protein levels are dynamic and can be affected by a range of biological and environmental factors. Understanding the relationships between these markers and the relevance that they have to brain ageing and disease may contribute towards improving brain health.

In Chapter 1, I introduce brain health and disease, describing approaches that can be used to measure brain health. I also detail the motivation for studying blood-based omics markers in relation to brain health. In Chapter 2, I introduce the circulating proteome and statistical approaches that can be used to link the proteome with genome-wide DNAm (the methylome). I also review statistical approaches that have been used to derive DNAm-based epigenetic biomarkers – epigenetic scores (or EpiScores) – of complex traits. In Chapter 3, I detail how Cox proportional hazards associations with incident disease outcomes can be used to identify early markers of disease risk. These markers can include either directly-measured proteins, or proxies for proteins (i.e. protein EpiScores). I also detail additional methodological considerations that are relevant to the approaches used in this thesis. In Chapter 4, I
summarise the five cohort studies used in this thesis: Generation Scotland (GS), the Lothian Birth Cohorts of 1921 and 1936 (LBC1921 and LBC1936), the Cooperative Health Research in the Region of Augsburg (KORA) cohort and the UK Biobank (UKB). In Chapter 5, I briefly outline the core aims and chosen methodologies that are specific to each empirical chapter of this thesis.

The first aim of this thesis is to conduct two studies that map the DNAm signatures of blood protein levels. Through this work, I test DNAm signatures for over 4000 proteins and uncover DNAm patterns that associate with protein markers of brain health traits. In Chapters 6-7, I conduct association studies to quantify the molecular signatures of circulating protein levels that are associated with brain health. The first study focuses on the epi/genetic architectures of S100β (a marker of neuroinflammation), whereas the second study maps the epigenetic signatures of 4,235 proteins and identifies those that are associated with brain health.

The findings from Chapters 6-7 may improve our understanding of protein regulation and the molecular signatures associated with brain health. Nevertheless, these studies do not quantify the extent that DNAm can proxy for protein signatures in the blood or be useful in stratifying individuals at high risk for neurological disease. Previous studies have developed a small number of EpiScores for blood protein levels. These studies have shown that for certain proteins, patterns of DNAm can proxy for the protein signal and these protein EpiScores may be useful indictors of health states. In the case of acute-phase inflammatory proteins whereby protein levels can fluctuate across multiple time point measurements, protein EpiScores may offer more stable measures than the proteins themselves. In instances whereby population studies do not have protein measures but do have Illumina-based DNAm measures, EpiScores for proteins can be projected into samples and potentially act as informative markers.

With the above rationale in mind, the second aim of this thesis is to perform a comprehensive assessment of the patterns of DNAm that can proxy for plasma protein levels. In Chapters 8-9, I create protein EpiScores for blood protein levels and assess whether the scores can serve as tools for disease risk stratification. I quantify how useful protein EpiScores may be as markers for neurological disease and brain health outcomes. I also study relationships between these EpiScores and diseases such as type 2 diabetes and heart disease, as these are risk factors for poorer neurological outcomes during ageing.
Finally, protein levels themselves are useful features that can augment risk stratification and serve as individual biomarkers for neurological diseases. The third aim of this thesis is therefore to assess blood protein signatures of incident neurological diseases and associated morbidities. In Chapter 10, I quantify individual relationships between 1,468 blood proteins and 23 incident diseases and mortality. I then develop protein-based scores (ProteinScores) for the incident outcomes, to understand the extent that information from multiple proteins present in the blood can be used to identify high risk individuals.

The protein EpiScore (Chapters 8-9) and ProteinScore (Chapter 10) studies I have conducted share a common objective, which is to develop blood-based markers that can identify high risk individuals prior to overt symptom presentation. These studies integrate electronic health data linkage for a range of diseases affecting the brain (Alzheimer’s dementia, amyotrophic lateral sclerosis (ALS), Parkinson’s disease, multiple sclerosis, ischaemic stroke), but also diseases that manifest peripherally that may elevate risk of adverse brain health during ageing (such as type 2 diabetes, heart disease, liver disease and inflammatory bowel disease).

In Chapter 11, I discuss the findings of this thesis with considerations of the limitations of the assays, methodologies and cohort populations used. I also recommend analyses that I believe would be relevant to extend this work in future.
1 Brain health

Brain health fluctuates across the life course. Declines in brain health occur during ageing and can be exacerbated by a range of genetic and environmental factors. In this chapter, I introduce central themes related to brain ageing and the burden of neurological disease. As this thesis is centred on the ageing brain, I focus on the onset of neurodegenerative diseases and stroke that typically occur from mid to later life. I detail the morphological and cognitive measures that are commonly used as indicators of brain health. Finally, I provide an overview of molecular measures that may be used as markers for brain health. In doing so, I describe the motivation for the use of blood-based information to study brain health as we age.

1.1 Central themes in brain health

Here, the onset of neurological disease from mid to later life and the burden that this places on society and healthcare systems is summarised. The complexity of studying brain health in the context of wider risk factors also presented, in addition to a brief rationale on the need for endophenotype measures of neurological diseases.

1.1.1 The ageing brain

The largest aggregated neuroimaging dataset (spanning a 100-year range, with 123,984 MRI scans across over 100 studies from 101,457 individuals) identified critical periods in the maturation of the human brain across the life course. These dynamic changes in brain morphology highlight that brain health is in flux, with declines in brain volume measures occurring from age 40 onwards (Figure 1-1). In conjunction with these observable declines in brain volume metrics, the risk of neurodegenerative diseases increases. Neurodegenerative diseases are hallmarked by selective neuronal loss, which results in a wide range of cognitive and motor symptoms. Alzheimer’s dementia is the most common neurodegenerative disease, followed by Parkinson’s disease. Other neurodegenerative diseases include ALS and Huntington’s disease. While multiple sclerosis is primarily mediated by autoimmune attacks on the nervous system, it is hallmarked by progressive neural loss. Dementia is characterised by impaired memory, language and cognitive functions and the most common form is Alzheimer’s dementia (60-70% of cases), followed by vascular, Lewy...
body and frontotemporal dementias. Parkinson’s disease involves progressive disruptions in movement including tremors, slowed movements, rigid muscles, impaired balance and a range of other motor symptoms. Ageing is the primary risk factor for most neurodegenerative diseases, but the age of onset can vary depending on disease subtype. For example, early onset Alzheimer’s dementia is more likely to include cases of autosomal dominant familial origin, whereas late onset dementia is predominantly sporadic, occurring after 65 years of age with no clear familial links. Parkinson’s disease risk increases with age and most frequently occurs from age 60 onwards. While stroke is characterised by acute vascular injury rather than neurodegeneration, it is highly associated with age. In the United States of America (USA), three-quarters of all strokes occur in individuals over 65 years old, with the risk of incident stroke doubling every decade after turning 55 years of age. Silent cerebrovascular disease is detectable through neuroimaging measures (brain infarcts, microbleeds and white matter lesions) that indicate vascular damage in the brain. These characteristics are associated with an elevated risk of future stroke events. Ischaemic stroke is the most common form of stroke (87% instances) that involves an occlusion to blood supply in a section of the brain, whereas haemorrhagic stroke (13% instances) involves bleeding within or around the brain.
Figure 1-1. Dynamic changes in brain morphology and the onset of neurological diseases across the life course.

Top panel shows trajectories of the median for MRI phenotypes as a function of log-scaled age. Triangles indicate the peak volume and circles indicate the peak growth rate for each measure. Bottom panel shows non-MRI defined milestones across the life course. Typical age-ranges of incident neurological diagnoses are shown as black boxes, with the age at diagnoses within the MRI dataset shaded in blue. Brown lines show developmental milestones from the literature (including the Tanner scale of physical development) and grey bars show growth charts for physical and ultrasonic variables. AD: Alzheimer’s dementia. ADHD: attention deficit hyperactivity disorder. ASD: autism spectrum disorder. ANX: anxiety or phobic disorders. BD: bipolar disorder. MDD: major depressive disorder. RMR: resting metabolic rate. SCZ: schizophrenia. Sourced from Bethlehem et al. (2022)¹. Copyright: Nature.
1.1.2 The burden of neurological diseases

The health of the ageing brain is a major concern for healthcare systems globally, given that life expectancy is projected to continue to rise in most developed countries through to 2030. A recent review states that although the global average lifespan has increased from 47 to 73 years in the seven decades since 1950, health-adjusted lifespan (which is the total number of years lived free from disease) has not followed suit. While the authors estimate an average nine year difference between lifespan and years lived disease free globally, this statistic varies considerably between countries. When ranking the World Health Organisation life expectancy data from 2019 by country, the difference between life expectancy at birth versus healthy life expectancy at birth ranged from 6.5 years (Lesotho, South Africa: life expectancy 50.7 years and healthy life expectancy 44.2 years) to 12.4 years (USA: life expectancy 78.5 years, health life expectancy 66.1 years). Across the 25 countries that were part of the European Union (EU) in 2005, it was estimated that between 16-20% of life was spent in late-life morbidity. Neurological diseases are a driver of this burden and are projected to increase in prevalence in the future. Having a neurological disorder is the leading cause of disability and second leading cause of mortality globally.

Stroke (67.3%), migraine (13.1%) and dementias (9.5%) were the leading causes of disability-adjusted life years due to neurological disease in 2015 across Europe. Between 2015 and 2035, the number of strokes in the United Kingdom (UK) per year is projected to increase by 60%. The number of individuals living with dementia is estimated to increase from 57 million as of 2019 to 153 million by 2050. Although the burden of neurological disease varies by population composition and country, there are consistent trends globally. An assessment on the global burden of disease dataset covering 369 diseases across 204 countries from 1990 to 2019 identified stroke as the leading cause of disability-adjusted life years (DALYs) across 19 of the 21 global regions studied (Figure 1-2a). Deuschl et al. (2020) observed that stroke and Alzheimer’s dementia are extremely pronounced as the two primary drivers of DALYs due to neurological diseases in mid to later life across Europe (Figure 1-2b). This is likely due to age as a driving risk factor for these diseases and Europe having some of the longest life expectancy estimates in the world.
Figure 1-2. Disability-adjusted life-years resulting from neurological diseases.

a, Ranking of age-standardised DALY rates of neurological diseases across 21 global regions in 2019. Lower rank numbering (red) indicates larger DALY rates, whereas diseases with the lowest DALY rate contributions (blue) have higher rank numbering. Adapted from Ding et al. (2022) 28. Copyright: Frontiers Public Health. b, Age-standardised DALY rates per age band in 2017 for the WHO European region. Adapted from Deuschl et al. (2020) 29. Copyright: Lancet Public Health. DALYs: disability-adjusted life-years.
1.1.3 Human complexity

Neurological disease risk involves a complex interplay between the genetic profile an individual is born with and their exposure to environmental and lifestyle factors that accumulate through the life course. Modifiable lifestyle factors are estimated to account for around 40% of dementia cases. Lifestyle factors that associate with poorer brain health outcomes include smoking, excessive alcohol consumption, physical inactivity, poorer sleep quality and being overweight. While these variables are normally recorded in cohort studies, they represent only a small number of the possible lifestyle and environmental influences that may account for disease risk. Environmental toxins have been linked to multiple neurodegenerative diseases.

For example, the neurotoxin β-methylamino-l-alanine (BMAA) is produced by cyanobacteria that can be present in food products and has been implicated in pathways associated with Parkinson’s disease. Vitamin D deficiency has been discussed as a potential risk factor in Alzheimer’s dementia, Parkinson’s disease and multiple sclerosis. Elevated exposure to fine particulate matter (diameter ≤2.5 μm) air pollution has been associated with increased rates of stroke and dementia. A growing body of evidence suggests that differences in dementia prevalence and incidence rates arise based on geographical regions. Clearly, the non-genetic component of neurological disease risk can be comprised of a range of known and unknown contributing factors.

Another aspect that requires consideration is the impact of pre-existing morbidities on the development of subsequent neurological diseases. Chronic, age-related diseases are increasing in prevalence. Diabetes, CVD, cancers and chronic respiratory diseases are responsible for a large proportion (around 80%) of deaths associated with chronic disease. Heart disease and type 2 diabetes have been associated with poorer cognitive function and accelerated brain ageing. The risk of stroke is intertwined with systemic factors like hypertension, heart disease, smoking and type 2 diabetes. Depression in earlier life is associated with dementia in later life. Methods that can generate an individual’s risk profile for a wide panel of such diseases simultaneously may therefore help to build a more complete view of their neurological risk profile. Prevention of these age-related morbidities may have an impact on the likelihood of developing neurological diseases in later life.
1.1.4 Characterising neurodegenerative diseases

Neurodegenerative diseases are characterised by a period of time whereby damage builds in the brain, before the symptomatic threshold is reached that prompts an individual to visit a clinician for diagnosis. Detecting individuals that are most likely to be experiencing the initial stages of disease pathogenesis in their brains would aid prevention strategies. While the diseases described in Section 1.1.1 can be grouped together and diagnosed based on symptomatic presentations, additional layers of information from neuroimaging, cognitive testing and molecular profiling can be used to augment and refine disease definitions. For example, Alzheimer’s dementia initially involves temporoparietal atrophy in the brain and is associated with the presence of amyloid-β (Aβ) plaques extracellularly and the formation of neurofibrillary tangles from abnormally hyperphosphorylated tau intracellularly. Parkinson’s disease involves a loss of neurons from the substantia nigra and the accumulation of alpha-Synuclein (α-Syn) aggregates known as Lewy bodies in neuronal cell bodies. While the described protein pathologies are localised in neuronal tissues, protein markers can leak into the cerebrospinal fluid (CSF) or blood. Neuroimaging and cognitive testing are also used to gain insight into brain health. These types of measures are sometimes referred to as ‘endophenotypes’.

Endophenotypes represent measureable biological traits that are thought to exist along the disease pathway, between molecular regulators and risk factors acting to produce the disease and the disease itself. A well-known example is the composite A/T/N criteria used to track cognitive impairment and dementia. This score is reliant on a combination of endophenotypes (A: amyloid-β from CSF or positron emission tomography (PET) scans, T: phosphorylated tau from CSF or PET scans, N: biomarkers of neurodegeneration or neuronal injury including fluorodeoxyglucose–PET, structural MRI, or CSF total tau). The presence of these markers does not guarantee translation to symptomatic presentation in all cases, but they can still be useful indicators that help to profile early risk patterns and unravel heterogeneity within individuals that would otherwise have similar symptomatic presentations.
1.2 Measurement of brain health

In this section, I provide an overview of neuroimaging and cognitive tests as two useful indictors that can track brain health.

1.2.1 Neuroimaging

As seen in (Figure 1-1), population-level patterns in structural neuroimaging markers can provide useful information to track the ageing brain. It is commonly recognised that reductions in brain volume (grey matter) occur during ageing, accompanied by reduced connectivity between regions of the brain (white matter). Reduced surface area of the brain, thinning of the cortex, and increases in the size of the ventricles are all indicators of this general phenomenon of global atrophy that have been reported and can be analysed by structural MRI. Diffusion MRI measures two indicators of axonal connectivity between regions of the brain: mean diffusivity (magnitude of water molecule diffusion) and fractional anisotropy (directional coherence of water molecule diffusion). Generally, fractional anisotropy decreases with age, whereas mean diffusivity tends to increase. A relative brain age metric (that scores a brain age acceleration measure based on a range of volumetric imaging features) trained on 3,377 healthy individuals ranging from 18 to 92 years old has been shown to associate with incident dementia (hazard ratio per standard deviation [SD] of relative brain age = 1.03 [95% CI = 1.02,1.04], $p < 0.0001$). Individuals with poorer performance in this brain age score at 45 years have also been associated with accelerated biological ageing and cognitive decline over longitudinal follow-up in the Dunedin Study (N=869). Therefore, such imaging-based scores may serve as potential means to target recruitment of high-risk individuals for intervention trials and could also serve as endpoints for monitoring changes in brain health in such trials.

It is important to note that older individuals do not experience the same morphological hallmarks or rate of brain ageing. Figure 1-3 (a-b) highlights that healthy, neurodegenerative disease-free individuals from LBC1936 aged within a year of one another (~73 years old, born in 1936) exhibit variability in global atrophy and white matter hyperintensities in the brain. Parts c-d of this figure show that while declines in total brain volume and increases in white matter hyperintensity volume are generally observed across repeat time point measures, there is variability in within-individual
trajectories. These differences may be due to a range of genetic and non-genetic factors. Cognitive ability, protein levels and epigenetic scores derived from blood and genetic loci have all been associated with brain imaging outcomes. These studies provide insight into how processes like systemic inflammation or genetic regulation associate with differences brain structure across populations as potential biomarkers for brain ageing. Whether changes to the brain that occur in healthy individuals represent subclinical manifestations of disease prior to overt symptom presentation has not been fully answered.
Figure 1-3. Individual differences in ageing hallmarks of the brain.

a, Participants are ordered from top left to bottom right by severity of global atrophy (total brain volume / intracranial volume). b, Severity of white matter hyperintensities from top left to bottom right. Individuals shown in parts a-b are from the LBC1936 (wave 2) and were all born in 1936 and were aged ~73 years. c, Longitudinal total brain volume in 3,189 participants from the UK Biobank. d, Longitudinal white matter hyperintensity volume measures for 3,152 participants from the UK Biobank. In parts c-d, local polynomial regression lines (loess) are shown for females (red) and males (blue). Sourced from Cox and Deary (2022) 61. Copyright: Aging Brain.
1.2.2 Cognitive tests

Declines in cognitive ability can serve as a warning signal for poorer brain health. A direct experience of change in cognitive ability is the primary symptom individuals will experience that prompts a dementia diagnosis. Although Parkinson’s disease, ALS and multiple sclerosis are typically characterised by motor symptoms, declines in cognitive function have been reported as leading non-motor symptoms in these diseases. Individuals that have had a stroke experience faster cognitive decline than stroke-free individuals from 1-3 years after event occurrence. In individuals without neurological disease diagnoses, cognitive decline occurs during ageing and predicts poorer health, reduced adherence to medications, impaired financial decision-making and risk of disease and death. Inter-individual variance in cognitive ability has been attributed to genetic factors, early life cognitive ability, educational attainment (EA), lifestyle factors and the impact of having morbidities such as CVD. Crystallised intelligence (prior learning and past experiences) remains stable during adulthood but declines after the age of 60, whereas fluid intelligence (reaction speed, working memory, processing speed, non-verbal reasoning and mental manipulation) declines across adulthood (Figure 1-4).
Figure 1-4. Crystallised and fluid intelligence over the human life course.

Mean scores for crystallised (white circles) and also fluid (black triangles) intelligence with standard errors of the mean values (vertical error bars). Dotted lines indicate tests that were completed as part of the Salthouse study. Continuous lines show scores from the Wechsler Adult Intelligence Scale (WAIS) IV intelligence test. Figure taken from Salthouse (2012) 78. Copyright: Annual Reviews.

The Mini-Mental State Examination (MMSE) is commonly used to measure cognitive decline and is based on a 30-part questionnaire assessing memory, visual construction, language capability, time and space awareness and attention capabilities 79. When assessing the distribution of cognitive ability across a population, a general factor of cognitive ability (g) can be calculated as a comparator measure between individuals in the sample. In 1904, Charles Spearman identified within-individual correlations in performance on various cognitive assessments in school children 80. This idea of the ‘positive manifold’ of interrelatedness between cognitive domains provided the basis for the development of the g factor, which indicates the shared variance in intelligence across cognitive test performances. This scoring structure is flexible to the inclusion of various cognitive scoring domains. The g factor has been associated with a range of structural brain imaging measures in the UKB cohort (n = 18,426) 81. The g factor was found to account for an average of 58.4% (SE = 4.8%) of the genetic variance in the seven individual cognitive traits considered as components of g in a recent study by de la Fuente et al. (2021) 82. Decline in the composite score for g has been shown to associate with three components of brain volumetric changes across the eighth decade of life in the Lothian Birth Cohort 1936, with cortex-wide atrophy representing the leading component 63.

1.3 Molecular indicators of brain health

In this section, I detail molecular indicators that can be used to profile brain health. I start by providing an overview of neuroinflammation, which is an endophenotype shared across multiple age-related neurological diseases that is driven by protein mediators in the body and brain. I then outline CSF and blood protein markers, which
typically have been profiled as indicators of changing disease risk across a range of neurological diseases. I then provide an overview of genetic and epigenetic measures, which can be used to better understand brain health and resolve the regulatory profiles of protein markers of brain health. Finally, I introduce considerations regarding the instability of single time point protein measures.

1.3.1 Neuroinflammation

Inflammaging is a state of low-grade inflammation that is associated with ageing. High and prolonged levels of inflammation have been linked to the development and exacerbation of many age-related diseases, chronic morbidity, frailty and premature death. While neuroinflammation encompasses a multitude of inflammatory events in the central nervous system, it is primarily mediated by a range of proteins that are typically released by microglia and astrocytes (such as cytokines, chemokines and secondary messengers). The peripheral immune system can promote dysfunction at the blood brain barrier (BBB), which has been associated with neuroinflammation. Inflammatory-associated disruption at the BBB and destruction or impairment of synapses and neurons are hallmarks common to all of the neurodegenerative diseases introduced in Section 1.1.1. Secondary neuroinflammation is also a major determinant of recovery and further injury after ischaemic stroke and is intensively studied to identify targetable, time-dependent cascades. Given that clear links between peripheral inflammation and the brain exist, markers of inflammation may provide insight into both the general inflammatory state of an individual and the likely state of neuroinflammation in the brain. As proteins represent active molecular drivers of inflammation within the body, they are useful indicators of these processes. The blood and CSF are typically used to profile these signatures and there are additional proteins (with roles beyond inflammation) that have been identified as biomarkers of neurological diseases, which are introduced in the following two sections.

1.3.2 Proteomic profiling of CSF

CSF circulates around the spinal cord and brain, removing waste and providing maintenance of the environment of the nervous system. Altered molecular composition of the CSF may inform on any molecular abnormalities in the brain, given the proximal nature of the CSF to nervous tissues. A range of proteins have been detected in CSF...
that are leading markers of neurological diseases. Elevated levels of hyperphosphorylated tau (p-tau) or total tau (t-tau) are accompanied by reduced levels of Aβ derivatives in the early stages of cognitive impairment and Alzheimer’s dementia. In a recent meta-analysis, total α-syn, Aβ 42 and neurofilament light (NEFL) were identified as promising CSF biomarkers for Parkinson’s disease. A systematic review of 47 studies found a panel of 27 proteins that were associated with diagnosis of Alzheimer’s dementia, with neurosecretory protein VGF and Chitinase 3 Like 1 (CH3L1) emerging as consistent proteins across the studies. A correlate of cognitive decline is synapse loss. In contrast, synaptic resilience is associated with maintained cognitive ability during ageing. In a recent systematic review that I co-screened studies for, we assessed 67 studies reporting associations between CSF markers of synapse loss and neuropsychological performance on cognitive tests. We identified some evidence supporting NEFL and neurogranin (NG) as markers of worse cognition in Alzheimer’s disease, frontotemporal dementia and typical cognitive ageing. However, this study highlights the challenges in consistency that the biomarker field faces, as heterogeneity in statistical reporting, methodologies and standardisation meant that conclusions could not be definitively drawn. While the lumbar puncture used to collect CSF is generally well-tolerated with few adverse events when performed correctly, it is still considered to be an invasive procedure.

1.3.3 Proteomic profiling of Blood

Blood can be measured with minimal invasiveness as compared to CSF sampling. Blood-based measures are also 1) relatively safe and convenient to collect – especially in older individuals who may be frail or in poor health, 2) scalable for population studies, 3) amenable to quantification of multiple molecular data types and 4) likely to reflect the overall health profile of an individual. While CSF is more closely reflective of the brain, blood-based signatures of inflammation (measured via protein markers) have been tracked across neurodegenerative diseases. A range of studies have found associations between inflammatory proteins in the blood and trends in MRI measurements and cognitive ability that are indicative of the ageing, disease-free brain. Understanding the relationship between blood markers and measures of brain health (MRI/cognitive) as well as prospective disease outcomes (time-to-event modelling) is essential. In some instances, blood-based protein
biomarkers correlate with CSF equivalents and serve as tools for disease discrimination. An example is the recent profiling of blood-based p-tau181 and p-tau231 as markers of Alzheimer’s dementia\textsuperscript{109,110}. The p-tau181 marker differentiated Alzheimer’s dementia from other neurodegenerative diseases, with Area Under the Curve (AUC) = 0.94–0.98. A one SD increase in the log p-tau181 measures were also associated with an elevated risk of future Alzheimer’s dementia in 119 individuals with mild cognitive impairments over an 8-year follow-up period, Hazard Ratio (HR) = 3.1 and P < 0.001\textsuperscript{110}. In instances such as these, blood protein measures can clearly provide value as potentially biomarkers for neurological disease. In addition to proteomic data, other ‘omics’ measures can be quantified from blood. In Sections 1.3.4 and 1.3.5, I outline genetic and epigenetic measures that can be derived from the blood, with considerations on the use of these markers for studying brain health and the molecular regulation of protein markers of brain health.

1.3.4 Genetic measures

Deoxyribonucleic acid (DNA) is comprised of adenine, cytosine, guanine and thymine nucleotide pairs. Variability in the genetic code provided by DNA can be introduced through gene insertions, deletions, substitutions or translocations. However, individual nucleotide base changes known as single nucleotide polymorphisms (SNPs) are the most common drivers of variability. Human genome reference panels (e.g. the 1000 Genomes Project) have documented maps capturing > 99% of SNPs with allele frequencies > 1\%\textsuperscript{111}. The variability in SNPs across individuals is thought to be \~1\% of the total genome, with more than 80 million SNPs profiled\textsuperscript{111}. Only \~1.5\% of the genome codes for proteins, but other sources of variation from non-coding regions are thought to potentially impact protein function and abundance\textsuperscript{111,112}. Genetic information is largely consistent across cells in the body, with minor exceptions in reproductive system cells and somatic mutations\textsuperscript{113}. Although \~20,000 proteins are encoded by genes present in the genetic code, up to 100 different proteoforms can be produced from a single gene through processes such as alternative splicing and post-translational modifications\textsuperscript{114}.

Genome-wide association studies (GWAS) test for statistically significant associations between hundreds of thousands of SNPs measured on chip-based microarrays and traits\textsuperscript{115}. The largest GWAS of general cognitive ability in 300,486 individuals found a
25% heritability (SE of 0.06%) that was attributable to genetic variance. This study identified 148 genome-wide significant (P < 5 × 10⁻⁸) independent SNPs that were associated with general cognitive function. The largest GWAS of late onset Alzheimer’s dementia (age at onset 65 or above) included 1,126,563 individuals, with 90,338 cases and 46,613 proxy cases (assessed via familial risk of the disease). This GWAS identified 38 genome-wide sentinel SNPs, with seven that were novel. Pathway enrichment suggested potential functional roles for the variants in microglial, immune cell and protein catabolism functions. A recent GWAS of stroke including 110,182 cases and 1,503,898 controls identified 89 independent genome-wide significant SNPs, with 61 that were novel.

1.3.5 DNA methylation measures

Environmental exposures, lifestyle behaviours, stochastic biological effects and genetic influences can result in epigenetic modifications across the genome. These can include histone modifications, chromatin remodelling, changes to non-coding RNAs, and DNAm at cytosine-guanine dinucleotides (CpGs). Epigenetic modifications are thought to change the accessibility of DNA, the conformation of chromatin structures and impact gene transcription. DNAm at CpGs is the most widely measured epigenetic modification in large-scale cohort populations and is therefore the focus of this thesis.

The most common approach to tracking DNAm across sites in the genome is through sodium bisulfite conversion. This approach distinguishes between methylated and unmethylated cytosines through quantification of a fluorescence signal (green = methylated, whereas red = unmethylated) based on binding of probes (Figure 1-5). This can be used to measure methylation at a single CpG site across multiple cells from bulk tissue (i.e. whole blood, or saliva) and sorted cell types. Illumina is the leading array-based technology for DNAm quantification (that employ the bisulfite approach), with the most recent Infinium MethylationEPIC (EPIC) BeadChip array able to test for DNAm at 863,904 CpG sites. In many cohorts, the older Infinium HumanMethylation450 (450K) BeadChip array has been used to measure DNAm, yielding measurements for up to 485,577 possible CpG sites. These platforms capture only around 1-2% of the CpGs in the genome, of which there are estimated to be approximately 28 million. DNA methylation probes are selected in locations enriched for
coding genes, promoter regions or enhancer elements. Two primary measures are used to record DNA methylation (DNAm) through Infinium probes. The Beta value is a ratio of methylated probe intensity against the sum of the methylated and unmethylated probe intensities, which ranges from 0 to 100%. A Beta value of 100% indicates that there is complete methylation of all the copies of the CpG site in the sample. M values represent the log₂ ratio of methylated versus unmethylated probe intensities. A positive M value implies that the CpG is more methylated than unmethylated across the copies of the CpG in the sample, whereas a negative value indicates the reverse. Beta values are more intuitive in terms of biological interpretation, with hypo- and hypermethylated CpGs (with Beta values ≤ 10% versus ≥ 90%, respectively).

Figure 1-5. Infinium DNA methylation detection example.

A) Infinium I probes use two beads: unmethylated (that recognise the unmethylated CpGs) and methylated (recognising methylated CpGs). Single base extension is used to extend from the bound bead by a single base, resulting in fluorescent green if cytosine or guanine is present, or red if thymine or adenine is present. The bisulfite
step selectively retains methylated cytosines, while converting unmethylated cytosines through deamination to thymines. This is why cytosines are only present in the methylated locus shown to the right of the panel. B) Infinium type II probes measure signals using the same bead type, which is designed such that it can match the bisulfite-converted DNA of both methylated and unmethylated loci. This bead type detects methylated cytosine, or bisulfite-converted thymine and single base extension can be run to incorporate a labelled nucleotide into the sequence of a given probe. This enables signal detection, with green fluorescence resulting from methylated sites and red for unmethylated sites. DNA, deoxyribonucleic acid. Figure adapted from Pidsley et al. (2016) \(^{120}\). Copyright: BioMed central.

The presence of methylation in promoter regions is most commonly thought to reduce gene expression through blocking the binding of transcription factors; however, some studies suggest that transcription factors can regulate DNAm \(^{123}\). The downstream effects of methylation have not been fully resolved and the functional role of DNAm is likely to be gene, site and disease specific. For example, smoking-induced hypomethylation across CpG sites in blood has been found to associate with subsequent risk of lung cancer \(^{124}\), whereas hypermethylation of gene body CpG islands has been associated with high dosage of oncogenes in liver cancer \(^{125}\). DNAm may also have no downstream impact on gene expression, acting predominantly as a record of the body’s response to exposures. This argument is compelling given that methylation is highly correlated with a range of environmental and lifestyle factors, such as smoking, which leaves a distinct signature on an individual’s blood-based methylome \(^{126}\). Recently, a study of methylation signatures for the inflammatory protein C-Reactive protein (CRP) provided evidence through Mendelian randomisation that DNAm was likely to be a consequence (rather than cause) of differential CRP levels in blood, with mediation from factors such as obesity and smoking \(^{127}\). As with GWAS, the DNAm signatures at individual CpGs (rather than SNPs) that associate with traits can be quantified. These studies are referred to as methylome-wide or epigenome-wide association studies (MWAS or EWAS) and are described further in Section 2.2.

1.3.6 Variability in DNAm measures

There are several factors that can be sources of variability in measures of DNAm (Figure 1-6). First, age is strongly correlated with changes in DNAm \(^{128,129}\). In a study
spanning eight UK-based cohort studies, DNAm was profiled in 7,037 individuals with an age range from birth to 98 years old. Age was associated with DNAm variability at 29,212 CpG sites, with a 1.3-fold increase in DNAm variability per year of the life course. Second, exposures are associated with changes to DNAm. Exposures include a spectrum of influences spanning environmental (e.g., pollution or social deprivation), lifestyle (e.g., smoking or physical activity) and biological factors (e.g., inflammation or disease pathology in the body). Third, DNAm can be influenced by genetic factors. Genetic variants that influence methylation directly are known as methylation quantitative trait loci (mQTLs). Min et al. (2021) scanned DNAm measures at 420,509 CpGs to create a database of over 270,000 mQTL associations that, on average, explained between 15-17% of the variance in DNAm. Fourth, unlike genetic information collected across blood, CSF and brain (that would be consistent), DNAm is known to be variable based on the tissue type sampled. Previous studies suggest that blood is unlikely to reflect the brain reliably for all CpGs, but that some sites more closely reflect brain DNAm than others. I have also previously shown poor translation in performance of DNAm-based scores for lifestyle factors from blood to brain, with variability across brain regions. Measurements at cg05575921 in the prefrontal cortex exhibited the largest correlation in relation to blood ($r = 0.61, n = 14, P = 0.02$) and smoking behaviour ($r = -0.65, n = 9, P = 0.06$). Given hypomethylation at cg05575921 represents the strongest hallmark of smoking detectable in blood, my findings support the notion that variability in DNAm across blood and brain is CpG-specific. Fifth, differences in DNAm exist across cell-types, which means that the proportion of cells within the bulk sample can alter measurements of DNAm. Finally, technical variability can be introduced through processing of samples across multiple assay plates, processing batches and measurement time points. The discussed sources of variability are important considerations that influence both study design and interpretation of statistical approaches that include DNAm. These factors are discussed in the context of DNAm-focused analyses methods in Sections 1.3.8, 2.2 and 2.3.

1.3.7 DNAm and brain health

DNAm changes could accumulate across the lifespan and impact pathways leading to neurological disease. In a study that examined prefrontal cortex tissue from 708 prospectively collected autopsied brains, DNAm at 71 of the 415,848 CpGs tested were associated ($P < 1.2\times10^{-7}$) with the burden of Alzheimer’s disease pathology in
the brain. The results from this MWAS implicated a CpG within the bridging integrator 1 (BIN1) gene locus, which has been implicated as a major risk susceptibility gene for Alzheimer's dementia. While brain DNA methylation (DNAm) is tissue-specific and informative, it is challenging to collect at scale and cannot be extracted in individuals that are alive (with the exception of biopsy samples extracted during surgeries in very few cases). This means that blood-based DNA methylation is a more widely profiled measure across cohort studies. Given that DNA methylation serves as a readout of the body’s response to exposures that are major risk factors for neurological diseases, the DNA methylation information from blood may still be of value for the stratification of disease risk and brain health.

Figure 1-6. Sources of variability in DNA methylation.

Sources of variability that impact the design and interpretation of DNA methylation-based analyses in large-scale, population-based studies with DNA methylation measured through array technologies such as those provided by Illumina are described. I created this figure as part of the review from Nabais et al, (2023), which I co-authored. The figure has been adapted for this thesis. Figure created using Biorender.com.
1.3.8 Molecular architecture of proteins

MWAS can identify differential methylation at CpG sites that associate with the levels of proteins. Protein MWAS are relatively new compared to protein GWAS, which identify protein quantitative trait loci (pQTLs) associations between SNPs and protein levels as the outcome. A seminal study by Sun et al. (2018) used linear regression to model 2,994 plasma protein levels measured using the SomaScan platform as outcomes, identifying 1,927 pQTLs for 1,478 proteins (Bonferroni-adjusted threshold P < 1.5×10^{-11}). More recent advances in pQTL mapping have profiled greater proteomic depth of coverage, extended comparisons to multiple protein assay platforms further and run analyses in increasingly large samples including 35,559 and 54,306 individuals. In MWAS, DNA-m-protein associations are known as protein quantitative trait methylation loci (pQTMs). These studies highlight that mapping the genetic and epigenetic architectures of protein levels can provide insight into their molecular regulation profiles and point towards potential pathways of interest. In protein GWAS, cis-acting pQTLs involve variants that are proximal (typically between 1Mb to 10Mb distance) to the transcriptional start site of the protein-coding gene on the same chromosome. Trans-acting pQTLs involve SNPs that are not proximal to the protein-coding gene and can have long-ranging effects from different chromosomes. Cis-acting SNPs are thought to have a higher chance of directly influencing the protein-coding gene, whereas trans-acting pQTLs may be indicative of more complex biological cascades. In protein MWAS, these principles can be applied to delineate cis- or trans-pQTMs. While a CpG involved in a cis-pQTM may affect the protein-coding gene expression, the directionality (i.e. between CpG and protein) cannot be inferred with the same certainty as a cis-pQTL (i.e. from SNP to protein).

1.3.9 Instability of protein measures

While sources of technical and biological variability can be introduced in genotype measurement, these effects are typically more pronounced in non-genetic omics
measures such as DNAm. Quantifying the reliability of markers over multiple time point measures is important to understand how useful markers could be for risk stratification. Test-retest reliability correlates measures of the same marker taken over multiple time points. The test-retest reliability has not been comprehensively profiled across the proteome. In most cases, test-retest reliability is calculated for specific immunoassay-derived protein biomarkers of interest from the blood. I am not aware of a test-retest assessment across the proteome using multiplexed protein measures of many hundreds or thousands of proteins simultaneously. However, current studies indicate that test-retest reliability varies on a protein-by-protein basis.

A recent systematic review and meta-analysis from Walsh et al. (2023) collated test-retest reliabilities for immunoassay-derived, blood-based inflammatory protein markers (50 studies assessing proteins on > 1 occasion, N = 48,674; 57% male; mean age 54 (range 13–79) years, time point range from 24 hours to 7+ years). The authors found that for the markers that had coverage across most time points (CRP, IL6, TNF-α and fibrinogen), stability was highest in the near-term of < 6 months (r’s = 0.80 to 0.61), poorer over 6 months to 3 years (r’s = 0.60-0.51), and even lower for > 3 years (r’s = 0.39-0.30) (Figure 1-7). The authors concluded that single time point measures of these key inflammatory markers provide reasonable insights in the short term (< 6 months), but are unreliable indicators of health in the longer term. The authors note that these findings are consistent with previous literature on stability of measures such as systolic blood pressure (range in r from 0.35 to 0.61 over 2 to 12 years) and blood lipids (range in r from 0.37 to 0.65 from 4 to 12 years).
<table>
<thead>
<tr>
<th>Biomarker</th>
<th>Interval</th>
<th>1 day - 3 months</th>
<th>6 months - 1 year</th>
<th>3 years</th>
<th>5 years</th>
<th>&gt;7 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRP</td>
<td>r = 0.789</td>
<td>r = 0.505</td>
<td>r = 0.530</td>
<td>r = 0.355</td>
<td>r = 0.347</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.646, 0.878]</td>
<td>[0.339, 0.641]</td>
<td>[0.399, 0.639]</td>
<td>[0.285, 0.420]</td>
<td>[0.234, 0.450]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>k = 8</td>
<td>k = 9</td>
<td>k = 14</td>
<td>k = 17</td>
<td>N = 11</td>
<td></td>
</tr>
<tr>
<td>IL-6</td>
<td>r = 0.607</td>
<td>r = 0.599</td>
<td>r = 0.574</td>
<td>r = 0.385</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.405, 0.753]</td>
<td>[0.419, 0.734]</td>
<td>[0.458, 0.670]</td>
<td>[0.284, 0.479]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>k = 8</td>
<td>k = 9</td>
<td>k = 9</td>
<td>k = 11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TNF-α</td>
<td>r = 0.802</td>
<td>r = 0.600</td>
<td>0.816*</td>
<td>r = 0.292*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.341, 0.952]</td>
<td>[0.288, 0.797]</td>
<td>[0.033, 0.978]</td>
<td>[0.106, 0.458]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>k = 5</td>
<td>k = 5</td>
<td>k = 4</td>
<td>k = 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fibrinogen</td>
<td>r = 0.883*</td>
<td>r = 0.614*</td>
<td>r = 0.586*</td>
<td>r = 0.533</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.313, 0.985]</td>
<td>[0.512, 0.699]</td>
<td>[0.472, 0.680]</td>
<td>[0.496, 0.568]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>k = 2</td>
<td>k = 4</td>
<td>k = 2</td>
<td>k = 6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 1-7. Test-retest reliability for four inflammatory blood markers.**

Meta-analysis of 50 studies assessing correlation between protein markers measures across one or more time point. Correlations are shown per time interval with 95% CIs shown inside square brackets. Number of independent study samples is given by k. CRP = C-reactive protein; IL = interleukin; TNF = tumor necrosis factor. Taken from Walsh et al. (2023) \(^{148}\). Copyright: Alzheimer’s Dementia.

A study assessed test-retest reliability of Alzheimer’s dementia candidate plasma biomarkers in 38 individuals from the Swedish BioFINDER-1 study \(^{149}\). A time lag of 6-10 weeks post-sampling was assessed (mean 7.4 ± SD 1.05 weeks). Test-retest variability was 4.1% for Aβ42/Aβ40, 20.0% for p-tau217, 23.7% for NEFL, and 25.0% for GFAP \(^{149}\) (Figure 1-8). In simulations, the authors reported that test-retest variability effects on stratification of Alzheimer’s dementia onset over 4 years could be largely neutralised by using all four plasma biomarkers simultaneously.
Figure 1-8. Test-retest variability of plasma protein markers of Alzheimer's dementia.

Individual-level test-retest variability for each biomarker with the mean and 95% CIs plotted. Relative percent change of biomarker values for each participant were calculated across the two samples (6-10 weeks from baseline sample). The SD of the relative percent change distribution was then used as the overall estimate of random error. Cullen et al. (2022) \(^{149}\). Copyright: Alzheimer's & Dementia.

As described in Section 1.3.6, DNAm assays can suffer from technical and biological variability across batches of samples analysed at multiple time points. A recent study of 71 individuals measured over two time points separated by a 6 week interval found that reliability ranged by CpG site \(^{139}\). The authors found that removal of unreliable probes during pre-processing could correct for the effect of unreliable probes. A further study assessing 41 DNAm-based predictor scores across 101 different normalisation strategies found that with adequate quality control and normalisation of DNAm, predictor scores were highly accurate and reliable across multiple time points \(^{150}\). Both DNAm and protein measures exhibit dynamic changes in response to exposures, but DNAm is likely to represent a more chronic and stable reflection of the body’s health state. The proportion of individuals classified as smokers based on their DNAm has
been shown to fall every year post smoking cessation (OR for smoker classification = 0.86 per year since cessation; 95% CI = 0.83–0.89; P < 2.0 × 10−16) 151. The authors found that a minimum of two years was required before individuals that had been high-dose smokers were assigned to the non-smoking cluster. This study highlights that while DNAm can be dynamic, it may be less reactive in the immediate term (i.e. a shorter half-life) than protein levels. This work is supported by two further studies that found persistent (over 2 years), stable DNAm differences in a subset of loci due to cigarette smoking 152. A study also found that an alcohol EpiScore was likely to be reflective of long-term exposures to alcohol, which may accumulate over the lifecourse 153. While some proteins are likely to be more stable than others, certain exposures can trigger rapid changes in protein levels. Acute-phase inflammatory proteins such as CRP are obvious examples that are highly variable. The healthy range for CRP is 3-10mg/L, which can rise to 40-200mg/L upon a bacterial infection 154 and 350-400mg/L within a 48 hour period of disease or bodily trauma 155–158. Once the inflammatory cascade resolves, CRP levels decrease exponentially, with a half-life of ~19 hours in plasma 158. An interesting hypothesis is that when tracking an index of an individual’s health state, assessing multiple sites of DNAm may be more reliable than the measurement of a single protein at a single time point. Tracking patterns of DNAm that associate with proteins could provide insight into key processes like inflammation, but also provide a means to proxy for the protein signal itself through DNAm. This is discussed in the rationale for generating protein EpiScores in Section 2.3.5.

1.4 Motivation for studying omics markers of brain health

In Section 1.1, I provided an overview of the central themes in brain health that included the ageing brain and the burden of neurological diseases in mid to later life. I detailed non-genetic factors associated with brain ageing and noted relationships between other morbidities and the onset of neurological diseases. In Section 1.2 I detailed neuroimaging and cognitive tests as measures of brain health. In Section 1.3, I gave an overview of molecular indicators of brain health, detailing the arguments for using blood measures as biomarkers. I introduced genetic and epigenetic measures that can be derived from blood, with brief examples of their relevance to brain health.
Finally, I highlighted that test-retest reliability pertaining to protein measures can be highly variable.

The measures discussed in this chapter can be integrated to study brain health and neurological disease (Figure 1-9). Identifying protein biomarkers from the blood that associate with cognitive, neuroimaging and disease traits may be helpful for risk stratification. Understanding the impact of non-genetic risk factors across the life course is desirable, but the depth of data required to study this across multiple phenotypes and time points is often unavailable. Given that DNAm is thought to track a more stable signature of exposure to lifestyle and environmental factors – that represent major risk factors for neurological diseases – it is likely to be a useful measure when studying neurological disease. DNAm may also represent a means to track protein-informed signatures from the blood, with less variability than single time point protein measures.

Recently, studies have emerged that conduct GWAS and MWAS with blood protein measures as outcomes. Conducting these analyses with protein markers of brain health as the outcomes may provide additional information regarding the molecular regulators of these proteins. These approaches map the epigenetic and genetic architectures of protein levels and are summarised in Section 2.2 in greater depth.

Over the past four years, there has been an emergence of DNAm-based predictors of blood protein levels, known as protein epigenetic scores (EpiScores). These approaches are documented in Section 2.3 in greater depth.
Figure 1-9. Motivation for integrating omics markers of brain health.

Identification of individuals with subclinical disease pathology or potential neuroinflammation would help to better-tailor prevention. Typical brain ageing and neurological disease onset are regulated by a combination of genetic and environmental/lifestyle factors that lead to heterogeneity across populations. From one blood sample, genetic, epigenetic (DNAm) and proteomic samples can be recorded. Key genetic loci and genes can be identified. DNAm differences have been associated with a range of non-genetic exposures (e.g. smoking, environment and disease states). DNAm is thought to act as a record of exposure, but in some instances can affect gene expression or associate with protein level differences in blood. Profiling the epigenetic and genetic patterns associated with protein markers of brain health may inform on the regulation of these proteins. Resolving molecular signatures (such as proteins) from the blood that associate with hallmarks of brain ageing (measured via cognitive testing and brain imaging) strengthens our understanding of the possible targets for brain health improvement. Figure created using Biorender.com.
1.5 Summary

Identifying the blood-based molecular correlates of brain health and neurological disease risk may aid the development of early preventative strategies. Mapping the molecular factors that associate with protein levels in the blood may uncover pathways occurring in the body that regulate brain health and disease. These molecular signatures from blood can also help to identify individuals that may have damage occurring in the brain, prior to overt symptom presentations. While the ability of DNAm to offer a more stable indicator of health state than proteins is not fully resolved, the test-retest instability of certain proteins suggests that this it is worthwhile to investigate this further.

In the following sections, I will present an introduction to proteomic measurement technologies that can be applied to serum/plasma from the blood. I then provide of the rationale for identifying DNAm patterns of protein levels, with examples of the statistical methods that can be used to achieve this.
2 Linking the proteome and methylome

In this chapter, I introduce the circulating proteome and common technologies used for quantification of protein levels. I introduce key methodologies that can profile relationships between DNA methylation (DNAm) patterns and plasma protein levels. In doing so, I cover the motivation for linking these molecular signatures. I also compare these approaches to the statistical methods that link genetic and protein data together, noting the similarities, differences and complementary nature of the methods.

2.1 The circulating proteome

In this section, I present an overview of the circulating proteome and the technologies that can be used to measure it.

2.1.1 Proteomic profile of blood

Plasma refers to the liquid component of blood, which comprises 55% of total blood volume. Plasma contains ~91% water and ~9% solids including: coagulants (mostly fibrinogen), proteins, electrolytes, immunoglobulins and other enzymes, hormones and vitamins. Serum is not synonymous with plasma, as it has clotting factors such as fibrinogen removed. Twenty proteins represent 99% of the fraction of the circulating proteome. Some of the most abundant proteins include ALB, IgG, IgA, IgM, SERPINA1, A2M, APOA1, APOA2, C3, TTR, and FGA. The Human Proteome Organisation (HUPO) Human Plasma Proteome Project (HPPP) includes the Human Plasma PeptideAtlas, (HPPA) which comprises 4,395 canonical and 1,482 additional human proteins with more ambiguous evidence detected in plasma in the 2021-07 release. The Human Secretome Project recently mapped proteins that are likely secreted in the human body via transcriptomics database measures. They identified 730 proteins that they predicted were secreted into the blood and assessed the abundance of 365 of these proteins that were detectable across various proteomic assays. Functional indexing of the proteins suggested that inflammatory proteins are a major component of those secreted into blood. Notably, variability was observed between 205 of the proteins that had both mass spectrometry (Mspec) and immunofluorescence assay measures available (overall Spearman rho = 0.78).
Figure 2-1. Proteins detected in human blood across three platforms.

A, Concentrations of 3,223 plasma proteins estimated through a Mspec assay approach that used spectral counting as part of the HPPA. Bars are coloured based on whether proteins were predicted to be: secreted to blood, secreted to other, membrane bound, or intracellular (example proteins annotated for each). B, Concentrations for 365 proteins actively secreted to blood measured via immunoassays. Bars show ranked median plasma concentrations and are coloured based on classification of proteins to functional categories (example proteins annotated for each). C, Correlation of 205 blood secretome proteins across Mspec and immunoassays. D, Number of the 730 proteins that were predicted to be secreted into blood that were detectable across the technologies studied. Adapted from Uhlén et al. (2019)\textsuperscript{162}. Copyright: Science.
2.1.2 Proteomic technologies.

Immunoassays measure the absolute quantity of single protein targets by incubating a sample with an antibody that is specific to the protein. Immunoflourescence quantifies protein concentration by recording light emitted by a fluorescent tag on the antibody, whereas immunoassays use colour change induced by antibody binding. While proteins can be measured from serum, plasma is typically used in high throughput, multiplexed assay designs. These approaches include Mspect and affinity-based assays. Here, I detail the generation of proteomics measures by SomaScan® and Olink® multiplexed platforms, which form the basis of the affinity-based assay samples included in this thesis.

Proximity-based extension assays (PEAs) rely on hybridisation of oligonucleotides with complementary sequences upon binding of both detection antibodies to the target protein. The bound oligonucleotides can then be amplified through the polymerase chain reaction (PCR). Through this dual-antibody recognition approach, the rates of cross-binding between antibodies and targets are minimised. This protocol is used by Olink® Bioscience (Uppsala, Sweden) (Figure 2-2). High-profile applications of this technology include the measurement (due to be released in 2023/2024) of approximately 3,000 plasma protein analyte levels in ~50,000 individuals in the UK Biobank as part of the UK Biobank Pharma Proteomics Project (UKB-PPP) Consortium. The most recent Olink explore platform measures up to 3,000 protein analytes selected for their functional significance in neurology, inflammatory, oncology or cardiometabolic panels.

In contrast, SomaLogic technology relies on protein-capture reagents called SOMAmers (Slow Offrate Modified Aptamers) that bind to the protein targets (Figure 2-3). SOMAmers are singular strands of DNA or RNA that are selected through systematic evolution of ligands by exponential enrichment (SELEX). SELEX involves generation of oligonucleotide databases that are then screened to select aptamers that are highly selective for specific protein targets. The coverage of the assay in 2009 was approximately 800 SOMAmers, increasing to 1,100 by 2012, 1,300 by 2015, 5,000 by 2018, and 7,000 protein measures in the most recent assay available since 2020. Although more agnostic regarding protein selection than Olink (that categorises proteins into functional panels), SomaScan incorporates protein targets spanning key
functionalities such as kinases, receptors, growth factors, hormones and a range of intracellular and extracellular secretions \(^{166}\).

Both the PEA and aptamer assays can measure proteins at low abundance in the blood, making them a favourable choice to quantify functionally important markers with low concentrations as compared to mass spectrometry approaches (that typically measure the most abundant proteins in a sample).

**Figure 2-2. Olink Proximity Extension Assay.**

**A,** Antibody pairs (light blue), labelled with DNA oligonucleotides (grey), bind to the target antigen (dark blue circle) in solution. **B,** Oligonucleotides (dark grey) in proximity as a result of dual antibodies (light blue) bound to the target (dark blue circle) hybridize and are then extended DNA polymerase. **C,** The created piece of DNA is amplified by PCR and can be readout by next-generation sequencing or quantitative PCR for measurement. Taken from Olink Proteomics: PEA – a high-multiplex immunoassay technology with qPCR or NGS readout. White Paper \(^{167}\). Copyright: Olink Proteomics.
Figure 2-3. Multiplexed SOMAmer affinity assay.

Step one, SOMAmer reagents (purple) are synthesised with fluorophores that act as photocleavable linkers and biotin. Step two, SOMAmers bound to streptavidin beads (light blue) are used to capture protein targets (tan) from a range of proteins present in solution. Step three, the unbound proteins are removed via washing, whereas the bound proteins stay tagged with the biotin in protein-specific complexes. Step four, UV light is used to break the photocleavable links and release the complexes into solution. Step five, protein-specific complexes remain bound, whereas non-specific complexes dissociate. Step six, a polyanionic competitor (green) inhibits the rebinding of any non-specific complexes. Step seven, new streptavidin beads (dark blue) are used to capture any complexes between biotinylated proteins and their SOMAmer reagent counterparts. Step eight, denaturing the proteins releases the SOMAmer reagents from the complexes and fluorophores can be measured after hybridisation to complimentary sequences on the microarray chip. Intensity of the fluorescence is detected and is proportionate to the available epitope in the sample. Adapted from visualisations provided in SomaScan® Assay v4.1 Technical Notes 168. Copyright: SomaLogic.
2.2 Mapping associations between the proteome and methylome

In this section I detail the statistical methods that can be used to map DNAm patterns associated with traits. I present literature reviews on how these approaches have been used to model DNAm patterns associated with circulating proteins. I discuss the motivations for using these methods in the context of brain health.

2.2.1 MWAS

MWAS can be run with DNAm as the predictor or the outcome. Most frequently, DNAm measures at CpGs are used as predictor variables. As with SNP-trait associations modelled through GWAS, the effect estimates for each DNAm-trait association can be estimated in turn through linear regressions. Typically, continuous or binary outcomes are modelled. Time-to-event analyses have also been conducted, although this approach is more computationally demanding than linear or logistic regression. Martingale residuals from Cox proportional hazards models on incident diseases can be included as dependent variables to model DNAm differences associated with time-to-event traits. As of 2022, the EWAS Atlas had a total of 1,002 studies recorded, with common outcomes including smoking status, aging, body mass index (BMI) and type 2 diabetes (Figure 2-4).

Figure 2-4. The rise in MWAS from 2009 to 2022.

Curated statistics on number of methylome-wide association studies (MWAS) publications (y-axis) per-year (b) and per-trait (excluding cancer) (c). These data were taken from EWAS Atlas database and used to create this figure in January 2023.
Comparatively fewer technical and biological considerations impact GWAS, since DNA is largely unchanged in the human lifespan. In GWAS, the genomic inflation factor is defined as the ratio of the observed median chi-squared test statistic to the expected median of the corresponding chi-squared distribution. A genomic inflation factor of 1 indicates there is little evidence of inflation. A similar approach can be taken to assess MWAS inflation, which can be influenced by various sources of variability in DNAm. Inflation of test statistics can lead to overestimated significance in associations, elevating the likelihood of false positives. Given the sources of DNAm variability outlined in Section 1.3.6, there are several considerations for the design and interpretation of MWAS studies. Adjusting for cell-type heterogeneity has been shown to bring MWAS inflation towards a value of 1 and is a widely accepted approach. However, this does not account for all sources of inflation, which are likely due to other technical or biological factors. Technical variation can occur between DNAm measured in different groups of individuals from a cohort at different time points and should therefore be accounted for. Biases can be introduced from the effects of unobserved or unadjusted for confounders. Adjusting for the principal components of the DNAm data that explain the largest proportion of variance was found to further reduce inflation (beyond cell-type adjustments) in work by Zhang et al. (2019). The authors ultimately found that a linear mixed model approach that estimated the effect of each DNAm probe while also adjusting for background genome-wide DNAm reduced the inflation statistic to a value of 1. When adjusting for additional covariates as fixed effects, this can also pose challenges regarding interpretation. For example, in a large-scale MWAS of educational attainment (EA) the authors note that in a basic model that did not account for smoking or BMI, associations between DNAm and EA may actually be due to these factors. However, as smoking and BMI are correlated with EA, models that adjusted for these factors may have reduced the potential to detect true associations between DNAm and EA. Staged models can therefore be run with increasingly complex sets of covariates to tease out the influences of observed confounders. Relatedness between individuals is another key confounder that should be accounted for. Genetic principal components can also be
regressed onto the outcome or included as covariates in MWAS to account for population structure \(^{174}\).

### 2.2.2 Rationale for MWAS of proteins

Modelling the DNAm signatures associated with intermediary markers (protein levels) may provide further context regarding the possible molecular contributions to complex traits. For example, in an MWAS of cognitive ability that I was a co-author on, despite 41% of the variance in the trait being attributable to DNAm, only three CpG sites were associated (posterior inclusion probability > 0.95 in Bayesian modelling) with the trait \(^{175}\). The previous associations reported for these sites in the EWAS catalog \(^{176}\) involved lung function, age, sex and metabolic traits – all of which are broad and provide limited insight into the specific biological pathways that may be related to cognitive ability. As proteins associate with cognitive ability and are effector molecules of brain disease, I hypothesise that mapping the DNAm profiles of protein markers of brain health traits may uncover additional insights beyond MWAS of the traits themselves. The high proportion of variance that was explained in the cognitive ability trait by DNAm (41%) suggests that there may be a large number of small, additive effects in complex biological pathways present, which is consistent with my hypothesis. This is expected, as cognitive ability (and many other traits examined as outcome phenotypes) are heterogeneous by nature and likely to be impacted by a range of lifestyle, environmental and disease effects.

### 2.2.3 MWAS of proteins

Mapping of the epigenetic and genetic architectures of proteins was introduced in Section 1.3.8. I conducted a literature review to search for MWAS of blood protein levels, focusing primarily on multiplexed protein platform measures. The following search criteria were used:

1. DNA methylation.mp OR MWAS.mp OR EWAS.mp OR epigenome-wide.mp OR methylome-wide.mp
2. Plasma protein.mp OR serum protein.mp OR plasma proteome.mp OR blood protein.mp OR circulating proteome.mp
3. 1 AND 2
Inclusion criteria were: (i) methylome/epigenome-wide association study, (ii) DNAm assessed via 450K or EPIC arrays, (iii) multiplexed proteins, (iv) DNAm and protein measurements from blood and (v) original research article. A search on PubMed and other platforms was also done manually to check and curate studies, identifying any studies that were not focused on multiplexed protein platform measures that may be relevant to this thesis. The outcome of the literature search is shown in Table 2-1.

Six studies were identified that ran MWAS of blood-based protein measures derived from multiplexed platforms. Zaghlool et al. (2020) had the largest coverage of plasma proteins to date (1,123 SomaScan measurements)\(^\text{177}\). This study performed linear regression MWAS in 944 individuals from the KORA population study using 470,837 CpG measures from the 450K Illumina array. A separate sample of 344 individuals from the Qatar Metabolomics Study on Diabetes (QMDiab) were used as a replication population. Staged MWAS were run, regressing out increasingly complex sets of covariates. In models that did not adjust for any covariates, 38,492 pQTMs were identified and 12,606 replicated (Bonferroni-adjusted \(P < 9.46 \times 10^{-1}\)). In the final model adjusting for age, sex, blood cell-type composition, cis-acting genetic variation via pQTLs for each protein, smoking, BMI and diabetes, 318 pQTMs were identified and 98 replicated across both populations (Figure 2-5). The most important covariate adjustment was for cell-type effects, which attenuated 24,011 of the original associations after adjustment. Seventy-two of the associations involved CpGs within the pappalysin-1, or pregnancy-associated plasma protein-A (PAPPA) gene locus. The authors note that PAPPA has a role in pregnancy, but is also implicated in inflammation, heart failure, tumour cell proliferation and various other functionalities\(^\text{177}\). When further adjustments were made for eosinophil counts in the blood, 70 of the 72 associations were attenuated to non-significance.

In the study by Zaghlool et al. (2020), fourteen of the 98 pQTMs included DNAm at CpG sites within the NLRC5 (NOD-like receptor family CARD domain containing 5) gene locus that were identified in pQTM involving immune-regulatory proteins (CD48, CD163, CXCL10, CXCL11, LAG3, FCGR3B, and B2M)\(^\text{177}\).
Figure 2-5. 98 pQTM that replicated across the KORA and QMDiab populations.

Outermost circular bars represent chromosomes of the human genome. Positions of CpGs and protein-coding genes involved in pQTM associations are shown. Connecting lines indicate the presence of a pQTM association. Figure taken from Zaghlool et al. (2020) 177. Copyright: Nature Communications.

The NLRC5 locus was therefore implicated as a potential mediator of the body’s response to inflammatory exposures. NLRC5 is an activator of the inflammasome that modulates the expression of receptors in the innate immune system 178. One of the key conclusions from this work is pQTM were particularly enriched for CpGs and proteins that had roles in inflammation. Thus, DNAm is likely to be useful as an
indicator of chronic, low-grade inflammatory signatures in the body. An important consideration is that Zaghlool et al. (2020) is the only protein MWAS study of the seven identified in this review that accounted for pQTL effects on proteins. The cis pQTLs were sourced from GWAS of the same 1,123 proteins performed previously and represent SNP effects that are likely to mediate genetic influences on protein levels (due to their proximity to the protein-coding transcriptional start site). Pre-adjusting protein levels for these SNP effects is thought to eliminate some of the quantifiable genetic effects on the protein, leaving a signal that is thought to be more reflective of non-genetic effects.

Four of the six multiplexed protein MWAS studies modelled MWAS of proteins alongside protein GWAS in the same samples using multiplexed array protein measures. In the first, Ahsan et al. (2017) ran linear mixed effects MWAS on 121 plasma protein measures derived from the Olink CVD I and Oncology I arrays in 729 individuals. These models were adjusted for age, sex, cell-type proportions, year-of-sampling and technical DNAm processing variables. GWAS were also performed. A total of 188 pQTMs were identified, with 88 CpGs associated with 44 proteins and thirteen CpGs associated with multiple protein levels. CpG sites in the NLRC5 gene were implicated in associations with IL12, IL18, CXCL11 and CXCL9. CpG sites in the AHRR ( Aryl-hydrocarbon receptor repressor) and F2RL3 (F2R-like thrombin or trypsin receptor 3) genes were associated with IL12 and WFDC2 protein levels, but were attenuated to non-significance after adjustment for smoking. Eighteen proteins had a CpG and SNP association. When adjusting for the SNP effect in the CpG-protein association, the 11 proteins that had signals in the same region (SNP and CpG within two megabases (Mb) of one another) remained genome-wide significant after adjustment for the SNP. However, the seven proteins that had CpGs involved in pQTMs that were not situated close to the SNP were unaffected by adjustment for the SNP effect and in some cases associations became more significant.

Hillary et al. performed three MWAS and GWAS studies of protein levels. Two of these studies included 92 neurology and 70 inflammatory Olink panel protein measures in individuals from the LBC1936 study. The neurology protein MWAS in 750 individuals was run via linear regression, accounting for age, sex, estimated immune cell proportions and technical DNAm processing covariates. This identified 26 epigenome-wide significant (P < 3.9 × 10^{-10}) pQTMs involving 9 proteins, with 17 cis
and 9 trans. Neural tissue was the most common tissue type that the gene for the proteins involved in pQTMs were expressed in. The genes corresponding to the CpGs were represented in immune functionalities. Three proteins had genome-wide SNPs identified in addition to pQTMs (NEP, MATN3 and MDGA1). Cross-checking the NEP pQTLs and pQTMs against mQTL databases revealed that the trans pQTL SNP (rs4687657, in the ITIH4 gene) was previously associated with differential DNAm levels at cg18404041 (ITIH4) and cg11645453 (ITIH1), which were the CpG sites implicated in pQTMs for NEP. These findings implied some level of co-regulation between NEP, ITIH4 and ITIH1. The inflammatory protein MWAS was conducted in 876 individuals using a Bayesian framework, with sensitivity analyses performed using linear mixed model approaches. Three pQTMs involving three proteins were identified, with DNAm able to account for up to 46% of variance in levels of CXCL10. The pQTMs included associations between the smoking-associated probe cg05575921 and CCL11, cg07839457 and CXCL9 and cg03938978 and IL18R1. These associations had posterior inclusion probabilities (PIP) of more than 95% in the Bayesian framework and $P < 5.14 \times 10^{-10}$ in the mixed model analysis.

The third study published by Hillary et al. (2022) ran MWAS/GWAS on a set of 282 plasma proteins from the SomaScan array in 778 individuals from Generation Scotland. The 282 proteins were selected due to previously-reported links between these proteins and Alzheimer’s dementia. Unlike all of the other protein MWAS studies (that use 450K array measures), this study used the EPIC array, with an expanded DNAm coverage to 771,619 CpG sites. Again, a Bayesian framework was employed, which 26 CpGs associated with the levels of 20 proteins (PIP > 95%), with 10 cis and 16 trans pQTMs. A novel pQTM between TREM2 levels and DNAm at cg02521229 that is positioned near the MS4A4A gene (implicated as a dementia risk locus) was identified, along with a trans-pQTL in this gene that was associated with higher TREM2 levels. MS4A4A has been implicated as a dementia risk gene. The findings suggest plasma assays may be able to capture risk mechanisms arising from MS4A4A polymorphisms and TREM2 levels and that the mechanism could involve differential methylation at the MS4A4A gene region.

Finally, Myte et al. (2019) analysed CpG associations with 161 plasma proteins (160 generated via the Olink platform, with immunoassay CRP measured also included) in 127 individuals. The study identified 85 pQTMs and found that the NLRC5 locus
had two CpG sites that were associated with multiple inflammatory protein levels (CXCL9 and CXCL11). This study is somewhat underpowered as compared to the other five studies identified that modelled multiplexed protein measures as outcomes in MWAS. The population is also enriched for ~50% cases of cancer, which is an important consideration that should be considered when interpreting the results.

Two additional MWAS were identified as part of the literature search. While these studies did not meet the inclusion criteria of utilising multiplexed protein panels, they did include large populations. Ligthart et al. (2016) and Wielscher et al. (2022) focused on MWAS studies of immunoassay-derived CRP levels, with increasing sample sizes of 8,863 and 22,774, respectively 127,183. A total of 1,511 pQTM were identified in the most recent meta-analysis performed by Wielscher et al. (2022), suggesting that pQTM associations in the original 2016 study may have been underpowered to detect many biological patterns of DNAm that are related to CRP levels 127. These studies reinforce the strong relationship between methylation and circulating immune proteins, as CRP is a primary mediator of the acute-phase immune response and is also thought to drive chronic inflammation 158.

This review highlights that there is heterogeneity of studies by protein panels, statistical approaches and protein coverage. Taken together, these studies suggest that there is a strong interconnectivity between the circulating proteome and DNAm. The connection between the methylome and inflammasome is especially pronounced in these studies. The links between DNAm and circulating protein levels may be important to understand in the context of chronic inflammation, neuroinflammation and the impact of these processes on neurological disease.
Table 2-1. Summary of key MWAS/EWAS studies profiling blood protein levels.

<table>
<thead>
<tr>
<th>Author (date)</th>
<th>Cohort</th>
<th>Age, % female</th>
<th>N individuals</th>
<th>N proteins</th>
<th>Metho d</th>
<th>Source</th>
<th>Proteomic platform</th>
<th>DNA platform</th>
<th>Key findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ligthart et al. (2016)</td>
<td>Meta-analysis across 11 cohorts</td>
<td>Mean 41-87 years across cohorts, 54% female</td>
<td>8,863</td>
<td>1</td>
<td>Linear mixed effects regression</td>
<td>Serum</td>
<td>High-sensitivity assay</td>
<td>450K</td>
<td>Meta-analysis of CRP EWAS. 218 pQTMs identified, 58 replicated in African Americans.</td>
</tr>
<tr>
<td>Ahsan et al. (2017)</td>
<td>Northern Sweden Population Health Study</td>
<td>Media 50.4 years, 53% female</td>
<td>698</td>
<td>121</td>
<td>Linear regression</td>
<td>Plasma</td>
<td>Olink Oncology I and CVD I</td>
<td>450K</td>
<td>188 CpG sites associated with 44 protein levels (P &lt; 1.06x10^-7). 13 CpG sites associated with multiple protein levels.</td>
</tr>
<tr>
<td>Myte et al. (2019)</td>
<td>Västerbotten intervention programme</td>
<td>Mean age 50 years, 42% female</td>
<td>127</td>
<td>161</td>
<td>Linear regression</td>
<td>Plasma</td>
<td>CRP assay, combined with Olink Oncology II and Inflammation panels</td>
<td>EPIC</td>
<td>85 pQTMs replicated results from Ligthart et al. (2016) and Ashan et al. (2017) MWAS studies. 4 CpGs implicated in &gt;1 pQTM, with 2 in the NLRC5 gene.</td>
</tr>
<tr>
<td>Study</td>
<td>Cohort</td>
<td>Mean Age (yrs)</td>
<td>% Female</td>
<td>Study Design</td>
<td>Platform</td>
<td>Array Size</td>
<td>GWAS/EWAS Details</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------------</td>
<td>--------------</td>
<td>----------------</td>
<td>----------</td>
<td>-------------------------------</td>
<td>------------------</td>
<td>------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hillary <em>et al.</em> (2019)</td>
<td>LBC1936</td>
<td>73 (72-75)</td>
<td>49.8%</td>
<td>Linear regression</td>
<td>Plasma Olink</td>
<td>450K</td>
<td>GWAS/EWAS on neurological proteins. In EWAS, 26 pQTMs for 9 proteins (P &lt; $3.9 \times 10^{-10}$).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hillary <em>et al.</em> (2020)</td>
<td>LBC1936</td>
<td>69.8 (66-74)</td>
<td>49%</td>
<td>Bayesian framework and mixed model approach</td>
<td>Plasma Olink</td>
<td>450K</td>
<td>GWAS/EWAS on inflammation proteins. In EWAS, 3 pQTMs, for 3 proteins (PIP ≥ 95%). $R^2 = 46%$ for CXCL10 levels.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zaghlool <em>et al.</em> (2020)</td>
<td>KORA</td>
<td>59.3 (52-68)</td>
<td>48.2%</td>
<td>Stepwise linear regression</td>
<td>Plasma SomaScan v3.2 array</td>
<td>450K</td>
<td>EWAS of protein levels. 318 pQTMs, 98 replicated in external cohort (QMDiab, N=344). NLRC5 signature of DNAm highlighted as inflammation-associated locus.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hillary <em>et al.</em> (2022)</td>
<td>Generation Scotland</td>
<td>59.9 (56-63)</td>
<td>59.1%</td>
<td>Bayesian framework and mixed model</td>
<td>Plasma SomaScan v4 array</td>
<td>EPIC</td>
<td>GWAS/EWAS of 282 Alzheimer’s dementia-related proteins. EWAS identified 26 pQTMs for 19 proteins.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Number of Cohorts</td>
<td>Age Range</td>
<td>Female (%)</td>
<td>Analytical Approach</td>
<td>Serum Assays</td>
<td>Specific Assays</td>
<td>Meta-analysis Remarks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------</td>
<td>-------------------</td>
<td>-----------</td>
<td>------------</td>
<td>---------------------</td>
<td>--------------</td>
<td>----------------</td>
<td>----------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wielscher et al. (2022)</td>
<td>30</td>
<td>16-76</td>
<td>43%</td>
<td>Linear regression</td>
<td>Serum</td>
<td>Specific assays for each cohort not stated.</td>
<td>450K or EPIC across cohorts may be largely a result of CRP differences. Smoking and obesity mediated CRP DNA methylation signatures.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* denotes studies that are pre-prints and not yet published. LBC1936: Lothian Birth Cohort 136. PIP: Posterior inclusion probability. KORA: Cooperative Health Research in the Region of Augsburg.
2.2.4 Challenges around causality in pQTMs

Integrating the methylome, proteome and phenome may help to resolve the challenges regarding DNAm-trait interpretation. Nevertheless, identifying causal pathways from MWAS remains complex as pQTMs (unlike pQTLs) have bidirectional possibilities. DNAm can both regulate transcription factor binding and therefore expression of genes, but can also be consequentially influenced by exposures such as protein levels. For example, the recent MWAS meta-analysis of CRP performed by Wilscher et al. (2022) included Mendelian randomisation between DNAm probes and the levels of CRP. Their results suggested that DNAm differences captured by pQTMs are a result of CRP, rather than being involved in the regulation of the protein. While the findings of this study agree with the consensus in the field - that DNAm largely represents the record of exposures (such as inflammation driven by CRP) on the body - the confidence that can be ascribed to causal testing is still debated. Querying publicly available databases can help to identify whether a CpG site involved in a pQTM association is likely influenced by a genetic variant, such as an mQTL. Examples of this approach were highlighted in the literature review through studies conducted by Hillary et al. (2019, 2020 and 2022).

2.3 Epigenetic scoring

Predictive scoring approaches can be used to model binary, continuous and time-to-event outcomes. Their primary aim is to develop scores that are generalisable, which means that overfitting the scores to characteristics that are specific to only the training population should be avoided. Ideally, features are selected and weights are derived in a training population with scores then tested in an out-of-sample, unseen population that has the outcome of interest measured. These unseen populations could represent either a completely independent cohort population, or a hold-out sample within the training cohort population. Once the signal captured by the score is validated in the unseen sample, scores can be projected into new populations providing they have measurements of the features used to derive the score.
In this section, I will outline the principles of creating DNAm-based, epigenetic scores (EpiScores) for traits. I highlight examples of EpiScores that have been generated using a diverse range of methods, describing how these methods build on the approaches used in the polygenic risk score (PRS) field. I then perform a literature review, summarising and discussing studies that have generated EpiScores for circulating protein levels.

2.3.1 EpiScores

A range of statistical methods can be used to generate DNAm-based, epigenetic scores (referred to as EpiScores). The types of approaches used are informed by work done in the PRS field, where scores are typically calculated using linear combinations of effect size coefficients assigned to SNPs in GWAS for an outcome. PRS have been derived for disease outcomes, lifestyle traits and molecular intermediaries such as protein levels \(^\text{185,188,189}\). They capture a genetic component of disease risk that is heritable, with predictive accuracy varying by disease \(^\text{188}\). Since PRS are reliant on genetic information, they are inherently limited in the context of disease risk stratification as they estimate liability to disease that is present from birth \(^\text{188}\). As detailed in Section 1.3.6, DNAm can be variable through the life course and is associated with various biological, environmental and lifestyle factors. Consequently, EpiScores are likely to reflect the accumulation profile of risk factors in a time-sensitive manner and may provide an additional layer of information that is relevant to disease risk stratification, beyond the standard PRS approach \(^\text{187}\).

2.3.2 Methods for EpiScore generation

To generate an EpiScore, informative CpG features should be identified in a training population. This is challenging given that the content on Illumina array-based technologies changes over time; typically, most cohort studies have measures at \(\sim450,000\) (450K array) or \(\sim800,000\) CpG sites (EPIC array) \(^\text{139}\). When generating EpiScores, there are several characteristics of the statistical design that must be considered. The extent to which data leakage may be occurring during training is important to consider. The presence of related individuals across training and test samples may inflate the performance of EpiScores, as DNAm variability can be
associated with genetic variation\textsuperscript{190}. Care must be taken to minimise batch effects arising from technical variation\textsuperscript{139}. The population composition of the training dataset is also an important determining factor that impacts performance in external test sets.

Table 2-2 summarises the methods that can be used to select features, with examples from the literature. As with PRS, the effect estimates for each CpG that are derived from linear MWAS approaches can be used to calculate scores based on the weighted linear sum of these effects. Typically, studies will apply a threshold to restrict CpG features to those with the strongest significance profile in an MWAS. For example, an EpiScore has been developed based on a linear mixed model MWAS of ALS diagnosis status in an Australian population (782 cases and 613 controls)\textsuperscript{191}. The scores for ALS were projected into a population from the Netherlands (1,159 cases, 637 controls) using various P-value thresholds for selection of CpG features. There was minimal variability in score performance with the best-performing score having an AUC of 0.69, with 95% CI = [0.66–0.71]\textsuperscript{191}. However, given the limitations around sources of inflation in linear MWAS discussed in Section 2.2.1, this method can be impacted by a phenomenon called the ‘winner’s curse’. The use of MWAS results from Bayesian MWAS have been used to generate EpiScores and may better-account for unobserved confounders when estimating effects for probes. However, Beta weighting coefficients are typically considered for all sites tested when building an EpiScore\textsuperscript{133,175}. This approach is considered to be more computationally expensive than other MWAS approaches. For translation of EpiScores in clinical practice, it may also be desirable to have a more parsimonious subset of CpGs, rather than using coefficients for all sites included on the array. As there is a high degree of collinearity in omics data, it is reasonable that an adequate predictive signal should be possible with fewer CpGs selected as features. In PRS studies, clumping is usually run on the SNPs identified through GWAS to isolate sentinel, independent signals, thereby streamlining the SNPs retained in a scoring approach\textsuperscript{192}. However, there is no equivalent practice used in relation to DNAm datasets.

An approach that has gained popularity in the EpiScore community is the use of penalised regression. This approach simultaneously considers all CpG measures but applies a penalty term such that CpGs that do not have large contributions to the model are shrunk towards zero or assigned zero\textsuperscript{193}. The penalty term is called lambda and
this term represents a hyperparameter that can be tuned through cross-validation, using different lambda values to identify the model with the optimal fit. Least absolute shrinkage and selection operator (LASSO) regression assigns some CpG sites with coefficients of 0, whereas ridge regression shrinks coefficients but does not assign any CpG sites with exact 0 values \(^{193}\). The LASSO approach has been used to create a range of EpiScores for smoking such as the Epigenetic Smoking status Estimator (EpiSmokEr) \(^{126}\), alcohol consumption \(^{194}\), incident type 2 diabetes \(^{195}\), major depressive disorder \(^{196}\) and a range of clinically-measured traits such as body mass index and cholesterol \(^{197}\). Elastic net regression is an intermediate method that sits between LASSO and ridge regression, which can handle multicollinearity between CpGs while creating sparsity in the features selected from the model \(^{198}\). Elastic net is not as harsh as LASSO, which will exclude all possible intercorrelated CpGs. However, it still reduces the selected features dramatically, as compared to ridge regression.

The application of the elastic net method has demonstrated optimal performance in the field of ‘epigenetic clocks’. Epigenetic clocks are EpiScores trained to capture aspects of ageing. The original epigenetic predictor of age was developed by Bocklandt et al. (2011) \(^{199}\) through LASSO regression. This was quickly followed up with epigenetic clocks developed through elastic net penalised regression by Horvath et al. (2013, using 8,000 samples across 51 tissues) \(^{200}\) and Hannum et al (2013, using blood measures from 656 individuals) \(^{201}\). A range of additional clocks have also been developed using elastic net penalised regression approaches: DNAm PhenoAge by Levine et al. (2018) \(^{202}\), DunedinPoAm – Pace of Ageing by Belsky et al (2020) \(^{203}\), DunedinPACE by Belsky et al. (2022) \(^{204}\); and GrimAge by Lu et al (2020) \(^{129}\). Each of these scores derive metrics that attempt to quantify biological age acceleration, which is calculated through regression of DNAm-estimated age onto chronological age. The Horvath, Hannum, DenedinPACE and DunedinPoAm scores were trained directly on DNAm at CpGs as features. The DNAm PhenoAge score was created by developing a measure of health state (termed PhenoAge) based on a range of lifestyle and clinical factors that included CRP levels (the only protein studied), which was used as the outcome in elastic net penalised regression. A more recent clock from Levine and colleagues uses principal components of the DNAm data as input features, which
has shown high reliability across multiple time points and population samples. Individuals with higher rates of biological aging than their chronological age are thought to be at greater risk of a range of adverse health outcomes. Perhaps the most widely-recognised, gold-standard measure of biological age acceleration is GrimAge, which has been linked to mortality, incident diseases and poorer brain health outcomes. As compared to the other epigenetic clock measures, GrimAge is unique in that it uses protein EpiScores as informative features (in addition to a DNAm-derived smoking score, age and sex). In Sections 2.3.4 and 2.3.5, I provide a detailed review of protein EpiScores and explore the motivations for generating them in further depth.
Table 2-2. Examples of common DNA-methylation based EpiScore approaches.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Features</th>
<th>Literature examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear regression summary statistics</td>
<td>DNAm at CpGs</td>
<td>BMI, height 208</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Schizophrenia 209</td>
</tr>
<tr>
<td>Linear mixed model summary statistics</td>
<td>DNAm at CpGs</td>
<td>ALS 191</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Alzheimer’s and Parkinson’s disease, ALS, schizophrenia, rheumatoid arthritis 210</td>
</tr>
<tr>
<td>Penalised linear regression</td>
<td>DNAm at CpGs</td>
<td>Smoking 126</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Alcohol consumption 153,194</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Incident diabetes 195</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Major depressive disorder 196</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Alcohol consumption, body fat percent, body mass index, lipoprotein cholesterol, waist-to-hip ratio. 197</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Epigenetic clocks (Horvath, Hannum, Levine) 200–202</td>
</tr>
<tr>
<td>PCA-based penalised regression using principal components of CpGs</td>
<td>PCA components</td>
<td>Epigenetic clocks (Levine) 205</td>
</tr>
<tr>
<td>Protein EpiScore based prediction using pre-trained protein EpiScores as features fed into penalised regression</td>
<td>Protein EpiScores</td>
<td>Epigenetic clocks (GrimAge) 129</td>
</tr>
<tr>
<td>Bayesian inference MWAS summary statistics</td>
<td>DNAm at CpGs</td>
<td>BMI, smoking 133</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cognitive ability 175</td>
</tr>
</tbody>
</table>

Adapted from a review I co-authored on, from Nabais et al. (2023) 142. PCA: principal components analysis.
2.3.3 EpiScore use case

EpiScores can be useful in a range of contexts. When incomplete health records exist for individuals, EpiScores can offer a means to impute measures. One of the clearest examples of this is smoking status. McCartney et al. (2018) developed a LASSO-based EpiScore for smoking status as a binary outcome including 2,523 never smokers and 921 current smokers. When tested in the LBC1936 cohort (418 non-smokers and 102 current smokers), a near-perfect discrimination was observed for the smoking EpiScore, with AUC = 0.98, 95% CI = [0.97-1.00]. Furthermore, smoking EpiScores have been found to be potentially more informative than self-report data. EpiScores could also be used to quantify the effects of clinical interventions, as the DNAm profile associated with smoking cessation has been shown to vary over the 5-year post-cessation timeline. The GrimAge measure of biological age acceleration is already being used as an endpoint in clinical trials. Epigenetic clocks for chronological age imputation are also useful tools in the context of forensic evidence assessment. Finally – and most relevant to this thesis – EpiScores can be used as biomarkers for risk prediction. In the following sections, I first outline a literature summary of protein EpiScore studies, followed by a rationale on the use-case for generating additional protein EpiScores.

2.3.4 Protein EpiScores

In Section 1.3.8, I introduced the concept of mapping the genetic and DNAm architectures for proteins. For proteins that have strong profiles of these molecular measures, scores can be trained using omics measures as input features and setting the protein as the outcome. In recent work, the OmicsPred platform has been developed that hosts genetic scores trained for 17,227 molecular traits (protein, metabolite and RNA sequencing measures). As datasets with both DNAm and protein measures have become available, DNAm-based protein EpiScores have been developed. These follow the same principles as the methodologies for EpiScore generation I introduced in Section 2.3.2.

I conducted a literature review to identify existing DNAm scores for blood protein levels. Studies were reported if they developed a minimum of one genome-wide
DNAm-based EpiScore with protein levels as the outcome. As this is a relatively new area of research, only six studies met the search criteria. These studies (summarised in Table 2-4) are introduced in this section and used to inform the rationale for generating additional protein EpiScores in the following section.

The GrimAge study by Lu et al. (2019) represents the first of the six studies identified in the literature search for protein EpiScores that generated EpiScores through elastic net penalised regression. This method was applied to 2,356 individuals from the Framingham Heart Study (FHS), with a 70% train 30% test split to develop protein EpiScores. Of the 88 immunoassay protein measures tested, 12 protein EpiScores had a correlation of greater than 0.35 with the measured protein in the test set. A Cox elastic net with time to mortality was then trained (219 deaths of the 1,729 total sample) with a DNAm-based score for smoking, age, sex and the 12 protein EpiScores as potential features. The test set included 88 deaths out of a total population of 625 individuals. The smoking score and seven protein EpiScores (B2M, ADM, GDF15, leptin, TIMP-1, PAI-1 and Cystatin-C) were selected as informative features. Validation was performed on five cohorts including over 6,000 individuals. This clock is one of the gold-standard measurements of biological age acceleration, which is the deviation from chronological age that is thought to reflect internal biological ageing. The GrimAge acceleration score was calculated by regressing GrimAge onto chronological age. This score was found to be highly associated with time to mortality (P=2.6x10^-53). Subsequent studies have found that GrimAge acceleration outperforms other epigenetic clock measures in associations with age-related diseases and mortality.

The GrimAge study reported that protein EpiScore measures outperformed their protein counterparts in associations with time-to-mortality in the test sample. Table 2-3 is taken from Supplementary Table 4 of the original study and highlights that in the test set comparisons this was true for PAI-1 plasma levels (EpiScore P=8.7x10^-4, versus protein P=0.074), TIMP-1 (EpiScore P=3.8x10^-4, versus P=0.017), B2M (EpiScore P=2.9x10^-3, versus protein P=1.3x10^-2) and cystatin C (EpiScore P=0.019, versus protein P=0.054).
Table 2-3. Comparison of protein EpiScores versus measured proteins in associations with time-to-mortality in the GrimAge.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Data</th>
<th>HR*</th>
<th>P-value</th>
<th>HR*</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ImmunoAssay</td>
<td>DNAm</td>
<td>ImmunoAssay</td>
<td>DNAm</td>
</tr>
<tr>
<td>Adrenomedullin</td>
<td>training</td>
<td>1.23</td>
<td>1.36</td>
<td>4.53E-4</td>
<td>3.85E-7</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>1.33</td>
<td>1.06</td>
<td>2.92E-2</td>
<td>7.23E-1</td>
</tr>
<tr>
<td>Beta-2 Microglobulin</td>
<td>training</td>
<td>1.31</td>
<td>1.51</td>
<td>7.42E-8</td>
<td>9.87E-8</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>1.22</td>
<td>1.94</td>
<td>1.33E-2</td>
<td>2.92E-3</td>
</tr>
<tr>
<td>Cystatin C</td>
<td>training</td>
<td>1.25</td>
<td>1.53</td>
<td>1.32E-4</td>
<td>1.44E-8</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>1.13</td>
<td>1.53</td>
<td>5.40E-2</td>
<td>1.92E-2</td>
</tr>
<tr>
<td>GDF-15</td>
<td>training</td>
<td>1.24</td>
<td>1.20</td>
<td>3.82E-5</td>
<td>2.58E-3</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>1.83</td>
<td>1.63</td>
<td>6.86E-13</td>
<td>2.32E-2</td>
</tr>
<tr>
<td></td>
<td>training male</td>
<td>1.28</td>
<td>2.23</td>
<td>1.03E-1</td>
<td>9.12E-6</td>
</tr>
<tr>
<td></td>
<td>test male</td>
<td>0.99</td>
<td>1.11</td>
<td>9.74E-1</td>
<td>6.83E-1</td>
</tr>
<tr>
<td></td>
<td>training female</td>
<td>0.93</td>
<td>0.97</td>
<td>4.81E-1</td>
<td>8.38E-1</td>
</tr>
<tr>
<td></td>
<td>test female</td>
<td>1.15</td>
<td>1.18</td>
<td>3.25E-3</td>
<td>5.91E-1</td>
</tr>
<tr>
<td>Leptin</td>
<td>training</td>
<td>1.14</td>
<td>1.31</td>
<td>3.95E-2</td>
<td>3.63E-6</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>1.19</td>
<td>1.37</td>
<td>7.42E-2</td>
<td>8.69E-4</td>
</tr>
<tr>
<td>PAI-1</td>
<td>training</td>
<td>1.33</td>
<td>2.61</td>
<td>1.38E-5</td>
<td>1.06E-8</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>1.29</td>
<td>3.30</td>
<td>1.71E-2</td>
<td>3.79E-4</td>
</tr>
</tbody>
</table>

* in units of one SD. Hazard ratios (HR) and the corresponding Cox regression p-values of the protein levels based on ImmunoAssay measure and protein EpiScore measure are provided. The p value is emboldened if the protein EpiScore was more significant than the protein. GDF-15=growth differentiation factor 15; PAI-1=plasminogen activation inhibitor 1; TIMP-1=tissue inhibitor metalloproteinase 1. The survival analysis was adjusted for chronological age, study center, and family structure. Cox models were run in both the training (219 deaths of 1,729 individuals) and testing (88 deaths of 625 individuals) samples. The analysis for leptin levels was stratified by sex. Taken from Lu et al. (2020)129. Copyright: Aging Cell.
The 12 protein EpiScores generated in the GrimAge study have since been tested as individual biomarkers for Parkinson’s disease (569 cases, 238 controls). In logistic regression models, a one SD decrease in EpiScore EFEMP-1 (OR= 0.83 per SD, 95% CI= 0.70, 0.98) and a one SD increase in EpiScore CD56 (OR =1.41, 95% CI= 1.11, 1.79) were associated with Parkinson’s disease.

Prior to GrimAge, Barker et al. (2018) utilised regression coefficients for CpGs derived from the MWAS of CRP conducted by Ligthart et al. (2016) (see Table 2-1 and Section 2.2.3 for details) to generate a CRP EpiScore. Weighting coefficients for seven CpGs that had the strongest evidence of a functional association with serum CRP measures were used to project the EpiScore for CRP into the ARIES (Accessible Resource for Integrated Epigenomics Studies) cohort. The CRP score correlated with CRP levels in blood samples collected at ages 15 and 17 (r > 0.1, P < 0.002 in both cases). When projected into samples derived at birth (n=758) and modelled in relation to cognitive outcomes across childhood and adolescence, higher levels of the CRP score were associated with poorer cognitive function. The same weighting coefficients were later utilised by Stevenson et al. (2020) to derive a CRP score in the LBC1936 (n=889) and GS (n=7,028) populations. Measured CRP declined over 9 years of follow-up in LBC1936 (Beta = − 0.014, SE = 0.005, P = 0.003), whereas EpiScore CRP increased by an average of 0.07 SD per year (SE = 0.004, P < 2 × 10^{-16}) (Figure 2-6). The inter-class correlations for CRP across the four LBC1936 waves of measurement was 0.72 (95% CI [0.69, 0.74], P < 2 × 10^{-16}). For the CRP EpiScore, the inter-class correlations across the four waves was 0.82 ([0.75, 0.86], P < 2 × 10^{-16}). When assessing the 7 CpGs that comprised the score, correlations ranged between 0.60 (cg27023597) and 0.94 (cg06126421) across the time point measures. The authors identified evidence for an enhanced temporal stability of the CRP EpiScore, which exhibited stronger correlations between waves, ranging from 0.53 (wave 1 versus wave 4) to 0.75 (wave 2 versus wave 3) (Figure 2-6). By comparison, the inter-wave correlations for measured CRP were from 0.3 (wave 1 versus wave 4) to 0.45 (wave 1 versus wave 2). Finally, the CRP EpiScore was associated with poorer cross-sectional cognitive ability in both the LBC1936 (Beta= − 0.08, SE = 0.03, FDR P = 0.04) and GS (Beta= − 0.04, SE = 0.01, FDR P = 0.04) populations. Neither measured CRP nor a genetic score for CRP had significant associations with cognitive ability across the populations tested (P > 0.16, the FDR-adjusted threshold). The
authors concluded that the CRP EpiScore had a higher test-retest reliability than measured CRP and that the EpiScore may represent a more stable signature of chronic inflammation.

Figure 2-6. CRP EpiScore in the LBC1936 population by age and inter-wave stability.

A, Trajectories of serum CRP and the CRP EpiScore over age in LBC1936. Grey lines show individual participant trajectories. Regression lines are plotted for females (blue) and males (purple). B, Inter-wave Pearson correlations for the CRP EpiScore score and serum CRP in the Lothian Birth Cohort 1936. Square brackets display 95% CIs. CRP: C-reactive protein; W: Wave. Adapted from Stevenson et al. (2020). Copyright: Clinical Epigenetics.
The second study by Stevenson et al. (2021) used elastic net penalised regression to train and test an EpiScore for IL6. I was involved in generation of the weights for the 35 CpG sites selected in the training set (875 individuals from LBC1936) that formed the basis of this EpiScore. In 7,028 individuals in GS, the EpiScore explained 4.4% of the variance in measured IL6. Both measured IL6 (417 individuals) and EpiScore IL6 (7,028 individuals) had associations with age (IL6 had Beta=0.02, SE =0.004, P =1.3×10^{-7} and EpiScore IL6 had Beta =0.02, SE=0.0009, P < 2×10^{-16}). Similarly to the CRP EpiScore findings, the IL6 EpiScore was associated with cognitive ability (n =7,028, Beta= −0.16, SE = 0.02, FDR P < 2×10^{-16}), whereas the measured IL6 was not (n =417, Beta= −0.06, SE = 0.05, P = 0.19).

The fourth protein EpiScore study is from Wielscher et al. (2021). This study ran an updated MWAS of CRP with the largest sample size to date (see Table 2-1 and Section 2.2.3 for details). The weighted linear combination of CpG effects from the MWAS was used to generate a CRP EpiScore, which was found to associate with other inflammatory markers (e.g. IL6), BMI, lung function, COPD and cardiometabolic traits. The increased relative risk associated with a 1% increase in the CRP EpiScore was found to vary (4.3% coronary artery disease, 2.9% for myocardial infarction, 1.7% for T2D, 1.007% for COPD and 0.2% for hypertension).

Finally, as part of a wider EpiScore assessment across linkage-derived traits and clinical measures, EpiScores were developed for blood assay biomarkers (e.g. urea, immune cell counts, HbA1c and Ferritin) by Thompson et al. (2022). However, these were developed in a relatively small population (n=831), were not tested in external populations and had limited performance (HbA1c was among the best-performing protein EpiScores but still had an R² value below 0.4%).
Table 2-4. Summary of studies generating EpiScores for blood protein levels identified by the literature search.

<table>
<thead>
<tr>
<th>Author (date)</th>
<th>Scores</th>
<th>Training cohort, N, Mean age, % female</th>
<th>Testing cohort, N, Mean age, % female</th>
<th>Key results</th>
<th>DNAm array</th>
<th>Method</th>
<th>Proteomics</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barker et al, 2018</td>
<td>CRP (i-ePGS)</td>
<td>EWAS Ligthart et al, 2020 – N=8863 from 11 cohorts, mean age ranged from 41-87, 54% female across studies</td>
<td>ARIES cohort: Ages 15 and 17 tested, 50% female, then projected to DNAm at ages 7 and birth.</td>
<td>CRP score (i-ePGS) was associated with CRP at age 15 ($r = 0.11$, $p &lt; 0.002$) and age 17 ($r = 0.181$, $p &lt; 0.0001$). Latent models suggested the CRP score was associated with poorer cognitive health in childhood to adolescence.</td>
<td>450K sites</td>
<td>Regression weights from EWAS</td>
<td>Training population: high-sensitivity assay</td>
<td>214</td>
</tr>
<tr>
<td>Lu et al, 2019</td>
<td>ADM, B2M, Cystatin C, Leptin, TIMP-1, PAI-1, GDF15</td>
<td>FHS, N=1731, 66, 54%</td>
<td>FHS, N=625, 67, 53%</td>
<td>Of 88 plasma proteins, 12 surrogates had $r &gt; 0.35$ (equivalent to $R^2$ 12%). Seven selected for GrimAge.</td>
<td>450K sites</td>
<td>Glmnet, 10cv, alpha = 0.5</td>
<td>Plasma: immunoasay measurements (pg/mL)</td>
<td>129</td>
</tr>
<tr>
<td>Stevens on et al, 2020</td>
<td>CRP</td>
<td>EWAS Ligthart et al, 2020 – N=8863 from 11 cohorts, mean age ranged from 41-87, 54%</td>
<td>LBC1936 Wave1, N=889, 69.5, 49% and GS, N=7,028, 50.9, 58%</td>
<td>DNAm CRP test set LBC1936 $r = 0.34$ (equivalent to $R^2$ 11%), GS $r = 0.28$ (equivalent to $R^2$ 7%). CRP score increased consistently with age and had higher</td>
<td>450K sites</td>
<td>Regression weights from EWAS</td>
<td>Serum: high-sensitivity assays (all studies except 215</td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Marker</td>
<td>Dataset</td>
<td>Sample Size</td>
<td>Methodology</td>
<td>Results</td>
<td>Notes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------</td>
<td>--------</td>
<td>---------</td>
<td>-------------</td>
<td>-------------</td>
<td>---------</td>
<td>-----------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stevens et al, 2021</td>
<td>IL6</td>
<td>LBC1936</td>
<td>Wave 1, N=875, 70, 49%</td>
<td>GS, N=7,028 (417 with measured IL6), 51, 58%</td>
<td>IL6 test set $R^2 = 4.4%$. Both DNAm and protein IL6 increased with age. DNAm IL6 had an inverse association with cognitive functioning, but not cognitive ability.</td>
<td>450K sites Glmnet 10cv, alpha = 0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wielscher et al, 2021</td>
<td>CRP</td>
<td>N=22,774 from 30 studies of diverse ancestries, mean age ranged from 16 to 76 across cohorts, 43% female across studies</td>
<td>Meta-analysis across AIRWAVE, KORA, FHS, SHIP, and YFS. No population demographics reported for specific subsets of cohorts used.</td>
<td>CRP score associated with a range of health outcomes, diseases and inflammatory proteins. No $R^2$ reported.</td>
<td>450K or EPIC across EWAS studies Regressions from EWAS Serum: assays used not stated for every cohort.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Measures</td>
<td>N</td>
<td>Subset Definitions</td>
<td>Performance</td>
<td>Filtered Sites</td>
<td>Protein Measures Derived From Clinical Linkage Tests</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------</td>
<td>---------------------------------</td>
<td>-----------</td>
<td>--------------------</td>
<td>-------------</td>
<td>----------------</td>
<td>-----------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thomson et al, 2022</td>
<td>Total protein, albumin, WBCs, ferritin, haemoglobin + other markers</td>
<td>831</td>
<td>train and test</td>
<td>Poor (0.1-0.6% $R^2$)</td>
<td>269,471 sites</td>
<td>From clinical linkage tests</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.3.5 Rationale for generating Protein EpiScores

The studies presented in Section 2.3.4 highlight that protein EpiScores may be useful tools for health stratification; however, prior to my results in Chapters 8 and 9, few proteins had been assessed. Additionally, protein EpiScores have not been trained using multiplexed protein assay measurements. In this section, I outline the specific rationale for generating additional protein EpiScores, drawing on the findings presented in the previous section. The six arguments that build the rationale for generating more protein EpiScores that are discussed in this section are summarised in Figure 2-7.

First, genetic scores (such as those developed by OmicsPred) \(^{189}\) neglect the non-genetic component of protein variability. As DNAm is reflective of a range of non-genetic factors that associate with risk of disease, it is well-positioned to capture the dynamic state of human health across the life course. This is evidenced by the GrimAge acceleration score \(^{129,206}\), which is able to track biological age acceleration. Additionally, a close interplay between inflammation and DNAm has been documented previously \(^{177}\). DNAm and proteins are implicated in processes such as inflammation, which means that they are ideal candidates to profile risk factor dependent changes in inflammatory signatures in the body. In contrast, genetic risk does not vary based on risk factor exposures.

Second, in Section 1.3.9 I described that while not extensively characterised across the proteome, there are examples indicating that test-retest reliability varies by protein. I also highlighted examples that indicate inflammatory proteins can exhibit variability across multiple time point measures. This is understandable given the rapid regulation of the inflammatory system that can occur based on exposure to pathogens and other triggers. In Section 2.2.3 I detailed that previous MWAS of proteins have found that there is a strong interconnectedness between inflammation, the methylome and the proteome. These studies suggest that DNAm signatures may be suited to proxying for inflammatory proteins. As DNAm is thought to record a more chronic record of exposures on the body, protein EpiScores may therefore be reflective of chronic, low-grade inflammatory exposure on the body and may be more stable than the measured protein. This question has not been comprehensively answered. However, the findings of work done by Stevenson et al. (2020 and 2021) \(^{215,216}\) that were discussed in
Section 2.3.4 suggest that EpiScores for CRP and IL6 capture protein-specific, inflammatory signals. The CRP EpiScore was also found to have greater test-retest stability than measured CRP, indicating that it may offer a more stable index of chronic inflammation across time points than measured CRP.

Third, there is evidence to suggest that protein EpiScores may have stronger phenotypic associations than the measured protein equivalents. In both studies from Stevenson et al (2020 and 2021) \(^{215,216}\), the CRP and IL6 EpiScores had stronger associations with cognitive ability than the protein counterparts. In the GrimAge study, the authors noted that individual associations with time-to-mortality in the test sample were stronger for four of the seven protein EpiScores versus measured proteins (Table 2-3). In recent work, the EpiScore for CRP generated by Stevenson et al. (2020) \(^{215}\) had on average 6.4-fold stronger associations with brain imaging phenotypes than serum CRP \(^{65}\). Taken together, these studies suggest that protein EpiScores may have stronger associations with endophenotypic measures (e.g. cognitive ability and MRI measures) than measured proteins. This may be due to their more stable and clearer inflammatory signal versus the measured protein. Thus, they may enhance the health-informative signal that is available from the blood.

Fourth, proteomic technologies are costly to implement and not presently available across all cohort studies. An EpiScore that can proxy for a key biomarker of disease risk and brain health such as CRP would allow any cohort with Illumina based DNAm to project this measure into the sample to study it across a range of cohort-specific health outcomes. Even in cohorts with protein data measured, often panel coverage is not complete. For example, Mspec typically measures the most abundant proteins present in blood (as described in Section 2.1). Therefore, training EpiScores for less abundant proteins that may still be highly relevant for disease risk stratification would enable proxies for these to be used in cohorts without the specific protein measured as part of the assays used.

Another reason for the development of protein EpiScores is that these scores may offer a different but complimentary signature to the original protein to augment risk stratification. As the GrimAge study demonstrated, not all of the 88 immunoassay proteins scanned had sufficiently detectable signal to develop EpiScores. Instead of pitching direct use of proteins against protein EpiScores, both biomarker measures
can be considered as useful tools for risk stratification approaches. Which of the measures is optimal will likely vary by protein and by disease or outcome studied.

Finally, if we take a view beyond single protein-disease biomarker relationships, the value of DNAm scores becomes more compounded. A portfolio of methylation scores (trained on lifestyle, disease, protein outcomes) can be aggregated together and projected simultaneously into a population with DNAm. This provides a holistic view of the health state, potential lifestyle behaviours, ageing rates and molecular profile (as indexed by protein EpiScores) for an individual. However, generation of an extensive range of epigenetic scores for protein levels has not been attempted to date. Profiling of EpiScores for protein levels as biomarkers for disease risk stratification has also not been quantified extensively. In the following section, I outline approaches that can be used to model protein and protein EpiScore markers for disease risk stratification. I also detail additional methodological considerations that are relevant to the work done in the empirical elements of this thesis.
Figure 2-7. Motivation for generating additional protein EpiScores.

Summary of six arguments that are central to the motivation for generating additional protein EpiScores. The workflow that is typically used to train, test and apply protein EpiScores is highlighted in the centre of the figure. Weighting coefficients are identified for CpGs in the training population, which are used to project the score into a test population. After validation of the protein EpiScore signal (against measured protein) in the test sample, it can be applied to any population that has DNAm measures at CpGs available. Figure created using Biorender.com.
2.4 Summary

In this chapter, I detailed statistical approaches that can be used to link DNAm with proteomic data. I described MWAS analyses and conducted a literature review on studies that have used MWAS to identify DNAm patterns associated with circulating protein levels. I provided examples whereby these techniques have identified insights that may be relevant to brain health. I then detailed statistical approaches that can be used to develop EpiScores. I described the rationale for linking DNAm and protein information through EpiScore methods and gave a literature overview of studies that have developed protein EpiScores to date, with implications for brain health.

As proteins are effectors of disease and typically the targets for therapeutic interventions, they are often the primary biomarkers for diseases. Taken together with the arguments outlined in Section 2.3.5, I hypothesise that protein EpiScores could be valuable tools for biomarker discovery and risk stratification. In the next chapter, I give a brief overview of approaches that can assess biomarkers (i.e. both measured proteins and protein EpiScores) for the stratification of incident outcomes. I also detail causal inference techniques that can be used to delineate causal versus consequential protein biomarkers, in addition to any remaining methodological considerations.
3 Protein-informed evaluation of brain health

In this Chapter, I outline how blood-based omics markers can be identified in Cox PH associations with incident diseases. This approach can be used to model both proteins and protein EpiScores. I then outline predictive scores that use omics features other than DNAm to stratify disease risk. Specifically, I introduce the concept of a ProteinScore – reliant on multiple protein features for risk stratification. I introduce Mendelian randomisation and colocalisation techniques as avenues to test for potential protein mediators of diseases. I then briefly note the importance of multiple testing correction, which is a key consideration for the empirical work in this thesis.

3.1 Evaluating biomarkers

In this section, I cover the rationale for early blood-based biomarker studies, introducing the Cox PH approach. I discuss how this is relevant for brain health and disease risk stratification. I introduce composite scoring approaches that rely on omics features that are not DNAm-based, in contrast to the EpiScores presented in Chapter 2. I then discuss methodological considerations for Cox PH analyses.

3.1.1 Rationale for early blood biomarker detection

Studies have shown that blood-based proteins such as p-tau217, p-tau181 and p-tau231 can differentiate Alzheimer’s dementia cases from controls and individuals with diagnoses of other neurodegenerative diseases \(^{218}\). Identifying blood markers in individuals with pre-existing disease diagnoses can be useful for diagnostic purposes. However, these markers may not be useful in the context of early identification of disease risk. Identifying individuals at the highest risk of developing a future disease event is a core goal of preventative medicine \(^{48,49}\). Advances in electronic health data linkage to GP and hospital records has facilitated the tracking of disease diagnoses in consenting individuals \(^{219}\). This means that for individuals that had a blood test as part of a cohort study, the diseases they went on to get post-baseline blood sampling (incident diseases) can be tracked through coding schemes \(^{220,221}\). Blood-based omics measurements from the original blood sample can provide a snapshot of the individual’s health state while they were alive and did not have overt symptoms. This is pertinent in the study of neurodegenerative diseases, as where damage is thought
to occur many years (in the case of Alzheimer’s dementia 2-3 decades) prior to symptom onset. Identifying those who may have underlying disease pathology through omics signatures in the blood may help pave the way for personalised, preventative targeting of interventions. Signatures and pathologies of diseases may be different in the earliest stages of progression versus the later stages where symptoms prompt a visit to healthcare provider. The integration of blood sampling in individuals who have not been diagnosed with diseases is therefore essential to fill in this often rarely-studied piece of the disease pathology timeline.

3.1.2 Cox PH analyses

The Cox proportional hazards (PH) model – first developed by David Cox in 1972 – is the basis for multivariable survival analyses. Cox PH analyses are useful when the goal is to identify the early risk factors or molecular signatures that associate with future disease events (incident disease). Therefore, cohort studies that have cross-sectional omics sampling, combined with electronic data linkage across multiple years of follow-up are well-positioned to identify early predictive biomarkers through Cox PH. Unlike linear and logistic regression, survival analyses model the association between a variable and the time taken for an event to occur. Typically studied outcomes include time to disease, relapse and progression and ultimately, death. In survival analyses, control individuals have not experienced the event during the study follow-up period. This is called censoring and occurs due to 1) the individual experiencing an event that makes the target event impossible (i.e. death), 2) the individual being lost-to-follow-up, 3) or the individual simply not having experienced the event yet. Right-censoring occurs when events happen during the observation window and is commonly-used in disease-onset modelling. The censoring window is typically set to the follow-up period whereby all disease linkage information is known, to ensure controls can be reliably classified as disease-free within the study period. Cox PH analyses typically map associations between predictive variables and time-to-event individually. The Cox PH elastic net penalised regression method can consider multiple predictive features simultaneously in relation to a time-to-event outcome. This was the approach used to develop the gold-standard epigenetic clock measure for biological age acceleration – GrimAge (see Section 2.3.4). Predictive scores can be trained and projected into a test population using this approach with a variety of omics input features.
3.1.3 Cox PH assumptions

There are some core assumptions that are made in the Cox PH model. First, is the independence of time-to-event, such that an individual’s time-to-event status is not dependent on the time-to-event of another individual in the population. Second, is that censoring is uninformative on the risk of developing the event of interest. Third, is the proportional hazards assumption that hazards are proportional and remain constant over time for individuals in the population. This assumption is central to the interpretation of Cox PH analyses and can be checked by extracting residuals – the most common is the Schoenfeld residual (the observed – expected values of covariates over each event time), which should be independent of time. In instances whereby violations occur, stratification across time-variable covariate bands and examination of the association over successive years of follow-up can be employed to understand variability in associations by time. Checking the assumption violations in regression models is critical as it helps to resolve the reliability and robustness of the associations, while also informing further sensitivity analyses. In all survival analyses, the outcome definition is critical.

3.1.4 Non-methylation omics composite scores

While genetic and methylation scores have become more prevalent over the past decade, other omics features can be used to derive scores. Multiplexed protein and metabolomics assays have become available in studies that are large enough to facilitate training and testing of scores using these features. In the UKB cohort, 249 metabolites are available in ~120,000 individuals. A recent study from Buergel et al. (2022) utilised these measures in combination with electronic health linkage to ICD9 and ICD10 codes within the cohort to create predictive scores for 24 incident outcomes. Protein measures have also recently been used as features for the predictive scoring of nine lifestyle and health traits, in addition to incident CVD and diabetes incident outcomes in 16,894 individuals with SomaScan protein measurements. These scores derived using proteomic features are be referred to as ‘ProteinScores’ in this thesis. The extension of predictive scoring methods to additional omics types will likely continue to expand, in-line with growing omics availability in large populations. I would hypothesise that the value of certain omics types will be highly disease-specific, as is evidenced from the results of the Beurgel et al. (2022) study.
that found varied performance of the metabolomic scores in the UKB test sample (Figure 3-1).

**Figure 3-1. Value of metabolomic risk scores is endpoint dependent.**

Discriminative performance of Cox PH models assessing scores trained on metabolomic features (MET) against three sets of variables (Age and Sex, the ASCVD risk score and a set of 34 clinically-relevant covariates termed PANEL). The sets of variables are also modelled combined with the MET score. Horizontal dashed lines indicate the median performance of the three sets of variables studied. ASCVD: atherosclerotic cardiovascular disease. Taken from Beurgel et al. (2022) 227. Copyright: Nature Medicine.

3.2 Methodological considerations

3.2.1 Causal inference

The biomarker signatures identified through individual Cox PH models or collections of features contributing to a Cox PH elastic net score are useful indicators for risk
stratification. Prevention approaches can use these insights to identify individuals that may benefit from interventions\textsuperscript{46,229}, but it is unlikely that all markers will be causal for a disease. Understanding the most likely molecular drivers of a disease can help to develop therapeutic interventions that may be disease-modifying\textsuperscript{230}. For example, blood and CSF levels of NEFL are markers of multiple neurodegenerative diseases\textsuperscript{231,232}. While highly useful for risk stratification and early identification of potential damage in the brain, the presence of this synaptic-associated protein in the blood may be a result of synapse loss. Mendelian randomisation is a technique that can be used to assess whether a protein associated with a disease may represent a causal mediator, or consequential marker\textsuperscript{233}. Colocalisation tests whether the same underlying variant is likely to affect two traits and can be used to triangulate evidence alongside MR\textsuperscript{234}. While the aims of this thesis are primarily centred on the risk stratification component of prevention in neurological disease, causal inference analyses is used in Chapter 6. These inference approaches are also complementary to the risk stratification approaches and are therefore summarised briefly for context.

### 3.2.2 Approaches to causal inference

Beyond pQTLs and pQTMs that were introduced in Section 1.3.8, genetic variants that regulate other molecular elements such as gene expression/transcript levels (eQTLs), messenger RNA splicing (sQTLs), DNAm (mQTLs) and histone activity (hQTLs) have also been identified. Mendelian randomisation uses these genetic variants (QTLs) as instrumental variables (IVs) to proxy for the effect of an exposure on an outcome. This can be thought of as analogous to a randomised control trial, as the inheritance of genetic variants that affect the exposure variable is randomised\textsuperscript{235} (Figure 3-2). While QTLs associated with complex traits can be used to test for evidence of causal relationships (i.e. BMI > heart disease), these are more unreliable than using QTLs associated with molecular factors. Proteins are well-suited as exposures, ascis-pQTLs are likely to directly regulate the protein. pQTLs are typically filtered to those acting in cis-effects on the protein (within the same genomic region), rather than trans-effects (on distal chromosomes, or beyond a defined genomic region proximal to the protein coding gene)\textsuperscript{145}. This approach using cis-pQTLs is more likely to satisfy the underlying assumptions of Mendelian randomisation. The three core assumptions are: 1) the instrument associates with the exposure, 2) the instrument
does not influence the outcome through another pathway that is not through the exposure and 3) the instrument does not associate with confounders (Figure 3-2). Violations in assumption (2) are known as horizontal pleiotropy (SNP > non-exposure pathway > outcome), which undermines the vertical pleiotropy that is tested in MR (SNP > exposure pathway > outcome). Software such as MR-Base allows for two-sample, bidirectional MR to be run. This can assess both the causal directionality of relationships from proteins to traits and vice versa.

In the field of neuroscience, Mendelian randomisation can help to understand whether a given protein is likely to be present as a marker in the blood as a consequence of pathology in the brain (such as breakdown of neurons and leakage of the blood-brain barrier) or a causal mediator in the disease. For example, a recent study scanned 1,827 immune and blood-brain barrier related protein biomarkers and identified 127 possible causal relationships with Alzheimer’s dementia that often implicated immune, amyloid and tau pathways.

**Figure 3-2. Mendelian randomisation principles and assumptions.**

The top panel shows the randomised control construct. The bottom panel shows how genetic variants associated with proteins can be used in Mendelian randomisation,
analogous to a randomised control trial. The red arrows show the causal effect being modelled. The assumptions of Mendelian randomisation are annotated in red circles: 1) the instrument associates with the exposure, 2) the instrument does not influence the outcome through another pathway that is not through the exposure and 3) the instrument does not associate with confounders. Adapted from Sanderson et al. (2022) \textsuperscript{235}. Copyright: Nature Reviews Methods Primers.

3.2.3 Colocalisation

Colocalisation relies on the central hypothesis that if a variant is associated with multiple related traits, the likelihood of it representing a causal locus is greater \textsuperscript{238,239}. Colocalisation can be used to quantify the likelihood that SNPs are causal for two traits; for example, whether an eQTL is also a pQTL, or whether a pQTL is also a QTL for a disease. Colocalisation can therefore be used to provide triangulation of Mendelian randomisation results. One widely-adopted approach to colocalisation is the coloc R package \textsuperscript{234,240}, which employs a Bayesian sampling methodology to test the following hypotheses:

- H0, no associations with either trait;
- H1, association with trait 1 only;
- H2, association with trait 2 only;
- H3, both traits have associations with separate causal variants;
- H4, both traits have associations at a shared causal variant.

A posterior probability (PP) near to 1 implies evidence supporting a specific hypothesis. Figure 3-3 summarises situational examples whereby colocalisation can be used to resolve variant signals for two traits \textsuperscript{241}. An example of colocalisation in the context of protein levels can be seen in work done as part of the MWAS and GWAS of circulating neurological protein levels by Hillary et al. (2019) \textsuperscript{146}. From the GWAS, a conditionally significant sentinel cis-pQTL was identified for plasma poliovirus receptor (PVR) levels. As PVR had been linked to Alzheimer’s dementia previously \textsuperscript{242}, the authors initially tested the relationship using Mendelian randomisation analyses and found evidence suggesting that PVR levels may be causally associated with Alzheimer’s dementia (Beta = 0.17, SE = 0.02, \( P = 5.2 \times 10^{-10} \) via a Wald ratio test). It was not possible to test for horizontal pleiotropy (assumption 2 of MR) owing to only
having one sentinel SNP available. The authors performed colocalisation to test for evidence that the same SNP variant may drive both PVR levels and Alzheimer’s dementia. This involved extraction of the 200kb region either side of the sentinel SNP, across GWAS summary statistics for PVR and Alzheimer’s dementia. The colocalisation test found that there was evidence suggesting that there were two distinct causal variants for the levels of the protein and Alzheimer’s dementia risk in the region (i.e. H3 had a PP of > 0.99). This is an example showing that colocalisation can be helpful in providing triangulation of MR evidence and assessment of the possibility of pleiotropy. The colocalisation did not support the argument for a single causal SNP for both PVR and Alzheimer’s dementia, but did indicate that the PVR gene region may be associated with development of the disease.

Figure 3-3. Causal variant situations that are tested in colocalisation approaches.

A, Two traits with causal variants in linkage disequilibrium. B, Two unrelated traits that have one shared causal variant. C, Two traits with shared causal variant, with the first trait influencing the second trait. D and E, one shared causal variant and another non-causal variant for two traits. B and C show colocalisation. D and E show colocalisation of a shared variant, but the presence of a distinct variant. Colocalisation does not
distinguish between situations where trait 1 and trait 2 are causally unrelated (B and D), versus those where trait 1 has a causal effect on trait 2 (C and E). Example regional association plots for each scenario show negative log_{10} p-value representations for the associations of variants by chromosomal location with each trait (blue = trait 1, red = trait 2). Taken from Zuber et al. (2022) \textsuperscript{241}. Copyright: AJHG.

3.2.4 Multiple testing considerations

A P-value represents the chance of observing the data or something more extreme under the null hypothesis of no association. Given that analyses can test thousands of associations, multiple testing correction is required \textsuperscript{243}. In GWAS, a guidance threshold of P < 5x10^{-8} has historically been used based on permutation analyses to adjust for multiple testing correction across SNPs \textsuperscript{244}. In MWAS, permutation analyses across five cohorts using the Illumina 450K array suggested that a genome-wide threshold of P < 3.6x10^{-8} may be appropriate \textsuperscript{245}. These thresholds have been applied to GWAS and MWAS models of single traits; however, when analyses are run for multiple outcomes – such as for multiplexed protein measures – further correction is required. For example, a study conducting MWAS of 1000 proteins may consider using a Bonferroni adjustment, with P < 0.05 / (number of outcomes × number of CpG sites tested). If the outcomes have a high degree of correlation then principal components analyses can be used to identify the number of independent signals that explain a majority of the variance in the outcomes. Another method used to adjust for multiple testing is the False Discovery Rate (FDR) approach proposed by Benjamini-Hochberg \textsuperscript{246}. This quantifies the likelihood of an incorrect rejection of a hypothesis occurring, through calculation of the expected proportion of false positives in the total number of rejected hypotheses. This method is considered less stringent than a Bonferroni adjustment.

3.3 Summary

In Section 3.1 I gave an overview of the rationale for identifying early, blood-based markers of incident disease events. The Cox PH approach discussed can be integrated to gain deeper insight into the biomarkers associated with brain health (Figure 3-4). In this thesis, Cox PH approaches are employed to model individual
biomarkers (proteins and protein EpiScores), in addition to multi-protein scores. The outcome in these models is time-to-event. In Section 3.2, I then described genetic inference methods that can be used to understand whether a protein is likely to be a driver of a neurological disease. While the work in this thesis is focused primarily on risk stratification markers (that can be causal and consequential), the genetic inference techniques can be applied to protein markers identified in Cox PH associations. Chapter 6 employs these genetic inference methods in the context of a neuroinflammatory protein marker and Alzheimer’s dementia. In detailing the approaches in this chapter, I also presented methodological considerations that are relevant to this thesis.
Figure 3-4. Motivation for biomarker approaches in brain health.

Longitudinal linkage can be integrated in Cox PH approaches to model omics signatures that associate with elevated incidence of disease onset in the years after baseline blood draw. Predictive scoring approaches can take omics features into consideration, deriving intermediary scores (i.e. between omics with proteomics as the outcome) that act as biomarkers, or generating scores directly for outcomes (i.e. between omics and a phenotype or disease outcome).

In the following chapter, I describe the five cohort studies used in the analyses that forms Chapters 6-10 (the Lothian Birth Cohorts 1936 and 1921, Generation Scotland, KORA and the UK Biobank cohorts).
# 4 Study cohorts and molecular samples

In this chapter, I outline the five study populations used in this thesis: Generation Scotland (GS), the Lothian Birth Cohorts of 1921 and 1936 (LBC1921 and LBC1936), the Cooperative Health Research in the Region of Augsburg (KORA) cohort and the UK Biobank (UKB). I summarise the measurement of relevant omics markers (genetic, DNAm, proteomic) within the sample populations. I also briefly detail the study populations used to generate the mQTL and eQTL datasets that I make use of in colocalisation and functional assessments in Chapters 6 and 7.

## 4.1 Generation Scotland

Generation Scotland: the Scottish Family Health Study (GS) is an extensive, family-structured, population-based cohort study of > 24,000 individuals from Scotland 247. Recruitment was conducted between 2006 and 2011 and mean age at baseline visit was 48 years (59% female), with family members invited to participate. Eligibility criteria included those aged between 35 and 65 years that had at least one sibling group and one first-degree relative who was aged 18 years old or above. There were 23,960 individuals who were recruited at baseline, with a total of 5,573 families that had a mean of 4 individuals for each family. The clinical visit involved collection of detailed health, cognitive, and lifestyle information, in addition to the collection of biological samples (blood, urine, saliva). A summary of the population recruited at baseline and the full cohort inclusion criteria is provided by Smith et al. (2006) 248 and Smith et al. (2013) 247. More than 98% of GS volunteers have now consented for researchers to access their electronic health records (EHRs) via data linkage to GP records (Read 2 codes) and hospital records via the International Classification of Diseases (ICD) codes. These data are available retrospectively and prospectively from the original time of blood draw, with up to 18 years of linkage post-baseline as of 2023.

The Stratifying Resilience and Depression Longitudinally (STRADL) cohort is a subset from the GS cohort that were re-contacted for an assessment primarily focused on the study of depression. Of the 21,525 individuals from the original GS sample that indicated their consent to be re-contacted for participation, 9,618 individuals responded. The cohort included 7,158 individuals, comprised of 2,460 families and 2,460 individuals that were unrelated. The protocol for the study with the full inclusion
criteria and baseline demographic information in STRADL is provided by Navrady et al. (2018)\(^{249}\). The baseline sample had 52% females and mean age 53 years. Of the individuals sampled, there were 1,188 that completed additional health assessments and gave a blood test used for sequencing of omics measures (including proteomic, DNAm and genetic measures). This sample was provided ~5 years after the GS study baseline\(^{249}\).

4.1.1 Ethics and funding

All aspects of GS and STRADL were awarded ethical approval from the National Health Service Scotland Tayside Committee on Medical Research Ethics (05/S1401/89 and 14/SS/0039). GS is supported by the Chief Scientist Office of the Scottish Government Health Directorates (CZD/16/6) and the Scottish Funding Council (HR03006). Research Tissue Bank status was awarded by the Tayside Committee on Medical Research Ethics (20-ES-0021). STRADL is funded by a Wellcome Strategic Award (104036/Z/14/Z). The Medical Research Council and the Wellcome Strategic Award (104036/Z/14/Z) funded genotyping in GS. The Medical Research Council and Wellcome supported DNAm typing (104036/Z/14/Z) in GS, in addition to the Brain and Behaviour Research Foundation (27404). Proteomic sampling in STRADL was funded by the Medical Research Council (MR/L023784/2) and Dementias Platform UK. All participants provided informed consent for their health data to be used for the purposes of biomedical research.

4.1.2 Genetic data

DNA was measured via the Invitrogen PicoGreen kit assay. A total of four microliters of sample was diluted to 50 nanograms for every microliter for the genotyping\(^{250}\). A first set of 9,863 samples were quantified using Illumina HumanOmniExpressExome-8 v1.0 BeadChips, with the remaining samples quantified via v1.2 of the BeadChip. PLINK v1.9b2\(^{251,252}\) was used to run quality controls, excluding of SNPs with a call rate < 0.98, minor allele frequency (MAF) ≤ 0.01 and Hardy-Weinberg equilibrium (HWE) test result of P ≤ 1 x 10\(^{-6}\). A total of 561,125 autosomal SNPs passed quality control. PCA of genotype data was performed to identify and remove any outliers more than 6 SDs from the mean of the first two principal components. There were 19,904 individuals with genotyped data (11,731 females and 8,173 males). Phasing of the genotyped SNPs was done via Segmented HAPlotype Estimation and Imputation Tool 2 (SHAPEIT2)\(^{253}\), with imputation using the Haplotype Reference Consortium.
reference panel (HRC.r1-1). Duplicate and monomorphic SNPs and SNPs with an imputation quality score of < 0.4 were excluded. A total of 24,161,581 SNPs were available for analyses.

4.1.3 DNA methylation data

There were 9,537 individuals in GS that had DNAm information available in 2019, which has since expanded in 2022 to DNAm sequencing in 18,413 individuals. Blood-based DNAm was generated in three sets using the Illumina EPIC array. DNAm data were generated in three sets, with 5,087 (Set 1), 4,450 (Set 2) and 8,877 (Set 3) individuals. Processing of these independent sets was done in 2017, 2019 and 2021, respectively. Sets 1 and 3 included related individuals within each set. All individuals in Set 2 were unrelated to each other and to individuals in Set 1 (using a threshold of < 0.05 from a genetic relationship matrix (GRM)). During quality control, CpG probes were filtered by removing those with low bead count (of <3) in ≥5% of samples or a high detection p-value (>0.05) in more than 5% of samples. Samples with a mismatch between predicted and recorded sex or ≥ 1% of CpGs with detection p-value > 0.05 were also removed in addition to saliva samples and genetic outliers. Non-specific, cross-hybridising and single nucleotide polymorphism (SNP) associated probes (with allele frequency >5%) and probes on the X and Y chromosomes were removed.

Measurements of blood DNAm in the STRADL subset of GS were processed in two sets on the Illumina EPIC array using the same methodology and quality control steps as those from the wider GS cohort. After processing, a total of 793,706 and 773,860 DNAm measurements at CpGs were available in sets 1 and 2, respectively. Set 1 included 504 individuals, while Set 2 included 306 individuals. These sets were truncated to include a total of 772,619 common probes. Of the 810 individuals available across the two sets, 778 had SomaScan protein data available for 4,235 protein analytes (56% female, mean age 66 years and SD of 2.8).

4.1.4 Proteomics data

SOMAscan® (version 4) technology was used to quantify plasma protein levels in STRADL. The aptamer-based assay (as described in Section 2.1.2) facilitates the simultaneous measurement of SOMAmers. Binding between plasma samples and
target SOMAmers occurred in incubation steps. Quantification was recorded from a fluorescent signal on microarrays. Quality control steps included hybridisation normalisation, signal calibration and median signal normalisation to adjust for inter-plate variation. The finalised dataset had 4,235 SOMAmer epitope measures that were available in 1,065 individuals. These corresponded to 4058 unique proteins.

In GS serum biomarker measurements of GDF15 and NT-proBNP were measured at a single (second) thaw of the stored sample aliquots. GDF-15 and NT-proBNP levels were measured via an immunofluorescence assay performed with a cobas e411 analyser (Roche Diagnostics, Basel, Switzerland) with the manufacturer’s advised reagents and quality controls.

4.2 The Lothian Birth Cohorts 1936 and 1921

The Lothian Birth Cohorts of 1921 (baseline LBC1921; N = 550) and 1936 (baseline LBC1936; N = 1091) are longitudinal studies of aging in individuals who reside in Scotland. Participants completed an intelligence test at age 11 years and were subsequently recruited by these cohorts at mean ages of 79 (LBC1921) and 70 (LBC1936). They have been followed up triennially for a series of cognitive, clinical, physical, and social data collected at the Wellcome Clinical Research Facility, Western General Hospital, Edinburgh. Blood samples have also been collected and used for genetic, epigenetic, and proteomic measurement. LBC1921 have been assessed five times, from ages 79 to 92 years. Five waves of assessment have been collected for LBC1936 (mean ages 70, 73, 76, 79 and 82 years). The sample size per wave is 1,091, 866, 697, 550 and 440. Sample collection for LBC1921 began in 1999 (Wave 1) and spanned to 2013 (Wave 5). Full details of the original study protocols can be found in work published by Deary et al. (2012). A more recent update detailing the additional waves of data collection is also available from Taylor et al. (2018).

4.2.1 Ethics and funding

Ethical approval for the LBC1921 and LBC1936 studies was obtained from the Lothian Research Ethics committee (LREC/1998/4/183; LREC/2003/2/29) and Multi-Centre Research Ethics Committee for Scotland (MREC/01/0/56). In both studies, all participants gave written informed consent for their data to be used for health research.
LBC1921 was funded by the UK’s Biotechnology and Biological Sciences Research Council (BBSRC) and the Chief Scientist Office (CSO) of the Scottish Government’s Health Directorates. LBC1936 is also funded by Age UK (Disconnected Mind project), and the Medical Research Council (G0701120, G1001245, MR/M013111/1, MR/R024065/1). Methylation typing in LBC1921 and LBC1936 populations was supported by the Centre for Cognitive Ageing and Cognitive Epidemiology (Pilot Fund award), The University of Edinburgh, The University of Queensland, Age UK and The Wellcome Trust Institutional Strategic Support Fund. Genotyping in LBC1921 and LBC1936 was funded by the Biotechnology and Biological Sciences Research Council (BB/F019394/1). Proteomic analyses in the cohorts were supported by the Age UK grant and the NIH Grants R01AG054628 and R01AG05462802S1.

4.2.2 Genetic data

DNA samples in LBC1921 and LBC1936 were genotyped at the Edinburgh Clinical Research Facility using the Illumina 610-Quadv1 array (San Diego). SNPs were imputed to the 1000 G reference panel (phase 1, version 3). Individuals were excluded based on whether there was a presence of sex mismatches, relatedness, SNPs with call rates of less than 0.95 and evidence of non-European ancestry. SNPs with a call rate of greater than 0.98, minor allele frequency greater than 0.01 and HEW test of $P \geq 0.001$ were retained in the analyses. SNPs with MAF > 0.05 and imputation quality > 0.6 were kept for analyses. The sample of 8,489,963 SNPs were subset to remove any with minor allele count < 25, leaving 7,307,523 SNPs.

4.2.3 DNA methylation data

DNA from whole blood was assessed in LBC1921 and LBC1936 using the Illumina 450 K methylation array at the Edinburgh Clinical Research Facility. Details of quality control have been described by Shah et al. (2014) \cite{263}. Raw intensity data were background-corrected and normalised via internal controls and methylation beta values were generated using the R package minfi \cite{264}. Manual inspection resulted in removal of low-quality samples that presented issues related to bisulphite conversion, staining signal, inadequate hybridisation, or nucleotide extension. Quality control was performed to exclude probes with low detection rates of < 95% at $P < 0.01$ and samples with low call rates (with < 450,000 probes detected at $P < 0.01$) were also
removed. Samples were removed if they had a poor match between genotype and SNP control probes, or incorrectly predicted sex using the DNAm information. DNAm was recorded at Waves 1, 3 and 4 of the LBC1921, at mean ages of 79 (n=515), 87 (n=181) and 90 (n=87) years of age. Sample collection for LBC1936 began in 2004 and is ongoing, with Wave 5 collected between 2017 and 2019. DNAm is available for Waves 1-4 of LBC1936, recorded at mean ages of 70 (n=906), 73 (n=801), 76 (n=619) and 79 (n=507).

4.2.4 Proteomic data

DNAm, genetic and proteomic (Olink inflammatory platform) data were available for individuals from LBC1936 Wave 1 (50% female, with n=875 at mean age 70 years [SD 0.8]). DNAm, genetic and proteomic (Olink neurology platform) data were available for individuals from LBC1936 Wave 2 (47% female, with n=706 at mean age 73 years [SD 0.7]) and LBC1921 Wave 3 (53% female, with n=162 at mean age 87 years [SD 0.4]). Plasma samples were analysed with either the Olink Inflammation 92-plex or the Olink Neurology 92-plex proximity extension assays (Olink Bioscience, Uppsala Sweden). One inflammatory panel protein (BDNF) failed quality control and was removed. A further 21 inflammatory proteins were removed, as over 40% of samples fell below the lowest limit of detection. Two neurology proteins (MAPT and HAGH) were excluded due to >40% of observations being below the lower limit of detection. This resulted in 90 neurology (LBC1936 Wave 2) and 70 inflammatory (LBC1936 Wave 1) proteins in LBC1936, with 92 neurology proteins available in LBC1921 Wave 3. Protein levels were rank-based inverse normalised and regressed onto age, sex, Olink array plate and four genetic components of ancestry derived from multidimensional scaling of the Illumina 610-Quadv1 genotype array.

Serum S100 calcium-binding protein β (S100β) samples were obtained during the main physical and cognitive testing appointment at Wave 2 of LBC1936. Samples were stored at −80 °C at the Wellcome Trust Clinical Research Facility, Western General Hospital, Edinburgh. When the assessment for Wave 2 was completed, they were transferred to the Department of Clinical Biochemistry, King's College London using cold-chain logistics. They were then stored at −20 °C. Assays were conducted using a chemiluminescence immunoassay S100β kit (catalogue number 314701, distributed by DiaSorin, Berks, UK) on a LIAISON chemiluminescence analyzer. The
lag between sample dispatch at the end of sampling and assay completion (i.e., time stored at −20 °C rather than −80 °C) was an average of 44 days (SD = 26) across four batches.

4.3 The KORA F4 cohort

The Cooperative Health Research in the Region of Augsburg (KORA) study is comprised of four cross-sectional surveys that were conducted from 1984 to 2001. The KORA study resource protocol is summarised by Wichmann et al. (2005). The cohort is centred on collecting resources that are useful to the fields of epidemiology, health care research and health economics investigations. A total of 17,602 individuals were randomly sampled from population registries in South Germany. KORA F4 is the follow-up study that occurred after the KORA S4 survey (conducted between 1999 and 2001) and features 4,261 individuals. The empirical work in this thesis utilises data from the KORA F4 follow-up study that was sampled between 2006 and 2008, when recruitment took place. Blood samples were collected during this period, with deep phenotyping. The F4 study includes a further subset of 3,080 individuals aged between 32 and 81 years at the time of baseline F4 recruitment. The F4 samples include 944 individuals with methylation, proteomic, and genetic data available (51% female, mean age 59 years [SD 7.8]).

4.3.1 Ethics and funding

All KORA participants have provided their written informed consent. The KORA study received approval from the Ethics Committee of the Bavarian Medical Association. The study was established and funded by the Helmholtz Zentrum München—German Research Center for Environmental Health. This institution is supported by the German Federal Ministry of Education and Research and the State of Bavaria. Additionally, KORA research was supported by the Munich Center of Health Sciences (MC-Health) and Ludwig-Maximilians-Universität, as part of LMUinnovativ.

4.3.2 Genetic data

The Affymetrix Axiom array was employed to quantify genotypes for 3,788 individuals in the original KORA S4 study. Quality control steps involved filtering to SNPs with a total genotyping rate of > 99.8%, while also filtering for those that had MAF > 0.01. A
total of 509,946 autosomal SNPs were used for imputation, which was performed using SHAPEIT2 \(^{253}\), with the 1000G phase 3 haplotype (build 37). Variants with certainty < 0.95, information metric < 0.7, HWE test P < 1 \(\times\) 10\(^{-6}\), missing genotype information, or MAF < 0.01 were excluded. After these steps, there were a total of 8,263,604 variants with a total genotyping rate of 0.97 that were kept for analyses.

4.3.3 DNA methylation data

Methylation data were generated for 1,814 individuals in KORA \(^{177}\). The Infinium HumanMethylation450 BeadChip was used to generate DNA\(\text{m}\) data for these individuals. SNP-associated CpG sites were initially excluded, which was followed by background correction performed using the minfi R package \(^{264}\). The detection rate was examined and any samples having a detection rate below 95% were excluded. Normalisation was carried out on methylation values using the minfi R package \(^{264}\). Estimation of white blood cell proportions was performed according to the Houseman method \(^{138}\). Methylation measurements for 485,512 CpG sites were available after these steps were complete. There were 35 CpG sites that were removed as they had fewer than 100 measures. Any CpG sites located on the X or Y chromosomes (11,231 and 416 CpGs, respectively) were also removed. There were 2,993 non-cg CpG sites that were also removed from the data, leaving 470,837 sites that were available.

4.3.4 Proteomic data

The SomaScan (v3.2) \(^{258}\) platform was used to quantify proteomic measurements from plasma samples in the KORA cohort \(^{179}\). This was performed for a total of 1,129 possible aptamer probes. Of the 1,000 plasma samples available for analyses, two samples were excluded due to biobank sampling errors. Another sample was excluded based on poor quality control measures. Of the 1,129 proteomic probe measurements, five failed quality control steps and 1,124 remained for analyses. Of the 997 remaining samples, there were 944 from individuals that had DNA\(\text{m}\) and genotyped data available in addition to the proteomic measurements \(^{177}\).

4.4 The UK Biobank

The UK Biobank (UKB) is a population-based cohort of \(~\)500,000 individuals that were between 40-69 years old at the time of recruitment (between 2006 and 2010).
Genome-wide genotyping, exome sequencing, electronic health record linkage, whole-body magnetic resonance imaging, physical and anthropometric measurements and blood and urine biomarkers are available. Further information regarding the full availability of measurements in UKB can be found at: https://biobank.ndph.ox.ac.uk/showcase/. UKB Pharma Proteomics Project (UKB-PPP) is a precompetitive consortium of 13 biopharmaceutical companies funding the generation of blood-based proteomic data from UKB volunteer samples.

4.4.1 Ethics and funding

Every participant provided informed consent for their data to be used in biomedical health research. UKB was established originally by the Wellcome Trust medical charity, Medical Research Council, Department of Health, the Northwest Regional Development Agency and the Scottish Government. UKB is funded by the Wellcome charity and the Medical Research Council (MRC) primarily. UKB has recieved funding from the Welsh Government, Cancer Research UK, the British Heart Foundation, and Diabetes UK. UKB is also supported by the National Health Service (NHS). Thirteen companies (Alnylam Pharmaceuticals, Amgen, AstraZeneca, Biogen, Bristol-Myers Squibb, Calico, Genentech, Glaxo Smith Klein, Janssen Pharmaceuticals, Novo Nordisk, Pfizer, Regeneron, and Takeda) funded the proteomic sampling for the UKB-PPP study.

4.4.2 Proteomic data

The UKB-PPP includes proteomic sampling for 54,306 UKB participants and 1,474 protein analytes 144. The panel consists of four Olink panels (Cardiometabolic, Inflammation, Neurology and Oncology). A total of 46,673 individuals were randomly selected from baseline, in addition to 6,385 individuals that were selected by the UKB-PPP consortium member companies. Another set of 1,268 individuals also had participated in a coronavirus disease 2019 (COVID-19) study and had both baseline and additional timepoint samples available for proteomic measurement. The randomised samples have previously been demonstrated to be highly representative of the wider UKB sample, whereas the consortium-selected individuals were specifically enriched for a set of 122 diseases that were of particular interest to members 144. Of the 54,309 individuals that had protein data available in the UKB-PPP
protein samples, 54,189 were available after quality control exclusions with 1,474 Olink protein analytes.

4.5 mQTL and eQTL dataset populations

The mQTL and eQTL datasets used for colocalisation analyses in Chapter 6 and functional assessment lookups of CpGs in Chapter 7 were sourced from analyses run in two previous studies. The eQTL database was utilised to run colocalisation in Chapter 6, whereby the dataset was subset to a 200kb region either 100kb upstream or downstream of the gene of interest (S100β). Both the eQTL and mQTL datasets were utilised in Chapter 7, to perform lookups of CpGs that were associated with protein levels, to ascertain whether there was evidence that the associations were driven by underlying mQTL or eQTL effects.

4.5.1 eQTL dataset

The eQTL dataset was sourced from eQTLGen Phase I, which made the results of the study by Võsa et al, (2018) available. This study included 31,684 individuals with predominantly European ancestry that were available via the eQTLGen Consortium. Over 11 million SNPs (with MAF ≥ 1%) were available across 19,942 genes studied and a range of assays were used to quantify gene expression levels (Illumina was the most commonly-used, with N=17,421 individuals). Meta-analyses were performed for cis-eQTL and trans-eQTL identification, with adjustments for technical effects and biological effects such as cell-type confounding. The majority of blood samples were whole-blood (80.4%), whereas the remainder were derived from peripheral blood mononuclear cells (19.6%). The cis-eQTLs were generated in analyses that assessed a 1Mb region upstream and downstream of a given SNP, whereas a required distance between the SNP and the centre of a gene was set to >5Mb for trans-eQTLs. An FDR-correction was applied, resulting in the identification of 16,987 cis-eQTL genes and 6,298 trans-eQTL genes.

4.5.2 mQTL dataset

The mQTL dataset was sourced from GoDMC (accessed August 2021), which made the results of the study by Min et al, (2021) available. The mQTL database was produced through collaboration across 36 cohort study populations and disease
datasets. GWAS of blood-based 450k Illumina methylation sites were run, with ~ 10 million genotypes imputed according to the 1000 Genomes Project reference panel 270. DNA methylation (DNAm) was measured from either whole blood or cord blood using the HumanMethylation450 or EPIC arrays. The effects of family structures and typical covariates (age, sex, estimated cell counts, predicted smoking and genetic principal components) were adjusted for. Associations that met the relaxed significance threshold of $P < 1 \times 10^{-5}$ were meta-analysed across 32,851 individuals, highlighting genetic variants that associated with the 420,509 CpG site measurements. Over 270,000 of the mQTL associations represented independent signals.

4.6 Summary

In this chapter, I provided an overview of the cohort populations and key omics data from each that have been used in this thesis. In Chapter 5, I present the core aims and chosen methodologies for each of the empirical Chapters 6-10 in this thesis.
5 Thesis aims and methods overview

Here, I first present the aims of each empirical chapter in this thesis that form Chapters 6-10 in Section 5.1. I then provide an overview of the key methodologies and cohort populations used for each empirical chapter in Section 5.2.

5.1 Thesis aims

The proteomic correlates of brain health and their molecular architectures have not been comprehensively characterised. Furthermore, no study has systematically generated epigenetic scores (EpiScores) for blood proteins at scale and related them to the onset of disease. Finally, proteomic signatures of incident disease may also be useful tools for risk stratification. With this in mind, there are five key aims and empirical chapters in this thesis.

**Aim 1 (Chapter 6):** To perform GWAS and MWAS of a known neuroinflammatory marker – S100β – and test for a causal relationship between this marker and Alzheimer's dementia (LBC1936).

**Aim 2 (Chapter 7):** To expand biomarker identification by conducting a brain health PheWAS of plasma levels of 4,235 proteins in 1,065 healthy older adults, and map epigenetic signatures of these proteins through MWAS (GS).

**Aim 3 (Chapter 8):** To generate epigenetic scores (EpiScores) for a comprehensive set of circulating protein levels (using several cohorts for training and testing) and test whether protein EpiScores predict incident neurological diseases and associated comorbidities in 9,537 adults (GS).

**Aim 4 (Chapter 9):** To directly compare the capability of EpiScore and measured GDF15 and Nt-proBNP levels for prediction of dementia and associated morbidities (Generation Scotland) and brain health characteristics (LBC1936).

**Aim 5 (Chapter 10):** To create and assess ProteinScores for incident disease outcomes and profile individual protein associations in 47,600 individuals (UKB-PPP).
5.2 Overview of key methods

Detailed methods sections can be found in Chapters 6-10. Here, I present brief rationales for methods employed in 1) the molecular association studies presented in Chapters 6 and 7; 2) the development and testing of protein EpiScores in Chapters 8 and 9; and 3) the proteomic assessment of incident disease in Chapter 10. Cohort populations and approaches referenced in this section are summarised in Figure 5-1.

In Chapter 6, GWAS (N=769) and MWAS (N=722) are performed on S100β levels in LBC1936. Linear regression GWAS (7,307,523 SNPs tested) are run in PLINK version 1.9. Independently associated variants are identified via conditional and joint association analyses (COJO) to isolate lead variants and mitigate against linkage disequilibrium. Colocalisation is used to test for evidence of shared variants underlying transcript and actual protein, using cis pQTLs and known eQTLs. Colocalisation and Mendelian randomisation (via the TwoSampleMR R package) are also used to test for evidence supporting a relationship between circulating S100β and Alzheimer’s dementia. The Alzheimer’s dementia GWAS from Jansen et al. (2019) that I used represented the largest available at the time of analyses (13,367,299 SNPs, with N=71,880 cases, and N=383,378 controls). When I queried the 200kb region for colocalisation either side of the sentinel SNP, there were 1,346 variants present in this region in the Alzheimer’s dementia GWAS results.

MWAS were run between 459,309 CpGs available through the Illumina 450k array and S100β levels. Confounders that are specific to DNAm (outlined in Section 1.3.6) can lead to intercorrelations between probes across the genome and may bias effect estimates. Therefore, I regressed DNAm measures onto a range of known covariates (age, sex, technical covariates, BMI, DNAm-derived immune cell proportions and an EpiSmokEr score – SSc method (based on 187 CpGs)). I also chose the omic-data-based complex trait analysis (OSCA) software, which is a leading resource for computing linear mixed model MWAS. I used the linear mixed model command MOMENT (multi-component MLM-based omic association excluding the target) in OSCA to run the MWAS. This approach identifies strongly- and weakly-associated CpG groups via linear regression with an epigenome-wide significance threshold for group status. In every MWAS association, these two groups are then modelled as random effects. All CpG sites > 50kb from the CpG of interest are fit within the random
effect components, which is thought to adjust for intercorrelations between distal probes that may arise from confounding factors. Zhang et al. (2019) demonstrated that the use of OSCA MOMENT commands, in combination with adjustment for DNAm-specific nuisance variables brings MWAS inflation towards 1.  

In Chapter 7 I perform MWAS of 4,235 SomaScan-derived plasma protein levels in the STRADL subset of GS (N=774). The STRADL sample included 120 related individuals with (> 0.05 threshold) that I calculated via a genetic relatedness matrix (GRM). To maximise the sample available but adjust for relatedness, MWAS were run with a GRM fit as a random effect. As this is not computationally viable using MOMENT, the MOA (MLM-based omic association) OSCA command was chosen for the analyses. This resulted in an approach that is equivalent to a linear model with adjustment for genetic relatedness structure. I chose to run staged MWAS with increasingly complex sets of covariates, to incrementally adjust for potential factors that may produce intercorrelated signatures between probes. Depression status was included as a covariate given the sample recruitment enriched for depression phenotypes. I regressed protein levels onto a range of relevant demographic and technical covariates, in addition to pQTLs specific to the protein. This approach mirrored previously-conducted MWAS in SomaScan by Zaghlool et al (2020). I sourced sentinel pQTLs from the seminal pQTL mapping study on the same SomaScan protein panel by Sun et al. (2018) and extracted measures of these variants in the STRADL genetic data. I chose to include pQTLs as I hypothesised that it would increase the likelihood that pQTMs captured through MWAS would represent non-genetic associations. Inflation statistics were calculated for each protein MWAS, providing a means to ascribe confidence to the interpretation of results.  

I also chose to scan the 4,235 protein levels in STRADL to identify markers of 15 traits associated with brain health (n=1,065, where the MWAS sample of 774 individuals represented a nested subset). Given the somewhat restricted sample size available, APOE status was binarised to a three-part variable and treated as a continuous predictor (0 = e2e2, e2e3; 1 = e3e3, 2 = e3e4, e4e4, with e2e4 individuals excluded). Intracranial volume was included as a covariate to mitigate against brain volume protein marker associations representing differences in head size. Across all associations tested in the study, a PCA-informed, Bonferroni adjustment for multiple testing was chosen, as there was an intercorrelated structure present in the 4,235
proteins. I also assessed how novel the MWAS and protein-brain trait associations were in the context of what was published in the field. Protein marker associations with brain health traits were explored in terms of direction of effect, as I wanted to map proteins that were associated with adverse or better brain health states. Finally, I used the results generated across both parts of the study to index the pQTMs that involved protein markers of brain health (pQTM ~ protein ~ brain) and explored the biological pathways associated with the components of these associations.

In Chapter 8, protein EpiScores are trained, tested and modelled as biomarkers for 12 incident diseases GS (n=9,537) – a cohort with extant, comprehensive EHR linkage that has DNAm but does not have protein measures. Unlike previous protein EpiScore studies, I used protein levels measured on multiplexed protein panels as protein outcomes. I tested a total of 953 proteins (160 Olink Inflammatory and Neurology proteins in >725 LBC1936 individuals, in addition to 793 SomaScan proteins in 944 KORA individuals). I chose to use elastic net penalised regression, as per the generation of protein EpiScores for CRP and IL6 from Stevenson et al. (2020 and 2021)\(^\text{215,216}\) and the GrimAge score by Lu et al. (2019)\(^\text{129}\). Elastic net penalised regression can simultaneously consider hundreds of thousands of DNAm measures at CpG sites as potentially-informative features. This makes it a more attractive method than taking the top CpG associations from MWAS of a protein, as pQTMs are assessed individually (hence, ignoring correlated structures of CpGs) and may be prone to winner’s curse. Elastic net represents a compromise between the penalties applied by LASSO and RIDGE regression, which is thought to be well-suited to DNAm measures that inherently have a high degree of intercorrelation. Therefore, an alpha parameter of 0.5 was chosen to align with the previous protein EpiScore approaches in the field, with 10 folds used to perform cross-validation in the training set. I evaluated score performance in external tests using incremental R\(^2\) (beyond age and sex) to gain insight into the variance in measured protein levels that could be explained by my scores. I also used Pearson’s correlations to set a cut-off for viable scores (r > 0.1, P < 0.05), which 109 of the 953 protein EpiScores met.

I ran Cox PH models to test individual associations between the 109 protein EpiScores and the incidence of 12 diseases, over 14 years of follow-up in GS (n=9,537). The primary R packages for performing Cox analyses between a given predictor and
outcome event in the context of covariates have been developed by Terry Therneau and include the survival \textsuperscript{273} and coxme \textsuperscript{274} packages. Cox PH analyses can be performed in the survival package, whereas coxme allows for modelling of random effects in conjunction with the fixed effects covariates included in the model. This is particularly useful in populations such as GS, that have a high degree of relatedness \textsuperscript{247}. A kinship matrix can be supplied to coxme, to adjust for relatedness as a by-subject random effect. I therefore chose this approach in GS. I ran models without relatedness adjustment using Coxph with the cox.zph function in the survival package. This allowed me to check the Cox PH assumption (HR stable over time) by calculating a P-value for each covariate at a local level, in addition to a global model value. A value of $P < 0.05$ indicates violation of the PH assumption. I ran a range of sensitivity analyses to understand whether the associations identified between protein EpiScores and incident diseases were attributable to cell-type differences, or could be attenuated by the GrimAge acceleration score (i.e. do my protein-specific scores add value beyond this general ageing score). There were large-scale multiplexed protein assessments of prevalent and incident type 2 diabetes available from previous studies. Therefore, I chose to check how many of my protein EpiScore associations with type 2 diabetes replicated previously-reported associations. I did this to understand whether my protein EpiScores were able to capture a disease-relevant, protein-biomarker signal in the absence of the measured protein.

In Chapter 9, EpiScores for serum GDF15 and NT-proBNP levels are developed, directly compared against protein measures and associated with brain health outcomes. I first run MWAS of each protein in GS ($n > 16,963$) using BayesR+ software. BayesR+ has been found to implicitly control for the unwanted effects of immune cell-type proportions, related participants and unknown confounders \textsuperscript{133}. I then decided to initially split the GS dataset available, to train and test EpiScores for these proteins. I did this to allow for direct Cox PH comparisons between proteins versus protein EpiScores as biomarkers for incident diseases in the test sample. This was a limitation of my original protein EpiScore study, which I wanted to address here. I chose to study incident dementia and ischaemic stroke, in addition to heart disease and type 2 diabetes (that both indirectly associate with the health of the brain during ageing). I then retrained the protein EpiScores on the full GS sample available, projecting them into external test sets in LBC1936 ($n = 322$ to 500). As LBC1836 has
multiple waves of cognitive ability and MRI scans recorded at 3-year approximate intervals, I modelled cross-sectional and longitudinal associations between the protein EpiScores and these variables from Wave 1 of the study.

In Chapter 10, I profile protein associations with 23 incident diseases and death (35,232 associations tested, between 1,468 proteins and 24 outcomes). I also consider the 1,468 proteins simultaneously in Cox PH elastic net regression models to create ProteinScores for the 20 incident outcomes that had a minimum of 150 cases. In Cox PH penalised regression, the onset period defined should be clinically relevant (i.e. greater than 10 years may not be helpful for immediate decision making) and must be suited to the time-to-event distribution for case onset in the population, as disease onset varies across the life course. Therefore, I chose to run 10-year onset models, as these have been used in clinically-implemented scores. Unlike GS, which has a family-based structure with high proportion of relatedness, the UKB-PPP sample includes very few related individuals. Therefore, 1,182 of the 52,744 individuals were excluded and Cox PH analyses were run without adjustment for a kinship matrix. The outcome for Cox PH elastic net models in the UK Biobank was time-to-event data for cases and controls, which were unbalanced (e.g. 2,822 cases of type 2 diabetes represented ~5% of the total population in the study). As performance statistics (AUC and PRAUC) are sensitive to highly skewed datasets, a randomised subsampling approach was taken to select a 1:3 case:control balance. The process of randomised ProteinScore sampling, training and testing was repeated over fifty iterations for each outcome, with the ProteinScore that added the median incremental difference beyond a null model selected. This approach was chosen to mitigate against unobserved population variation being randomly selected for and driving divergence in score performance. Fold sizes that were proportionate to the number of cases available for training were chosen. ProteinScores were evaluated in incremental models to acquire performance due to the score that was added beyond other risk factors. Due to the scale of the sample (~50,000 individuals) and the availability of health data linkage, I was able to model multiple incident neurological diseases (including Alzheimer’s dementia, ALS, Parkinson’s disease and stroke). I also included cardiometabolic, inflammatory, autoimmune and cancer diagnoses as well as mortality. The strategy was chosen to progress towards a set of ProteinScores that could be used as a panel to gain a holistic insight into risk of multiple diseases that an individual may have.
Figure 5.1. Cohort study populations used in Chapters 6-10 of this thesis.

An overview of the cohort populations and datasets used in the empirical chapters of this thesis. Key methods used for each section have been highlighted in green.

5.3 Summary

I outlined the major aims of this thesis and their novelty. I also provided an overview of the chosen methods for each empirical chapter of this thesis, with a summary of the cohort populations and samples used in each chapter (Figure 5-1). In the next chapter, I present my first study, which conducts GWAS and MWAS on a marker of neuroinflammation S100β and test for a relationship between this protein and Alzheimer’s dementia through Mendelian randomisation and colocalisation techniques.
6 Molecular architecture of neuroinflammatory protein S100β and its relationship to Alzheimer’s disease

6.1 Introduction

S100β has a role in neuroinflammation and is implicated as a candidate factor linked to the pathogenesis of Alzheimer’s dementia 275-278. In this chapter, I first perform GWAS and MWAS of serum S100β in the LBC1936 cohort (N=769, mean age 72.5 years, SD = 0.07). I then use COJO to identify independent genetic loci, which are used as genetic instruments to test for evidence of a causal relationship between S100β and Alzheimer’s dementia. Colocalisation analyses were first run between S100β and eQTLs, followed by S100β and Alzheimer’s dementia loci.

This study has been published in Wellcome Open Research in 2021 279 and is included in full in Section 6.2. My role in the analyses was to co-supervise a rotation student and facilitate the running of the initial GWAS and EWAS analyses. I then re-ran the work done, adding the COJO, Mendelian randomisation and colocalisation analyses and drafted the manuscript.

Code used for this work is available in the GitHub repository: https://github.com/DanniGadd/s1888864_Supplementary_Material/tree/main/Chapter_6. The summary statistics for the GWAS and MWAS are hosted at the Zenodo repository: https://zenodo.org/record/5591776. No additional Supplementary files were included in this publication.
6.2 The genetic and epigenetic profile of serum S100β in the Lothian Birth Cohort 1936 and its relationship to Alzheimer's disease
The genetic and epigenetic profile of serum S100β in the Lothian Birth Cohort 1936 and its relationship to Alzheimer’s disease [version 2; peer review: 2 approved]

Danni A Gadd, Robert I McGeachan, Robert F Hillary, Daniel L McCartney, Sarah E Harris, Roy A Sherwood, N Joan Abbott, Simon R Cox, Riccardo E Marioni

1Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, Other (Non-U.S.), EH4 2XU, UK
2Centre for Discovery Brain Sciences, University of Edinburgh, Edinburgh, Other (Non-U.S.), EH8 9JZ, UK
3Lothian Birth Cohorts, University of Edinburgh, Edinburgh, Other (Non-U.S.), EH8 9JZ, UK
4Department of Psychology, University of Edinburgh, Edinburgh, Other (Non-U.S.), EH8 9JZ, UK
5Department of Clinical Biochemistry, King’s College Hospital NHS Foundation Trust, London, Other (Non-U.S.), SE5 9RS, UK
6Institute of Pharmaceutical Science, King’s College London, London, Other (Non-U.S.), WC2R 2LS, UK

* Equal contributors

First published: 10 Nov 2021, 6:306
https://doi.org/10.12688/wellcomeopenres.17322.1
Latest published: 27 Jan 2022, 6:306
https://doi.org/10.12688/wellcomeopenres.17322.2

Abstract

Background: Circulating S100 calcium-binding protein (S100β) is a marker of brain inflammation that has been associated with a range of neurological conditions. To provide insight into the molecular regulation of S100β and its potential causal associations with Alzheimer’s disease, we carried out genome- and epigenome-wide association studies (GWAS/EWAS) of serum S100β levels in older adults and performed Mendelian randomisation with Alzheimer’s disease.

Methods: GWAS (N=769, mean age 72.5 years, sd = 0.7) and EWAS (N=722, mean age 72.5 years, sd = 0.7) of S100β levels were performed in participants from the Lothian Birth Cohort 1936. Conditional and joint analysis (COJO) was used to identify independent loci. Expression quantitative trait locus (eQTL) analyses were performed for lead loci that had genome-wide significant associations with S100β. Bidirectional, two-sample Mendelian randomisation was used to test for causal associations between S100β and Alzheimer’s disease. Colocalisation between S100β and Alzheimer’s disease GWAS loci was also examined.

Results: We identified 154 SNPs from chromosome 21 that associated (P<5x10^{-8}) with S100β protein levels. The lead variant was located in the S100β gene (rs8128872, P=5.0x10^{-17}). We found evidence that two independent causal variants existed for both transcription of S100β and S100β protein levels in our eQTL analyses. No CpG sites were
associated with S100β levels at the epigenome-wide significant level (P<3.6x10^{-8}); the lead probe was cg06833709 (P=5.8x10^{-6}), which mapped to the LG11 gene. There was no evidence of a causal association between S100β levels and Alzheimer’s disease or vice versa and no evidence for colocalisation between S100β and Alzheimer’s disease loci.

Conclusions: These data provide insight into the molecular regulators of S100β levels. This context may aid in understanding the role of S100β in brain inflammation and neurological disease.

Keywords
Epigenetic, Genetic, S100β, Inflammation, EWAS, GWAS, Alzheimer’s disease

This article is included in the Generation Scotland gateway.
Introduction

The calcium-binding protein S100 beta (S100β) has been suggested as a biomarker for central nervous system disease.1,2 Expressed most commonly in astrocytes, its cytoplasmic location and calcium-binding capability allows S100β to mediate calcium homeostasis, cell proliferation and survival intracellularly, while also triggering the RAGE-associated inflammatory response when secreted extracellularly.2 Part of pro-inflammatory danger-associated molecular patterns (DAMPs), elevated S100β is linked to cytokine cascades in the brain.3

Although the exact pathophysiology is still unknown, a number of small-scale studies have reported elevated circulating or cerebrospinal fluid (CSF) S100β levels in individuals with nervous system injury, neuroinflammatory conditions, white matter ageing and Alzheimer’s dementia and delirium.4–9. Follicerebrospinal fluid (CSF) S100β of small-scale studies have reported elevated circulating or

Whether S100β has a direct involvement in the pathogenesis of Alzheimer’s disease is still unclear. Further to its potential role in inflammatory exacerbation in the brain, research suggests that at nanomolar concentrations, S100β can have protective and neurotrophic effects.10–14. Despite the widely discussed biological importance of S100β, the possible epigenetic regulators of the protein have not been investigated. One study using blood spots taken at birth has previously identified two genetic associations in relation to circulating S100β levels; rs62224256 on chromosome 21, 21kb downstream of the pericentrin gene (PCNT) and rs28397289 on chromosome 6, within the human leukocyte antigen (HLA) region.15. Elucidating the mechanisms that determine inter-individual variation in circulating S100β levels, in healthy individuals, may therefore provide insight into S100β’s role in health and disease. Further genetic mapping of S100β may also facilitate causal association tests with disease endpoints.

Here, we perform genome- and epigenome-wide association studies (GWAS/EWAS) of S100β in relatively healthy older adults from the Lothian Birth Cohort 1936 (LBC1936). We then use genetic instruments identified for S100β to test for bidirectional causal associations with Alzheimer’s disease, via two-sample Mendelian randomisation. Demographic information for the GWAS (N=769) and EWAS (N=722) sample groups are summarised in Table 1.

Methods

The Lothian Birth Cohort 1936

The Lothian Birth Cohort 1936 (LBC1936) is a longitudinal study of cognitive ageing. Cohort members were born in 1936 and took part in the Scottish Mental Survey 1947 at age 11 years. Approximately 60 years later, those individuals that were living mostly within the Edinburgh area were re-contacted (n = 1,091, recruited at mean age 70 years). Recruitment and testing of the LBC1936 cohort have been described previously.16,17. Briefly, the data collection has included detailed phenotypic, biological

The S100β gene was also identified as a site of differential DNA methylation (DNAm) relating to Braak staging in a previous epigenome-wide association study (EWAS) of cortical post-mortem tissues (n=159).18.

Table 1. Demographics summary for the EWAS and GWAS sample populations.

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>EWAS n (%)</th>
<th>Mean (sd)</th>
<th>GWAS n (%)</th>
<th>Mean (sd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum sample (n)</td>
<td>722</td>
<td>769</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S100β (μg/L)</td>
<td>0.086 (0.035)</td>
<td>0.085 (0.034)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex (Female)</td>
<td>340 (47.1)</td>
<td>367 (47.7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age at S100β (years)</td>
<td>72.5 (0.7)</td>
<td>72.5 (0.7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Body Mass Index (kg/m²)</td>
<td>27.9 (4.3)</td>
<td>27.9 (4.3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EpiSmokEr</td>
<td>0.84 (5.2)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
and cognitive sampling, over a series of follow-up waves (approximately every 3–4 years since recruitment). Body mass index (BMI) was measured at clinic visits and recorded in kg/m². The S100β data available in this study were from the second wave of testing, at around three years after the Wave 1 visit to the clinic. DNA methylation data were also collected at the same time-point as S100β sampling. Genotyping was performed on DNA from blood samples collected at Wave 1.

S100β measurement

Serum samples were obtained from participants during the main physical and cognitive testing appointment at Wave 2. After collection, samples were stored at −80 °C at the Wellcome Trust Clinical Research Facility, Western General Hospital, Edinburgh, until the conclusion of the wave. They were then transferred to the Department of Clinical Biochemistry, King’s College London using cold-chain logistics, where they were stored at −20 °C until assays were conducted using a chemiluminescence immunoassay S100β kit (catalogue number 314701, distributed by DiaSorin, Berks, UK) on a LIAISON chemiluminescence analyzer. The lag between sample dispatch at the end of sampling and assay completion (i.e., time stored at −20 °C rather than −80 °C) was an average of 44 days (SD = 26) for four batches. The minimal detectable concentration of the assay was 0.02 μg/L.

Genotyping

LBC1936 DNA samples were genotyped at the Edinburgh Clinical Research Facility using the Illumina 610-Quadv1 array (Wave 1: n = 1005; mean age: 69.6 ± 0.8 years; San Diego). Preparation and quality control steps have been reported previously. SNPs were imputed to the 1000 G reference panel (phase 1, version 3). Briefly, individuals were excluded on the basis of sex mismatches, relatedness, SNP call rate of less than 0.95, and evidence of non-European ancestry. SNPs with a call rate of greater than 0.98, minor allele frequency in excess of 0.01, and Hardy-Weinberg equilibrium test with P ≥ 0.001 were included in analyses. Only SNPs with a minor allele frequency > 0.05 and imputation quality > 0.6 were retained. The remaining 8,489,963 SNPs were filtered to remove those that had a minor allele count < 25. A total of 7,307,523 SNPs were available for GWAS analyses.

DNA methylation

DNA methylation from whole blood at Wave 2 of the Lothian Birth Cohort 1936 was measured using the Illumina 450 K methylation array at the Edinburgh Clinical Research Facility. Complete details of quality control steps taken to process the dataset have previously been described. Briefly, raw intensity data were background-corrected and normalised using internal controls. Manual inspection facilitated the removal of low quality samples presenting issues relating to bisulphite conversion, staining signal, inadequate hybridisation or nucleotide extension. Further quality control analyses were performed to exclude probes with low detection rate <95% at P < 0.01 and samples with a low call rate (<450,000 probes detected at p-values of less than 0.01) were also excluded. Finally, samples were removed if there was a poor match between genotype and incorrect DNA methylation-predicted sex, or SNP control probes. DNA methylation at Wave 2 was processed in three sets (n=256, 461 and 5, for sets 1, 2, and 3, respectively). In total, there were 459,309 CpG sites used in EWAS analyses. Lothian Birth Cohort Wave 2 DNA methylation data were used to generate an epigenetic score for smoking – known as EpiSmokEr in the sample. This score utilises previously derived weights calculated in an independent sample and has been previously shown to robustly reflect smoking status.

S100β sample preparation

There were 834 individuals with S100β concentrations recorded at Wave 2 of the Lothian Birth Cohort study. Six measurements greater than four standard deviations from the mean were excluded, as per previous analyses that utilised this sample. There were 769 individuals with genome-wide genetic data (mean 72.5 years, sd = 0.7) and 722 individuals with epigenome-wide DNA methylation data (mean 72.5 years, sd = 0.7) available. Table 1 summarises demographic information for these sample populations. In the maximum sample available in GWAS and EWAS (N=722), serum S100β levels were higher in females (beta = 0.26, SE = 0.07, P = 2.2×10⁻⁴) and older individuals (correlation of 0.18 and beta = 0.16 per year, SE = 0.03, P = 2.9×10⁻⁶ in linear models) (Figure 1).

Body mass index (BMI) and smoking are common lifestyle covariates that have well-documented DNAm signatures; we therefore tested whether these traits should be adjusted for in our analyses. S100β levels were positively associated with body mass index (BMI) (beta = 0.10 per kg/m², SE = 0.03, P = 2×10⁻⁶ in linear models) (Figure 1). S100β protein levels were transformed by rank-based inverse normalisation and regressed onto age, sex, BMI at Wave 2 (kg/m²) and four genetic principal components of ancestry in separate analyses groups (EWAS N=722, GWAS N=769). Standardised residuals (mean = 0, variance = 1) from these linear regression models were brought forward as the protein level variable for the respective analyses.

Data quality control and preparation was conducted in R (Version 4.0.3)

Genome-wide association study (GWAS)

Linear regression was used to assess the effect of each of the 7,307,523 available SNPs on the levels of S100β via PLINK (Version 1.9). Genome-wide stepwise conditional analysis was performed through GCTA-COJO using the ‘cojo-slct’ option to identify independent variants. Individual level genotype data were used for the reference linkage disequilibrium (LD) structure along with default settings of the software. The variance (r²) in S100β levels that could be explained by this variant was calculated as follows: r² = 2 × MAF × (1-MAF) × beta², where
beta = effect size of the SNP and MAF = the effect allele frequency.

Expression quantitative trait loci (eQTL) colocalisation
We cross-referenced sentinel *cis* pQTLs that were selected by GCTA-COJO analyses with publicly available *cis* eQTL data taken from the eQTLGen consortium\(^{26}\). The *cis* eQTLs were subset to the same chromosome as the *cis* pQTL. A 200 kb region (either upstream or downstream) was extracted from our GWAS summary statistics for S100\(\beta\) to capture *cis* effects within 100 kb of the target gene\(^{27}\). eQTLs for this region were then extracted from the eQTLGen consortium summary statistics for the S100\(\beta\) region. The shared SNPs across transcripts for S100\(\beta\) and S100\(\beta\) protein levels were then tested for colocalisation using the coloc package\(^{28}\) (Version 5.1.0) in R, with five hypotheses in Bayesian tests with default priors\(^{28}\). In addition to the null hypothesis (no causal variant), hypothesis 1 indicated a causal variant was present for S100\(\beta\) protein levels only. Hypothesis 2 indicated that there was a causal variant for the S100\(\beta\) transcript only. Hypothesis 3 indicated that there were independent causal variants for both S100\(\beta\) transcription and S100\(\beta\) protein levels. Hypothesis 4 indicated that there were two association signals that contributed to both S100\(\beta\) gene expression and S100\(\beta\) protein levels. A posterior inclusion probability (PP) > 0.95 was taken as the threshold for hypothesis testing.

Mendelian randomisation
Two-sample, bidirectional Mendelian Randomisation (MR) was used to test for potentially causal associations between S100\(\beta\) protein levels and Alzheimer’s disease. Associations from separate GWAS were used as genetic instruments. As allele assignment is randomised, the SNPs associated with the exposure are randomised to the effects of confounders and likely to be causally upstream of the exposure\(^{29}\). Summary statistics from a GWAS performed by Jansen *et al.*\(^{30}\) were used as the Alzheimer’s disease dataset (13,567,299 SNPs, with N=71,880 cases, and N=383,378 controls). The GWAS summary statistics for S100\(\beta\) were sourced from the analyses in this study that used samples from the Lothian Birth Cohort 1936 (N=769). Importantly, the Alzheimer’s disease summary statistics were based on a meta-analysis of cohorts that were independent of the Lothian Birth Cohort 1936. All analyses were performed using the TwoSampleMR package (Version 0.5.6) in R\(^{29}\). One assessment quantified the association between S100\(\beta\) levels as the exposure and Alzheimer’s disease as the outcome. A second assessment then quantified the association between Alzheimer’s disease as the exposure and S100\(\beta\) levels as the outcome. In each of the MR analyses, clumping was used to prune SNPs for linkage disequilibrium (LD) at \(r^2 < 0.001\). When testing the association with S100\(\beta\) as the exposure and Alzheimer’s disease as the outcome, only one of the 154 SNPs with \(P<5 \times 10^{-8}\) (rs8128872; the sentinel variant identified by GCTA-COJO analyses in our GWAS of S100\(\beta\)) remained after LD pruning. The effect estimate for S100\(\beta\) to Alzheimer’s disease was therefore determined using the Wald ratio test (a ratio of effect per risk allele on trait to effect per risk allele on protein levels). An F statistic for the strength of the association between the sentinel SNP and the exposure was calculated using the method: \(F = ((N-k-1) / k) \times (r^2 / (1-r^2)), \) where \(N\) = sample size,
k = number of SNPs and $r^2 = \text{variance explained in S100}\beta$ levels by the genetic instruments. The $r^2$ statistic was calculated as follows: $r^2 = 2 \times \text{MAF} \times (1-\text{MAF}) \times \beta^2$, where $\beta$ = effect size of the SNP and MAF = the effect allele frequency. When testing causal associations with Alzheimer’s disease as the exposure and S100\beta as the outcome, 30 of the 2,357 SNPs with $P<5\times10^{-8}$ remained after LD pruning and 29 were present in the S100\beta summary statistics. Multi-method MR was then performed using the 29 SNPs from the Jansen et al.\textsuperscript{30} summary statistics. As multiple independent variants were identified, a multi-method MR approach was chosen\textsuperscript{29}. Unity between the estimates from these methods indicates that the results are more likely to be robust. The MR Egger approach did not find strong evidence of horizontal pleiotropy present (non-significant MR-Egger intercept).

Colocalisation

Colocalisation analysis can be used to derive the probability that common genetic variants are shared between two phenotypes in a given region of the genome. The colocal package (Version 5.1.0) was used to conduct colocalisation analyses for the sentinel SNP in the S100\beta region and the Jansen et al.\textsuperscript{30} summary statistics for Alzheimer’s disease GWAS. Each dataset was subset to a 200 kb section (upstream or downstream) surrounding the sentinel SNP on chromosome 21. Rare variants with MAF $< 0.01$ and variants with missing MAF were excluded from the analysis. A total of 1,346 variants were included for Alzheimer’s disease and 1,010 variants were included for S100\beta. A single causal variant assumption is made in the analysis that there is one causal variant per trait and the probability of colocalisation between loci can be derived. Four hypotheses were used in Bayesian tests with default priors\textsuperscript{28}, as per the eQTL colocalisation tests, but for the presence of Alzheimer’s disease causal variants in the same region as S100\beta. In addition to the null hypothesis (no causal variants in the region assessed), hypothesis 1 indicated a causal variant was present for Alzheimer’s disease only. Hypothesis 2 indicated that there was a causal variant for S100\beta levels only. Hypothesis 3 indicated that there were independent causal variants for both Alzheimer’s disease and S100\beta levels. Hypothesis 4 indicated that a common variant contributed to both Alzheimer’s disease and S100\beta levels. A posterior inclusion probability (PP) > 0.95 was taken as the threshold for hypothesis testing.

Epigenome-wide association study (EWAS)

DNA methylation data were regressed onto age, sex, DNAm set, DNAm batch, BMI at Wave 2 (kg/m$^2$), the DNAm-based smoking score EpiSmokEr\textsuperscript{29}, four genetic principal components and the measured levels of five immune cells (eosinophils, basophils, lymphocytes, neutrophils and monocytes). EWAS was conducted using OmicS-data-based complex trait analysis (OSCA)\textsuperscript{30}. The MOMENT method was used to test for associations between S100\beta levels and DNAm at individual CpG sites. MOMENT is a mixed-linear-model-based method that is able to account for unobserved confounders and the correlation between distal probes that may be introduced by these confounders. CpG sites were the independent variables and the dependent variable was the S100\beta protein residuals.

Results

Genetic profiling of S100\beta

The linear regression genome-wide association study identified 154 SNPs (Figure 2, full summary statistics are available in the Extended Data) on chromosome 21 that were associated with S100\beta levels at $P<5\times10^{-8}$ (N=769). There was no evidence of genomic inflation (lambda = 0.99; Figure 3). Conditional and joint analysis (GCTA-COJO) resulted in the identification of one independent pQTL rs8128872 (COJO beta = -0.46, SE = 0.05, $P = 3.2\times10^{-10}$) associated with S100\beta levels. The rs8128872 variant was found to explain 9% of the variance in S100\beta levels. The pQTL was a cis variant (located 1,419,889 base pairs downstream of the transcription start site of the S100\beta gene on chromosome 21). There was strong evidence (posterior probability (PP) > 0.95) that two independent causal variants existed for both transcription of S100\beta and S100\beta protein levels in our eQTL analyses for the S100\beta locus (Hypothesis 3: Posterior Probability (PP) = 1.0, Hypotheses 1, 2 and 4 = 0).

Two-sample Mendelian randomisation (MR) was used to test for a causal association between S100\beta serum levels (using our cis GWAS data) and Alzheimer’s disease. As only one significant SNP remained after LD pruning ($P = 76.67$, indicating a strong effect of the instrument on the S100\beta exposure), the causal effect estimate was determined using the Wald ratio. There was no evidence of an effect of S100\beta serum levels on risk of Alzheimer’s disease ($P = 0.95$); (Table 2). Similarly, there was no evidence to suggest that a causal relationship was present between Alzheimer’s disease and serum S100\beta ($P > 0.05$). Colocalisation analyses provided further evidence that the loci for Alzheimer’s disease and S100\beta were not localised together (Hypothesis 2: Posterior Probability (PP) = 0.99, Hypothesis 3: PP = 0.01, Hypotheses 1 and 4 = 0).

Epigenetic profiling of S100\beta

No CpGs were significantly associated ($P<3.6\times10^{-8}$) with S100\beta levels in the EWAS study (N=722) (Figure 2, full summary statistics are available in the Extended Data). The site with the lowest p-value ($P = 5.8\times10^{-10}$) was cg06833709, which is located within the LGII region, known to encode the leucine-rich glioma inactivated 1 protein (known as epitempin). There was no evidence of genomic inflation (lambda = 0.94, Figure 3).

Discussion

We have characterised the genetic and epigenetic profiles of S100\beta, a circulating protein that has been associated with brain inflammation and neurological disease pathology. We identified a genome-wide significant cis-pQTL (rs8128872, $P = 5.0\times10^{-15}$) that was associated with inter-individual variability in circulating S100\beta levels and found evidence that this pQTL was likely to be distinct from the eQTL for S100\beta transcription. Mendelian randomisation suggested no evidence of a causal association between S100\beta and Alzheimer’s disease or vice versa. Furthermore, there were no CpG probes that had epigenome-wide significant associations with S100\beta.
Figure 2. Miami plot of the GWAS (upper panel; N=769) and EWAS (lower panel, N=722) of S100β. Blue lines indicate a suggestive threshold of $P<1\times10^{-5}$; red lines indicate genome-wide thresholds of $P<5\times10^{-8}$ (GWAS) and $P<3.6\times10^{-8}$ (EWAS).

Figure 3. QQ Plots and Genomic Inflation statistics for the GWAS (A) and EWAS (B) of S100β in the LBC1936 sample. Lambda ($\lambda$) values are annotated in each case.

Whereas the study size was modest for our GWAS and EWAS, previous investigations of similar sample sizes have identified genome-wide SNP and epigenome-wide CpG correlates of protein levels\textsuperscript{12-15}. These include proteomic analyses in the age homogeneous Lothian Birth Cohort 1936 sample that we use in our study\textsuperscript{14,15}. Despite this, larger GWAS and EWAS
efforts may help to identify additional loci, which could be used for genetic correlation and more detailed Mendelian randomisation analyses.

Our identification of a single sentinel SNP for S100β suggests that though limited, there is some evidence for genetic regulation of this protein in blood. The cis-pQTL that we identify (rs8128872) adds to the genetic profile of S100β generated in a previous GWAS analysis, which identified two SNPs in the PCNT and HLA regions\(^a\). While our EWAS suggested that there is no epigenome-wide signature of differential DNA methylation at this CpG is associated with S100β and Alzheimer’s disease. Studies implicating S100β as a candidate marker for brain inflammation and neurological disease are often performed in cerebrospinal fluid\(^3,7,8\), which may provide a closer reflection of brain pathology and neurological disease biology is limited when using blood measurements, many blood-based biomarkers have been found to predict and offer insight into Alzheimer’s disease\(^39\). Therefore, approaches that seek to triangulate between blood, cerebrospinal fluid and brain tissues may strengthen the identification of biomarker signals in future.

The lack of evidence for a causal relationship between S100β and Alzheimer’s disease suggests either that 1) S100β levels in the blood are not directly related to the disease, or 2) any associations are more modest than this study is powered to reliably detect. The GWAS sample size was relatively small and it is imperative that our GWAS and Mendelian randomisation results are independently validated by future cohorts that have S100β measurements available. This will further elucidate the likelihood of a causal relationship between S100β and Alzheimer’s disease. Studies implicating S100β as a candidate marker for dementia are often performed in cerebrospinal fluid\(^3,8\), which may provide a closer reflection of brain pathology and comparisons between blood and CSF S100β levels may therefore yield differing conclusions. However, in the Lothian Birth Cohort 1936 sample we use in this study, serum S100β levels have cross-sectionally been associated with poorer general fractional anisotropy\(^9\), a marker of brain ageing that is associated with increased risk of cognitive decline and dementia\(^9\). Given its role as a mediator in inflammatory cascades within the brain\(^9,10\), it is likely that S100β serum levels may be modulated by multiple factors that could be independent of Alzheimer’s disease, or indirect from the dementia-associated pathology occurring in the brain. This may be evidenced by our lack of causal association to Alzheimer’s disease and these pathways should be explored to delineate targets for therapeutic interventions that may alter neuroinflammation through S100β mediation.

There are several limitations to our study. First, the Lothian Birth Cohort 1936 are of European ancestries, have little variation in age and selection biases may exist in this cohort, who are considered to be of higher socioeconomic class to the wider Scottish population\(^10\). Therefore, these findings may not generalise to individuals of different ethnic backgrounds, age profiles or socioeconomic groups, though this also means that we are not largely reliant on statistical adjustment for these confounders. Second, our data are from relatively healthy individuals, none of whom reported a diagnosis of dementia at recruitment. It is plausible that analyses in individuals in specific diagnoses groups may yield differing findings. Finally, while the amount of information regarding brain pathology and neurological disease biology is limited when using blood measurements, many blood-based biomarkers have been found to predict and offer insight into Alzheimer’s disease\(^39\). Therefore, approaches that seek to triangulate between blood, cerebrospinal fluid and brain tissues may strengthen the identification of biomarker signals in future.

**Conclusion**

We have established evidence for modest genetic, but not epigenetic contributions to the levels of S100β, a protein marker for brain inflammation and neurological disease. We found no evidence for a causal relationship between serum S100β and Alzheimer’s disease. Future studies should seek to corroborate these findings across blood, cerebrospinal fluid and brain tissue.

**Data availability**

**Underlying data**

Lothian Birth Cohort 1936 data are not publicly available due to them containing information that could compromise participant consent and confidentiality. Lothian Birth Cohort 1936 data are available on request from the Lothian Birth Cohort Study, University of Edinburgh.

If you are interested in working with the Lothian Birth Cohort 1936 data, you must complete a Data Request Form, indicating the variables you wish to access from the Data Dictionaries.

### Table 2. Mendelian Randomization summary statistics for two-sample, bidirectional tests between S100β serum levels and Alzheimer’s disease.

<table>
<thead>
<tr>
<th>Exposure</th>
<th>Outcome</th>
<th>Method</th>
<th>N SNP</th>
<th>Odds Ratio</th>
<th>Beta</th>
<th>SE</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>S100β</td>
<td>Alzheimer’s disease</td>
<td>Wald ratio</td>
<td>1</td>
<td>1.0002</td>
<td>3×10^-4</td>
<td>4.9×10^-3</td>
<td>0.95</td>
</tr>
<tr>
<td>Alzheimer’s disease</td>
<td>S100β</td>
<td>MR Egger</td>
<td>29</td>
<td>N/A</td>
<td>0.33</td>
<td>0.51</td>
<td>0.52</td>
</tr>
<tr>
<td>Alzheimer’s disease</td>
<td>S100β</td>
<td>Weighted median</td>
<td>29</td>
<td>N/A</td>
<td>-0.21</td>
<td>0.53</td>
<td>0.69</td>
</tr>
<tr>
<td>Alzheimer’s disease</td>
<td>S100β</td>
<td>Inverse variance weighted</td>
<td>29</td>
<td>N/A</td>
<td>-0.53</td>
<td>0.38</td>
<td>0.16</td>
</tr>
<tr>
<td>Alzheimer’s disease</td>
<td>S100β</td>
<td>Simple mode</td>
<td>29</td>
<td>N/A</td>
<td>-1.02</td>
<td>0.98</td>
<td>0.31</td>
</tr>
<tr>
<td>Alzheimer’s disease</td>
<td>S100β</td>
<td>Weighted mode</td>
<td>29</td>
<td>N/A</td>
<td>-0.03</td>
<td>0.49</td>
<td>0.96</td>
</tr>
</tbody>
</table>

\(^a\)Underlying data

Whole genome association results were available on request from the Lothian Birth Cohort 1936 data are not publicly available due to them containing information that could compromise participant consent and confidentiality. Whole genome association results were available on request from the Lothian Birth Cohort Study, University of Edinburgh.
Data Dictionaries and Data Request Forms are freely accessible at the following website: https://www.ed.ac.uk/lothian-birth-cohorts/data-access-collaboration. Completed forms must then be sent to Dr Simon Cox for approval.

All code is available at the following Gitlab repository: https://github.com/DanniGadd/GWAS-and-EWAS-of-S100-. Data Dictionaries and Data Request Forms are freely accessible at the following website: https://www.ed.ac.uk/lothian-birth-cohorts/data-access-collaboration.

Extended data

This project contains the following extended data files:

- s100b_EWAS_output.mlma (EWAS summary statistics for the epigenetic association study of S100β levels, for each of the 459,309 CpG probes tested)

- s100b_GWAS_output.txt (GWAS summary statistics for the genetic association study of S100β levels, including the 154 variants that had $P < 5 \times 10^{-8}$)

References


Reporting guidelines

Data are available under the terms of the Creative Commons Zero “No rights reserved” data waiver (CC0 1.0 Public domain dedication).

Ethical approval and consent
Ethical approval for LBC1936 was obtained from the Multi-Centre Research Ethics Committee for Scotland (MREC/01/0/56) and the Lothian Research Ethics committee (LREC/1998/4/183; LREC/2003/2/29). All participants provided written informed consent and the study was performed in accordance with the Helsinki declaration.

Acknowledgements
The authors thank the participants of the Lothian Birth Cohort 1936 study for their valuable contributions to this work.


6.3 Conclusion

This was the first study to characterise the DNAm and genetic architectures associated with S100β within the same population. It was also the first study to test for the presence of a causal association between S100β and Alzheimer’s dementia. The GWAS identified 154 SNPs, with one sentinel genome-wide significant cis-pQTL (rs8128872, \(P = 5.0 \times 10^{-17}\)) that was associated with serum S100β levels. When using this variant as an instrument in Mendelian randomisation, there was no evidence of a causal relationship – in either direction – between S100β and Alzheimer’s dementia. There were no associations in the MWAS of S100β that had \(P < 3.6 \times 10^{-8}\). This suggests that there is not a strong, blood-based DNAm profile associated with serum S100β levels.

Larger sample sizes will aid in resolving the full molecular architecture of S100β. A subsequent study has since used the GWAS summary statistics provided in my study to perform a Mendelian randomisation study assessing causal relationships between S100β and a more extensive set of neurological disease traits. This study has replicated my findings in relation to Alzheimer’s and identified a potentially-causal association from S100β levels to major depression.

In this study, I provided insight into the epi/genetic architectures of S100β. One limitation with approaches such as this is that a singular protein is honed in on, which can miss information represented by the circulating proteome more widely. Therefore, in the next chapter, I take a different approach – starting with measures of 4,235 protein levels from the blood. I perform MWAS on each protein measure and assess those that associate with brain health traits. This represents a more data-driven approach to exploring the molecular signatures that be important factors for brain health across the circulating proteome.
7 Characterisation of DNAm profiles associated with protein markers of brain health

7.1 Introduction

Although it is difficult to tease apart causation when mapping pQTM, they identify patterns of DNAm across the genome that are relevant to our molecular understanding of protein levels (as discussed in Section 2.2.4). The largest MWAS of the circulating proteome was conducted using ~1000 plasma protein levels measured on the SomaScan V3 platform and identified 98 pQTM. Given that the SomaScan V5 platform measures over 4,000 plasma protein levels, the DNAm signatures of many proteins are uncharacterised. Studies have begun to map the proteome-wide signatures of brain health traits such as APOE status, cognitive decline and dementia. Recent work identified 15 blood protein markers that could stratify future risk of Alzheimer’s dementia (HR ranging from 1.08 to 1.64). These studies suggest that while blood is proximal to brain, circulating protein signatures (either produced in the bloodstream, or secreted from the brain and other organs) can stratify brain health. Understanding the protein markers of brain health in a population that does not have neurodegenerative diagnoses may also highlight proteins that maintain or perturb brain health during healthy ageing. Resolving DNAm patterns associated with such proteins may provide insight into candidate pathways associated with brain health.

In this chapter, I first use OSCA software to conduct MWAS studies of 4,235 proteins in 774 individuals from the GS cohort. I also assess the relationship between each of the 4,235 proteins and 15 brain health traits in linear mixed models, using a larger sample of GS where the MWAS population is a nested subset (N=1,065). I adjust these analyses for the relatedness structure between individuals. I then integrate these findings to identify pQTM signatures specific to protein markers that associate with brain health traits. I explore the pathways of interest that these signatures point towards that may be of value in understanding and stratifying brain health.

This study was published in Nature Communications in 2022 and is included in full in Section 7.2. I led the conception, analyses and manuscript publication for this work. The full supplementary files and R code used for this study are made available at: https://github.com/DanniGadd/s1888864_Supplementary_Material/tree/main/Chapter_7.
7.2 Integrated methylome and phenome study of the circulating proteome reveals markers pertinent to brain health
Integrated methylome and phenome study of the circulating proteome reveals markers pertinent to brain health

Characterising associations between the methylome, proteome and phenome may provide insight into biological pathways governing brain health. Here, we report an integrated DNA methylation and phenotypic study of the circulating proteome in relation to brain health. Methylome-wide association studies of 4058 plasma proteins are performed (N = 774), identifying 2928 CpG-protein associations after adjustment for multiple testing. These are independent of known genetic protein quantitative trait loci (pQTLs) and common lifestyle effects. Phenome-wide association studies of each protein are then performed in relation to 15 neurological traits (N = 1,065), identifying 405 associations between the levels of 191 proteins and cognitive scores, brain imaging measures or APOE e4 status. We uncover 35 previously unreported DNA methylation signatures for 17 protein markers of brain health. The epigenetic and proteomic markers we identify are pertinent to understanding and stratifying brain health.
The health of the ageing brain is associated with risk of neurodegenerative disease\(^1\). Relative brain age—a measure of brain health calculated using multiple volumetric brain imaging measures—has recently been shown to predict the development of dementia\(^2\). Structural brain imaging and performance in cognitive tests are well-characterised markers of brain health\(^3\), which clearly associate with potentially modifiable traits such as body mass index (BMI), smoking and diabetes\(^4\). Understanding the interplay between environment, biology and brain health may therefore inform preventative strategies.

Multiple layers of omics data indicate the biological pathways that underlie phenotypes. Proteomic blood sampling can track peripheral pathways that may impact brain health, or record proteins secreted from the brain into the circulatory system. Although proteome-wide characterisation of cognitive decline and dementia risk\(^5\)–\(^10\) have been facilitated at large scale by SOMAscan\textsuperscript{®} protein measurements, there is a need to further integrate omics to characterise brain health phenotypes. Epigenetic modifications to the genome record an individual's response to environmental exposures, stochastic biological effects, and genetic influences. Epigenetic changes include histone modifications, non-coding RNA, chromatin remodelling, and DNA methylation (DNAm) at cytosine bases, such as 5-hydroxymethylcytosine. These are implicated in changes to chromatin structure and the regulation of pathways associated with neurological diseases\(^11\),\(^12\). However, DNAm at cytosine-guanine (CpG) dinucleotides is the most widely profiled blood-based epigenetic modification at large scale.

Modifications to DNAm at CpG sites play differential roles in influencing gene expression at the transcriptional level\(^13\). Additionally, DNAm accounts for inter-individual variability in circulating protein levels\(^14\)–\(^16\). Recently, through integration of DNAm and protein data, we have shown that epigenetic scores for plasma protein levels—known as Episcores—associate with brain morphology and cognitive ageing markers\(^8\) and predict the onset of neurological diseases\(^17\). These studies highlight that while datasets that allow for integration of proteomic, epigenetic and phenotypic information are rarely-available, they hold potential to advance risk stratification. Integration may also uncover candidate biological pathways that may underlie brain health.

Associations between protein levels and DNAm at CpGs are known as protein quantitative trait methylation loci (pQTLs) and can be quantified by methylome-wide association studies (MWAS) of protein levels. The largest MWAS of protein levels to date assessed 1123 SOMAmers in 1065 individuals in Generation Scotland cohort using EPIC array DNAm at 772,619 CpG sites. We then identify which of the 4058 protein levels associate with one or more of 15 neurological traits (seven structural brain imaging measures, seven cognitive scores and APOE e4 status) in 1065 individuals from the same cohort where the pQTM data are a nested subset. By overlapping these datasets, we probe the epigenetic signatures of proteins that are related to brain health. For these signatures, we map potential underlying genetic components and chromatin interactions that may play a role in protein level regulation.

**Results**

**Methylome-wide studies of 4058 plasma proteins**

We conducted MWAS to test for pQTM associations between 772,619 CpG sites and 4058 circulating protein levels (corresponding to 4235 SOMAmers; Supplementary Data 1). The MWAS population included 774 individuals from Generation Scotland (mean age 60 years [SD 8.8], 56% Female; Supplementary Data 2). 143 principal components explained 80% of the cumulative variance in the 4235 measurements (Supplementary Fig. 1 and Supplementary Data 3). A threshold for multiple testing based on these components was applied across all MWAS (\(P < 0.05/(143 \times 772,619) = 4.5 \times 10^{-7}\)).

In our basic model adjusting for age, sex and available genetic pQTL effects from Sun et al.\(^20\) 238,245 pQTM (2107 cis and 236,138 trans; representing 0.005% of tested associations) had \(P < 4.5 \times 10^{-10}\) (Supplementary Data 4). In our second model that further adjusted for Houseman-estimated white blood cell proportions\(^21\), there were 3,213 associations (453 cis and 2760 trans) that had \(P < 4.5 \times 10^{-10}\) (Supplementary Data 5). Smoking status and BMI are known to have well-characterised DNAm signatures\(^22\)–\(^23\); fully-adjusted models were

---

**Fig. 1** Methylome and phenotype study of the plasma proteome in relation to brain health study design. A total of 4058 plasma proteins (corresponding to 4235 SOMAmers) were measured in 1065 individuals in Generation Scotland. A methylome-wide association study (MWAS) of each plasma protein level was conducted in 774 individuals that represented a nested subset of the full sample that had DNAm measurements available. A phenome-wide protein association study between DNAm at the immune-associated locus NLRCS and seven immune-related proteins (\(P < 2.5 \times 10^{-7}\)). This suggested that DNAm not only reflects variability in the proteome but is closely related to chronic systemic inflammation. Hillary et al. have also assessed epigenetic signatures for 281 SOMAmers protein measurements that were previously associated with Alzheimer’s disease, in the Generation Scotland cohort that we utilised in this study\(^7\). However, proteome-wide assessment of pQTM has not been tested against a comprehensive spectrum of brain health traits.

Here, we conduct an integrated methylome- and phenome-wide assessment of the circulating proteome in relation to brain health (Fig. 1), using 4058 protein level measurements (Annotation information provided in Supplementary Data 1). We characterise CpG-protein associations (pQTM) for these proteins in 774 individuals from the Generation Scotland cohort using EPIC array DNAm at 772,619 CpG sites. We then identify which of the 4058 protein levels associate with one or more of 15 neurological traits (seven structural brain imaging measures, seven cognitive scores and APOE e4 status) in 1065 individuals from the same cohort where the pQTM data are a nested subset. By overlapping these datasets, we probe the epigenetic signatures of proteins that are related to brain health. For these signatures, we map potential underlying genetic components and chromatin interactions that may play a role in protein level regulation.

---

https://doi.org/10.1038/s41467-022-32319-8
therefore further adjusted for these factors. There were 2,928 associations (451 cis and 2,477 trans) in the fully-adjusted models (Supplementary Data 6). 2,847 pQTM associations were significant in all models. There were 191 unique proteins with associations in the fully-adjusted models, corresponding to 195 SOMAmer measurements (two SOMAmers were present for CLEC11A, GOLM1, ICAM5 and LRP11). Figure 2 summarises these findings. Genomic inflation statistics for these 195 SOMAmer measurements (fully-adjusted MWAS) are presented in Supplementary Data 7. In a sensitivity analysis, restriction of the threshold for cis pQTMs from 10 Mb to 1 Mb from the transcription start site of the gene encoding the protein yielded 409 cis pQTMs (a reduction of 42 pQTMs) in the fully-adjusted MWAS. A summary of known pQTLs and a record of whether these were available for adjustment is provided in Supplementary Data 8. Characterising the genomic location of the findings, 46% of cis and 29% of trans pQTMs in the fully-adjusted MWAS involved CpGs positioned in either a CpG Island, Shore or Shelf (Supplementary Data 6).

Pleiotropic pQTM associations in the fully-adjusted MWAS
Pleiotropy was observed for both CpG sites and protein levels (Fig. 3). Nineteen proteins had 10 or more pQTMs in the fully-adjusted MWAS (Supplementary Data 9). Of the 2,928 pQTMs in the fully-adjusted MWAS, 987 involved Pappalysin-1 (PAPPA) and there were a further 1,116 pQTMs that involved the Proteoglycan 3 Precursor protein. The remaining 825 pQTMs involved 189 unique protein levels, with 434 cis and 391 trans associations (Fig. 2). Principal components analyses indicated high correlations between CpGs associated with the pleiotropic proteins PAPPA and PRG3, whereas the CpGs involved in the remaining 825 pQTMs were largely uncorrelated (Supplementary Fig. 2). pQTM frequencies for the 1,837 unique CpGs selected in the fully-adjusted models, with their respective genes and EWAS catalog lookup of epigenome-wide significant (P < 3.6 × 10^{-8}) phenotypic associations is presented in Supplementary Data 10. Of these CpGs, sites within the NLRC5, SLC7A11 and PARP9 gene regions exhibited the highest levels of pleiotropy (Fig. 3).

The pleiotropic findings for PAPPA and cg07839457 (NLRC5 gene) replicated previous MWAS results from Zaghloul et al. (944 individuals, with 1,123 protein SOMAmers). Of the 98 pQTMs identified by Zaghloul et al., 81 were comparable (both the protein and CpG sites from the 98 pQTMs were available across both MWAS). Of these 81 pQTMs, 26 replicated at our significance threshold (P < 4.5 × 10^{-10}) with the same direction of effect, a further 16 replicated at the epigenome-wide significance threshold (P < 2.2 × 10^{-4}).

**Fig. 2 | Methylome-wide association studies (MWAS) of 4058 plasma proteins.** a Summary of MWAS results for 4058 protein levels in Generation Scotland (N = 774). The number of protein quantitative trait methylation loci (pQTMs) that had P < 4.5 × 10^{-10} (Bonferroni threshold for multiple testing adjustment) in the basic, white blood cell proportion (WBC)-adjusted and fully-adjusted models. Cis associations (purple) and trans associations (green) are summarised for each model. Covariates used to adjust DNAm are described for each model. Normalised protein levels were adjusted for age, sex, 20 genetic principal components (PCs), protein quantitative trait loci (pQTLs) and technical variables and scaled to have a mean of 0 and standard deviation of 1. Results were generated through linear regression models. Full summary statistics with full P values can be accessed in Supplementary Data 6. Created with BioRender.com. b Flow diagram showing the distinction between the highly pleiotropic PAPPA and PRG3 protein pQTMs and the 825 pQTMs that involved the levels of a further 189 proteins. TSS: transcriptional start site of the gene. The 434 cis pQTMs (purple) lay on the same chromosome and ≤10 Mb from the transcriptional start site (TSS) of the gene protein, whereas the 391 trans pQTMs (green) lay >10 Mb from the TSS of the gene protein or on a different chromosome. Created with BioRender.com. c Genomic locations for 825 of the 2,928 fully-adjusted pQTMs, excluding highly pleiotropic associations for PAPPA and PRG3 protein levels, with cis pQTMs in purple and trans pQTMs in green. Chromosomal location of CpG sites (x-axis) and protein genes (y-axis) are presented. A list of the full association counts for each protein and CpG site can be found in Supplementary Data 8, 9.
wide significance threshold ($P < 3.6 \times 10^{-8}$)\cite{25} and a further 39 replicated at nominal $P < 0.05$ (Supplementary Data 11 and Supplementary Fig. 3). When accounting for 26 pQTM$\mathrm{s}$ that were previously reported by Zaghlool et al. and 10 pQTM$\mathrm{s}$ that were previously reported by Hillary et al.\cite{14,19,28} 2892 of the 2928 fully-adjusted pQTM$\mathrm{s}$ were previously unreported. Of these 2892 pQTM$\mathrm{s}$, 1109 involved the levels of 41 proteins that were measured by Zaghlool et al. (973 pQTM$\mathrm{s}$ for PAPPA and 136 additional pQTM$\mathrm{s}$ for the levels of 40 proteins), whereas 1783 pQTM$\mathrm{s}$ involved the levels of proteins that were previously unmeasured (1116 pQTM$\mathrm{s}$ for PRG3 and 667 further pQTM$\mathrm{s}$ for 148 proteins).

**Proteome associations with brain health phenotypes**

We next conducted a proteome-wide association study of brain health characteristics (protein PhEWAS of brain imaging, cognitive scoring and APOE e4 status, alongside age and sex; Fig. 4). Distribution plots for the seven cognitive scores and seven brain imaging phenotypes are presented in Supplementary Figs. 4, 5. A maximum sample of 1065 individuals was available (mean age 59.9 years [SD 9.6], 59% Female; Supplementary Data 2); all 774 individuals from the pQTM study were included in these analyses. A threshold for multiple testing adjustment was calculated based on 143 independent components that explained >80% of the 4235 SOMAmer levels (Supplementary Data 3 and Supplementary Fig. 1). This equated to $P < 0.05/(143) = 3.5 \times 10^{-4}$. The levels of 587 plasma proteins were associated with age and 545 were associated with sex, with 222 proteins common to both phenotypes (Supplementary Data 12). When comparable associations from three studies (with $N > 1000$) were tested\cite{20,26,27} 97% of age and 98% of sex associations replicated in one or more of studies (Supplementary Data 12).

There were 191 unique protein markers that had a total of 405 associations with brain health characteristics (Supplementary Fig. 6 and Fig. 4a). These consisted of 95 brain imaging (Supplementary Data 2) and 100 cognitive scores (Supplementary Data 3). When compared to a previous proteome-wide association study\cite{28} that investigated 8682 pQTM$\mathrm{s}$ from 3000 individuals using the same threshold, the current study identified 231 additional pQTM$\mathrm{s}$ that were associated with brain health characteristics, and 71 of these were also associated with both age and sex (Supplementary Data 13).

In conclusion, this study provides a comprehensive understanding of the proteome-wide associations with brain health characteristics, offering potential insights into the complex interplay between proteins and brain health.
Data 13), 296 cognitive test score (Supplementary Data 14) and 14 APOE e4 status (Supplementary Data 15) associations. Supplementary Data 16 stratifies these associations by direction of effect and Supplementary Data 17 provides full summary statistics for all 405 associations. Of the seven brain morphology traits, Relative Brain Age and General Fractional Anisotropy (gFA) had the largest number of associations, with 24 and 22 protein markers identified, respectively. Of the cognitive score traits, Processing Speed and General Cognitive Ability scores were associated with the highest number of protein markers (102 and 73, respectively). The underlying data for the 14 APOE e4 status associations are plotted in Supplementary Fig. 7.

Stratifying the 405 associations by direction of effect revealed that the majority (89%) of associations involved higher levels of the proteins that were associated with less favourable brain health (Supplementary Data 16). Eighty-seven of the 405 associations involved protein levels that were associated with more favourable brain health; this signature included the levels of SLITRK1, NCAN and COL11A2. Higher levels of ASB9, RBL2, HEXB and SMPD1 were associated with poorer brain health. Protein interaction network analyses for the genes corresponding to the 191 protein markers (Supplementary Fig. 8) indicated that many of the proteins clustered together, implying shared underlying functions. An inflammatory cluster including CRP, ITIH4, C3, C5, COL11A2 and SIGLEC2 was present and higher levels of these markers were associated with poorer brain health outcomes. Gene set enrichment analyses on the 191 genes corresponding to the protein markers (Supplementary Fig. 9) supported the link between many of the proteins associated with poorer brain health and the innate immune system, while also implicating extracellular matrix, lysosomal, metabolic and additional inflammatory pathways. Tissue expression profiles of the 191 genes (Supplementary Fig. 10) indicated that many of the markers were expressed in non-neurological tissues; however, some proteins were expressed in nervous tissues. Markers

**Fig. 4 | Phenome-wide associations studies (PheWAS) of 4058 plasma proteins and brain health.** a Number of protein marker associations with $P < 3.5 \times 10^{-4}$ for each of the 15 traits related to brain health in the phenome-wide protein association studies (protein PheWAS). These studies included a maximum sample of 1065 individuals with protein measurements from Generation Scotland and tested for associations between 15 phenotypes and the levels of 4058 plasma proteins via linear mixed effects regression. Cognitive score (green), brain imaging (light blue) and APOE e4 status (dark blue) associations are summarised. Full summary statistics for the 405 associations with $P$ values are presented in Supplementary Data 17. All associations were generated through linear regression and were adjusted for multiple testing correction. b Heatmap of standardised beta coefficients for 77 of the 405 protein PheWAS associations ($P < 3.5 \times 10^{-4}$ indicated by an asterisk). These include three proteins that had associations with both APOE e4 status and one or more cognitive scores, in addition to 22 proteins that had associations with both a brain imaging measure and a cognitive score. Negative and positive beta coefficients are shown in blue and red, respectively. A heatmap describing the full 405 associations for APOE e4 status, cognitive scores and brain imaging measures is available in Supplementary Fig. 6. All associations were generated through linear regression and were adjusted for multiple testing correction.
such as ASB9 and NCAN were found to be consistently identified across multiple brain imaging traits as markers of poorer and better brain health, respectively (Supplementary Data 16). While many of the associations for brain imaging measures identified proteins that were distinct from those found for cognitive scores and APOE e4 status, 22 protein markers were associated with both a cognitive score and a brain imaging trait (Fig. 4b and Supplementary Data 18). A principal component analysis of the 22 protein levels was conducted. The first five components had an eigenvalue >1 and a cumulative variance of >80% was explained by the first 10 components. These are both commonly-used thresholds for deciding how many principal components to retain\(^6\) (Supplementary Fig. 11). Three APOE e4 status markers (ING4, APOB and CRP) were also associated with cognitive scores (Fig. 4b).

### Replication of protein PheWAS associations

Six of the 14 APOE e4 status associations replicated previous SOMAmer protein findings (\(N = 4785\) and \(N = 227\))\(^6\), and eight previously unreported relationships involved NEFL, ING4, PAF, MENT, TMCC3, CRP, FAM20A and PEFI. Several of the markers for cognitive function were identified in previous work relating Olink proteins to cognitive function (such as CPM)\(^7,8\) and work that characterised SOMAmer proteins for their potential involvement in Alzheimer’s disease (such as SVEP1)\(^9\). No studies have performed SOMAmer-based, whole prostate PheWAS studies of the brain imaging and cognitive score traits we have profiled in a healthy ageing population that were not enriched for neurodegenerative diseases. However, replication of associations from several studies\(^9,10,11\) was found for a small subset of associations (Supplementary Data 19).

### Integration of the brain health proteome with our pQTM dataset

Differential DNAm signatures were explored for the 191 protein markers that had \(P < 3.5 \times 10^{-4}\) in associations with either cognitive scores, brain imaging measures or APOE e4 status in the protein PheWAS. Of the 191 proteins, 17 had pQTMs in the fully-adjusted MWAS. Higher levels of 15 of these proteins were associated with poorer brain health, while AMY2A and CST5 were associated with more favourable brain health. There were a total of 35 pQTMs involving 31 unique CpGs that were located within 20 distinct genes (Supplementary Data 20), with 15 trans (Fig. 5) and 20 cis associations. All pQTMs were previously unreported. The 20 cis pQTMs involved the levels of CHI3L1, IL18R1, SIGLEC5, OLFM2, UGDH, CRHBP, AMY2A and CFIHR1 proteins. The trans pQTMs involved the levels of SCUBE1, RBL2, TNFRSF1B, CST5, HEBX, ACY1, CRTAM, SMPDI and RB5P proteins.

Of the 20 cis pQTMs, 11 involved CpGs in different genes to the protein-coding gene on the same chromosome, whereas the remaining 9 pQTMs involved CpGs located within the protein-coding gene. Several CpG sites were associated with multiple protein levels in the trans pQTMs (Fig. 5). DNAm at site cg06690545 in the SLCA21 gene was associated with RB5P, ACY1 and SCUBE1 levels. The cg12924350 site in the CHPT1 gene was associated with HEBX and SMPDI levels. The cg07335657 site in the NLRCS gene was associated with the levels of CRTAM and TNFRSF1B. There was also a protein that had several trans associations with multiple CpG sites; pQTMs were identified between circulating RBL2 levels and cg0132052, cg0539861, cg18481976, cg192774008 and cg18481976, with the NEK/ITIH3/ITIH gene region of chromosome 3.

### Functional mapping of neurological pQTMs

A lookup that integrated information from the GoDMC and eQTLGen databases assessed whether pQTMs were partially driven by an underlying genetic component. This identified methylation quantitative trait loci (mQTLs) for CpGs that were associated with CHI3L1, IL18R1 and SIGLEC5 levels and were also expression quantitative trait loci (eQTLs) for the respective proteins (Supplementary Data 20). Further visual inspection of the distributions for the 35 pQTMs indicated that trinodal distributions—suggestive of unaccounted SNP effects—were present for CpGs involved in seven of the pQTMs (Supplementary Fig. 12).

Tissue expression profiles for the 33 genes that were linked to either CpGs or proteins in the 35 neurological pQTMs were summarised in Supplementary Fig. 13. Gene set enrichment for these 33 genes identified enrichment for immune effector pathways in a subset of 11 genes, whereas a cluster of four genes (SMPDI, HEBX, AMI24 and AMI20) were enriched for amylase and hydrolase activity (Supplementary Fig. 14).

Of the 35 pQTMs, seven had CpGs that were located in either a CpG shore or Shelf position and there were 13 that were located either 1500 bp or 200 bp from the TSS of the protein-coding gene (Supplementary Data 20). Fifteen pQTMs involved CpGs that were located in the gene body and 7 were located in either the first exon or UTR regions (Supplementary Data 20).

Promoter-capture Hi-C and ChIP-sequencing integration were used to assess the interactions and chromatin states of our pQTMs and associated CpG loci. This analysis focused on 11 of the 20 cis pQTMs that involved CpGs on the same chromosome as the protein-coding gene, but was located in a different gene. Mapping information is provided for the seven cis pQTMs in Supplementary Figs. 15–21. In all instances, we found evidence of spatial co-localisation of these genes using promoter-capture Hi-C data from brain hippocampal tissue. We attempted to contextualise these sites further with ChIP-seq (ENCODe project) analyses of active chromatin marks H3K27ac and H3K4me1 and repressive chromatin H3K4me3 and H3K27me3 in both peripheral blood mononuclear cells (PBMCs) and brain hippocampal. ChIP-seq data suggested that in many instances there were shared regulatory regions that existed across both blood and hippocampal samples that were hubs for local promoter interactions. For example, promoter loops were found linking the S100Z and CRHBP genes, with a signature of activating (H3K4me1 and H3K27ac) and silencing (H3K27me1 and H3K4me3) marks (normally considered bivalent chromatin) that may form the basis for shared regulation of this gene locus.

### Discussion

We have conducted a large-scale integration of the circulating proteome with indicators of brain health and blood-based DNA methylation. We characterised 191 protein markers that were associated with either brain imaging measures, cognitive scores or APOE e4 status in an ageing population. We also report methylome-wide characterisations for the SOMAmer panel V.4 (4038 protein measurements) in a nested subset of this population. By overlapping these datasets, we uncovered 35 methylation signatures for 17 protein markers of brain health. We delineated pQTM CpGs that had evidence of underlying genetic influence and characterised the potential for chromatin interactions for genes involved in cis pQTMs. As this population consists of older individuals that were not enriched for neurodegenerative diseases, the markers we identify are likely indicators of healthy brain ageing.

Many of the 191 proteins identified in the protein PheWAS were part of inflammatory clusters with shared functions in acute phase response, complement cascade activity, innate immune activity and cytokine pathways. Tissue expression analyses suggested that a large proportion of the 191 protein markers were not expressed in the brain; this supports work suggesting that sustained peripheral inflammation influences general brain health\(^11\),\(^13\),\(^14\) and accelerates cognitive decline\(^15\),\(^16\). However, a subset of proteins were expressed in the central nervous system. Given that leakage at the blood-brain-barrier interface has been hallmarkd as a part of healthy brain ageing\(^17\),\(^18\), there is a possibility that brain-derived proteins may enter the bloodstream as biomarkers. SLIT and NTRK-like Family Member 1 (SLITRK1),
Neurocan (NCAN) and IgLON family member 5 (IGLON5) were examples of proteins expressed in brain for which higher levels associated with either larger grey matter volume, larger whole brain volume, or higher general fractional anisotropy. SLITRK1 localises at excitatory synapses and regulates synapse formation in hippocampal neurons. NCAN is a component of neuronal extracellular matrix and is linked to neurite growth. IGLON5 has been implicated in maintenance of blood–brain barrier integrity and an anti-IGLON5 antibody disease involves the deterioration of cognitive health. Taken together, the protein markers identified in the PheWAS may, therefore, reflect pathways that could be targeted to improve brain health.

Integration of our fully-adjusted protein MWAS dataset revealed 35 associations between DNAm and 17 protein markers of brain health (Fig. 6; Supplementary Data 20). All 35 associations were previously unreported. While this study is focused on blood DNAm—limiting generalisation to brain DNAm—many of the 35 pQTMs involved CpGs and proteins that have been previously implicated in neurological processes. DNAm at site cg06690548 (located in the SLC7A11 gene) was of particular interest; differential DNAm at this CpG in blood has been identified as a causal candidate for Parkinson’s disease (N > 900 cases and N > 900 controls). Xc- is the cystine-glutamate antiporter encoded by SLC7A11, which facilitates glutamatergic transmission, oxidative stress defence and microglial response in the brain and is a target for the neurodegeneration-associated environmental neurotoxin β-methylamino-L-alanine. Analyses in the wider Generation Scotland cohort suggests that cg06690548 is a site associated with alcohol consumption. The proteins associated with cg06690548 in the subset of this cohort that we assessed (ACY1, SCUBE1 and RBP5)

Fig. 5 | Trans pQTMs involving protein markers of brain health. Circular plot showing 15 trans pQTM associations between DNAm at 11 CpG sites and the levels of nine protein markers of brain health that had $P < 4.5 \times 10^{-10}$ (Bonferroni threshold for multiple testing adjustment) in the fully-adjusted MWAS. Chromosomal positions are given on the outermost circle. Details of the full set of 35 pQTMs for protein markers of brain health are provided in Supplementary Data 20 with $P$ values. Results were generated through linear regression models.
have known links to liver function. DNA methylated at cg06690548 in blood has also been recently implicated in the largest methylome-wide association study (MWAS) of amyotrophic lateral sclerosis (ALS) to date (6763 cases, 2943 controls). Given that ACY1, SCUBE1 and RBP5 were markers for either lower processing speed and higher relative brain age, the CpG sites we identify in this study—such as cg06690548—may be important plasma markers for mediation of environmental risk on brain health that merit further exploration. cg06690548 lies within the first intron of SLC7A11, indicating that this site is of potential functional significance.

The presence of NLRC5-associated CpGs and various other inflammatory proteins in our neurological protein pQTM suggests that the methylome may capture an inflammatory component of brain health. Many of the genes corresponding to CpGs and proteins involved in the 35 pQTM were enriched for immune effector processes and were not expressed in brain. However, some genes did show evidence for brain-specific expression, such as acid sphingomyelinase (SMPD1) and Hexosaminidase Subunit Beta (HEXB). The HEXB and SMPD1 proteins associated with DNA methylation at cg11294350 (in the CHPT1 gene), are involved in neuronal lipid degradation in the brain and have been associated with the onset of a range of neurodegenerative conditions. RBL2 is another protein that had partial expression signals across brain regions; the NEK4/ITIH3/ITIH1 region was the location for five CpGs with differential DNA methylation linked to RBL2 levels. This region is implicated in schizophrenia and bipolar disorder by several large-scale, genome-wide association studies (GWAS). Similarly, the RBL2 locus has been associated with intelligence.

Fig. 6 | Exploration of trans pQTM for protein markers of brain health. a Three trans associations with the CpG site cg06690548 in the SLC7A11 gene, which encodes a synaptic protein that is involved in glutamate transmission and oxidative stress. cg06690548 has been implicated in methylome-wide studies of Parkinson’s disease and Amyotrophic lateral sclerosis (ALS) risk. b Five trans associations between CpGs in the ITIH3/ITIH1/NEK4 region on chromosome 3 and the levels of RBL2, which was associated with lower Global Grey Matter Volume. The RBL2 gene has been implicated in genome-wide association studies (GWAS) of cognitive ability, intelligence and educational attainment. The ITIH3/ITIH1/NEK4 region has been implicated in GWAS of Schizophrenia and Bipolar disorder. c Two trans associations between DNA methylation at cg11294350 in the CHPT1 gene and two proteins with lysosomal-associated function (SMPD1 and HEXB) that were associated with higher Relative Brain Age and lower General Fractional Anisotropy. Associations with a positive beta coefficient are denoted as red connecting lines, whereas associations with a negative beta coefficient are denoted as blue connecting lines. The full 35 pQTM for protein markers of brain health (15 trans and 20 cis) can be found in Supplementary Data 20. All associations were generated using linear regression and were adjusted for multiple testing. Created with BioRender.com.
cognitive function and educational attainment in GWAS (n > 260,000 individuals)⁶⁵,⁶⁶. Given that this study utilised CpGs from the Illumina EPIC array, 15 of the 31 unique CpGs did not have mQTL characterisations in public databases, which primarily comprise results from the earlier 450 K array. However, our plots showing pQTM associations suggested that for several CpGs (such as cg12945530 that associated with SMPD1 and HEXB), there may be a partial genetic component influencing DNAm. As mQTLs tend to explain 15–17% of the additive genetic variance of DNAm⁶⁸, it is possible that the signals we isolate in these instances are partially driven by genetic loci, but are also likely driven by unmeasured environmental and biological influences. In the case of SIGLEC5, IL18R1 and CHI3L, mQTLs were identified that were also eQTLS, providing evidence that mQTLs for these CpG sites were possible regulators of protein expression.

Integration of promoter-capture Hi-C chromatin interaction and ChIP-seq databases⁷⁰ provided evidence for long-range interaction relationships for cis pQTMs with CpGs in different gene regions that are proximal to the protein-coding gene of interest. This suggests that in such instances, the pQTMs may reflect regulatory relationships in the 3-dimensional genomic neighbourhood. The pQTMs therefore direct us towards pathways that can be tested in experimental contexts that advance our understanding of disease. Positional information suggested that many CpGs involved in the 3-dimensional genomic neighbourhood. The pQTMs therefore direct us towards pathways that can be tested in experimental contexts that advance our understanding of disease. Positional information suggested that many CpGs involved in neurological pQTMs lay within 1500 bp of the TSS of the respective protein-coding gene. While positional information of CpGs is thought to infer whether DNAm is likely to play a role in the expression regulation of nearby genes, this is still somewhat disputed. Some studies suggest that transcription factors regulate DNAm⁷¹ and differential methylation at gene body locations predicts dosage of functional genes⁷². Additionally, the DNAm signatures of proteins we quantify represent widespread differences across blood cells that are related to circulating protein levels and are therefore not derived from the same cell-types as proteins. Despite this limitation, previous work supports DNAm scores for proteins as useful markers of brain health, suggesting there is merit in integrating DNAm signatures of protein levels in disease stratification⁷².

Our study has several limitations. First, though full replication of our results was not possible, our replication of pQTMs identified by Zaghloul et al.⁷³ reinforces inflammation signalling as intrinsic to the methylome signature of blood proteins. This also suggests that pQTMs may be common across ancestries. Second, we observed a substantial inflation for PAPPA and PRG3 proteins. While comprehensive adjustment for estimated immune cells was performed and the remainder of CpGs involved in pQTMs did not show high correlations (Supplementary Fig. 2), concurrently measured blood components such as haemoglobin, red blood cells and platelets were not available. Future studies should seek to resolve signals with more detailed blood cell phenotyping and immune cell estimates⁷⁴. Third, 89% of the proteins identified in our protein PhEWA did not have epigenetic pQTMs; this may be due to 1) the presence of pathways relating to neurological disease that are not reflected by blood immune cell DNAm, 2) underpowered analyses, or 3) the presence of indirect associations that are not captured by our MWAS approach. Fourth, the extent to non-specific and cross-aptamer binding with SOMAmer technology has not been fully resolved⁷⁵. Fifth, there are likely unknown genetic influences on pQTMs. Further characterisation of pQTLS and advances in multimodelling techniques will aid in the separation of genetic and environmental influences on epigenetic signatures. Sixth, differences in blood and brain DNAm and pQTLS are emerging; these indicate that blood-based markers may not fully align to biology of brain degeneration⁷⁶. However, our ChIP-seq analysis of chromatin regulation suggested that some regulatory states may persist between blood and brain. Seventh, profiling DNAm signatures alone cannot capture the full role of the epigenome in brain health. Integration of more diverse epigenetic markers will be critical to further resolve these relationships. Finally, though we have incorporated a wide portfolio of brain health measures, we recognise that these are not extensive. Increasing triangulation across modalities, as we have shown here, will be useful in identifying candidate markers.

In conclusion, by integrating epigenetic and proteomic data with cognitive scoring, brain morphology and APOE e4 status, we identify 132 protein markers of brain health. We characterise DNAm signatures for all 4058 proteins included in the study, uncovering 35 associations between differential DNAm and the levels of 17 of the protein markers of brain health. These data identify candidate targets for the preservation of brain health and may inform risk stratification approaches.

Methods

The Generation Scotland sample population

A YouTube video providing an overview of this study and detailing how summary statistics can be accessed is available at: https://www.youtube.com/channel/UCxQrFFTIItF25YKfJTXuumQ. The Stratifying Resilience and Depression Longitudinally (STRADL) cohort used in this study is a subset of N = 1188 individuals from Generation Scotland: The Scottish Family Health Study (GS). Generation Scotland constitutes a large, family-structured, population-based cohort of >24,000 individuals recruited to GS between 2006 and 2011. During a clinical visit detailed health, cognitive, and lifestyle information was collected in addition to biological samples. Of the 21,525 individuals contacted for participation, N = 1188 completed additional health assessments and biological sampling -5 years after GS baseline⁷⁶. Of these, N = 1,065 individuals had proteomic data available and N = 778 of these had DNAm data available. Four individuals from this subset were excluded from the DNAm sample due to having incomplete depression status information, leaving 774 individuals available for analyses. Supplementary Data 2 summarises the demographic characteristics across the two groups, with descriptive statistics for phenotypes.

Proteomic measurement

SOMAscan® V.4 technology was used to quantify plasma protein levels. This aptamer-based assay facilitates the simultaneous measurements of multiple Slow Off-rate Modified Aptamers (SOMAmers)⁷⁷. SOMAmers were processed for 1065 individuals from the STRADL subset of Generation Scotland. Brieﬂy, binding between plasma samples and target SOMAmers was achieved during incubation and quantitation was recorded using a ﬂuorescent signal on microarrays. Quality control steps included hybridisation normalisation, signal calibration and median signal normalisation to control for inter-plate variation. Full details of quality control stages are provided in Supplementary Information. In the final dataset, 4235 SOMAmer epitope measurements were available in 1065 individuals and these corresponded to 4058 unique proteins (classiﬁed by common Entrez gene names). Supplementary Data 1 provides annotation information for the 4235 SOMAmer measurements that were available.

DNAm measurement

Measurements of blood DNAm in the STRADL subset of GS subset were processed in two sets on the Illumina EPIC array using the same methodology as those collected in the wider Generation Scotland cohort⁷⁸,⁷⁹. Quality control details are provided in Supplementary Information. Briefly, samples were removed if there was a mismatch between DNAm-predicted and genotype-based sex and all nonspecific CpG and SNP probes (with allele frequency >5%) were removed from the methylation file. Probes which had a beadcount of less than 3 in more than 5% of samples and/or probes in which >1% of samples had a detection P > 0.01 were excluded. After quality control, 793,706 and 773,860 CpG were available in sets 1 and 2, respectively. These sets were truncated to include a total of 772,619
common probes and were joined together for use in the MWAS, with 476 individuals included in set 1 and 298 individuals in set 2. DNA methylome specific technical variables (measurement batch and set) were adjusted in all MWAS models.

**Phenotypes in Generation Scotland**

All phenotypes in Generation Scotland MWAS and PheWAS samples are summarised in Supplementary Data 2. An epigenetic score for smoking exposure, EpiSmokEr was calculated for all individuals with DNA methylation data. The methylH implementation of the Houseman method was used to calculate estimated white blood cell proportions for Sets 1 and 2. Blood reference panels were sourced from Reinius et al. with accession GSE35069. The blood gse35069 complete panel was used to imputed measures for Monocytes, Natural Killer cells, B cells, Granulocytes, CD4+ T cells and CD8+ T cells. Eosinophil and Neutrophil estimates were also sourced through the blood gse35069 panel. Body mass index (body weight in kilograms, divided by squared height in metres) was available for all individuals, alongside depression status (defined using a research version of the Structured Clinical Interview for DSM disorders (SCID) assessment), which was coded as a binary variable of no history of depression (0) or lifetime episode of depression (1). Five individuals did not have depression status information and were excluded from the MWAS and PheWAS analyses, where appropriate. APOE e4 status was available for 1050 individuals. APOE e4 status was coded as a numeric variable (e2e2 = 0, e2e3 = 0, e3e3 = 1, e3e4 = 2, e4e4 = 2). Fifteen e2e4 individuals were excluded due to small sample size.

Scores from five cognitive tests (Supplementary Fig. 4; Supplementary Data 2) measured at the clinic visit for the STRADL subset of GS were considered. Cognitive scores were measured at the baseline clinic visit and full details are provided in Supplementary Information. Briefly, these included the Wechsler Logical Memory Test (maximum possible score of 50), the Wechsler Digit Symbol Substitution Test (maximum possible score of 133), the verbal fluency test (based on the Controlled Oral Word Association task), the Mill Hill Vocabulary test (maximum possible score of 44) and the Matrix Reasoning test (maximum possible score of 15). Outliers were defined as scores >3.5 standard deviations above or below the mean and were removed prior to analysis. The first unrotated principal component combining logical memory, verbal fluency, vocabulary and digit symbol tests was calculated as a measure of general cognitive ability (g). General fluid cognitive ability (gf) was extracted using the same approach, but with the vocabulary test (a crystallised measure of intelligence) excluded from the model. While highly similar to g, the gf score is exclusive to measures such as memory and processing capability that are considered fluid. gf may therefore be of greater relevance for assessing cognitive decline in ageing individuals.

The derived brain volume measures (Supplementary Fig. 5; Supplementary Data 2) were recorded at two sites (Aberdeen and Edinburgh). Data processing used the resources provided by the Edinburgh Compute and Data Facility (http://www.ecdf.ed.ac.uk/). Brain volume data included total brain volume (ventricle volumes excluded), global grey matter volume, white matter hyperintensity volume and total intracranial volume. Intracranial volume was treated as a covariate to adjust for head size in all tests including brain volume associations. The derived global white matter integrity measures included gFA and global mean diffusivity. The protocols applied to derive the brain volume measures from T1-weighted scans, and white matter integrity measures from diffusion tensor imaging scans were measured at baseline and follow-up, and full details are provided in Supplementary Information. Brain Age was estimated using the software package brainageR (Version 2.1. https://doi.org/10.5281/zenodo.3476365, available at https://github.com/james-cole/brainageR), which uses machine learning and a large training set to predict age from whole-brain voxel-wise volumetric data derived from structural T1 images. This estimate was regressed on chronological age to produce a measure of Relative Brain Age (residuals from the linear model). Outliers for all imaging variables were defined as measurements >3.5 standard deviations above or below the mean and were removed prior to analyses.

**Phenome-wide association analyses**

Prior to running protein PheWAS analyses, protein levels were transformed by rank-based inverse normalisation and scaled to have a mean of zero and standard deviation of 1. Models were run using the lme4knin function in the coxme R package (Version 2.2–16)77. This modelling strategy allows for mixed-effects linear model structure with adjustment for relatedness between individuals. Models were run in the maximum sample of 1065 individuals, with the 4235 protein levels as dependent variables and phenotypes as independent variables. Continuous variables were scaled to mean of zero and variance one and missing data were excluded from lmekin models. Each model adjusted for age and sex (male = 1, reference female = 0). A random intercept was fitted for each individual and a kinship matrix was included as a random effect to adjust for relatedness. Diagnosis of depression (case = 1, reference control = 0) at the clinic visit was included as a covariate in all models, due to known selection bias for depression phenotypes in STRADL. Clinic study site and protein lag group (storage time before proteomic sequencing) were included as covariates in all models. For the analyses with age and sex as the predictors of interest, two beta coefficients for age and sex were extracted from the same model structure. In the remaining PheWAS models, either numerical APOE e4 status variable (e2 = 0, e3 = 1, e4 = 2), cognitive test scores or brain imaging phenotypes were included in addition to the described covariates as scaled predictors. The beta coefficients were extracted for the phenotype in each protein-phenotype association. All analyses of brain volume measures included further adjustment for intracranial volume (ICV) and study site as main effects, in addition to the interaction between these variables. ICV was used to account for head size. Processing batch, and presence or absence of manual intervention during quality control were also included as covariates for volumetric brain imaging associations. The lrmcomp function in the stats R package (Version 3.6.2) was used to generate principal components for the 4,235 SOMAmers measurements (N = 1065). 143 components explained >80% of the cumulative variance in protein levels (a commonly-used threshold for the retention of principal components2; Supplementary Fig. 1 and Supplementary Data 3). These 143 components were used to derive the PheWAS multiple testing adjustment threshold of P < 0.05 / 143 = 3.5 × 10^-4. This method was chosen due to the presence of high intercorrelations within the protein data.

**Epigenome-wide association study of protein levels**

Prior to running the MWAS, protein levels for 774 individuals with complete phenotypic information were log transformed and regressed on age, sex, study site, lag group, 20 genetic principal components (generated from multidimensional scaling of genotype data from the Illumina 610-Quadv1 array) and known pQTL effects (from a previous genome-wide association study of 4034 SOMAmers targeting 3622 proteins from Sun et al.)79. Residuals from these models were then rank-based inverse normalised and taken as forward as protein level data. Methylation data were in M-value format and were pre-adjusted for age, sex, processing batch, methylation set and depression status7. A second model further adjusted for estimated white blood cell proportions (Monocytes, CD4+ T cells, CD8+ T cells, B cells, Natural Killer cells, Granulocytes and Eosinophils). While Neutrophil estimates were available, they were excluded due to high correlation (r > 0.95) with Granulocyte proportions (Supplementary Fig. 22). Finally, the fully-
adjusted model further regressed DNAm onto an epigenetic score for smoking. EpismokeR 2 and BMI.

Omics-data-based complex trait analysis (OSCA) 3 Version 0.41 was used to run EWAS analyses. Within OSCA, a genetic relationship matrix (GRM) was constructed for the STRADL population. A threshold of 0.05 was used to identify 120 individuals likely to be related based on their genetic similarity. For this reason, the MOA method was used to calculate associations between individual CpG sites and protein levels, with the addition of the GRM as a random effect to adjust for relatedness between individuals. 4, 5 CpG sites were the dependent variables and the 4235 proteins were the independent variables.

Four fully-adjusted models did not converge (NAGLU, CFHR2, MST1, PILRA) and were excluded. A threshold for multiple testing correction (P < 4.5 × 10^-16) was based on 143 independent protein components with cumulative variance >80% (Supplementary Fig. 1 and Supplementary Data 3) (P < 0.05/(143 × 722,619) CpGs). A more conservative threshold based on total number of SOMAMers was also considered (P < 0.05/(4235 × 722,619) = 1.5 × 10^-14) and is detailed in Supplementary Data 4-6. pQTM subsets were classified as cis if the CpG was on the same chromosome as the protein-coding gene and fell within 10 Mb of the transcriptional start site (TSS) of the protein gene. pQTM subsets involving a CpG located on a different chromosome to the protein-coding gene, or >10 Mb from the TSS of the protein gene were classed as trans.

Circos plots were created with the circovize package (Version 0.4.12). BioRender.com was used to create Figs. 1, 2, 3 and 6. All analyses were performed in R (Version 4.0). 7 Functional mapping and tissue expression analyses

Functional mapping and annotation 8, 9 gene set enrichment and tissue expression analyses were conducted for genes corresponding to protein markers that were identified through the PheWAS study, in addition to genes linked to either CpGs or proteins in the neurological pQTM subset. Protein-coding genes were selected as the background set and ensemble v92 was used with a false discovery rate adjusted P < 0.05 threshold for gene set testing. For the genes corresponding to protein markers in the PheWAS a minimum overlapping number of genes was set to 3. The STRING 8 database was queried to build a protein interaction network based on all proteins that had associations in the PheWAS, mQTL and eQTL lookups were performed using the GoDMC 9 and eQTLGen databases 10, respectively. UCSC database searches were used to profile the positional information relating to CpGs in the pQTMs.

Although inter-chromosomal chromatin interactions are unlikely to be stable and persistent, seven proteins with cis pQTM involving CpGs located intra-chromosomally to the proximal protein-coding gene were considered for ChIP-seq and promoter-capture Hi-C mapping to interrogate local chromatin interactions and states that might form the basis for co-regulation of these loci. ChIP-seq data from PBMCs and brain hippocampus were selected from the ENCODE project 11, with accession identifiers available in Supplementary Data 21. Processed promoter-capture Hi-C data for brain hippocampal tissue was integrated from Jung et al. 12, and are available at NCBI Geo with accession GSM61589. Data concerning both promoter-promoter interactions and promoter-other interactions were concatenated and all regions subsequently visualised on the WashU epigenome browser 13.

Ethics declarations

All components of GS received ethical approval from the NHS Tayside Committee on Medical Research Ethics (REC Reference Number: 05/S1401/89). GS has also been granted Research Tissue Bank status by the East of Scotland Research Ethics Service (REC Reference Number: 20/ES/0021), providing generic ethical approval for a wide range of uses within medical research. All participants included in the current study provided informed consent for the use of their data for biomechanical research.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The fully-adjusted MWAS summary statistics for 4231 protein levels generated in this study have been deposited in the MRC-IEU EWAS catalogue 4. These files are also available through a Zenodo repository at https://doi.org/10.5281/zenodo.6801458.

Datasets generated in this study are made available in Supplementary Data files 1–21. The raw data from Generation Scotland are not available due to them containing information that could compromise participant consent and confidentiality. Generation Scotland is run as a Resource for the research community. Requests to use the Resource are made from: Academic collaborators: employees who are party to the Generation Scotland Collaboration Agreement, or researchers or employees of an academic institution or the NHS. Commercial organisations: specific arrangements have been defined to allow commercial organisations to access Generation Scotland resources. Data can be obtained from the data owners. Instructions for accessing Generation Scotland data can be found here: https://www.ed.ac.uk/generation-scotland/for-researchers/access; the GS Access Request Form can be downloaded from this site. Completed request forms must be sent to access@generationscotland.org to be approved by the Generation Scotland Access Committee.

For any further correspondence and material requests please contact Dr Riccardo Marioni at riccardo.marioni@ed.ac.uk. Source data are provided with this paper.

Code availability

All R code used in this study is available with open access at the following GitHub repository: https://github.com/DanniGadd/Epigenome-phenotype-studies-of-brain-health-outcomes.

References

42. Fournier, M. et al. Implication of the glutamate-cystine antiporter xCT in schizophrenia cases linked to impaired GSH synthesis. npj Schizophr. 3, 31 (2017).
54. Witt, S. H. et al. Investigation of manic and euthymic episodes identifies state-and trait-specific gene expression and stab1 as a

Acknowledgements
This research was funded in whole, or in part, by the Wellcome Trust [104036/Z/14/Z, 108890/Z/15/Z, 221890/Z/20/Z]. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. Wellcome Trust 4-year PhD in Translational Neuroscience—training the next generation of basic neuroscientists to embrace clinical research et1108890/Z/15/Z. D.A.G., R.F.H. Wellcome Trust PhD for clinicians, Edinburgh Clinical Academic Track for Veterinary Surgeons. [225442/Z/22/Z]. R.I.M. Alzheimer’s Research UK major project grant ARUK-PG2017B–10. D.L.M., R.E.M. Alzheimer’s Society major project grant AS-19B-010. R.E.M. Chief Scientist Office of the Scottish Government Health Directorate (CZD/16/6) and the Scottish Funding Council (HR03006). Generation Scotland: D.J.P., A.M.M. Wellcome Trust award 1048890/Z/15/Z. D.A.G., R.F.H. Wellcome Trust for clinicians, [220857/Z/20/Z]. A.M.M. Medical Research Council. MC_PC_17209. A.M.M. National Institutes of Health. R01MH124873. A.M.M. The European Union’s Horizon 2020 research and innovation programme under grant agreement No 84776. A.M.M. Medical Research Council [MR/L023784/2]; Dementias Platform UK–L.S., A.N.H. MRC QTl in Health and Disease Programme Grant MC_UU_00007/10 C.H. Medical Research Council Award to the University of Oxford (grant no. MC_PC_17215). L.S. NIHR Biomedical Research Centre at Oxford Health NHS Foundation Trust. L.S. Medical Research Council (MR/R024065/1). S.R.C. U.S. National Institutes of Health (RO1AG054628). S.R.C. Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (Grant Number 221890/Z/20/Z). S.R.C. Joint grant from the Biology and...

Author contributions

Competing interests
R.E.M. has received a speaker fee from Illumina and is an advisor to the Epigenetic Clock Development Foundation. A.M.M. has previously received speaker’s fees from Illumina and Janssen and research grant funding from The Sackler Trust. R.F.H. has received consultant fees from Illumina. All other authors declare no competing interest.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-32319-8 (2022).
Correspondence and requests for materials should be addressed to Riccardo E. Marioni.

Peer review information Nature Communications thanks Lucas Salas, Benjamin Sun and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022
7.3 Conclusion

This study conducted the largest MWAS in terms of plasma proteome coverage to date. MWAS summary statistics for 4,231 protein levels are provided open-access for the research community. My MWAS findings replicated the strong DNAm signature associated with the *NLRC5* locus identified by Zaghlool *et al.* (2020), which supports the link between DNAm and chronic, low-grade inflammation. Of the 2,928 pQTMs identified through my Bonferroni-adjusted epigenome-wide significance threshold ($P < 4.5 \times 10^{-10}$), 1,783 involved the levels of 149 proteins that had not been previously quantified in MWAS.

Of the 405 protein associations that met my Bonferroni-adjusted threshold ($P < 3.5 \times 10^{-4}$) in associations with one or more of the 15 brain health traits, 89% involved proteins whose elevated levels in the blood were associated with less favourable brain health. Pathway analyses that I conducted for this set of proteins identified a large inflammatory cluster (including CRP, ITIH4, C3, C5, COL11A2 and SIGLEC2). As part of the study, I performed a replication assessment in the context of existing literature. A total of 97% of the age and 98% of the sex associations that I characterised for the 4,235 proteins were replicated across one or more of three previous studies (N>1000). When comparing the brain health traits to previous literature, I identified eight previously unreported associations between proteins and *APOE* status. A selection of protein associations with the 14 cognitive and brain imaging measures were consistent with previous findings, but the associations I identified were largely novel. A replication assessment for the cognitive and brain imaging trait associations is available at: [https://github.com/DanniGadd/s1888864_Supplementary_Material/blob/main/Chapter_7/Supplementary_Data_19.xlsx](https://github.com/DanniGadd/s1888864_Supplementary_Material/blob/main/Chapter_7/Supplementary_Data_19.xlsx).

By integrating the results from both parts of the study, I was able to identify 35 pQTMs involving 17 proteins that were markers of one or more of the brain health traits. All 35 pQTMs were previously unreported. Interestingly, none of the 14 protein markers that were associated with *APOE* e4 allele status had pQTMs. A CpG site within the *SLC7A11* region (cg06690548) was associated with three protein markers of brain health (ACY1, SCUBE1 and RBP5). This CpG was of particular interest as differential DNAm at this site had been identified as a causal candidate for Parkinson’s disease.
(N > 900 cases and N > 900 controls) and is linked to the environmental neurotoxin β-methylamino-L-alanine. Another interesting signature was identified between DNAm at the cg11294350 site (CHPT1) and two proteins with lysosomal-associated roles (SMPD1 and HEXB) that associated with lower General Fractional Anisotropy and higher Relative Brain Age. Five associations were identified between probes in the ITIH3/ITIH1/NEK4 region and the levels of RBL2, which was associated with lower Global Grey Matter Volume.

A key limitation of this study is that the MWAS was performed in 774 individuals and may be underpowered to characterise true DNAm patterns that are present in the blood and associate with protein levels. This may be why we detect an enriched set of pQTMs around loci such as NLRC5, as inflammatory influences on DNAm and protein measures may be strong and detectable in this sample size. However, there is also the possibility that some proteins will not have a true biological DNAm signature in blood that is detectable. This is also informative for our understanding of the potential molecular regulators of these proteins, which may be associated with other molecular characteristics and exposures that do not associate with DNAm patterns. MWAS using larger population sizes will help to resolve pQTM signals further. Similarly, though we were able to replicate 81 of the comparable pQTMs reported by Zaghlool et al. in 2020 (P < 0.05, or above, with the same direction of effect in each instance) replication across cohorts of diverse ages and ancestries are required when data become available.

Chapters 6-7 have mapped DNAm signatures of protein levels in two populations. In the next two chapters, I harness DNAm patterns to generate protein EpiScores and evaluate these scores as tools for disease prediction and brain health stratification.
8 Creation of EpiScores for the circulating proteome and assessment as biomarkers for incident disease

8.1 Introduction

Protein MWAS – conducted by Zaghloul et al\textsuperscript{177} in 2020 and more recently expanded on by myself in 2022 (see Chapter 7)\textsuperscript{287} – have reported that low-grade inflammation is particularly well-captured through DNAm signatures of circulating protein levels. From the results of this MWAS, I observed that while not every protein had a detectable DNAm signature, a proportion did. Harnessing the DNAm patterns associated with plasma protein levels may lead to the development of novel biomarker scores for disease risk stratification. Test-retest reliability varies by protein, with acute phase inflammatory proteins exhibiting high variability across multiple time point measures (as discussed in Section 1.3.9). DNAm proxies for proteins (i.e. protein EpiScores) could therefore serve as more reliable readouts of the chronic inflammatory state of the body. Protein EpiScores could also be used as surrogate measures of protein levels in populations that have DNAm measures available but do not have proteins quantified.

Existing studies have predominantly used elastic net penalised regression to generate protein EpiScores. Seven protein EpiScores were first reported as components of GrimAge, the gold-standard biological age predictor\textsuperscript{129}. Protein EpiScores for IL6 and CRP have since been developed\textsuperscript{65,215,216,288}. The EpiScore for CRP was found to have a more stable test-retest profile as compared to longitudinal measures of the protein\textsuperscript{215}. The IL6 EpiScore had 6.4-fold stronger effect estimates versus the measured protein in associations with a range of brain imaging traits\textsuperscript{65}. Before publication of my results in Chapter 8, no study had attempted a proteome-wide assessment of protein EpiScore generation and tested whether these scores can serve as biomarkers for incident disease stratification.

In this chapter, I use elastic net penalised regression to train EpiScores for a maximum of 953 protein levels, testing performance of the scores in two independent cohorts. I take forward the most robust EpiScores and project them into the GS cohort (N=9,537). Taking advantage of the extant data linkage available in GS to GP and hospital records, I assess whether the novel EpiScore biomarkers I have created can be tools for incident disease risk stratification. In Cox PH models, I examine
neurological diseases (Alzheimer’s dementia, ischaemic stroke, major depression) as well diseases known to increase risk of neurodegenerative diseases in later life (diabetes, ischaemic heart disease, inflammatory diseases).

This study was published in *eLife* in 2022. A correction was published in 2023 and the corrected final version is included in full in **Section 8.2.** A collaborator (Shaza Zaghlool) used code that I provided to train EpiScores in the German KORA cohort, due to data restrictions. I trained EpiScores in the LBC1936 population. I then integrated the weighting coefficients, ran the downstream analyses and drafted the manuscript. Robert Hillary and Daniel McCartney contributed towards the Cox PH analyses performed in GS.

The supplementary files and code for this study are available at: https://github.com/DanniGadd/s1888864_Supplementary_Material/tree/main/Chapter 8.
8.2 Epigenetic scores for the circulating proteome as tools for disease prediction
Epigenetic scores for the circulating proteome as tools for disease prediction


1Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, United Kingdom; 2Department of Physiology and Biophysics, Weill Cornell Medicine-Qatar, Education City, Doha, Qatar; 3Computer Engineering Department, Virginia Tech, Blacksburg, United States; 4Department of Psychology, University of Edinburgh, Edinburgh, United Kingdom; 5Lothian Birth Cohorts, University of Edinburgh, Edinburgh, United Kingdom; 6Centre for Clinical Brain Sciences, Chancellor’s Building, University of Edinburgh, Edinburgh, United Kingdom; 7Department of Psychiatry, University of Oxford, Oxford, United Kingdom; 8Department of Psychology, The University of Texas at Austin, Austin, United States; 9Population Research Center, The University of Texas at Austin, Austin, United States; 10Research Unit Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany; 11Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany; 12German Center for Cardiovascular Research (DZHK), partner site Munich Heart Alliance, Munich, Germany; 13German Center for Diabetes Research (DZD), Neuherberg, Germany; 14Scientific Service Group Biomolecular Mass Spectrometry, Max Planck Institute for Heart and Lung Research, W.G. Kerckhoff Institute, Bad Nauheim, Germany; 15German Centre for Cardiovascular Research (DZHK), Partner Site Rhine-Main, Max Planck Institute of Heart and Lung Research, Bad Nauheim, Germany; 16Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia; 17Medical Research Council Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, United Kingdom; 18Division of Psychiatry, University of Edinburgh, Royal Edinburgh Hospital, Edinburgh, United Kingdom

Abstract Protein biomarkers have been identified across many age-related morbidities. However, characterising epigenetic influences could further inform disease predictions. Here, we leverage epigenome-wide data to study links between the DNA methylation (DNAm) signatures of the circulating proteome and incident diseases. Using data from four cohorts, we trained and tested epigenetic scores (EpiScores) for 953 plasma proteins, identifying 109 scores that explained between 1% and 58% of the variance in protein levels after adjusting for known protein quantitative trait loci (pQTL) genetic effects. By projecting these EpiScores into an independent sample (Generation Scotland; n = 9537) and relating them to incident morbidities over a follow-up of 14 years, we uncovered...
130 EpiScore-disease associations. These associations were largely independent of immune cell proportions, common lifestyle and health factors, and biological aging. Notably, we found that our diabetes-associated EpiScores highlighted previous top biomarker associations from proteome-wide assessments of diabetes. These EpiScores for protein levels can therefore be a valuable resource for disease prediction and risk stratification.

Editor’s evaluation
This is an important study that demonstrates the potential utility of the circulating proteome for disease prediction and risk stratification.

Introduction
Chronic morbidities place longstanding burdens on our health as we age. Stratifying an individual’s risk prior to symptom presentation is therefore critical (NHS England, 2016). Though complex morbidities are partially driven by genetic factors (Fuchsberger et al., 2016; Yao et al., 2018), epigenetic modifications have also been associated with disease (Lord and Cruchaga, 2014). DNA methylation (DNAm) encodes information on the epigenetic landscape of an individual and blood-based DNAm signatures have been found to predict all-cause mortality and disease onset, providing strong evidence to suggest that methylation is an important measure of disease risk (Hillary et al., 2020a; Lu et al., 2019; Zhang et al., 2017). DNAm can regulate gene transcription (Lea et al., 2018), and epigenetic differences can be reflected in the variability of the proteome (Hillary et al., 2019; Hillary et al., 2020b; Zaghloul et al., 2020). Low-grade inflammation, which is thought to exacerbate many age-related morbidities, is particularly well captured through DNAm studies of plasma protein levels (Zaghloul et al., 2020). As proteins are the primary effectors of disease, connecting the epigenome, proteome, and time to disease onset may help to resolve predictive biological signatures.

Epigenetic predictors have utilized DNAm from the blood to estimate a person’s ‘biological age’ (Lu et al., 2019), measure their exposure to lifestyle and environmental factors (McCartney et al., 2018c; McCartney et al., 2018a; Peters et al., 2021), and predict circulating levels of inflammatory proteins (Stevenson et al., 2020; Stevenson et al., 2021). A leading epigenetic predictor of biological aging, the GrimAge epigenetic clock incorporates methylation scores for seven proteins along with smoking and chronological age, and is associated with numerous incident disease outcomes independently of smoking (Hillary et al., 2020a; Lu et al., 2019). This suggests there is predictive value gained in utilising DNAm scores relevant to protein levels as intermediaries for predictions. Methylation scores also point towards the pathways that may act on health beyond the protein biomarker that they are trained on. A portfolio of methylation scores for proteins across the circulating proteome could therefore aid in the prediction of disease and offer a different, but additive signal beyond methylation or protein data alone. Generation of an extensive range of epigenetic scores for protein levels has not been attempted to date. The capability of specific protein scores to predict a range of morbidities has also not been tested. However, DNAm scores for interleukin-6 (IL-6) and C-reactive protein (CRP) have been found to associate with a range of phenotypes independently of measured protein levels, show more stable longitudinal trajectories than repeated protein measurements, and, in some cases, outperform blood-based proteomic associations with brain morphology (Stevenson et al., 2021; Conole et al., 2021). This is likely due to DNAm representing the accumulation of more sustained effects over a longer period of time than protein measurements, which have often been shown to be variable in their levels when measured at multiple time points (Koenig et al., 2003; Liu et al., 2015; Moldoveanu et al., 2000; Shah et al., 2014). DNAm scores for proteins could therefore be used to alert clinicians to individuals with high-risk biological signatures, many years prior to disease onset.

Here, we report a comprehensive association study of blood-based DNAm with proteomics and disease (Figure 1). We trained epigenetic scores – referred to as EpiScores – for 953 plasma proteins (with sample size ranging from 706 to 944 individuals) and validated them using two independent cohorts with 778 and 162 participants. We regressed out known genetic pQTL effects from the protein levels prior to generating the EpiScores to preclude the signatures being driven by common SNP data that are invariant across the lifespan. We then examined whether the most robust predictors (n = 109 EpiScores) associated with the incidence of 12 major morbidities (Table 1), over a follow-up period of
eLife digest  Although our genetic code does not change throughout our lives, our genes can be turned on and off as a result of epigenetics. Epigenetics can track how the environment and even certain behaviors add or remove small chemical markers to the DNA that makes up the genome. The type and location of these markers may affect whether genes are active or silent, this is, whether the protein coded for by that gene is being produced or not. One common epigenetic marker is known as DNA methylation. DNA methylation has been linked to the levels of a range of proteins in our cells and the risk people have of developing chronic diseases.

Blood samples can be used to determine the epigenetic markers a person has on their genome and to study the abundance of many proteins. Gadd, Hillary, McCartney, Zaghlool et al. studied the relationships between DNA methylation and the abundance of 953 different proteins in blood samples from individuals in the German KORA cohort and the Scottish Lothian Birth Cohort 1936. They then used machine learning to analyze the relationship between epigenetic markers found in people’s blood and the abundance of proteins, obtaining epigenetic scores or ‘EpiScores’ for each protein. They found 109 proteins for which DNA methylation patterns explained between at least 1% and up to 58% of the variation in protein levels.

Integrating the ‘EpiScores’ with 14 years of medical records for more than 9000 individuals from the Generation Scotland study revealed 130 connections between EpiScores for proteins and a future diagnosis of common adverse health outcomes. These included diabetes, stroke, depression, various cancers, and inflammatory conditions such as rheumatoid arthritis and inflammatory bowel disease.

Age-related chronic diseases are a growing issue worldwide and place pressure on healthcare systems. They also severely reduce quality of life for individuals over many years. This work shows how epigenetic scores based on protein levels in the blood could predict a person’s risk of several of these diseases. In the case of type 2 diabetes, the EpiScore results replicated previous research linking protein levels in the blood to future diagnosis of diabetes. Protein EpiScores could therefore allow researchers to identify people with the highest risk of disease, making it possible to intervene early and prevent these people from developing chronic conditions as they age.

up to 14 years in the Generation Scotland cohort (n = 9537). We also tested for associations between EpiScore levels and COVID-19 disease outcomes. We regressed out the effects of age on protein levels prior to training and testing; age was also included as a covariate in disease prediction models. We controlled for common risk factors for disease and assessed the capacity of EpiScores to identify previously reported protein-disease associations.

Our MethylDetectR shiny app (Hillary and Marioni, 2020) has CpG weights for the 109 EpiScores integrated such that it automates the process of score generation for any DNAm dataset and is available at: https://www.ed.ac.uk/centre-genomic-medicine/research-groups/marioni-group/methyl-detectr. A video on how to use the MethylDetectR shiny app to generate EpiScores is available at: https://youtu.be/65Y2Rv-4tPU.

Results

Selecting the most robust EpiScores for protein levels

To generate epigenetic scores for a comprehensive set of plasma proteins, we ran elastic net penalised regression models using protein measurements from the SOMAscan (aptamer-based) and Olink (antibody-based) platforms. We used two cohorts: the German population-based study KORA (n = 944, mean age 59 years [SD 7.8], with 793 SOMAscan proteins) and the Scottish Lothian Birth Cohort 1936 (LBC1936) study (between 706 and 875 individuals in the training cohort, with a total of 160 Olink neurology and inflammatory panel proteins). The mean age of the LBC1936 participants at sampling was 70 (SD 0.8) for inflammatory and 73 (SD 0.7) for neurology proteins. Full demographic information is available for all cohorts in Supplementary file 1A.

Prior to running the elastic net models, we rank-based inverse normalised protein levels and adjusted for age, sex, cohort-specific variables and, where present, cis and trans pQTL effects identified from previous analyses (Hillary et al., 2019; Hillary et al., 2020b; Suhre et al., 2017)
Train EpiScores
793 SomaScan proteins
KORA (N=944)

Select robust scores

(r > 0.1, P < 0.05)
GS:STRADL n=778
LBC1921 n=162

Train EpiScores
160 Olink proteins
LBC1936 (N≤875)

Apply EpiScores to Generation Scotland (N=9,537)
Integrate 14 years of electronic health data linkage

130 EpiScore - morbidity associations

34 SOMAscan - diabetes associations

28 highlighted previously reported protein - diabetes associations

Figure 1. EpiScores for plasma proteins as tools for disease prediction study design. DNA methylation scores were trained on 953 circulating plasma protein levels in the KORA and LBC1936 cohorts. There were 109 EpiScores selected based on performance (r > 0.1, p < 0.05) in independent test sets. The selected EpiScores were projected into Generation Scotland, a cohort that has extensive data linkage to GP and hospital records. We tested whether levels of each EpiScore at baseline could predict the onset of 12 leading causes of morbidity, over a follow-up period of up to 14 years; 130 EpiScore-disease associations were identified, for 10 morbidities. We then assessed whether EpiScore associations reflected protein associations for diabetes, which is a trait that has been well characterised using SOMAscan protein measurements. Of the 34 SOMAscan-derived EpiScore-diabetes associations, 28 highlighted previously reported protein-diabetes associations.

(Materials and methods). Of a possible 793 proteins in KORA, 84 EpiScores had Pearson r > 0.1 and p < 0.05 when tested in an independent subset of Generation Scotland (The Stratifying Resilience and Depression Longitudinally [STRADL] study, n = 778) (Supplementary file 1B). These EpiScores were selected for EpiScore-disease analyses. Of the 160 Olink proteins trained in LBC1936, there were 21 with r > 0.1 and p < 0.05 in independent test sets (STRADL, n = 778, Lothian Birth Cohort 1921: LBC1921, n = 162) (Supplementary file 1C). Independent test set data were not available for four Olink proteins. However, they were included based on their performance (r > 0.1 and p < 0.05) in a holdout sample of 150 individuals who were left out of the training set. We then retrained all selected predictors on the full training samples.

A total of 109 EpiScores (84 SOMAscan-based and 25 Olink-based) were brought forward (r > 0.1 and p < 0.05) to EpiScore-disease analyses (Figure 2 and Supplementary file 1D). There were five EpiScores for proteins common to both Olink and SOMAscan panels, which had variable correlation
strength (GZMA $r = 0.71$, MMP1 $r = 0.46$, CXCL10 $r = 0.35$, NTRK3 $r = 0.26$, and CXCL11 $r = 0.09$). Predictor weights, positional information, and cis/trans status for CpG sites contributing to these EpiScores are available in Supplementary file 1E. The number of CpG features selected for EpiScores ranged from 1 (lyzozyme) to 395 (aminoacylase-1 [ACY1]), with a mean of 96 (Supplementary file 1F).

The most frequently selected CpG was the smoking-related site cg05575921 (mapping to the AHRR gene), which was included in 25 EpiScores. Counts for each CpG site are summarised in Supplementary file 1G. This table includes the set of protein EpiScores that each CpG contributes to, along with phenotypic annotations (traits) from the MRC-IEU EWAS catalog (MRC-IEU, 2021) for each CpG site having genome-wide significance (p < 3.6 × 10^{-8}) (Saffari et al., 2018). GeneSet enrichment analysis of the original proteins used to train the 109 EpiScores highlighted pathways associated with immune response and cell remodelling, adhesion, and extracellular matrix function (Supplementary file 1H).

**Table 1. Incident morbidities in the Generation Scotland cohort.**

Counts are provided for the number of cases and controls for each incident trait in the basic and fully adjusted Cox models run in the Generation Scotland cohort (n = 9537). Mean time-to-event is summarised in years for each phenotype. Alzheimer’s dementia cases and controls were restricted to those older than 65 years. Breast cancer cases and controls were restricted to females.

<table>
<thead>
<tr>
<th>Morbidity</th>
<th>Basic model</th>
<th>Fully adjusted model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N cases</td>
<td>N controls</td>
</tr>
<tr>
<td></td>
<td>(mean, SD)</td>
<td>(mean, SD)</td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>63</td>
<td>9289</td>
</tr>
<tr>
<td>Alzheimer’s dementia</td>
<td>69</td>
<td>3764</td>
</tr>
<tr>
<td>Bowel cancer</td>
<td>78</td>
<td>9398</td>
</tr>
<tr>
<td>Depression</td>
<td>95</td>
<td>8317</td>
</tr>
<tr>
<td>Lung cancer</td>
<td>100</td>
<td>9433</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>131</td>
<td>5356</td>
</tr>
<tr>
<td>Inflammatory bowel disease</td>
<td>194</td>
<td>9114</td>
</tr>
<tr>
<td>Stroke</td>
<td>313</td>
<td>9026</td>
</tr>
<tr>
<td>COPD</td>
<td>322</td>
<td>8960</td>
</tr>
<tr>
<td>Ischaemic heart disease</td>
<td>385</td>
<td>8649</td>
</tr>
<tr>
<td>Diabetes</td>
<td>429</td>
<td>8757</td>
</tr>
<tr>
<td>Pain</td>
<td>1329</td>
<td>5480</td>
</tr>
</tbody>
</table>

COPD: chronic obstructive pulmonary disease.

**EpiScore-disease associations in Generation Scotland**

The Generation Scotland dataset contains extensive electronic health data from GP and hospital records as well as DNAm data for 9537 individuals. This makes it uniquely positioned to test whether EpiScore signals can predict disease onset. We ran nested mixed effects Cox proportional hazards models (Figure 3) to determine whether the levels of each EpiScore at baseline associated with the incidence of 12 morbidities over a maximum of 14 years of follow-up. The correlation structures for the 109 EpiScore measures used for Cox modelling are presented in Figure 2—figure supplement 1.

There were 286 EpiScore-disease associations with a false discovery rate (FDR)-adjusted p < 0.05 in the basic model. After further adjustment for common risk factor covariates (smoking, social deprivation status, educational attainment, body mass index [BMI], and alcohol consumption), 130 of the 286 EpiScore-disease associations from the basic model had p < 0.05 in the fully adjusted model (Supplementary file 1I–J). Ten of the 130 fully adjusted associations failed the Cox proportional hazards assumption for the EpiScore variable (p < 0.05 for the association between the Schoenfeld residuals and time; Supplementary file 1K). When we restricted the time-to-event/censor period by each year
Figure 2. Test performance for the 109 selected protein EpiScores. Test set correlation coefficients for associations between protein EpiScores for (a) inflammatory Olink, (b) neurology Olink, and (c) SOMAmer protein panel EpiScores and measured protein levels are plotted. 95% confidence intervals are shown for each correlation. The 109 protein EpiScores shown had $r > 0.1$ and $p < 0.05$ in either one or both of the GS:STRADL ($n = 778$) and LBC1921 ($n = 162$) test sets, wherever protein data was available for comparison. Data shown corresponds to the results included in Supplementary file Figure 2 continued on next page.
of possible follow-up, there were minimal differences in the EpiScore-disease hazard ratios between follow-up periods that did not violate the assumption and those that did (Supplementary file 1L). The 130 associations were therefore retained as the primary results.

The 130 associations found in the fully adjusted model comprised 70 unique EpiScores that were related to the incidence of 10 of the 12 morbidities studied. Diabetes and chronic obstructive pulmonary disease (COPD) had the greatest number of associations, with 38 and 37, respectively. Figure 4 presents the EpiScore-disease relationships for COPD and the remaining nine morbidities: stroke, lung cancer, ischaemic heart disease (IHD), inflammatory bowel disease (IBD), rheumatoid arthritis (RA), depression, bowel cancer and pain (back/neck). There were 16 EpiScores that associated with the onset of three or more morbidities. Figure 5 presents relationships for these 16 EpiScores in the fully adjusted Cox model results. Of note is the EpiScore for Complement 5 (C5), which associated with four outcomes: stroke, diabetes, RA and COPD. Of the 34 SOMAscan-derived EpiScore associations with incident diabetes, 28 replicated previously reported SOMAscan protein associations (Elhadad et al., 2020; Gudmundsdottir et al., 2020; Ngo et al., 2021) with incident or prevalent diabetes in one or more cohorts (Figure 6 and Supplementary file 1M).

### Immune cell and GrimAge sensitivity analyses

Correlations of the 70 EpiScores that were associated with incident disease (P < 0.05 in the fully-adjusted cox proportional hazards models) with covariates suggested interlinked relationships with both estimated white blood cell proportions and GrimAge acceleration (Figure 3—figure supplement 1). These covariates were therefore added incrementally to the fully-adjusted Cox models (Figure 3). There were 99 associations that remained statistically significant (FDR p < 0.05 in the basic model.

![Diagram showing the number of associations in basic and adjusted models.](image)

**Figure 3.** Nested Cox proportional hazards assessment of protein EpiScore-disease prediction. Mixed effects Cox proportional hazards analyses in Generation Scotland (n = 9537) tested the relationships between each of the 109 selected EpiScores and the incidence of 12 leading causes of morbidity (Supplementary file 1I-J). The basic model was adjusted for age and sex and yielded 286 associations between EpiScores and disease diagnoses, with false discovery rate (FDR)-adjusted p < 0.05. In the fully adjusted model, which included common risk factors as additional covariates (smoking, deprivation, educational attainment, body mass index (BMI), and alcohol consumption), 130 of the basic model associations remained significant with p < 0.05. In a sensitivity analysis, the addition of estimated white blood cells (WBCs) to the fully adjusted models led to the attenuation of 31 of the 130 associations. In a further sensitivity analysis, 78 associations remained after adjustment for both immune cell proportions and GrimAge acceleration.

The online version of this article includes the following figure supplement(s) for figure 3:

**Figure supplement 1.** Phenotypic trait and estimated white blood cell proportion correlations with EpiScores.
and p < 0.05 in the fully adjusted model) after adjustment for immune cell proportions, of which 78 remained significant when GrimAge acceleration scores were added to this model (Supplementary file 1J). In a further sensitivity analysis, relationships between both estimated white blood cell (WBC) proportions and GrimAge acceleration scores with incident diseases were assessed in the Cox model structure independently of EpiScores. Of the 60 possible relationships between WBC measures and the morbidities assessed, three were statistically significant (FDR-adjusted p < 0.05) in the basic model and remained significant with p < 0.05 in the fully adjusted model (Supplementary file 1N). A higher
A proportion of natural killer cells was linked to decreased risk of incident COPD, RA and diabetes. The GrimAge acceleration composite score was associated with COPD, lung cancer, IBD, diabetes and RA in the fully adjusted models (p < 0.05) (Supplementary file 1O). The magnitude of the GrimAge effect sizes was comparable to the EpiScore findings.

**Relationship between EpiScores and subsequent COVID-19**

Two previous studies including pilot proteomic measurements from the Generation Scotland cohort (N = 199 controls) as part of wider analyses found that several proteins corresponding to our EpiScores were associated with COVID-19 outcomes (Demichev et al., 2021; Messner et al., 2020). These included proteins such as CRP, C9, SELL, and SHBG, all of which were associated with one or more incident diseases in this study. Two subsets (N = 268 and N = 173) of the Generation Scotland sample who contracted COVID-19 were therefore used to test the hypothesis that EpiScores would associate with COVID-19 outcomes (acquired >9 years after the blood draw for DNAm analyses). No significant associations were identified that delineated differences between the development of long-covid (duration >4 weeks) or hospitalisation from COVID-19 (associations that had p < 0.05 did not withstand Bonferroni adjustment for multiple testing) (Supplementary file 1P).

**Discussion**

Here, we report a comprehensive DNAm scoring study of 953 circulating proteins. We define 109 robust EpiScores for plasma protein levels that are independent of known pQTL effects. By projecting these EpiScores into a large cohort with extant data linkage, we show that 70 EpiScores associate with the incidence of 10 leading causes of morbidity (130 EpiScore-disease associations in total), but do not associate with COVID-19 outcomes. Finally, we show that EpiScore-diabetes associations highlight previously measured protein-diabetes relationships. The bulk of EpiScore-disease associations are independent of common

---

Figure 5. Protein EpiScores that associated with the greatest number of morbidities. EpiScores with a minimum of three relationships with incident morbidities in the fully adjusted Cox models. The network includes 16 EpiScores as dark blue (SOMAscan) and grey (Olink) nodes, with disease outcomes in black. EpiScore-disease associations with hazard ratios < 1 are shown as blue connections, whereas hazard ratios > 1 are shown in red. COPD: chronic obstructive pulmonary disease. IHD: ischaemic heart disease. Data shown corresponds to the results included in Supplementary file 1J.
lifestyle and health factors, differences in immune cell composition and GrimAge acceleration. EpiScores therefore provide methylation-proteomic signatures for disease prediction and risk stratification.

The consistency between our EpiScore-diabetes associations and previously identified protein-diabetes relationships (Elhadad et al., 2020; Gudmundsdottir et al., 2020; Ngo et al., 2021) suggests that epigenetic scores identify disease-relevant biological signals. In addition to the comprehensive lookup of SOMAscan proteins with diabetes, several of the markers we identified for COPD and IHD also reflect previous associations with measured proteins (Ganz et al., 2016; Serban et al., 2021). The three studies used for the diabetes comparison represent the largest candidate protein characterizations of type 2 diabetes to date and the top markers identified included aminoacylase-1 (ACY-1), sex hormone-binding globulin (SHBG) and growth hormone receptor (GHR) (Elhadad et al., 2020; Gudmundsdottir et al., 2020; Ngo et al., 2021). Our EpiScores for these top markers were also associated with diabetes, in addition to EpiScores for several other protein markers reported in these studies. A growing body of evidence suggests that type 2 diabetes is mediated by genetic and epigenetic regulators (Kwak and Park, 2016) and proteins such as ACY-1 and GHR are thought to influence a range of diabetes-associated metabolic mechanisms (Kim and Park, 2017; Pérez-Pérez et al., 2012). Proteins that we identify through EpiScore associations, such as NTR domain-containing protein 2 (WFIIKN2), have also been causally implicated in type 2 diabetes onset.

Figure 6. Replication of known protein-diabetes associations with protein EpiScores. EpiScore-incident diabetes associations in Generation Scotland (n = 9537). The 34 SOMAscan (top panel) and four Olink (bottom panel) associations shown with p < 0.05 in fully adjusted mixed effects Cox proportional hazards models. Of the 34 SOMAscan-derived EpiScores, 28 associations were consistent with protein-diabetes associations (pink) in one or more of the comparison studies that used SOMAscan protein levels. Six associations were novel (blue). Data shown corresponds to the results included in Supplementary files 1J and M.
through Mendelian randomisation analysis (Ngo et al., 2021). In the case of diabetes, EpiScores may therefore be used as disease-relevant risk biomarkers, many years prior to onset. Validation should be tested when sufficient data become available for the remaining morbidities.

With modest test set performances (e.g., SHBG \( r = 0.18 \) and ACY-1 \( r = 0.25 \)), it is perhaps surprising that such strong synergy is observed between EpiScores for proteins that associated with diabetes and the trends seen with measured proteins. Nonetheless, DNAm scores for CRP and IL-6 have previously been shown to perform modestly in test sets (\( r \approx 0.2 \), equivalent to \( \approx 4\% \) explained variance in protein level), but augment and often outperform the measured protein related to a range of phenotypes (Stevenson et al., 2020; Stevenson et al., 2021). Compared to scores utilising DNAm for the prediction of singular diseases, our EpiScores enable the granular study of individual protein predictor signatures with clinical outcomes.

Our large-scale assessment of EpiScores provides a platform for future studies, as composite predictors for traits may be created using our EpiScore database. These should be tested in incident disease predictions when sufficient case data are available. Our results indicated that the set of 109 EpiScores are likely to be heavily enriched for inflammatory, complement system and innate immune system pathways, in addition to extracellular matrix, cell remodelling, and cell adhesion pathways. This reinforces previous work linking chronic inflammation and the epigenome (Zaghlool et al., 2020). It also suggests that EpiScores could be useful in the prediction of morbidities that are characterised by differential inflammatory states. An example of this is the EpiScore for Complement Component 5 (C5), which was associated with the onset of four morbidities (Figure 5). The EpiScore for C5 is likely to reflect the biological pathways occurring in individuals with heightened complement cascade activity and could be utilised to alert clinicians to individuals at high risk of multimorbidity. Elevated levels of C5 peptides have been associated with severe inflammatory, autoimmune, and neurodegenerative states (Ma et al., 2019; Mantovani et al., 2014; Morgan and Harris, 2015) and a range of C5-targeting therapeutic approaches are in development (Alawieh et al., 2018; Brandolini et al., 2019; Hawksworth et al., 2017; Hernandez et al., 2017; Morgan and Harris, 2015; Ort et al., 2020).

Though EpiScores such as C5 – which occupy central hubs in the disease prediction framework – may provide evidence of early methylation signatures common to the onset of multiple diseases, we did not observe associations between EpiScores and COVID-19 hospitalisation or long-COVID status. This is perhaps surprising, given that many of the morbidities that our EpiScores predicted are also known risk factors for increased risk of death due to COVID-19 (Williamson et al., 2020). Many of the proteins corresponding to EpiScores in our study were also associated with COVID-19 severity and progression in two previous studies that included a pilot sample (\( N = 199 \)) from the Generation Scotland cohort at baseline as control data (Demichev et al., 2021; Messner et al., 2020). COVID-19 likely has multiple intersecting risk factors that impact severity and recovery, and the lack of associations we observe is likely to be in part due to the limited number of COVID-19 cases available in Generation Scotland. Additionally, there is a large lag time between baseline biological measurement and COVID-19 in our analyses, whereas the two studies that found protein marker associations integrated protein measurements longitudinally and from samples extracted during COVID-19 progression. With increased power available through continued data linkage, EpiScore relationships with COVID-19 outcomes may be observed in future work.

This study has several limitations. First, we demonstrate that EpiScores carry disease-relevant signals that may be clinically meaningful to delineate early disease risk when comparing relative differences within a cohort. However, projecting a new individual onto a reference set is complicated from absolute differences in methylation quantification resulting from batch and processing effects. Second, future studies should assess paired protein and EpiScore contributions to traits, as inference from EpiScores alone, while useful for disease risk stratification, is not sufficient to determine mechanisms. This may also highlight EpiScores that outperform the measured protein equivalent in disease. Third, the epitope nature of the protein measurement in the SOMAscan panel may incur probe cross-reactivity and non-specific binding; there may also be differences in how certain proteins are measured across panels (Pietzner et al., 2020; Sun et al., 2018). Comparisons of multiple protein measurement technologies on the same samples would help to explore this in more detail. Fourth, there may be pQTLs with small effect sizes that were not regressed from the proteins prior to generating the EpiScores. Fifth, while training and testing was performed across multiple cohorts, it is likely that further development of EpiScores in larger proteomic samples with diverse ancestries will improve power to generate robust scores. Upper bounds for DNAm prediction of complex traits,
such as proteins, can be estimated by variance components analyses (Hillary et al., 2020b; Trejo Banos et al., 2020; Zhang et al., 2019). Finally, associations present between EpiScore measures and disease incidence may have been influenced by external factors such as prescription medications for comorbid conditions and comorbid disease prevalence.

We have shown that EpiScores for circulating protein levels predict the incidence of multiple diseases, up to 14 years prior to diagnosis. Our findings suggest that DNAm phenotyping approaches and data linkage to electronic health records in large, population-based studies have the potential to (1) capture inter-individual variability in protein levels; (2) predict incident disease risk many years prior to morbidity onset; and (3) highlight disease-relevant biological signals for further exploration. The EpiScore weights are publicly available, enabling any cohort with Illumina DNAm data to generate them and to relate them to various outcomes. Given the increasingly widespread assessment of DNAm in cohort studies (McCartney et al., 2020; Min et al., 2021), EpiScores offer an affordable and consistent (i.e., array-based) way to utilise these signatures. This information is likely to be important in risk stratification and prevention of age-related morbidities.

Materials and methods
The KORA sample population
The KORA F4 study includes 3080 participants who reside in Southern Germany. Individuals were between 32 and 81 years of age when recruited to the study from 2006 and 2008. In the current study, there were 944 individuals with methylation, proteomics, and genetic data available. The Infinium HumanMethylation450 BeadChip platform was used to generate DNAm data for these individuals. The Affymetrix Axiom array was used to generate genotyping data and the SOMAscan platform was used to generate proteomic measurements in the sample.

DNAm in KORA
Methylation data were generated for 1814 individuals (Petersen et al., 2014); 944 also had protein and genotype measurements available. During preprocessing, 65 SNP probes were excluded and background correction was performed in minfi (Aryee et al., 2014). Samples with a detection rate of less than 95% were excluded. Next, the minfi R package was used to perform normalisation on the intensity methylation measures (Aryee et al., 2014), with a method consistent with the Lumi:QN + BMIQ pipeline. After excluding non-cg sites and CpGs on sex chromosomes or with fewer than 100 measures available, 470,837 CpGs were available for analyses.

Proteomics in KORA
The SOMAscan platform (Version 3.2) (Gold et al., 2010) was used to quantify protein levels in undepleted plasma for 1129 SOMAmer probes (Suhre et al., 2017). Of the 1000 samples provided for analysis, two samples were excluded due to errors in bio-bank sampling and one based on quality control (QC) measures. Of the 997 samples available, there were 944 individuals with methylation and genotypic data. Of the 1129 probes available, five failed the QC, leaving a total of 1124 probes for the subsequent analysis. Protein measurements were transformed by rank-based inverse normalisation and regressed onto age, sex, known pQTLs, and 20 genetic principal components of ancestry derived from the Affymetrix Axiom Array to control for population structure. pQTLs for each protein were taken from a previous GWAS in the sample (Suhre et al., 2017).

The LBC1936 and LBC1921 sample populations
The Lothian Birth Cohorts of 1921 (LBC1921; N = 550) and 1936 (LBC1936; N = 1091) are longitudinal studies of aging in individuals who reside in Scotland (Deary et al., 2012; Taylor et al., 2018). Participants completed an intelligence test at age 11 years and were recruited for these cohorts at mean ages of 79 (LBC1921) and 70 (LBC1936). They have been followed up triennially for a series of cognitive, clinical, physical, and social data, along with blood donations that have been used for genetic, epigenetic, and proteomic measurement. DNAm, proteomic (Olink platform), and genetic data for individuals from Waves 1 (n=875 at mean age 70 years and sd 0.8) and 2 (n=706 at mean age 73 years and sd 0.7) of the LBC1936 and Wave 3 of the LBC1921 (n=162 at mean age 87 years and sd 0.4) were available.
**DNAm in LBC1936 and LBC1921**

DNA from whole blood was assessed using the Illumina 450 K methylation array. Details of QC have been described elsewhere (Shah et al., 2014; Zhang et al., 2018). Raw intensity data were background-corrected and normalised using internal controls. Manual inspection resulted in the removal of low-quality samples that presented issues related to bisulphite conversion, staining signal, inadequate hybridisation, or nucleotide extension. Probes with low detection rate <95% at p < 0.01 and samples with low call rates (<450,000 probes detected at p < 0.01) were removed. Samples were also removed if they had a poor match between genotype and SNP control probes, or incorrect DNAm-predicted sex.

**Proteomics in LBC1936 and LBC1921**

Plasma samples were analysed using either the Olink neurology 92-plex or the Olink inflammation 92-plex proximity extension assays (Olink Bioscience, Uppsala Sweden). One inflammatory panel protein (BDNF) failed QC and was removed. A further 21 proteins were removed, as over 40% of samples fell below the lowest limit of detection. Two neurology proteins, MAPT and HAGH, were excluded due to >40% of observations being below the lower limit of detection. This resulted in 90 neurology (LBC1936 Wave 2) and 70 inflammatory (LBC1936 Wave 1) proteins in LBC1936 and 92 neurology proteins available in LBC1921. Protein levels were rank-based inverse normalised and regressed onto age, sex, four genetic components of ancestry derived from multidimensional scaling of the Illumina 610-Quadv1 genotype array and Olink array plate. In LBC1936, pQTLs were adjusted for, through reference to GWAS in the samples (Hillary et al., 2019; Hillary et al., 2020b).

**Generation Scotland and STRADL sample populations**

Generation Scotland: the Scottish Family Health Study (GS) is a large, family-structured, population-based cohort study of >24,000 individuals from Scotland (mean age 48 years) (Smith et al., 2013). Recruitment took place between 2006 and 2011 with a clinical visit where detailed health, cognitive, and lifestyle information was collected along with biological samples (blood, urine, saliva). In GS, there were 9537 individuals with DNAm and phenotypic information available. The STRADL cohort is a subset of 1188 individuals from the GS cohort who undertook additional assessments approximately 5 years after the study baseline (Navrady et al., 2018).

**DNAm in Generation Scotland and STRADL**

In the GS cohort, blood-based DNAm was generated in two sets using the Illumina EPIC array. Set 1 comprised 5190 related individuals whereas Set 2 comprised 4583 individuals, unrelated to each other and to those in Set 1. During QC, probes were removed based on visual outlier inspection, bead count <3 in over 5% of samples, and samples with detection p-value below adequate thresholds (McCartney et al., 2018b; Seeboth et al., 2020). Samples were removed based on sex mismatches, low detection p-values for CpGs and saliva samples and genetic outliers (Amador et al., 2015). The quality-controlled dataset comprised 9537 individuals (n_{Set1} = 5087, n_{Set2} = 4450). The same steps were also applied to process DNAm in STRADL.

**Proteomics in STRADL**

Measurements for 4235 proteins in 1065 individuals from the STRADL cohort were recorded using the SOMAscan technology; 793 epitopes matched between the KORA and STRADL cohorts and were included for training in KORA and testing in STRADL. There were 778 individuals with proteomics data and DNAm data in STRADL. Protein measurements were transformed by rank-based inverse normalisation and regressed onto age, sex, and 20 genetic principal components (derived from multidimensional scaling of genotype data from the Illumina 610-Quadv1 array).

**Electronic health data linkage in Generation Scotland**

Over 98% of GS participants consented to allow access to electronic health records via data linkage to GP records (Read 2 codes) and hospital records (ICD codes). Data are available prospectively from the time of blood draw, yielding up to 14 years of linkage. We considered incident data for 12 morbidities. Ten of the diseases are listed by the World Health Organization (WHO) as leading causes of either morbidity or mortality (Hay et al., 2017; World Health Organization, 2018). Inflammatory bowel disease (IBD) (Kassam et al., 2014) and RA (James et al., 2018) are also included as traits.
as they have been reported as leading causes of disability and morbidity and the global burdens of these diseases are rising (Alatab et al., 2020; Safiri et al., 2019). Prevalent cases (ascertained via retrospective ICD and Read 2 codes or self-report from a baseline questionnaire) were excluded. For IBD prevalent cases were excluded based on data linkage alone. Included and excluded terms can be found in Supplementary files 1Q-1B1. Alzheimer’s dementia was limited to cases/controls with age of event/censoring ≥65 years. Breast cancer was restricted to females only. Recurrent, major and moderate episodes of depression were included in depression. Diabetes was comprised of predominantly type 2 diabetes codes and additional general diabetes codes such as diabetic retinopathy and diabetes mellitus with renal manifestation that often occur in individuals with type 2 diabetes. Type 1 and juvenile diabetes cases were excluded.

Elastic net protein EpiScores
Penalised regression models were generated for 160 proteins in LBC1936 and 793 proteins in KORA using Glmnet (Version 4.0-2) (Friedman et al., 2010) in R (Version 4.0) (R Development Core Team, 2020). Protein levels were the outcome and there were 428,489 CpG features per model in the LBC1936 training and 397,630 in the KORA training. An elastic net penalty was specified (alpha = 0.5) and cross validation was applied. DNAm and protein measurements were scaled to have a mean of zero and variance of one.

In the KORA analyses, 10-fold cross validation was applied and EpiScores were tested in STRADL (n = 778). Of 480 EpiScores that generated ≥1 CpG features, 84 had Pearson r > 0.1 and p < 0.05 in STRADL. As test set comparisons were not available for every protein in the LBC1936 analyses, a holdout sample was defined, with two folds set aside as test data and 10-fold cross validation carried out on the remaining data (ntrain = 576, ntest = 130 for neurology and ntrain = 725, ntest = 150 for inflammatory proteins). We retained 36 EpiScores with Pearson r > 0.1 and p < 0.05. New predictors for these 36 proteins were then generated using 12-fold cross validation and tested externally in STRADL (n = 778) and LBC1921 (n = 162, for the neurology panel). Twenty-one EpiScores had r > 0.1 and p < 0.05 in at least one of the external test sets. Four EpiScores did not have external comparisons and were included based on holdout performance.

Functional annotations for each of the proteins used to train the finalised set of 109 EpiScores were sourced from the STRING database (Jensen et al., 2009). GeneSet enrichment analysis against protein-coding genes was performed using the FUMA database, to quantify which canonical pathways were most commonly implicated across the 109 genes corresponding to the proteins used to train the 109 EpiScores (Watanabe et al., 2017). The background gene-set was specified as protein coding genes and a threshold of FDR p < 0.05 was applied for enrichment status, with the minimum overlapping genes with gene-sets set to ≥2.

The 109 selected EpiScores were then applied to Generation Scotland (n = 9537). DNAm at each CpG site was scaled to have a mean of zero and variance of one, with scaling performed separately for GS sets.

Associations with health linkage phenotypes in Generation Scotland
Mixed effects Cox proportional hazards regression models adjusting for age, sex, and methylation set were used to assess the relationship between 109 EpiScores and 12 morbidities in Generation Scotland. Models were run using coxme (Therneau, 2020b) (Version 2.2-16) with a kinship matrix accounting for relatedness in Set 1. Cases included those diagnosed after baseline who had died, in addition to those who received a diagnosis and remained alive. Controls were censored if they were disease free at time of death, or at the end of the follow-up period. EpiScore levels were rank-base inverse normalised. Fully adjusted models included the following additional covariates measured at baseline: alcohol consumption (units consumed in the previous week); deprivation assessed by the Scottish Index of Multiple Deprivation (GovScot, 2016); BMI (kg/m²); educational attainment (an 11-category ordinal variable); and a DNAm-based score for smoking status (Bollepalli et al., 2019). A false discovery rate multiple testing correction p < 0.05 was applied to the 1308 EpiScore-disease associations (109 EpiScores by 12 incident disease traits). Proportional hazards assumptions were checked through Schoenfeld residuals (global test and a test for the protein-EpiScore variable) using the coxph and cox.zph functions from the survival package (Therneau, 2020a) (Version 3.2-7). For each association failing to meet the assumption (Schoenfeld residuals p < 0.05), a sensitivity analysis was run across yearly follow-up intervals.

Fully adjusted Cox proportional hazards models were run with Houseman-estimated white blood cell proportions as covariates (Houseman et al., 2012). A further sensitivity analysis added GrimAge
acceleration (Lu et al., 2019) as an additional covariate. Basic and fully adjusted Cox models were also run with estimated monocyte, B-cell, CD4T, CD8T, and natural killer cell proportions as predictors, in addition to models with GrimAge acceleration as the predictor of incident disease.

Correlation structures for EpiScores, DNA-m estimated white blood cell proportions, and phenotypic information were assessed using Pearson correlations and pheatmap (Kolde, 2019) (Version 1.0.12) and ggcorrplot packages (Version 0.1.3) (Kassambara, 2019). The psych package (Version 2.0.9) (Revelle, 2020) was used to perform principal components analysis on EpiScores. Figures 1 and 2 were created with BioRender.com. Associations for EpiScores that were related to a minimum of three morbidities were subset from the fully adjusted Cox proportional hazards results and were visualised using the ggraph package (Version 2.0.5) (Pedersen, 2021). This network representation was used (Figure 5) to highlight protein EpiScores that were connected with multiple morbidities.

**Consistency of disease associations between EpiScores and measured proteins**

Comparisons were conducted between EpiScore-diabetes associations and type 2 diabetes associations with measured proteins using three previous large-scale proteomic studies (Elhadad et al., 2020; Gudmundsdottir et al., 2020; Ngo et al., 2021). In these studies, six cohorts were included (Study 1: KORA n = 993, HUNT n = 940 [Elhadad et al., 2020], Study 2: AGES-Reykjavik n = 5438 and QMDiab n = 356 [Gudmundsdottir et al., 2020], Study 3: Framingham Heart Study n = 1618 and the Malmo Diet and Cancer Study n = 1221). Study 1 included the KORA dataset, which we use in this study to generate SOMAscan EpiScores. We characterised which SOMAscan-based EpiScore-diabetes associations from our fully adjusted results reflected those observed with measured protein levels. We included basic (nominal p < 0.05) and fully adjusted results (with either FDR or Bonferroni-corrected p < 0.05), wherever available, across the lookup cohorts (Supplementary file 1M).

**Relationship between EpiScores and COVID-19 outcomes**

Associations between each of the 109 selected protein EpiScores and subsequent long-COVID or COVID-19 hospitalisation were tested in the Generation Scotland population. A binary variable was used for long-COVID based on self-reported COVID-19 duration from the CovidLife study survey 3 questionnaire (N = 2399 participating individuals) (Fawns-Ritchie et al., 2021). Participants were asked about the total overall time they experienced symptoms in their first/only episode of illness, as well as their COVID-19 illness duration. The dataset is correct as of February 2021 when the survey 3 was administered. Of the 9537 individuals with DNA-m that were included in incident disease analyses, 173 indicated that they had COVID-19 and 56 of these individuals reported having long-COVID (>4 weeks duration of symptoms after infection). The mean duration from DNA-m measurement to long-COVID for these 56 individuals was 11.2 years (sd 1.2). Hospitalisation information, derived from the Scottish Morbidity Records (SMR01), was used to obtain COVID-19 hospital admissions using ICD-10 codes U07.1 (lab-confirmed COVID-19 diagnosis), and U07.2 (clinically diagnosed COVID-19). This data linkage identified 268 of the 9537 individuals that had COVID-19 diagnoses and 29 had been recorded as being hospitalised due to COVID-19. The mean duration from DNA-m measurement to hospitalisation for these 29 individuals was 11.9 years (sd 1.4). Logistic regression models with either hospitalisation or long-COVID status as binary outcomes were used, with the 109 scaled protein EpiScores as the independent variables. Sex and age at COVID testing were included as covariates. The latter was defined as the age at positive COVID-19 test or 1 January 2021 if COVID-19 test data were not available.

**Acknowledgements**

We are grateful to all study participants of KORA, LBC1936, LBC1921, and GS for their invaluable contributions to this study. This research was funded in whole, or in part, by the Wellcome Trust [104036/Z/14/Z, 220857/Z/20/Z, 108890/Z/15/Z, 203771/Z/16/Z, 216767/Z/19/Z]. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.
## Additional information

### Competing interests
Robert F Hillary: has received consultant fees from Illumina. Riccardo E Marioni: has received speaker fees from Illumina and is an advisor to the Epigenetic Clock Development Foundation. The other authors declare that no competing interests exist.

### Funding

<table>
<thead>
<tr>
<th>Funder</th>
<th>Grant reference number</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wellcome Trust</td>
<td>108890/Z/15/Z</td>
<td>Danni A Gadd, Robert F Hillary</td>
</tr>
<tr>
<td>Wellcome Trust</td>
<td>203771/Z/16/Z</td>
<td>Anna J Stevenson</td>
</tr>
<tr>
<td>Alzheimer’s Research UK</td>
<td>ARUK-PG2017B−10</td>
<td>Daniel L McCartney, Riccardo E Marioni</td>
</tr>
<tr>
<td>Qatar Foundation</td>
<td>Biomedical Research Program at Weill Cornell Medicine</td>
<td>Shaza B Zaghlool, Karsten Suhre</td>
</tr>
<tr>
<td>Qatar National Research Fund</td>
<td>NPRP11C-0115-180010</td>
<td>Shaza B Zaghlool, Karsten Suhre</td>
</tr>
<tr>
<td>Bundesministerium für Bildung und Forschung</td>
<td>Helmholtz Zentrum München</td>
<td>Christian Gieger, Annette Peters, Melanie Waldenberger, Johannes Graumann</td>
</tr>
<tr>
<td>Munich Center of Health Sciences</td>
<td>LMUinnovativ</td>
<td>Christian Gieger, Annette Peters, Melanie Waldenberger, Johannes Graumann</td>
</tr>
<tr>
<td>Bavarian State Ministry of Health and Care</td>
<td>DigiMed Bayern</td>
<td>Christian Gieger, Annette Peters, Melanie Waldenberger, Johannes Graumann</td>
</tr>
<tr>
<td>NIHR Biomedical Research Centre, Oxford</td>
<td>MR/L023784/A2</td>
<td>Liu Shi</td>
</tr>
<tr>
<td>Dementias Platform UK</td>
<td>MR/L023784/A2</td>
<td>Liu Shi</td>
</tr>
<tr>
<td>Medical Research Council</td>
<td>MC_UU_00007/10</td>
<td>Caroline Hayward</td>
</tr>
<tr>
<td>Wellcome Trust</td>
<td>104036/Z/14/Z</td>
<td>Ian J Deary, David J Porteous, Andrew M McIntosh</td>
</tr>
<tr>
<td>Wellcome Trust</td>
<td>220857/Z/20/Z</td>
<td>Andrew M McIntosh</td>
</tr>
<tr>
<td>Wellcome Trust</td>
<td>216767/Z/19/Z</td>
<td>Chloe Fawns-Ritchie, Cliff Nangle, Archie Campbell, Robing Flaig, Ian J Deary, David J Porteous, Caroline Hayward, Andrew M McIntosh, Riccardo E Marioni</td>
</tr>
<tr>
<td>Chief Scientist Office of the Scottish Government Health Directorates</td>
<td>CZD/16/6</td>
<td>David J Porteous</td>
</tr>
<tr>
<td>Scottish Funding Council</td>
<td>HR03006</td>
<td>David J Porteous</td>
</tr>
<tr>
<td>Australian Research Council</td>
<td>Fellowship FT200100837</td>
<td>Allan F McRae</td>
</tr>
<tr>
<td>Funder</td>
<td>Grant reference number</td>
<td>Author</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>------------------------</td>
<td>-------------------------------</td>
</tr>
<tr>
<td>Australian Research Council</td>
<td>DP160102400</td>
<td>Peter M Visscher</td>
</tr>
<tr>
<td>National Health and Medical Research Council</td>
<td>1113400</td>
<td>Peter M Visscher</td>
</tr>
<tr>
<td>Medical Research Council and Biotechnology and Biological Sciences Research Council</td>
<td>MR/K026992/1</td>
<td>Ian J Deary</td>
</tr>
<tr>
<td>Biotechnology and Biological Sciences Research Council</td>
<td></td>
<td>Ian J Deary</td>
</tr>
<tr>
<td>Royal Society</td>
<td>Wolfson Research Merit Award</td>
<td>Ian J Deary</td>
</tr>
<tr>
<td>Chief Scientist Office (CSO) of the Scottish Government’s Health Directorates</td>
<td></td>
<td>Ian J Deary</td>
</tr>
<tr>
<td>Age UK</td>
<td>(Disconnected Mind project)</td>
<td>Sarah E Harris, Ian J Deary, Simon R Cox</td>
</tr>
<tr>
<td>Medical Research Council</td>
<td>G0701120</td>
<td>Ian J Deary</td>
</tr>
<tr>
<td>Biotechnology and Biological Sciences Research Council</td>
<td>BB/F019394/1</td>
<td>Ian J Deary</td>
</tr>
<tr>
<td>Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society</td>
<td>221890/Z/20/Z</td>
<td>Simon R Cox</td>
</tr>
<tr>
<td>National Institutes of Health</td>
<td>RF1AG073593</td>
<td>Elliot M Tucker-Drob</td>
</tr>
<tr>
<td>National Institutes of Health</td>
<td>R01AG054628</td>
<td>Elliot M Tucker-Drob, Ian J Deary, Simon R Cox</td>
</tr>
<tr>
<td>Health Data Research UK</td>
<td>substantive site award</td>
<td>Archie Campbell</td>
</tr>
<tr>
<td>Medical Research Council</td>
<td>MRC Human Genetics Unit core support</td>
<td>Caroline Hayward</td>
</tr>
<tr>
<td>Medical Research Council</td>
<td>MR/R024065/1</td>
<td>Ian J Deary</td>
</tr>
<tr>
<td>Medical Research Council</td>
<td>MR/M013111/1</td>
<td>Ian J Deary</td>
</tr>
<tr>
<td>Medical Research Council</td>
<td>G1001245</td>
<td>Ian J Deary</td>
</tr>
<tr>
<td>Australian Research Council</td>
<td>FL180100072</td>
<td>Peter M Visscher</td>
</tr>
<tr>
<td>National Health and Medical Research Council</td>
<td>1010374</td>
<td>Peter M Visscher</td>
</tr>
<tr>
<td>National Institutes of Health</td>
<td>P30AG066614</td>
<td>Elliot M Tucker-Drob</td>
</tr>
<tr>
<td>National Institutes of Health</td>
<td>P2CHD042849</td>
<td>Elliot M Tucker-Drob</td>
</tr>
<tr>
<td>Alzheimer’s Society</td>
<td>AS-PG-19b-010</td>
<td>Riccardo E Marioni</td>
</tr>
</tbody>
</table>
Author contributions
Danni A Gadd, Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review and editing; Robert F Hillary, Conceptualization, Formal analysis, Investigation, Methodology, Software, Visualization; Daniel L McCartney, Shaza B Zaghloul, Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation; Anna J Stevenson, Investigation, Methodology; Yipeng Cheng, Chloe Fawns-Ritchie, Data curation, Formal analysis; Cliff Nangle, Archie Campbell, Robin Flaig, Sarah E Harris, Rosie M Walker, Liu Shi, Elliot M Tucker-Drob, Christian Gieger, Annette Peters, Melanie Waldenberger, Johannes Graumann, Allan F McRae, Ian J Deary, David J Porteous, Caroline Hayward, Peter M Visscher, Simon R Cox, Kathryn L Evans, Andrew M McIntosh, Data curation, Investigation; Karsten Suhre, Conceptualization, Data curation, Investigation, Methodology; Riccardo E Marioni, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Writing – original draft, Writing – review and editing

Author ORCIDs
Danni A Gadd  http://orcid.org/0000-0001-6398-5407
Cliff Nangle  http://orcid.org/0000-0001-5432-1158
Sarah E Harris  http://orcid.org/0000-0002-4941-5106
Karsten Suhre  http://orcid.org/0000-0001-9638-3912
Riccardo E Marioni  http://orcid.org/0000-0003-4430-4260

Ethics
Human subjects: All KORA participants have given written informed consent and the study was approved by the Ethics Committee of the Bavarian Medical Association. All components of GS received ethical approval from the NHS Tayside Committee on Medical Research Ethics (REC Reference Number: 05/S1401/89). GS has also been granted Research Tissue Bank status by the East of Scotland Research Ethics Service (REC Reference Number: 20/ES/0021), providing generic ethical approval for a wide range of uses within medical research. Ethical approval for the LBC1921 and LBC1936 studies was obtained from the Multi-Centre Research Ethics Committee for Scotland (MREC/01/0/56) and the Lothian Research Ethics committee (LREC/1998/4/183; LREC/2003/2/29). In both studies, all participants provided written informed consent. These studies were performed in accordance with the Helsinki declaration.

Decision letter and Author response
Decision letter  https://doi.org/10.7554/eLife.71802.sa1
Author response  https://doi.org/10.7554/eLife.71802.sa2

Additional files
Supplementary files
• Supplementary file 1. Demographic information and supplementary datasets. (A) Demographic and array information for the cohorts and samples used in the study. (B) SomaScan panel EpiScore performance in the Stratifying Resilience and Depression Longitudinally (STRADL) test set. (C) Performance of Olink panel EpiScores in holdout, STRADL, and LBC1921 test sets. (D) Annotations for the proteins corresponding to the 109 selected EpiScores. (E) Predictor weights for the 109 EpiScores from Olink and SomaScan panels which passed testing in independent cohorts. (F) CpG feature counts for the 109 selected EpiScores. (G) Frequency of CpG sites selected for EpiScores with EWAS catalog annotations to phenotypic traits. (H) FUMA canonical pathway Gene set
enrichment for the genes encoding the 109 proteins EpiScores were trained on. (I) Basic Cox proportional hazards model results in Generation Scotland. (J) Fully adjusted and sensitivity analyses results for Cox proportional hazards models in Generation Scotland. (K) Schoenfeld residual Cox sensitivity analyses. (L) Schoenfeld residual Cox sensitivity analyses split by year of follow-up. (M) SOMAscan-EpiScore diabetes association lookup against three large-scale plasma protein-diabetes studies. (N) White blood cell sensitivity analyses. (O) GrimAge sensitivity analyses. (P) COVID-19 analyses. Q-1B1 Primary and secondary diagnosis codes for each of the 12 morbidities in this study that were used to assign case/control status of participants.

• Transparent reporting form

Data availability
Datasets generated in this study are made available in Supplementary file 1; this file includes the protein EpiScore weights for the 109 EpiScores we provide for future studies to use. Our MethylDetectR shiny app (Hillary and Marioni, 2020) has CpG weights for the 109 EpiScores integrated such that it automates the process of score generation for any DNAm dataset and is available at: https://www.ed.ac.uk/centre-genomic-medicine-research-groups/marioni-group/methyldetectr. A video on how to use the MethylDetectR shiny app to generate EpiScores is available at: https://youtu.be/65Y2Rv-4tPU. All datasets used to create figures are included in Supplementary file 1 and specific locations for these are noted in figure legends. All code used in the analyses is available with open access at the following Gitlab repository: https://github.com/DanniGadd/EpiScores-for-protein-levels (copy archived at swh:1:rev:a5130f2b3895a0d95f0d8f8826a9f5e80aaf86). The source datasets analysed during the current study are not publicly available due to them containing information that could compromise participant consent and confidentiality. Data can be obtained from the data owners. Instructions for Lothian Birth Cohort data access can be found here: https://www.ed.ac.uk/lothian-birth-cohorts/data-access-collaboration. Instructions for accessing Generation Scotland data can be found here: https://www.ed.ac.uk/generation-scotland/for-researchers/access; the ‘GS Access Request Form’ can be downloaded from: https://www.ed.ac.uk/generation-scotland/data-access-collaboration. Dr Simon Cox must be contacted to obtain a Lothian Birth Cohort ‘Data Request Form’ by email: . Instructions for accessing Generation Scotland data can be found here: https://www.ed.ac.uk/generation-scotland/for-researchers/access; the ‘GS Access Request Form’ can be downloaded from this site. Completed request forms must be sent to access@generationscotland.org to be approved by the Generation Scotland access committee. Data from the KORA study can be requested from KORA-gen: https://www.helmholtz-munich.de/en/kora-for-scientists/cooperation-with-kora/index.html. Requests are submitted online and are subject to approval by the KORA board.

References


Hawksworth OA, Li XX, Coulthard LG, Wolvetang EJ, Woodruff TM. 2017. New concepts on the therapeutic control of complement anaphylatoxin receptors. Molecular Immunology 89:36–43. DOI: https://doi.org/10.1016/j.molimm.2017.05.014, PMID: 30574040


Kwak SH, Park KS. 2016. Recent progress in genetic and epigenetic research on type 2 diabetes. Experimental & Molecular Medicine 48:e220. DOI: https://doi.org/10.1038/emmm.2016.7, PMID: 26964836


et al. 2017. Connecting genetic risk to disease end points through the human blood plasma proteome. Nature Communications 8:14357. DOI: https://doi.org/10.1038/ncomms14357, PMID: 28240269
8.3 Conclusion

This was the first study to generate a large portfolio of protein EpiScores and assess their utility as individual indicators of incident diseases. Of 953 proteins tested, there were 109 protein EpiScores that had Pearson coefficient $r > 0.1$ and $P < 0.05$ in correlations with measured proteins. This suggests that some proteins may have stronger DNAm profiles that are detectable in blood using these sample populations. With increasing sample size, it may be possible to detect further DNAm patterns associated with proteins. The 109 protein EpiScores are enriched for inflammatory proteins such as Complement component five (C5). Given that test-retest rates are variable and inflammatory proteins may exhibit variability across time point measures, EpiScores such as the one I have generated for C5 may be more stable measures of chronic inflammation. I observed replication of known protein biomarker associations that were reported in previous studies for type 2 diabetes in my EpiScore-diabetes associations. This suggests that EpiScores capture not only a disease-relevant signal, but that the signal is somewhat representative of the original measured protein. One limitation to the Cox PH analyses is that I did not have measured proteins available in GS (n=9,537) to directly compare protein EpiScores with measured protein equivalents in Cox PH analyses. This should be investigated further to understand whether EpiScores can be comparable, or even outperform measured proteins for disease prediction.

My results suggest that protein EpiScores may represent attractive options for feature reduction – condensing many hundreds of thousands of CpG features to a set that are relevant for protein prediction. The 109 protein EpiScores rely on ~10,000 CpGs. They could therefore be used as biologically-informed signatures that carry a large amount of disease-relevant information in relatively few measures (as compared to the full ~450,000 sites on a typical DNAm array). The protein EpiScores I have created have already begun to be used as predictive features in Cox PH elastic net models to generate composite predictors for type 2 diabetes $^{195}$ and CVD $^{289}$, augmenting clinical risk prediction. The protein EpiScores were also used in the development of bAge – an updated biological age score (termed bAge) that outperforms GrimAge in some instances $^{212}$. In the next chapter, I examine the relationship between EpiScores for two peripheral markers that are typically associated with CVD risk (GDF15 and Nt-proBNP) and brain health traits. I also address the limitation of the current study by
performing direct EpiScore comparisons with the measured protein equivalents in associations with incident dementia and three diseases that increase the risk of neurodegenerative diseases in later life: ischaemic heart disease, ischaemic stroke and type 2 diabetes.
9 Creation of GDF15 and NT-proBNP EpiScores and assessment of their relevance to brain health

9.1 Introduction

Growth differentiation factor 15 (GDF15) and N-terminal pro-B-type natriuretic peptide (NT-proBNP) are well-characterised cardiovascular risk biomarkers that also associate with type 2 diabetes and dementia onset. These blood-based markers therefore represent interesting candidates for 1) EpiScore development and 2) assessing the relationship between EpiScores trained on peripherally secreted protein markers and brain health traits. Additionally, the DNAm profiles for these proteins have not been comprehensively characterised in MWAS. A GDF15 EpiScore was developed originally as part of the GrimAge biological age score \(^{129}\). As this GDF15 EpiScore was trained in 2,356 individuals from the Framingham Heart Study, there is scope to develop a GDF15 score using a larger training sample population.

In this chapter, I make use of serum GDF15 and Nt-proBNP measurements for over 16,000 individuals in the GS cohort. I characterise associations between these proteins and four incident diseases (dementia, ischaemic stroke, ischaemic heart disease and type 2 diabetes). I also conduct EWAS (equivalent to MWAS) studies of each marker. Initially, EpiScores for each marker are trained and tested subsets of the population that are unrelated to each other, while maximising cases in the test set (\(N_{\text{train}} \sim 8,000, N_{\text{test}} \sim 2,000\)). In the test set, this allows me to directly compare the EpiScores to measured proteins in Cox PH associations, using the same diseases and modelling structure as profiled initially in the wider population. I then utilise the LBC1936 cohort to replicate the performance of EpiScores versus measured proteins as an external test set, before examining associations between the EpiScores and brain health traits. LBC1936 has cognitive testing and brain imaging available across multiple waves of collection and is therefore well-placed for this assessment of the relationship between protein EpiScore markers and the ageing brain. This study has been submitted for publication and is detailed in full from Section 9.2 to 9.7. The code and supplementary materials for this work are available at: https://github.com/DanniGadd/s1888864_Supplementary_Material/tree/main/Chapter_9.
9.2 Methylation-based predictors of serum GDF15 and NT-proBNP track onset of clinical morbidities affecting the body and brain

Authors: Danni A. Gadd¹, Hannah M. Smith¹, Donncha Mullin²,³, Ola Chybowska¹, Robert F. Hillary¹, Dorien M Kimenai⁴, Elena Bernabeu¹, Yipeng Cheng¹, Chloe Fawns-Ritchie¹,⁵, Archie Campbell¹, Danielle Page⁶, Adele Taylor⁶, Janie Corley⁶, Maria Del C. Valdés-Hernández⁶,⁷,⁸, Susana Muñoz Maniega⁶,⁷,⁸, Mark E. Bastin⁶,⁷,⁸, Joanna M. Wardlaw⁶,⁷,⁸,⁹, Rosie M. Walker⁶,¹⁰, Kathryn L. Evans¹, Andrew M. McIntosh², Caroline Hayward¹,¹³, Tom Russ²,³, Sarah E. Harris⁶, Paul Welsh¹⁴, Naveed Sattar¹⁴, Simon R. Cox⁶, Daniel L. McCartney¹, Riccardo E. Marioni¹,†

¹ Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK
² Division of Psychiatry, University of Edinburgh, Royal Edinburgh Hospital, Edinburgh, UK
³ Alzheimer Scotland Dementia Research Centre, University of Edinburgh, UK
⁴ BHF Centre for Cardiovascular Science, University of Edinburgh, UK
⁵ Division of Psychology, University of Dundee, Dundee, Scotland, UK
⁶ Lothian Birth Cohorts, Department of Psychology, University of Edinburgh, UK
⁷ Scottish Imaging Network, A Platform for Scientific Excellence (SINAPSE) Collaboration, Edinburgh, UK
⁸ Centre for Clinical Brain Sciences, Edinburgh Imaging and UK Dementia Research Institute, University of Edinburgh, UK
⁹ Neurovascular Imaging Research Core, UCLA, Los Angeles, CA
¹⁰ School of Psychology, University of Exeter, Perry Road, Exeter, UK
¹¹ British Heart Foundation Data Science Centre, Health Data Research UK, London, UK
¹² Health Data Research UK, London, UK
¹³ Medical Research Council Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK
¹⁴ School of Cardiovascular and Metabolic Health, BHF Glasgow Cardiovascular Research Centre, University of Glasgow, 126 University Place, Glasgow, UK
Abstract:

Background: Plasma growth differentiation factor 15 (GDF15) and N-terminal pro-B-type natriuretic peptide (NT-proBNP) are cardiovascular biomarkers that associate with a range of diseases. Epigenetic scores (EpiScores) for GDF15 and NT-proBNP may provide new routes for risk stratification.

Results: In the Generation Scotland cohort (N ≥ 16,963), GDF15 levels were associated with incident dementia, ischaemic stroke and type 2 diabetes, whereas NT-proBNP levels were associated with incident ischaemic heart disease, ischaemic stroke and type 2 diabetes (all P_FDR < 0.05). Bayesian Epigenome-wide association studies (EWAS) identified 12 and 4 DNA methylation (DNAm) CpG sites associated (Posterior Inclusion Probability [PIP] > 95%) with levels of GDF15 and NT-proBNP, respectively. EpiScores for GDF15 and NT-proBNP that were trained in a subset of the population. The GDF15 EpiScore replicated protein associations with incident dementia, type 2 diabetes and ischaemic stroke in the Generation Scotland test set (Hazard Ratios (HR) range 1.36 – 1.41, P_FDR <0.03). The EpiScore for NT-proBNP replicated the protein association with type 2 diabetes, but failed to replicate an association with ischaemic stroke. EpiScores explained comparable variance in protein levels across both the Generation Scotland test set and the external LBC1936 test cohort (R^2 range of 5.7-12.2%). In LBC1936, both EpiScores were associated with indicators of poorer brain health. Neither EpiScore was associated with incident dementia in the LBC1936 population.

Conclusions: EpiScores for serum levels of GDF15 and NT-proBNP associate with body and brain health traits. These EpiScores are provided as potential tools for disease risk stratification.

Keywords: GDF15, NT-proBNP, epigenetic, DNA methylation, dementia, diabetes, cardiovascular, stroke, risk stratification, brain
9.3 Background

Delaying or preventing the onset of chronic diseases is a major challenge. Traditional risk factor models provide a foundation to achieve this \cite{290,291}, but can be augmented by molecular-level data. Plasma growth differentiation factor 15 (GDF15) and N-terminal pro-B-type natriuretic peptide (NT-proBNP) are biomarker candidates that have yielded promising results as indicators of a range of morbidities. GDF15 is associated with low-grade inflammation and age-related immunosuppression \cite{292}. Higher levels of GDF15 have been found, through Mendelian randomisation, to causally associate with increased risk of cardiometabolic stroke, atrial fibrillation, coronary artery disease and myocardial infarction \cite{293}. A recent proteome-wide studies that assessed 1,301 \cite{294} proteins identified GDF15 as the top marker of multimorbidity. This was also confirmed by my work in Chapter 10, whereby I assessed 1,468 \cite{295} proteins and identified GDF15 as the top marker of multimorbidity. NT-proBNP is a metabolite of pro B-type natriuretic peptide (BNP), which is a natriuretic and diuretic hormone released by heart muscle in response to wall stretch \cite{296}. An inverse relationship between the levels of NT-proBNP in blood and incident diabetes has been reported \cite{297}, whereas lower levels of NT-proBNP have been associated with more favourable cardiovascular outcomes in randomised control trials \cite{298,299,300}. Elevated levels of GDF15 and NT-proBNP in individuals diagnosed with COVID-19 have been linked to more severe outcomes \cite{301,302}. Both protein markers have also been found to associate with vascular brain injury, poorer neurocognitive performance and incident dementias \cite{282,303}.

DNAm-based epigenetic scores (EpiScores) for blood proteins have been found to serve as markers of incident diseases \cite{287} and augment clinically-used risk factors for risk stratification \cite{195,289}. DNAm reflects the body’s chronic response to low-grade inflammation, environmental and biological exposures \cite{65,177}. A study that directly compared an EpiScore for C-Reactive protein (CRP) to measured CRP found that the EpiScore had greater test-retest reliability over time point measures \cite{215}. This suggests that EpiScores may be more stable indictors than measured proteins in some instances. An EpiScore for GDF15 levels based on changes to DNAm at CpG sites across the genome is one of seven protein EpiScores that contribute to GrimAge, a leading epigenetic predictor of biological age acceleration, healthspan and lifespan.
However, the performance of protein EpiScores against within-sample protein measurements in relation to incident diseases has not been comprehensively investigated. Additionally, EpiScores have typically been trained in samples of restricted size, with training sets typically ranging from 775 to 2,356 individuals.

Here, we assess the viability of EpiScores for serum GDF15 and NT-proBNP as markers of disease outcomes and brain health. Using GDF15 and NT-proBNP measures available in Generation Scotland (N ≥ 16,963), we first profile associations between GDF15 and NT-proBNP and four incident diseases (type 2 diabetes, ischaemic heart disease, ischaemic stroke and dementia), in addition to COVID-19 outcome severity. These diseases were chosen for the study as they have been linked to GDF15 and NT-proBNP and were available through electronic health linkage. We next map the epigenetic architectures of the two proteins, before training and testing protein EpiScores for them in subsets of Generation Scotland. In the test set, direct biomarker comparisons between measured proteins and the EpiScore equivalents are performed in relation to the four incident diseases assessed in the full Generation Scotland sample initially. EpiScores are then retrained in the full sample available and tested externally in the Lothian Birth Cohort 1936 (LBC1936), where associations with brain health traits are also profiled cross-sectionally and longitudinally.
Figure 9.1. Study design for this assessment of GDF15 and NT-proBNP
EpiScores as biomarkers.

Disease associations and epigenome-wide association studies (EWAS) for each protein were first characterised in the full Generation Scotland sample. EpiScores for each protein were initially trained and tested in subsets of the population (test set N ≥ 2,808). This allowed EpiScores to be compared with measured proteins in association with the four incident diseases profiled in the test set. EpiScores were then retrained on the full sample and tested externally in the LBC1936 Wave 4 population, which had measures of both proteins and DNAm available. EpiScores were projected into the larger LBC1936 Wave 1 population (that has DNAm but no protein measures) and profiled in associations with brain health traits, cross-sectionally and longitudinally. Consent for dementia linkage was available from Wave 2 of the LBC1936; therefore, we also tested whether EpiScores were associated with time-to-dementia. EpiScores
were modelled with polygenic risk scores (PRS) for the proteins. CpG: cytosine-phosphate-guanine. IHD: ischaemic heart disease.

9.4 Results

9.4.1 Sample populations

There were 18,413 Generation Scotland participants (59% female) that had DNAm measurements, with a mean age of 48 years (SD 15), a minimum age of 17 years and maximum age of 98 years (Supplementary Table 1)\textsuperscript{248,305}. Of these, 17,489 had GDF15 measurements and 16,963 had NT-proBNP measurements. Subsets of this sample that were unrelated to one another were used to initially train and test EpiScores for GDF15 ($N_{\text{train}} = 8,207$, $N_{\text{test}} = 2,954$) and NT-proBNP ($N_{\text{train}} = 8,002$, $N_{\text{test}} = 2,808$) (Supplementary Table 1). Measurements of serum GDF15 and NT-proBNP levels were available at Wave 4 (mean age 79 years, with 0.6 SD) of the LBC1936 study. These samples were used as external test sets for EpiScores trained on the full Generation Scotland sample. Of 507 individuals at Wave 4, 322 had GDF15 measures (48% female) and 500 had NT-proBNP measures (49% female). LBC1936 has successive Waves of measurements (Waves 1-5, collected at mean ages of 70, 73, 76, 79 and 82 years old, with SD < 1 at each Wave)\textsuperscript{261,262}. EpiScores were projected into Wave 1 (895 individuals with DNAm, but no protein measures) and evaluated in relation to cross-sectional and longitudinal brain health traits. As consent to dementia linkage was available from Wave 2, associations between EpiScores and time-to-dementia were also tested in LBC1936.

9.4.2 GDF15 and NT-proBNP disease associations

Six associations (Figure 9-2) were identified in Cox proportional hazards (PH) mixed effects models between protein levels and incident diseases in Generation Scotland ($N \geq 16,963$). These associations had False Discovery Rate (FDR) $P < 0.05$ in basic (age and sex adjusted) models and $P < 0.05$ in fully-adjusted models (that further adjusted for smoking, alcohol intake, body mass index (BMI), social deprivation and years of education) (Supplementary Table 2). Counts for cases, controls and mean time-to-onset for cases are provided in Supplementary Table 2. In basic logistic regression models, GDF15 was associated with subsequent hospitalisation due to
COVID-19 (odds ratio (OR) per SD = 2.0, 95% confidence interval (CI) = [1.2, 3.4], FDR P = 0.037), as opposed to having COVID-19 without hospitalisation. An inverse association was identified between a one SD increase in NT-proBNP levels and COVID-19 hospitalisation (OR = 0.59, 95% confidence interval (CI) = [0.38, 0.93], FDR P = 0.046). No associations in relation to long-COVID as a binary outcome had FDR P < 0.05. Full summary statistics are provided in Supplementary Table 3.

Figure 9-2. Disease associations for GDF15 and NT-proBNP in Generation Scotland (N ≥ 16,963).

Fully-adjusted hazard ratios from Cox PH mixed effects regression models between protein levels and incident diseases are plotted with 95% confidence intervals. The six associations in red had FDR P < 0.05 in basic and P < 0.05 in fully-adjusted models, whereas associations that had P > 0.05 are shown in black. Hazard ratios are plotted per 1 SD increase in the rank-base inverse normalised levels of each marker. Fully-adjusted models controlled for age, sex, relatedness and common health and lifestyle factors (smoking, alcohol intake, BMI, social deprivation and years of education).

9.4.3 GDF15 and NT-proBNP epigenetic associations

In variance components analysis of GDF15 and NT-proBNP in Generation Scotland (N ≥ 16,963), DNAm explained 36% of the variance in GDF15 levels (lower and upper confidence intervals [CIs] = 32% to 39%) and 32% of the variance in NT-proBNP levels (lower and upper CIs = 27% to 36%). In the EWAS, there were 12 and 4 associations (Bayesian Posterior Inclusion Probability [PIP] > 95%) between differential DNAm at 14 unique CpG sites and the levels of GDF15 and NT-proBNP, respectively. The CpG sites cg03546163 (FKBP5) and cg13108341 (DNAH9) were associated with both GDF15 and NT-proBNP.
Table 9-1 summarises the CpG sites, the biomarkers they associated with, the genes the CpGs are annotated to, and a selection of traits that DNAm at these CpGs have previously been associated with in EWAS studies. The full index of MRC IEU EWAS Catalogue associations (available as of August 2023) for these 14 CpG sites are available in Supplementary Table 4.

<table>
<thead>
<tr>
<th>CpG</th>
<th>PIP</th>
<th>Biomarker</th>
<th>CpG Gene</th>
<th>CpG trait associations (MRC-IEU EWAS Catalogue)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg03546163*</td>
<td>0.98</td>
<td>GDF15, NT-proBNP</td>
<td>FKB5</td>
<td>Chronic kidney disease, fetal vs adult liver, body mass index, waist circumference, mortality, age, neurodegenerative disorders.</td>
</tr>
<tr>
<td>cg13108341*</td>
<td>1.00</td>
<td>GDF15, NT-proBNP</td>
<td>DNAH9</td>
<td>Cancer treatment</td>
</tr>
<tr>
<td>cg00757033</td>
<td>1.00</td>
<td>NT-proBNP</td>
<td>WDR51B</td>
<td>Crohn's disease, inflammatory bowel disease, age</td>
</tr>
<tr>
<td>cg05412028</td>
<td>0.99</td>
<td>NT-proBNP</td>
<td>ABCC4</td>
<td>Age, ageing, primary Sjogrens syndrome</td>
</tr>
<tr>
<td>cg19693031</td>
<td>1.00</td>
<td>GDF15</td>
<td>TXNIP</td>
<td>Fetal vs adult liver, triglycerides, sex, hbA1c, alcohol consumption, blood pressures, hepatic fat, waist circumference, cholesterol measures, age.</td>
</tr>
<tr>
<td>cg02650017</td>
<td>1.00</td>
<td>GDF15</td>
<td>PHOSPHO1</td>
<td>Type 2 diabetes, primary Sjogrens syndrome, C-reactive protein, body mass index, serum high-density lipoprotein cholesterol, Crohn's disease, body mass index, coagulation factor VIII, eosinophilia, age.</td>
</tr>
<tr>
<td>cg06918740</td>
<td>1.00</td>
<td>GDF15</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>cg08900409</td>
<td>1.00</td>
<td>GDF15</td>
<td>PGPEP1</td>
<td>Age</td>
</tr>
<tr>
<td>------------</td>
<td>------</td>
<td>-------</td>
<td>--------</td>
<td>-----</td>
</tr>
<tr>
<td>cg25460262</td>
<td>1.00</td>
<td>GDF15</td>
<td>GDF15</td>
<td>Fetal vs adult liver</td>
</tr>
<tr>
<td>cg21088460</td>
<td>1.00</td>
<td>GDF15</td>
<td>GDF15</td>
<td>N/A</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>cg05575921</th>
<th>1.00</th>
<th>GDF15</th>
<th>AHRR</th>
<th>Extensive set of smoking-associated traits, lung function/cancer traits, body mass index, serum cotinine, C-reactive protein, IgG glycosylation measures, educational attainment, cognitive ability, statin use, urinary cadmium, mortality, post-traumatic stress disorder, age, acute myocardial infarction.</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg25410121</td>
<td>1.00</td>
<td>GDF15</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>cg15058033</td>
<td>0.97</td>
<td>GDF15</td>
<td>PLXNB2</td>
<td>N/A</td>
</tr>
<tr>
<td>cg16993186</td>
<td>0.97</td>
<td>GDF15</td>
<td>CELF2</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Table 9.1. EWAS of GDF15 and NT-proBNP levels in Generation Scotland (N ≥ 16,963).

Posterior inclusion probabilities (PIPs) are provided for all CpG-protein associations (PIP > 0.95) in the BayesR EWAS. * Two CpGs were associated with both GDF15 and NT-proBNP. A selection of traits implicated in associations (P<3.8x10^{-6}, with n>100) with the CpGs from the MRC-IEU EWAS Catalogue (as of August 2023) are shown. HbA1c: glycated haemoglobin. IgG: immunoglobulin G.

9.4.4 EpiScores for GDF15 and NT-proBNP in Generation Scotland

EpiScores for GDF15 and NT-proBNP were initially trained and tested in subsets of Generation Scotland that were unrelated to one another. Predictor weights for EpiScores are available in Supplementary Table 5. Performance in the test set was modelled through the incremental variance ($R^2$) in protein levels that scores could explain beyond a null linear regression model that adjusted for age and sex. The
EpiScore for GDF15 trained using the full set of EPIC array probes had an $R^2$ of 12.2%, whereas the NT-proBNP EpiScore had an $R^2$ of 5.7%. Similar performance was observed when comparing with EpiScores trained using sites available on the 450k array subset (Supplementary Fig. 1). When modelling EpiScores and polygenic risk scores (PRS) derived for each protein (see Methods), additive effects beyond the null model were observed for GDF15 ($R^2$ of 15.5%) and NT-proBNP ($R^2$ of 6.9%). A full summary of the results is provided in Supplementary Table 6.

9.4.5 EpiScore replication of protein biomarker associations

The same Cox PH model structure (as shown in Figure 9-2) was used to directly compare protein levels and EpiScores in the Generation Scotland test. All protein-disease associations – except the association between NT-proBNP and ischaemic stroke – were replicated by EpiScores with consistent direction of effect in fully-adjusted models (Figure 9-3, Supplementary Table 7). Mean time-to-onset, counts for cases and controls and full summary statistics are available in Supplementary Table 7. Mean attenuation in the absolute log of the HR due to the additional adjustment for lifestyle factors beyond age and sex was 6% for protein associations and 12% for EpiScore associations. Of the four protein EpiScore associations identified in fully-adjusted models, three withstood further adjustment for estimated immune cell proportions (attenuation in the absolute log of the HR ranging from 0 – 9%). The association between the GDF15 EpiScore and dementia had $P = 0.064$, with 5% attenuation in the absolute log of the HR. Supplementary Table 7). As only five instances of COVID-19 hospitalisation and nine instances of long-COVID were reported in the test population, we did not conduct protein EpiScore and protein comparisons for these outcomes.
Figure 9.3. Comparison of EpiScores versus measured protein equivalents in fully-adjusted associations with incident diseases in the Generation Scotland test sample (N ≥ 2,808).

For each disease, the protein-disease association is plotted, with the equivalent protein EpiScore-disease association shown directly beneath it for comparison. Hazard ratios are plotted per 1 SD increase in the rank-based inverse normalised levels of each marker. GDF15 results are shown in the top panel, whereas NT-proBNP results are shown in the bottom panel. Nine associations (red) had FDR P < 0.05 in basic and P < 0.05 in fully-adjusted Cox proportional hazards mixed effects models in the test samples. Fully-adjusted models adjusted for age, sex, relatedness and common lifestyle risk factors (smoking, alcohol intake, BMI, social deprivation and years of education). Associations that were non-significant (P > 0.05 in fully-adjusted models) are shown in black.

9.4.6 EpiScore application to the LBC1936 external cohort

EpiScores for each protein were then retrained in the entire Generation Scotland sample (NGDF15 = 17,489 and NNT-proBNP = 16,963). Although we make predictor weights for EpiScores trained on the full EPIC array probes and the subset of probes on the
older 450k array available (Supplementary Table 8), the LBC1936 external test cohort in this study measured DNAm using the 450k array. Thus, the EpiScores trained on the 450k probe subset were projected into this population for external validation.

In the LBC1936 test sample (N_{GDF15} = 322 and N_{NT-proBNP} = 500), incremental R^2 values of 8.9% for GDF15 and 8.1% for NT-proBNP EpiScores were observed, beyond age and sex-adjusted linear regression models (Figure 9-4a). When a PRS for each protein was modelled together with the EpiScores, incremental variance explained rose to 13.7% and 9.1% for GDF15 and NT-proBNP, respectively (Supplementary Table 9). Finally, the GDF15 EpiScore generated previously by Lu et al as part of the GrimAge biological age acceleration predictor^{129} was projected into the Wave 4 GDF15 test set (322 individuals) and evaluated. It explained 5.6% of the variance in GDF15 beyond age and sex, as compared to the 8.9% observed modelling our updated GDF15 score. The two GDF15 EpiScores had a Pearson correlation \( r = 0.32 \) in the test sample.

9.4.7 EpiScore assessment in LBC1936

The EpiScores that were validated against protein measures in the LBC1936 Wave 4 external test set were then projected into methylation measured at Wave 1 (a time point nine years prior), which has a larger DNAm sample available but no protein measures. Structural equation models were then run to characterise associations between the protein EpiScores and five brain health traits (cognitive ability and four structural brain imaging measures). This allowed for EpiScore relationships with both cross-sectional brain health (Wave 1, N=895 individuals with EpiScores, total model N=1,091) and longitudinal change in brain health (Waves 1-5 for cognitive change and Waves 2-5 for brain imaging changes) to be tested (five brain health traits x two EpiScores x cross-sectional and longitudinal associations = 20 hypothesis tests).

Seven of the twenty basic model associations tested had FDR P < 0.05 (Supplementary Table 10). All seven associations involved cross-sectional brain health phenotypes and had negative effect estimates (standardised betas ranging from -0.05 to -0.19). Higher GDF15 and NT-proBNP EpiScores were associated with lower general cognitive ability and lower brain volumes. None of the ten slope associations assessing relationships between the EpiScores and longitudinal change...
in the five brain health phenotypes were significant at FDR $P < 0.05$. In models that further adjusted for additional health and lifestyle factors, five associations had $P < 0.05$ (Figure 9-4b, Supplementary Table 10). A one standard deviation increase in GDF15 EpiScore levels was associated with lower normal appearing white matter volume ($\text{Beta} = -0.07$, $\text{SE} = 0.02$, $P = 2.2 \times 10^{-3}$), poorer general cognitive ability ($\text{Beta} = -0.09$, $\text{SE} = 0.04$, $P = 9.1 \times 10^{-3}$) and lower total brain volume ($\text{Beta} = -0.05$, $\text{SE} = 0.02$, $P = 3.5 \times 10^{-3}$). A one standard deviation increase in NT-proBNP EpiScore levels was associated with lower normal appearing white matter volume ($\text{Beta} = -0.05$, $\text{SE} = 0.02$, $P = 0.02$) and lower total brain volume ($\text{Beta} = -0.03$, $\text{SE} = 0.01$, $P = 0.03$). In a sensitivity analysis that further adjusted for immune cell proportions, the two NT-proBNP associations were attenuated ($P > 0.07$) and the associations between NT-proBNP and lower cognitive ability was found to be significant ($\text{Beta} = -0.10$, $\text{SE} = 0.04$, $P = 6.0 \times 10^{-3}$). In the sensitivity analysis, the three GDF15 associations remained significant ($P < 0.05$), with a mean attenuation of 11% in Beta effect magnitude (Supplementary Table 10).

Individuals consented to share disease information from electronic health records from Wave 2 of the study onwards. In Cox regression models that utilised Wave 2 as the baseline and modelled incident dementia as the outcome ($N_{\text{cases}} = 108$, $N_{\text{controls}} = 672$, mean time-to-event for cases = 8.6 years [SD 3.42] and maximum follow-up of 14.3 years), no associations were identified for either EpiScore (Supplementary Table 11).
Figure 9.4. External assessment of the GDF15 and NT-proBNP EpiScores in LBC1936.

a, Correlation plots between measured protein levels and GDF15 (orange) and NT-proBNP (red) EpiScores in the LBC1936 Wave 4 external test set (N_{GDF} = 322, N_{NT-proBNP} = 500). Pearson correlation coefficients are annotated in each instance. 

b, Standardised beta coefficients derived from structural equation models (SEMs) between EpiScore levels at LBC1936 Wave 1 (N=895 with DNAm, N=1,091 total) and cross-sectional measures of brain health traits that had FDR P < 0.05 in basic (age and sex adjusted) models and P < 0.05 after adjustment for further lifestyle covariates. All associations had a negative beta coefficient (blue). Twenty EpiScore-trait associations were tested in total: 10 cross-sectionally and 10 assessing longitudinal change in brain traits.

9.5 Discussion

Here, biomarker-disease associations for GDF15 and NT-proBNP were first observed in Generation Scotland, prior to developing EpiScores for these proteins. EpiScores replicated protein associations with incident diseases in the Generation Scotland test sample. In the LBC1936 external test population, the GDF15 and NT-proBNP
EpiScores explained 9% and 8% of the variance in the protein levels, respectively, with higher levels of the EpiScores associated with poorer brain health cross-sectionally. EWAS of each protein highlighted 14 CpGs with differential DNAm.

This study provides EpiScores for GDF15 and NT-proBNP trained in the largest samples to date as tools for health stratification. Despite the LBC1936 test set being older than the Generation Scotland cohort (mean age of 79 versus 48 years), the EpiScores had $R^2$ values comparable to those observed in the Generation Scotland test set. In the external LBC1936 test set, the GDF15 EpiScore had improved performance (an additional $R^2$ of 3.3%) when compared to the GDF15 EpiScore derived by Lu et al in 2019 as part of the GrimAge biological age acceleration predictor. This is likely due to differences in the sample sizes used for training the two GDF15 scores (2,356 individuals as compared to 17,489 individuals in our study). It may also be due to our training and testing populations having more homogeneous ancestry (Scottish) than the populations used to train the original GrimAge GDF15 score (mixed white European ancestry). No other EpiScores for either GDF15 or NT-proBNP exist in the literature to our knowledge; these EpiScores can therefore be utilised as new tools for risk stratification and can be projected into any cohort with Illumina-based DNAm. We provide EpiScore weights trained on both the 450k and EPIC arrays for use in future research.

Generation Scotland is one of the world’s largest single cohort resources with DNAm, protein measures, and extant data linkage to electronic health records. This allowed for direct comparisons between protein and EpiScore measures in the context of incident disease analyses, which has only recently been possible owing to the expansion of the cohort’s epigenetic resource. As DNAm may record chronic exposure to a range of environmental risk factors and biological processes such as inflammation, EpiScores may be reflective of a range of biological pathways that occur upstream of disease diagnoses. Given that GDF15 and Nt-pro-BNP are promising biomarkers for a range of diseases, our EpiScores are well-positioned candidates with many potential use-cases. The results of inclusion of the PRS for proteins in incremental variance models suggested that EpiScore signals were largely independent of genetic architectures on the proteins, as additive improvements in incremental variance observed when PRS and EpiScores were modelled together. This is in concordance with previous studies that found additive epi/genetic heritability
estimates for plasma protein levels $^{146,147}$. While we have previously regressed out protein quantitative trait loci (pQTLs) from proteins prior to training EpiScores $^{287}$, there is an argument that EpiScores capturing a combination of genetic and epigenetic signatures may enhance the disease-predictive signal available. Both approaches are likely viable for the creation of new biomarkers.

The higher proportion of variance explained by the GDF15 EpiScore as compared to the NT-proBNP EpiScore suggests that GDF15 was better-characterised by DNAm differences across the genome. This may be due to its association with chronic inflammation, as we have observed particularly strong DNAm signatures associated with inflammatory proteins in previous work $^{284,287}$. A stronger DNAm signature was also observed for GDF15 in our EWAS analyses. To our knowledge, this represents the first EWAS of NT-proBNP. The only other EWAS of GDF15 levels was performed by us previously, using GDF15 measures from the SomaLogic assay $^{284}$, where we identified no associations for GDF15 passing Bonferroni correction. The improved power to detect associations in the present study (17,489 rather than 774 individuals) may have facilitated identification of associations in the present study. There were two CpG sites associated with both GDF15 and NT-proBNP (cg03546163 in FKBP5 and cg13108341 in DNAH9), which suggests a partially-shared DNAm signature across the proteins. FK506-binding protein 5 (FKBP5) is implicated in cellular stress response $^{307}$. One previous study found cg03546163 to be differentially methylated in 107 individuals with type 2 diabetes that went onto develop end stage renal disease, versus 253 controls who did not $^{308}$.

The lack of associations with incident ischaemic heart disease in the Generation Scotland test set may be due to limited sample size, as an association between protein NT-proBNP and ischaemic heart disease was observed in the full Generation Scotland sample. Additionally, the GDF15 EpiScore association with incident dementia observed in Generation Scotland did not replicate in LBC1936. This may be due to differences in the way the phenotypes were defined across LBC1936 (consensus committee) versus Generation Scotland (Read and ICD codes only), or different cohort sampling frames and recruitment strategies.

Our findings support previous work identifying associations between GDF15 and Nt-proBNP protein levels and severe COVID-19 outcomes in hospitalised individuals $^{301,302}$. Although very few hospitalisation cases were available (n=28), both proteins
(sampled a mean of 11 years prior to COVID-19 diagnoses) associated with subsequent hospitalisation due to COVID-19. GDF15 is likely to be elevated in individuals with multiple morbidities that may contribute towards greater risk of hospitalisation due to viral illnesses. Diabetes has been associated with increased risk of hospitalisation and adverse outcomes in COVID-19 \(^{309,310}\). Both proteins (and equivalent EpiScores) should be investigated in populations that have DNAm quantified nearer to, or at COVID-19 diagnoses to further resolve these signals.

This study has several limitations. First, whereas it was advantageous to have common ancestry across both (Scottish) training and validation datasets, future studies should test EpiScores across larger populations that include non-European ancestries and larger age ranges as data become available. Second, emerging evidence has quantified differences in genetic associations with the measurements of the same proteins across panels that use antibody-based versus aptamer-based quantification methods \(^{311}\). A particular example of interest from this study was GDF15 levels, which was highlighted as a protein that may have different conformational shapes (isoforms) that are targeted by the two assay methods \(^{311}\). While it is likely that increased training sample size led to improved performance of our GDF15 score versus the GrimAge GDF15 score in the LBC1936 test set, it is possible that technical or biological variability across protein assays may also underlie differences in performance of scores. EpiScores trained on protein measurements from different panels should be therefore be compared further when data become available. Similarly, differences in the protein assay method across the previous EWAS of GDF15 (aptamer-based) that we ran and the present study (immunochemiluminescence) may also introduce variability and EWAS across multiple protein panels should be compared when samples are available.

In conclusion, EpiScores for blood-based GDF15 and NT-proBNP levels are generated in this study and have been found to be useful indictors of disease risk stratification, with disease-specific use-cases. The EpiScores can be derived in any population with Illumina-based DNAm measurements and may be integrated into epigenetic screening panels in future to better-identify high-risk individuals.
9.6 Methods

9.6.1 Generation Scotland

Generation Scotland is a population-based cohort study that includes ~8,000 families from across Scotland. Study recruitment of 23,960 participants occurred between 2006 and 2011, while participants were aged between 18 and 99 years. In addition to completing health and lifestyle questionnaires, participants donated blood samples for biomarker and omics measurement. Details on DNAm quality control in Generation Scotland are provided in Supplementary Information. The quality-controlled DNAm dataset comprised a total of 18,413 individuals with 760,838 CpG sites available on the EPIC array. GDF15 and NT-proBNP measurement details are provided in Supplementary Information. There were 18,413 individuals with GDF15 measures (Supplementary Table 1) and these were subset to 17,489 individuals that had DNAm (mean 1038.7 pg/mL [SD 928]). There were 17,863 individuals with NT-proBNP measures, with 16,963 that had DNAm available (mean 94.6 pg/mL [SD 211.2]). Electronic health records via data linkage to GP records (Read 2 codes) and hospital records (10th revision of the International Classification of Diseases codes [ICD-10 codes]) were assessed prospectively from the time of blood draw. Incident data for all-cause dementia, type 2 diabetes, ischaemic stroke and ischaemic heart disease were considered with censoring date October 2020. Dementia cases were defined as per a previous review of dementia linkage codes, whereas code lists for all other diseases are available in Supplementary Tables 12-14. Prevalent cases (ascertained via retrospective linkage or self-report from a baseline questionnaire) were excluded from each disease trait, leaving only incident diagnoses. Dementia analyses were limited to cases/controls with age of event/censoring ≥ 65 years. Type 1 and juvenile diabetes cases were treated as control observations in the type 2 diabetes analyses. Death was treated as a censoring event.

9.6.2 Lothian Birth Cohort 1936

The Lothian Birth Cohort 1936 (LBC1936) is a longitudinal study of ageing of people residing in Edinburgh and surrounding areas in Scotland (N = 1,091). Individuals were born in 1936 and completed an intelligence test when they were 11 years old. They were later recruited to the cohort at a mean age of 70 years old and have been
followed up triennially for a series of cognitive, clinical, social and physical measurements in five Waves (mean ages 70, 73, 76, 79 and 82 – all with SD below 1 for measures at each Wave). Blood samples were taken and used to derive protein, epigenetic and genetic measurements. Sample measurement details for the DNAm measures available in LBC1936 are provided in Supplementary Information. DNAm is available at the four successive waves of the study (N=895, 787, 619 and 507 in Waves 1, 2, 3 and 4, respectively). Both GDF15 (N=322) and NT-proBNP (N=500) serum levels were measured at Wave 4 of the study (mean age 79 years, SD 0.6) and were used to externally test EpiScore performance. From Wave 2 of the LBC1936, individuals consented for linkage to health records for research. Dementia cases were defined by a consensus committee that completed decisions in August 2022. Potential cases were identified through a combination of electronic health record linkage, death certificate data and clinician visits to individuals that were suspected of having cognitive impairments or dementia. Of the 865 individuals who had provided linkage consent at Wave 2, 118 were confirmed as having dementia.

9.6.3 Epigenome-wide association studies in Generation Scotland

The xBayesR+ software implements penalised Bayesian regression on complex traits. The BayesR method has been found to outperform linear and mixed model approaches and implicitly adjusts for probe correlations, data structure (such as relatedness) and unobserved confounders. Prior mixture variances for the methylation data (760,838 CpG sites) were set to 0.001, 0.01 and 0.1 and epigenome-wide associations studies (EWAS) were run for GDF15 (N=17,489) and NT-proBNP (N=16,963) levels in Generation Scotland. Protein measurements were transformed by rank-based inverse normalisation, regressed onto age, sex and 20 genetic principal components and scaled to have a mean of 0 and variance of 1. DNAm measurements in beta format were regressed onto age, sex and processing batch and scaled to have a mean of 0 and variance of 1. Houseman immune cell estimates were included as fixed effect covariates. Effect size estimates were obtained through Gibbs sampling over the posterior distribution, conditional on input data. The Gibbs protocol had 10,000 samples, with 5,000 samples of burn-in followed by a thinning of 5 samples to reduce autocorrelation. Methylation probes that had a posterior inclusion probability of ≥ 95% were deemed to be significant for each protein.
9.6.4 EpiScore development

Elastic net penalised regression was used to train EpiScores for GDF15 and Nt-pro-BNP levels. As Generation Scotland has extensive phenotyping and extant linkage to primary care and hospital records, EpiScores were first trained and tested in subsets of the full sample that were unrelated to one another to facilitate direct comparisons between EpiScore and protein levels in associations with incident diseases. EpiScores were then retrained on the full Generation Scotland sample and tested in LBC1936 – an external cohort. For both analyses, DNAm beta values were considered with missing CpG measurements mean imputed. To generate alternative versions of the EpiScores that can be applied to existing cohort studies with older Illumina array data (450k array), CpGs were filtered to the intersection of the 450k and EPIC array sites. A total of 760,838 EPIC array probes and 390,461 450k probes were available. CpG measurements were scaled to have a mean of 0 and variance of 1, prior to training. Protein measurements in training samples were transformed by rank-based inverse normalisation, regressed onto age, sex and 20 genetic principal components and scaled to have a mean of 0 and variance of 1. Penalised regression models were performed using Big Lasso (Version 1.5.1) in R (Version 4.0) \(^\text{315}\). GDF15 and NT-proBNP protein levels were the outcomes. An elastic net penalty was specified (alpha=0.5). In the within-Generation Scotland analyses 10-fold cross validation was applied to select the lambda value that minimised the mean prediction error, whereas 20-fold cross validation was applied when training EpiScores in the full Generation Scotland sample.

A summary of the individuals with protein measurements available that were used to train and test EpiScores in the initial, within-Generation Scotland analyses is provided in Supplementary Fig. 2. Briefly, individuals that were part of the same family as disease cases in the test sample were excluded from the training sample. In the test subset of Generation Scotland, control individuals that were related to those in the training sample were excluded. A total of 8,207 individuals with GDF15 and 8,002 individuals with NT-proBNP measurements were therefore used to train EpiScores, while 2,954 individuals with GDF15 and 2,808 individuals with NT-proBNP measurements comprised the test samples. When retraining the EpiScores on the full Generation Scotland sample, there were 17,489 and 16,963 individuals available for the GDF15 and NT-proBNP EpiScores, respectively. Supplementary Fig. 3
summarises the training and testing samples used, which included 500 individuals with NT-proBNP and 322 individuals with GDF15 measures in the external LBC1936 test set.

9.6.5 EpiScore testing

To test EpiScores, the additional variance in protein levels that the EpiScores explained over a null model was quantified by running the following models:

Model 1: protein ~ age + sex

Model 2: protein ~ age + sex + protein EpiScore

The incremental variance ($R^2$) in protein levels explained due to the protein EpiScore was calculated by subtracting the $R^2$ derived from model 1 from that in model 2. In these models, scaled, rank-based inverse normalised protein levels were used for testing. Pearson correlation coefficients were also calculated between GDF15 and NT-proBNP levels and their respective EpiScores in the test set and plotted. Protein EpiScores were tested using the described approach in both the Generation Scotland test subset (N≥2,808) and the individuals in Wave 4 of the LBC1936 external cohort that had measures of the proteins available (NGDF = 322, NNT-proBNP = 500). To assess the incremental variance that could be attributed to genetic architectures of the proteins, polygenic risk scores (PRS) for the proteins were calculated using genome-wide association study (GWAS) summary statistics generated in the Generation Scotland population via BOLT-LMM 316 (see Supplementary Information). A summary of sentinel protein quantitative trait loci (pQTLs) identified by conditional and joint analyses (COJO) via Genome-wide Complex Trait Analysis (GCTA) software 271 for the GWAS results are available in Supplementary Table 15. PRS were derived using PRSice software 317. The PRS utilised pQTLs that had $P < 5 \times 10^{-8}$, with clumping (parameters: $R^2 = 0.25$, distance = 250kb, p1 = 1). PRS were modelled in incremental variance assessments singularly and additively with the EpiScores in the test sets in relation to measured proteins.

9.6.6 Cox proportional hazards analyses in Generation Scotland

Cox proportional hazards mixed effects regression models were used to assess the relationship between measured levels of GDF15 (N=17,489) and NT-proBNP (N=16,693) levels in the baseline Generation Scotland sample and four incident
morbidities. The same model structure was also used in the test subset of the Generation Scotland sample where proteins and EpiScores were available for direct comparisons. All models were run using coxme \(^{318}\) (Version 2.2-16) with a kinship matrix accounting for relatedness. Cases included those diagnosed after baseline who had died, in addition to those who received a diagnosis and remained alive. Controls were censored if disease free at time of death, or at the end of the follow-up period. Date of censoring was set to October 2020, which was the latest date of the GP data linkage information. Protein levels were rank-based inverse normalised and scaled to have a mean of 0 and variance of 1 prior to analyses. Basic models were run adjusting for age and sex. Fully-adjusted models further controlled for alcohol consumption (units consumed in the previous week); social deprivation (assessed by the Scottish Index of Multiple Deprivation \(^{319}\)); body mass index (kilograms/height in metres squared); educational attainment (an 11-category ordinal variable) and a DNAm-based score for smoking status \(^{320}\). Each of these covariates was sampled at baseline.

An FDR multiple testing correction \(P < 0.05\) was applied to basic model associations across all diseases tested. Basic associations were considered to be significant if they had FDR \(P < 0.05\). Associations in fully-adjusted models were considered to be significant if they had unadjusted \(P < 0.05\). Proportional hazards assumptions were checked through Schoenfeld residuals (global test and a test for the protein variable) using the coxph and cox.zph functions from the survival package \(^{321}\) (Version 3.2-7). For each association failing to meet the assumption (Schoenfeld residuals \(P < 0.05\)), a sensitivity analysis was run across yearly follow-up intervals. There were minimal differences in hazard ratios between follow-up periods that did not violate the assumption and those that did. All associations were therefore retained.

9.6.7 COVID-19 analyses in Generation Scotland

Associations between measured levels of GDF15 and NT-proBNP and subsequent long-COVID (derived through CovidLife study survey 3 questionnaire \(^{322}\)) or COVID-19 hospitalisation (derived through hospital linkage) were tested in the full Generation Scotland population. The preparation of the two binary outcome variables (long-COVID or hospitalisation due to COVID-19) is detailed in Supplementary Information. Logistic regression models with either hospitalisation (28 of 491 possible individuals) or long-COVID status (87 of 269 possible individuals) were run, with
standardised (measured) proteins as the independent variables. Controls were defined as individuals that had COVID-19 but did not experience hospitalisation or long-COVID. Sex and age at COVID testing were adjusted for in the models. The latter was defined as the age at positive COVID-19 test or 1st January 2021 if COVID-19 test data were not available.

9.6.8 EpiScore associations with brain health traits in LBC1936

As longitudinal cognitive testing and brain morphology measures are available in LBC1936, structural equation models (SEM) were used to examine the relationship between each EpiScore and brain health traits (cross-sectionally and longitudinally). Outcomes included: general cognitive ability \( (g) \), total brain volume, normal-appearing white matter volume, global grey matter volume and white matter hyperintensity volume. Cognitive test data were available at all measurement Waves (mean ages 70, 73, 76, 79 and 82) and brain magnetic resonance imaging (MRI) data were available from the second Wave (mean ages 73, 76, 79 and 82). Information on how the SEM analyses were constructed, with information on the number of individuals with cognitive and brain imaging measures at each Wave is included in Supplementary Information. Basic models were run with adjustment for age and sex, whereas fully-adjusted models included further covariates: DNAm-derived immune cell proportion estimates, DNAm-derived smoking score\(^{320}\), self-reported alcohol consumption, BMI and the Scottish Index of Multiple Deprivation\(^{323}\). Intercept (cross-sectional associations) and slope (longitudinal change) coefficients were extracted. A total of 1,091 individuals were modelled as part of the SEM analyses, with 895 individuals that had EpiScore measures available at Wave 1.

Individuals consented to share disease information from electronic health records from Wave 2 of the study onwards. Cox PH models were run to test associations between Wave 2 GDF15 and NT-proBNP EpiScores and incident dementia diagnoses after Wave 2 baseline, with basic adjustments for age and sex. The test population included 780 individuals who had dementia ascertainment and EpiScore information available at Wave 2, with 108 of these individuals having received a dementia diagnosis post-baseline (mean time-to-event 8.6 years [SD 3.4]). For the 108 incident cases, time-to-event was calculated using age at diagnosis. For controls who had died age at death
was used for censoring, whereas age at the date of the dementia consensus meeting decision was taken forward for controls that remained alive.
9.7 Conclusion

This study trained EpiScores for protein levels in GS – one of the largest DNAm resources in existence. I was able to show direct comparisons between EpiScore and measured protein levels for incident disease prediction, highlighting the potential disease-specific use-case for scores developed for serum GDF15 and NT-proBNP levels. I provided evidence that protein EpiScores were able to replicate the measured protein equivalents in Cox PH analyses. This was particularly evident for type 2 diabetes associations with both EpiScores. The variance in protein levels explained by the EpiScores was reasonably well-replicated in the external LBC1936 test sample. EpiScores were also associated with cross-sectional cognitive and brain imaging traits in LBC1936, in fully-adjusted models (FDR P < 0.05). Neither EpiScore was associated with longitudinal changes in brain health traits.

Taken together, these results suggest that both GDF15 and Nt-proBNP EpiScores could be useful measures to track risk of morbidities affecting the body and brain. The EpiScores trained in the full Generation Scotland sample available (Ntrain >= 16,000) should be projected into other cohorts with Illumina-based DNAm for further assessment with clinical and health outcomes. Additionally, while it was not possible to perform direct protein versus EpiScore comparisons in relation to brain imaging and cognitive test traits in LBC1936, this should be assessed in future, using approaches similar to Conole et al. (2021) 65. This study was also limited by a lack of external datasets where EpiScores could be tested across diverse ancestries and age groups.

In Chapters 8 and 9, I have created EpiScores for proteins and shown that they can be useful measures for augmenting disease risk stratification. I have modelled diseases affecting the brain directly, while also studying leading causes of morbidity that have known impacts on brain health in later life. In Chapter 10, I test whether protein measurements in the UK Biobank can be used to stratify the risk individuals have of developing a range of incident diseases affecting the brain and body. I do this by mapping individual protein associations with incident diseases, in addition to training and testing ProteinScores for the 10-year onset of incident outcomes.
10 Plasma protein profiles associated with the onset of neurological diseases and associated morbidities

10.1 Introduction

As proteins are typically the effector molecules of disease, characterising direct relationships between protein biomarkers and incident diseases is informative. Recently ~1,500 Olink protein analytes have been quantified in over 50,000 individuals in the UKB-PPP project, representing the largest resource of its kind. As the UKB has data linkage to GP and hospital records from baseline (between 2006 to 2010), it is now possible to scan the circulating proteome to identify markers of multiple incident diseases. When selecting individuals from the wider UKB cohort (>500,000 individuals) for inclusion in the UKB-PPP project (~50,000 individuals), the consortium enriched the population for incident cases of neurological diseases. These diseases include incident cases of Alzheimer’s dementia (n=499), ALS (n=262) and Parkinson’s disease (n=699), providing a valuable resource for the study of early blood-based signatures of neurodegenerative diseases.

In this chapter, I assess individual relationships between 1,468 Olink plasma protein analyte levels and the incidence of 24 incident outcomes (23 leading age-related diseases and all-cause mortality). I then supply the 1,468 protein measures to elastic net penalised regression models as potentially-informative features, to train and test ProteinScores for the 10-year onset of every incident outcome that has a minimum of 150 incident cases available. I assess ProteinScores against increasingly complex sets of stages covariates that include age and sex, six additional lifestyle factors and 18 additional clinically-relevant biomarker and physical measures. I take forward the type 2 diabetes ProteinScore for further exploration in the context of the clinically-used gold-standard diagnostic marker for type 2 diabetes glycated haemoglobin (HbA1c) and a PRS for type 2 diabetes. I also explore the comparability between metabolomic and proteomic scores for type 2 diabetes and all-cause mortality. This study has been submitted for publication and is detailed in full from Sections 10.2 to 10.7. Supplementary information and code used for this study is available at: https://github.com/DanniGadd/s1888864_Supplementary_Material/tree/main/Chapter_10.
10.2 Blood protein assessment of leading incident diseases and mortality in UK Biobank

Authors: Danni A. Gadd\textsuperscript{1,2}, Robert F. Hillary\textsuperscript{1,2}, Zhana Kuncheva\textsuperscript{1,3}, Tasos Mangelis\textsuperscript{1,3}, Yipeng Cheng\textsuperscript{2}, Manju Dissanayake\textsuperscript{1,3}, Romi Admanit\textsuperscript{4}, Jake Gagnon\textsuperscript{4}, Tinchi Lin\textsuperscript{4}, Kyle Ferber\textsuperscript{4}, Heiko Runz\textsuperscript{5}, Biogen Biobank Team, Riccardo E. Marioni*\textsuperscript{1,2}, Christopher N. Foley*\textsuperscript{1,3}, Benjamin B. Sun\textsuperscript{5,6*}

\textsuperscript{1} Optima Partners, Edinburgh, EH2 4HQ, UK.

\textsuperscript{2} Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, EH4 2XU, UK.

\textsuperscript{3} Bayes Centre, The University of Edinburgh, Edinburgh, EH8 9BT, UK.

\textsuperscript{4} Global Analytics and Data Science, Research and Development, Biogen Inc., Cambridge, MA, USA.

\textsuperscript{5} Translational Sciences, Research and Development, Biogen Inc. Cambridge, MA, USA.

\textsuperscript{6} Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, CB1 8RN, UK.

* Equal contributions.

Correspondence: and .
Abstract

The circulating proteome offers insights into the biological pathways that underlie disease. Here, we test relationships between 1,468 Olink protein levels and the incidence of 23 age-related diseases and mortality in the UK Biobank (N=47,600). We report 3,209 associations between 963 protein levels and 21 incident outcomes, identifying markers of multiple morbidities. Next, protein-based scores (ProteinScores) are developed using penalised Cox regression. When applied to test sets, six ProteinScores improve Area Under the Curve (AUC) estimates for the 10-year onset of incident outcomes beyond age, sex and a comprehensive set of 24 lifestyle factors, clinically-relevant biomarkers and physical measures. Furthermore, the ProteinScore for type 2 diabetes outperforms a polygenic risk score and HbA1c – a clinical marker used to monitor and diagnose type 2 diabetes. Performance of scores reliant on metabolomic and proteomic features is also compared. These data characterise early proteomic contributions to major age-related disease and demonstrate the value of the plasma proteome for risk stratification.

10.3 Background

Identifying individuals that are at high risk of age-related morbidities may aid in personalised medicine. Circulating proteins can discriminate disease cases from controls and delineate risk of incident diagnoses. While singular protein markers offer insight into the mediators of disease, harnessing multiple proteins simultaneously may improve clinical utility. Clinically-available non-omics scores such as QRISK typically profile 10-year onset risk of disease. Proteomic scores have recently been trained on diabetes, cardiovascular and lifestyle traits as outcomes in 16,894 individuals. Proteomic and metabolomic scores have also been developed for time-to-event outcomes including all-cause mortality.
Here, we demonstrated how large-scale proteomic sampling can identify candidate protein targets and facilitate the prediction of leading age-related incident outcomes in mid-to-later life (study design summary in Extended Data Fig. 1). We used 1,468 Olink plasma protein measurements in 47,600 individuals (aged between 40 and 70 years old) available as part of the UK Biobank Pharma Proteomics Project (UKB-PPP) \(^{144}\). Cox proportional hazards (PH) models were used to characterise associations between each protein and 24 incident diseases and all-cause mortality, ascertained via electronic health data linkage. Next, the dataset was randomly split into training and testing subsets to train proteomic scores (ProteinScores) and assess their utility for modelling either 5-year or 10-year onset of the 19 incident outcomes that had a minimum of 150 cases available. We modelled ProteinScores alongside clinical biomarkers, polygenic risk scores (PRS) and metabolomics measures to investigate how these markers may be used to augment risk stratification.

10.4 Results

10.4.1 The UKB-PPP sample

In this study, 1,468 protein analytes (Supplementary Table 1) measured at baseline in 47,600 unrelated individuals ranging between 40-70 years old (Supplementary Table 2) were used. Further details on the protein samples and the preparation pipeline applied are summarised in Extended Data Fig. 2 and Supplementary Information. Principal components analyses indicated that the first 678 components explained a cumulative variance of 90% in the protein levels (Supplementary Table 3).
10.4.2 Protein associations with incident outcomes

We identified differential plasma protein levels that were associated with the onset of 23 diseases (that included leading causes of disability and reductions in healthy life expectancy) \(^{24,342,343}\) and all-cause mortality (Table 1). The maximal follow-up period was 15 years across the 24 outcomes.

In minimally-adjusted (age- or age- and sex-adjusted) models, there were 5,273 significant associations between 1,211 unique proteins and 23 outcomes (Bonferroni-adjusted P threshold = 3.1x10^-6) (Supplementary Table 4). Upon further adjustment for health and lifestyle risk factors (body mass index (BMI), alcohol consumption, social deprivation, education status, smoking status and physical activity) there were 3,209 associations with \(P < 3.1x10^{-6}\) (Fig. 1a, Supplementary Table 5).

These 3,209 associations involved 963 unique protein analytes and 21 outcomes, ranging from one association for amyotrophic lateral sclerosis, cystitis and multiple sclerosis, to 652 and 663 for mortality and liver disease, respectively (Supplementary Table 6).

Fifty-four proteins had significant associations with eight or more incident morbidities (Fig. 1b); in all instances, higher levels of the proteins at baseline were associated with a higher risk of disease or death over the study period (i.e. HR > 1). Of the 54 proteins, GDF15 had the largest number of associations (11 incident outcomes), followed by IL6 and PLAUR (10 incident outcomes). These candidate markers of multiple morbidities were also identified in logistic regression models run between the protein levels and multimorbidity status over the 15-year maximal follow-up period (Supplementary Table 7, Supplementary Information).
A sensitivity analysis modelled each of the 35,232 Cox PH associations tested over increasing yearly case follow-up intervals. Of the 3,209 associations, 2,915 and 1,957 had $P < 3.1 \times 10^{-6}$ (the Bonferroni-adjusted threshold) when restricting cases up to 10-year and 5-year onset, respectively (Supplementary Tables 8-9, Supplementary Information). These results can be examined in a Shiny app available at: https://protein-disease-ukb.optima-health.technology. The app also includes an interactive network that can be manipulated to view the 3,209 associations across multiple morbidities.

A second sensitivity analysis explored the potential impact of medication in a subset of the population that had this information available (35,073 individuals). Ischaemic heart disease was chosen, as a range of blood-pressure lowering medications are commonly used to delay or prevent this disease. Of the 371 protein-ischaemic heart disease associations that had $P < 3.1 \times 10^{-6}$ in the fully-adjusted models in this subset, 336 remained statistically significant at the same $P$-value threshold after adjusting for blood-pressure lowering medication use at baseline (Supplementary Table 10, Supplementary Information).
<table>
<thead>
<tr>
<th>Incident diagnosis</th>
<th>Incident cases (N)</th>
<th>Controls (N)</th>
<th>Mean years to incident case diagnosis (sd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schizophrenia</td>
<td>54</td>
<td>47,449</td>
<td>6.5 (3.4)</td>
</tr>
<tr>
<td>Brain/CNS cancer</td>
<td>82</td>
<td>47,507</td>
<td>5.5 (2.8)</td>
</tr>
<tr>
<td>Multiple sclerosis</td>
<td>96</td>
<td>47,165</td>
<td>5.6 (3.2)</td>
</tr>
<tr>
<td>Major depression</td>
<td>111</td>
<td>47,229</td>
<td>4.2 (3.1)</td>
</tr>
<tr>
<td>Systemic lupus erythematosus</td>
<td>134</td>
<td>47,096</td>
<td>5.1 (2.6)</td>
</tr>
<tr>
<td>Endometriosis a</td>
<td>157</td>
<td>24,768</td>
<td>4.8 (3.3)</td>
</tr>
<tr>
<td>Vascular dementia b</td>
<td>195</td>
<td>33,907</td>
<td>8.1 (3)</td>
</tr>
<tr>
<td>Gynaecological cancer a</td>
<td>256</td>
<td>25,185</td>
<td>5 (3)</td>
</tr>
<tr>
<td>Amyotrophic lateral sclerosis</td>
<td>264</td>
<td>47,269</td>
<td>5.4 (2.7)</td>
</tr>
<tr>
<td>Inflammatory bowel disease</td>
<td>275</td>
<td>46,727</td>
<td>5.9 (3.3)</td>
</tr>
<tr>
<td>Lung cancer</td>
<td>403</td>
<td>47,158</td>
<td>5.9 (3.2)</td>
</tr>
<tr>
<td>Liver disease</td>
<td>432</td>
<td>47,104</td>
<td>7 (3.3)</td>
</tr>
<tr>
<td>Alzheimer’s dementia b</td>
<td>446</td>
<td>33,642</td>
<td>7.8 (2.8)</td>
</tr>
<tr>
<td>Colorectal cancer</td>
<td>508</td>
<td>46,890</td>
<td>5.8 (3.1)</td>
</tr>
<tr>
<td>Cystitis a</td>
<td>531</td>
<td>24,160</td>
<td>4.1 (3)</td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>593</td>
<td>46,310</td>
<td>6.8 (3.2)</td>
</tr>
<tr>
<td>Parkinson’s disease</td>
<td>659</td>
<td>46,802</td>
<td>5.4 (3.2)</td>
</tr>
<tr>
<td>Ischaemic stroke</td>
<td>765</td>
<td>46,657</td>
<td>6.8 (3.4)</td>
</tr>
<tr>
<td>Breast cancer a</td>
<td>772</td>
<td>24,086</td>
<td>5.2 (3.1)</td>
</tr>
<tr>
<td>Prostate cancer a</td>
<td>1,001</td>
<td>20,628</td>
<td>5.7 (3.1)</td>
</tr>
<tr>
<td>Chronic obstructive pulmonary disease</td>
<td>1,998</td>
<td>44,948</td>
<td>6.3 (3.4)</td>
</tr>
<tr>
<td>Type 2 diabetes</td>
<td>2,822</td>
<td>43,370</td>
<td>6 (3.3)</td>
</tr>
<tr>
<td>Ischaemic heart disease</td>
<td>3,338</td>
<td>41,341</td>
<td>6.3 (3.4)</td>
</tr>
<tr>
<td>Death</td>
<td>4,445</td>
<td>43,155</td>
<td>7.9 (3.5)</td>
</tr>
</tbody>
</table>

**Table 10-1. The 24 incident outcomes profiled over a maximum of 15 years of follow-up in the UK Biobank (N=47,600).**

Counts for incident cases and controls are provided, with mean years to diagnosis for incident cases. These data were used in individual Cox PH models to identify protein levels that were associated with incident outcomes. a Sex-stratified traits. b Alzheimer’s and vascular dementia were restricted to individuals aged 65 years or above at the time of diagnosis for cases, or at the time or censoring for controls. CNS: central nervous system.
Figure 10.1. Individual protein associations with incident outcomes in the UK Biobank (N=47,600).

a, Number of associations between protein analytes and time-to-onset for 20 outcomes that had $P < 3.1 \times 10^{-6}$ (Bonferroni-adjusted threshold) in both basic and fully-adjusted Cox PH models. There were 3,209 associations in total involving 963 protein analytes. b, Hazard ratios (HR) per a one SD higher level of the transformed protein analytes (compared within individuals at baseline) are plotted. Fifty-four protein analytes that were associated with eight or more outcomes in the individual Cox PH models are shown. Each association is represented by a rectangle. Cox PH models were adjusted for age, sex and six lifestyle factors (BMI, alcohol consumption, social deprivation, educational attainment, smoking status and physical activity). Every association identified for these proteins had HR > 1 (red) and associations are shaded based on HR effect size (darkest colouration indicating larger magnitude of effect). The largest HR shown is for the association between GDF15 levels and liver disease (HR = 3.7). COPD: chronic obstructive pulmonary disease.
10.4.3 ProteinScore development

We developed ProteinScores via Cox PH elastic net regression for 19 diseases that had a minimum of 150 incident cases available. Of fifty randomised iterations tested (see Methods), ProteinScores with the median difference in AUC beyond a minimally-adjusted model were selected for each outcome (Supplementary Table 11). Summaries of protein features selected for the 19 ProteinScores are available in Supplementary Tables 12-13, ranging from five features selected for endometriosis to 201 for all-cause mortality (Extended Data Fig. 3). Cumulative time-to-onset distributions for cases (Extended Data Figs. 4-5) indicated that amyotrophic lateral sclerosis, endometriosis and cystitis were better-suited to 5-year onset assessments (80% of cases diagnosed by year 8 of follow-up). All remaining ProteinScores were evaluated for 10-year onset.

Selected ProteinScores were modelled alongside various combinations of covariates (Extended Data Fig. 6). The differences in AUC resulting from the addition of the ProteinScores into the three models with increasingly complex sets of covariates are summarised in Fig.2a. A tabular summary of the AUC statistics are available in Supplementary Table 14. Singular inclusion of the ProteinScores had either equal or higher performance than the maximal set of 26 covariates in eight instances. Tests for significant differences between receiver operating characteristic (ROC) curves for the sets of covariates with/out the ProteinScores were performed. Eleven ProteinScores had ROC P < 0.0026 (the Bonferroni-adjusted P-value threshold) beyond minimally-adjusted covariates. When adding ProteinScores to models that included both minimally-adjusted and lifestyle covariates, nine ProteinScores had P < 0.0026 in ROC model comparison tests. When adding ProteinScores to models that further adjusted for an additional 18 clinically-measurable covariates, six ProteinScores (type 2
diabetes, COPD, death, Alzheimer’s dementia, ischaemic heart disease and Parkinson’s disease) had \( P < 0.0026 \) in model comparisons with/out the ProteinScore (Fig. 2b).
Figure 10-2. Value offered by ProteinScores for incident outcomes in the UK Biobank.

a, Differences in AUC resulting from the addition of the 19 ProteinScores to models with increasingly extensive sets of covariates: 1) minimally-adjusted (age and sex where traits were not sex-stratified), 2) minimally-adjusted with the addition of a core set of six lifestyle covariates and 3) further adjustment for an extended set of 18 covariates that are measured in clinical settings (physical and biochemical measures). AUC plots are ordered by increasing AUC differences in the minimally-adjusted models. All ProteinScore performance statistics shown correspond to 10-year onset, except those for ALS, endometriosis and cystitis that were assessed for 5-year onset.

b, A breakdown of the AUC values achieved by different combinations of risk factors with/out the ProteinScores are shown for the six incident outcomes whereby the ProteinScore contributed statistically significant beyond a model including all 24 minimal, lifestyle and extended set variables (ROC P < 0.0026, the Bonferroni-adjusted threshold). All six of the best-performing ProteinScores shown were

10.4.4 Exploration of the type 2 diabetes ProteinScore

Type 2 diabetes was chosen as a case study for exploration. Glycated haemoglobin (HbA1c) averages long-term glucose over 2-3 months and is employed to monitor pre-clinical diabetes risk (42-47mmol/mol) and diagnose the disease (with two repeated measurements >48mmol/mol)\(^ {344,345}\). As the ProteinScore for type 2 diabetes added value beyond the extended set of covariates that included HbA1c, we directly compared performance of HbA1c and the ProteinScore in the test sample alongside a polygenic risk score (PRS) for type 2 diabetes. In the test set, 1,105 cases (with mean time to onset 5.4 years [SD 3.0]) and 3,264 controls had all three measures available. The rank-based inverse normal transformed levels of the ProteinScore and HbA1c discriminated incident case and control distributions similarly (Fig. 3a) and HbA1c levels tended to be higher across ProteinScore risk deciles (Fig. 3b). In incremental Cox PH models for the 10-year onset of type 2 diabetes (Fig. 3c) the singular use of the ProteinScore (AUC = 0.89) outperformed both HbA1c (AUC = 0.85) and the PRS (AUC = 0.68). In ROC model comparisons between HbA1c alone and HbA1c with the ProteinScore added, a statistically significant improvement due to the ProteinScore was identified (ROC P < 0.0026). When the PRS was added to this model (including HbA1c and the ProteinScore), the AUC remained unchanged (0.91) (Supplementary Table 15).
Figure 10-3. Exploration of the type 2 diabetes ProteinScore.

a, Case (red) and control (blue) discrimination for HbA1c and the type 2 diabetes ProteinScore in the test set (1,105 cases, 3,264 controls, mean time to case onset 5.4 years [SD 3.0]). Both markers were rank-based inverse normalised and scaled to have a mean of 0 and standard deviation of 1. b, HbA1c (mmol/mol) per decile of the type 2 diabetes ProteinScore in the test set. The shaded rectangle indicates the type 2 diabetes HbA1c screening threshold (42-47 mmol/mol). c, ROC curves for incremental 10-year onset models incorporating HbA1c, the type 2 diabetes ProteinScore and a polygenic risk score (PRS) for type 2 diabetes individually and concurrently.
In a sensitivity analysis, we considered metabolomic and proteomic features for score generation. Type 2 diabetes and all-cause mortality were chosen as case studies as: 1) they had a large number of cases available; 2) ProteinScores for these traits were amongst the top-performing and 3) there is evidence that both traits may be stratified by metabolomic features. A total of 12,050 of the 47,600 individuals with protein data had metabolomics measures (Supplementary Information). Test sets used for ProteinScores were subset to those with metabolomics data, for type 2 diabetes (N\textsubscript{cases\_train} = 377, N\textsubscript{controls\_train} = 1,002, N\textsubscript{cases\_test} = 309, N\textsubscript{controls\_test} = 898) and mortality (N\textsubscript{cases\_train} = 616, N\textsubscript{controls\_train} = 1,680, N\textsubscript{cases\_test} = 410, N\textsubscript{controls\_test} = 1,048). Performance of a MetaboScore (considering metabolite features), ProteinScore (considering protein features) and MetaboProteinScore (combining metabolomic and proteomic features) is summarised for both traits in Extended Data Fig. 7, with further output provided in Supplementary Table 16. The selected features and weights for each score are available in Supplementary Table 17. For all-cause mortality, the ProteinScore (AUC = 0.82) outperformed the MetaboScore (AUC = 0.69), with an AUC of 0.83 when both individual scores were modelled concurrently. For type 2 diabetes, the ProteinScore (AUC = 0.87) and MetaboScore (AUC = 0.85) were more comparable in performance, with an additive AUC of 0.89 when both individual scores were modelled concurrently.
10.5 Discussion

This study quantified circulating proteome signatures that are reflective of multiple incident diseases in mid-to-later life. These data suggest that augmenting traditional risk factors with proteomic, metabolomic and genetic data types may further hone risk stratification.

We demonstrated that relatively few circulating proteins can add value to risk stratification, up to a decade prior to formal diagnoses. ProteinScores for incident type 2 diabetes, COPD, ischaemic heart disease, Alzheimer’s dementia, Parkinson’s disease and death demonstrated value beyond a comprehensive set of 26 covariates; equal or higher AUCs were observed for models including all covariates compared to those with just the ProteinScore. This suggests that ProteinScores can absorb a large proportion – if not all – of the typical covariate signal. The scores minimise the need for extensive recording of lifestyle, physical and biomarker measures, offering a streamlined set of metrics to proxy for an individual’s health status.

While much interest is currently devoted to employing PRS for disease prediction, they neglect environmental components of disease risk and may therefore be limited in the context of complex age-related disease. Our ProteinScore for type 2 diabetes outperformed the PRS, which is likely due to proteins representing an interface that captures genetic, environmental and lifestyle contributions to disease risk. The improvement in AUC resulting from concurrent modelling of HbA1c and the type 2 diabetes ProteinScore suggests that the latter provides additional predictive value.

Our results also suggest that jointly considering ProteinScores with scores generated using metabolomic features may further augment risk stratification. An additive improvement resulting from the addition of the MetaboScore to the ProteinScore
A subset of the individual protein-disease associations we report likely represent direct mediators of disease. We encourage further exploration of this via techniques such as Mendelian randomisation and colocalisation. Modelling that takes into account multimorbidity trajectories over the life course would also aid in understanding the role of prevalent diseases and medication use on future disease risk. The largest number of associations and strongest effect sizes (by magnitude of the absolute log of the hazard ratio) were observed for liver disease. For neurological diseases and cancers, where fewer associations were identified, it is possible that bulk blood is less able to
capture the full spectrum of disease pathogenesis, which may be localised to distal or
more refined tissues. Similarly, the panel of proteins available may reflect certain
diseases better than others. Despite having relatively few individual protein
associations, the Alzheimer’s dementia ProteinScore was both one of the best-
performing and largely unchanged upon addition of covariates. As therapeutic
interventions for neurodegenerative diseases have greater efficacy when
implemented earlier in the disease pathogenesis, ProteinScores such as this
may help with trial recruitment strategies. Correlations between the covariates and
ProteinScores used in this work (Supplementary Table 18) suggest that the former
reflect a range of lifestyle, physiological and health measures – indicating that they
may be useful measures that can proxy for health status.

Of the 720 proteins that were identified as indicators of multimorbidity status, 716 were
associated with age (Bonferroni-adjusted \( P < 1.7 \times 10^{-5} \), with 648 having positive effect
sizes) in a previous analysis of the same dataset (See Table S5 from Sun et al 2023
\(^{144}\)). Future studies could explore their possible causal contributions to disease and
whether they have differential effects across the lifecourse. Examples of such proteins
include growth differentiation factor 15 (GDF15), Interleukin-6 (IL6) and plasminogen
activator urokinase receptor (PLAUR) – three proteins that had the largest number of
associations with individual incident diseases in our study. GDF15 was previously
identified as the top marker of future multimorbidity from 1,301 plasma proteins tested
\(^{294,350}\). IL6 mediates chronic, low-grade inflammation, is a key biomarker of ageing \(^{351}\),
with anti-IL6 developed for a range of inflammation-associated diseases \(^{352,353}\).
PLAUR has previously been associated with incident cancer, cardiovascular disease,
diabetes \(^{354}\). Similarly, increased levels of neurofilament light (NEFL) – a marker
indicative of synapse degeneration \(^{232,355}\) – associated with incidence of multiple
neurological traits (Parkinson’s disease, Alzheimer’s dementia, multiple sclerosis, amyotrophic lateral sclerosis and ischaemic stroke).

This study has several limitations. First, the assessment of scores via regression within a test sample, followed by the calculation of an AUC is not direct prediction and cannot translate easily to new populations. Second, non-random selection of disease cases through the UKB-PPP consortium may have introduced biases. Third, it was not possible to source an external test set for the ProteinScores with sufficient incident case counts to enable a meaningful replication assessment. Fourth, variation in protein analyte levels across measurement technologies has been reported. Fifth, the protein measured were recorded in relative scale, which limits translation of scores to new populations. Sixth, death was treated as a censoring event; competing risks and multi-state modelling approaches may provide a more nuanced analytical strategy. Finally, the UK Biobank population is largely comprised of individuals with European ancestry and a restricted age range (40-71 years, with a mean of 57 years); future studies in equally well-characterized cohorts will be needed to assess translation to other populations, age ranges and ethnicities.

Taken together, this study identified proteomic signatures of disease and highlighted the value of the plasma proteome for 10-year risk stratification.

10.6 Methods

10.6.1 The UK Biobank sample population

UK Biobank (UKB) is a population-based cohort of around 500,000 individuals aged between 40-69 years that were recruited between 2006 and 2010. Genome-wide
genotyping, exome sequencing, electronic health record linkage, whole-body magnetic resonance imaging, blood and urine biomarkers and physical and anthropometric measurements are available. More information regarding the full measurements can be found at: https://biobank.ndph.ox.ac.uk/showcase/. The UK Biobank Pharma Proteomics Project (UKB-PPP) is a precompetitive consortium of 13 biopharmaceutical companies funding the generation of blood-based proteomic data from UKB volunteer samples.

10.6.2 Proteomics in the UK Biobank

The UKB-PPP sample includes 54,219 UKB participants and 1,474 protein analytes measured across four Olink panels (Cardiometabolic, Inflammation, Neurology and Oncology: annotation information provided in Supplementary Table 1)\(^{144}\). A randomised subset of 46,595 individuals were selected from baseline UKB, with 6,376 individuals selected by the UKB-PPP consortium members and 1,268 individuals included that participated in a COVID-19 study. The randomised samples have been shown to be highly representative of the wider UKB population, whereas the consortium-selected individuals were enriched for 122 diseases\(^{144}\). Details on sample selection for UKB-PPP, in addition to processing and quality control information for the Olink assay are provided in Supplementary Information. Of 54,219 individuals that had protein data measured, there were 52,744 that were available after quality control exclusions (as per Sun et al 2023\(^{144}\)) with 1,474 Olink protein analytes measured (annotations in Supplementary Table 1)\(^{144}\). The sample is predominantly white/European (93%), but also has individuals with black/black British, Asian/Asian British, Chinese, mixed, other and missing ethnic backgrounds (7%). Sun et al (2023) includes associations between the protein levels studied here and age, sex, lifestyle and health factors\(^{144}\).
Extended Data Fig. 2 summarises the processing steps applied to this dataset to derive a complete set of measurements for use. Briefly, of 107,161 related pairs of individuals (calculated through kinship coefficients > 0 across the full UKB cohort), 1,276 pairs were present in the 52,744 individuals. After exclusion of 104 individuals in multiple related pairs, in addition to one individual randomly selected from each of the remaining pairs, there were 51,562 individuals. A further 3,962 individuals were excluded due to having >10% missing protein measurements. Four proteins that had >10% missing measurements (CTSS.P25774.OID21056.v1 and NPM1.P06748.OID20961.v1 from the neurology panel, PCOLCE.Q15113.OID20384.v1 from the cardiometabolic panel and TACSTD2.P09758.OID21447.v1 from the oncology panel) were then excluded. The remaining 1% of missing protein measurements were imputed by K-nearest neighbour (k=10) imputation using the impute R package (Version 1.60.0)\(^\text{356}\). The final dataset consisted of 47,600 individuals and 1,468 protein analytes. Assessments of protein batch, study centre and genetic principal components suggested that these factors had minimal effects on protein levels (lowest correlation between protein levels and residuals of 0.94) (Supplementary Information). Therefore, protein levels were not adjusted for these factors.

10.6.3 Phenotypes in the UK Biobank

Demographic and phenotypic information for the 47,600 individuals with complete protein data for 1,468 analytes are available in Supplementary Table 2. Lifestyle covariates included: BMI (weight in kilograms divided by height in metres squared), alcohol intake frequency (1 = Daily or almost daily, 2 = Three-Four times a week, 3 = Once or twice a week, 4 = One-Three times a month, 5 = Special occasions only, 6 = Never), the Townsend index of deprivation (higher score representing greater levels
of deprivation) and smoking status (0 = Never, 1 = Previous, 2 = Current), physical activity (0 = between 0-2 days/week of moderate physical activity, 1 = between 3-4 days/week of moderate physical activity, 2 = between 5-7 days/week of moderate physical activity) and education status (1 = college/university educated, 0 = all other education). Of the 47,600 individuals with complete protein data, there were 52, 52, 236, 56 and 59 missing entries for alcohol, smoking, BMI, physical activity and deprivation, respectively. No imputation of missing data was performed for the inclusion of these variables in individual Cox PH analyses. There were an additional 2,556, 188 and 59 individuals that answered ‘prefer not to answer’ and were excluded from physical activity, smoking and alcohol variables, respectively.

10.6.4 Electronic health data linkage in the UK Biobank

Electronic health linkage to NHS records was used to collate incident diagnoses. Death information was sourced from the death registry data available through the UK Biobank. Cancer outcomes were sourced from the cancer registry (ICD codes), whereas non-cancer diseases were sourced from first occurrence traits available in the UK Biobank. The first occurrence traits integrate GP (read2/3), ICD (9/10) with self-report and ICD codes present on the death registry to identify the earliest date of diagnoses. These data sources are linked to 3-digit ICD trait codes. A summary of codes used to extract each of the outcomes included in the present study are detailed in Supplementary Information. The following 23 diseases were included: liver disease, systemic lupus erythematosus, type 2 diabetes, amyotrophic lateral sclerosis, Alzheimer’s dementia, endometriosis, chronic obstructive pulmonary disease (COPD), inflammatory bowel disease, rheumatoid arthritis, ischaemic stroke, Parkinson’s disease, vascular dementia, ischaemic heart disease, major depressive disorder, schizophrenia, multiple sclerosis, cystitis and lung, prostate, breast, gynaecological,
brain/CNS and colorectal cancers. These represent a selection of leading age-related causes of morbidity, mortality and disability. In all analyses involving sex-specific diseases, the population was stratified to males or females and sex was not included as a covariate in incremental Cox PH assessments. Traits that were stratified included gynaecological cancer, breast cancer, endometriosis and cystitis (all female-stratified) and prostate cancer (male-stratified).

10.6.5 Incident disease calculation in the UK Biobank

Dates of diagnoses for each disease were ascertained through electronic health linkage. Using the date of baseline appointment, time-to-first-onset for each diagnoses in years was calculated. Time-to-onset for controls was defined as the time from baseline to censoring date (Supplementary Information). Death was treated as a censoring event. Time-to-censor date was calculated for the controls that remained alive, whereas if a control individual had died during follow-up time-to-death was taken forward for Cox PH models. Any cases that were prevalent at baseline were excluded. Alzheimer’s and vascular dementias were restricted to age at onset (or censoring) of 65 years or older in all analyses. Sex-specific traits were stratified across all analyses.

10.6.6 Individual Cox proportional hazards analyses

Cox proportional hazards models were run between each protein and each incident disease using the ‘survival’ package (Version 3.4-0) in R (Version 4.2.0). Protein levels were rank-based inverse normalised and scaled to have a mean of 0 and standard deviation of 1 prior to analyses. Minimally-adjusted Cox PH models for sex-stratified traits included age at baseline as a covariate, whereas the remaining models adjusted for age and sex. Lifestyle-adjusted models further controlled for education status, BMI, smoking status, social deprivation rank, physical activity and alcohol
intake frequency. A Bonferroni-adjusted P-value threshold for multiple testing based on the 678 components that explained 90% of the cumulative variance in the 1,468 protein analyte levels (Supplementary Table 3) and 24 outcomes tested was applied across all Cox PH models ($P < 0.05/(678 \times 24) = 3.1 \times 10^{-6}$ used as the Bonferroni-adjusted P-value threshold). Proportional hazards assumptions were checked through examination of protein-level Schoenfeld residuals.

A sensitivity analysis was performed for each of the 35,232 fully-adjusted associations tested, restricting cases to successive years of follow-up. These sensitivity analyses were visualised using the Shiny package (Version 1.7.3) in R. The magnitude of change in hazard ratios for individual associations can be examined by year of case follow-up to assess consistency of effect sizes. Whether marker associations are stronger or weaker when restricting to cases occurring in the near-term (1-5 years of follow-up) can also be examined. A network visualisation was also created within the Shiny interface to highlight the fully-adjusted associations that had $P < 3.1 \times 10^{-6}$ using networkD3 (Version 3.0.4) and igraph (Version 1.3.5) R packages. To further verify the markers of multiple morbidities identified in individual Cox PH analyses, logistic regression models were also run between each of the 1,468 protein analyte levels and multimorbidity status (defined as 1,454 individuals that received 3 or more of the 23 disease diagnoses over the 15-year follow-up period). A sensitivity analyses was also run for ischaemic heart disease associations with/out adjustment for blood-pressure lowering medication reported at baseline in a subset of individuals (35,073 of 47,600) that had medication information available. Supplementary Information provides details on the classification of medications as per the anatomical therapeutic chemical (ATC) classification categories. A total of 14,074 individuals (of the 35,073) indicated they were taking one or more of the above blood-pressure lowering
medications at baseline. This was treated as a binary variable and the comparison with/out adjustment for this variable was performed for ischaemic heart disease Cox PH associations in the subset of 35,073 individuals. Adjustments for age, sex and six lifestyle factors were included in both sets of analyses, with 2,456 cases, 27,468 controls.

10.6.7 ProteinScore development

MethylPipeR is an R package with accompanying user interface that we have previously developed for systematic and reproducible development of incident disease predictors. Using MethylPipeR, ProteinScores that considered 1,468 Olink protein levels were trained using Cox PH elastic net regression via the R package Glmnet (Version 4.1-4). Penalised regression minimises overfitting by the use of a regularisation penalty and the best shrinkage parameter (λ) was chosen by cross-fold validation with alpha fixed to 0.5. Of the 24 outcomes featured in the individual Cox PH analyses, 19 that had a minimum case count of 150 were selected for ProteinScore development. The chosen strategy for ProteinScore development included training ProteinScores for each trait across fifty randomised iterations (with each iteration including a different combination of cases and controls in train and test sets). This strategy quantifies the stability of the ProteinScore performance, which is critical given that unobserved confounders that may be enriched during random selection of individuals from the wider population. The ProteinScore training strategy is summarised in Extended Data Figure 9. Briefly, 50 iterations of each ProteinScore were performed that randomised sample selection by 50 randomly sampled seeds (values between 1 and 5000). For each iteration, cases and controls were randomly split into 50% groups for training and testing. From the 50% training control population, a subset of controls were then randomly sampled to give a case:control ratio of 1:3 in
order to balance the datasets. For traits with over 1000 cases in training samples 10 folds were used. For traits with between 500 and 1000 cases in training, five folds were used. Three folds were used when there were fewer than 500 cases in the training sample. Protein levels were rank-based inverse normalised and scaled to have a mean of 0 and standard deviation of 1 in the training set. The linear combination of weighting coefficients for selected protein features from cross-validation within the folds of the training set were then used to generate a ProteinScore for each individual in the test samples. Of the 50 training iterations tested, models that had no features selected were documented (Supplementary Table 11).

10.6.8 Assessment of ProteinScore performance

Cumulative time-to-onset distributions for cases (Extended Data Figs. 5-6) indicated that amyotrophic lateral sclerosis, endometriosis and cystitis were better-suited to 5-year onset assessments in the test sample (80% of cases were diagnosed at 8-years post-baseline). All remaining ProteinScores were tested in the context of 10-year onset (80% of cases were not diagnosed 8-years post-baseline). Across the 50 ProteinScore iterations for each trait, 50% of cases and controls that were not randomly selected for training were reserved for testing. For a visualisation of the test set sampling and assessment strategy, see Extended Data Fig. 9. In the test set, cases that had time-to-event up to or including the 5-year or 10-year thresholds used for onset prediction were selected, while cases beyond the threshold were placed with the control population, which was then randomly sampled in a 1:3 ratio. Weighting coefficients for features selected during ProteinScore training were used to project scores into the test sample. Incremental Cox PH models were run in the test sample to obtain cumulative baseline hazard and onset probabilities, which were used to derive AUC estimates. The test set sampling strategy ensured that while the majority of cases occurred up to
the onset threshold, there were a small proportion (~3%) of cases included in Cox PH models with onset times after the 10- or 5-year threshold, to simulate a real-world scenario for risk stratification. If cases fell beyond the 5-year or 10-year threshold for onset, they were recoded as controls in the AUC calculation. Cumulative baseline hazard probabilities were calculated using the Breslow estimator available in the ‘gbm’ R package (Version 2.1.8.1) \(^{362}\). Survival probabilities were then generated through taking the exponential of the negative cumulative baseline hazard at 5 or 10 years to the power of the Cox PH prediction probabilities. ProteinScore onset probabilities were calculated as one minus these survival probabilities. AUC and ROC statistics were extracted for the survival probabilities using the calibration function from the ‘caret’ R package (Version 6.0-94) \(^{363}\) and the evalmod function from the ‘MLmetrics’ R package (Version 1.1.1) \(^{364}\).

ProteinScores that yielded the median incremental difference to the AUC of a minimally-adjusted model (adjusting for age- or age- and sex) were selected from the fifty possible ProteinScores for each trait. If no features were selected during training, models were weighted as performance of 0 in the median model selection. In some instances, features were selected during training and incremental Cox PH models were run successfully, but the random sampling of the test set did not include a case with time-to-event at or after the 5-year or 10-year onset threshold. Therefore, these models were excluded as cumulative baseline hazard distributions did not reach the onset threshold and could not be extracted for AUC calculations. The number of models, with minimum and maximum performance was documented (Supplementary Table 11). Taking this approach mitigated against the presence of extreme case:control profiles driving ProteinScore performance and minimised the possibility
of bias being introduced by selecting train and test samples based on matching for specific population characteristics.

Selected ProteinScores for each trait were then evaluated to quantify the additional value (in terms of increases in AUC) that resulted from the addition of ProteinScores. Minimally-adjusted models included age and sex (if traits were not sex-stratified). Lifestyle-adjusted models then further accounted for common lifestyle covariates (education status, BMI, smoking status, social deprivation rank, physical activity and alcohol intake frequency). Finally, models included covariates from the minimally-adjusted, lifestyle-adjusted and an extended set of clinically-measured variables were then assessed (see Extended Data Fig. 7). In each case, the difference in AUC resulting from the addition of the ProteinScore was reported. ROC P-value tests were used to ascertain whether the improvements offered by selected ProteinScores for each outcome were statistically significant, beyond each set of increasingly saturated covariates. A Bonferroni-adjusted P-value threshold for ROC P tests was used based on the 19 ProteinScore traits (P < 0.05/19 = 0.0026). The ‘precrec’R package (Version 0.12.9) was used to generate ROC and Precision-Recall curves for each ProteinScore. A series of models that included only the ProteinScore were also considered for each outcome, to quantify whether protein data alone could absorb much of the predictive performance achieved by the covariates.

A set of 26 possible covariates used across the minimally-adjusted, lifestyle-adjusted and extended set analyses were assessed for missingness, imputed (where missingness was < 10%) and utilised in ProteinScore evaluation as a maximal, extended set of covariates. Further details of variable selection and preparation are supplied in Supplementary Information. Additional covariates (considered in
addition to age, sex and six lifestyle traits that were used in individual Cox PH analyses) included: leukocyte counts (10^9 cells/Litre), erythrocyte counts (10^12 cells/Litre), haemoglobin concentration (grams/decilitre), mean corpuscular volume (femtolitres), platelet count (10^9 cells/Litre), cystatin C (mg/L), cholesterol (mmol/L), alanine aminotransferase (U/L), creatinine (umol/L), urea (mmol/L), triglycerides (mmol/L), LDL (mmol/L), CRP (mg/L), aspartate aminotransferase (U/L), glycated haemoglobin – HbA1c – (mmol/mol), albumin (g/L), glucose (mmol/L) and systolic blood pressure (mmHg). After covariate processing steps were complete, a population of 43,437 individuals was available with complete information for ProteinScore testing. Phenotypic summaries of the additional covariates for this population are summarised in Supplementary Table 2.

10.6.9 Further assessment of the type 2 diabetes ProteinScore

Glycated Haemoglobin (HbA1c) is a blood-based measure of chronic glycemia that is highly predictive of type 2 diabetes events and is recommended as a test of choice for the monitoring and diagnosis of type 2 diabetes. HbA1c (mmol/mol) measurements (fieldID 30750) and the type 2 diabetes polygenic risk score (PRS) available in UK Biobank (fieldID 26285) were extracted. A contour plot showing both variables grouped by those who went on to be diagnosed with type 2 diabetes over a 10-year period was created. HbA1c levels were also plotted against ProteinScore risk deciles. HbA1c and the ProteinScore levels were rank-based inverse normalised and assessed individually and concurrently in incremental models for 10-year onset of type 2 diabetes in the ProteinScore test set. A Pearson correlation coefficient (r) between the transformed HbA1c and ProteinScore levels was calculated. The 10-year incremental Cox PH models were used to derive onset probabilities for calculation of AUCs after adding the ProteinScore to models adjusting for HbA1c and the type 2
diabetes PRS. Model comparisons were used (test of the difference in ROC curves) to quantify the value added by the ProteinScore beyond the PRS and HbA1c.

10.6.10 Metabolomic score comparison

Metabolomics measures were available in 12,050 of the 47,600 individuals with proteomic data included in the study (see Supplementary Information for details on data preparation). Type 2 diabetes and death were chosen as case studies for further exploration. The train and test sets used to develop the main ProteinScores were subset to those with metabolomics available for type 2 diabetes (N cases\_train = 377, N controls\_train = 1,002, N cases\_test = 309, N controls\_test = 898) and death (N cases\_train = 616, N controls\_train = 1,680, N cases\_test = 410, N controls\_test = 1,048). Scores that considered only metabolomic features (MetaboScore), only proteomic features (ProteinScore) and joint omics features (MetaboProteinScore) were trained and tested in these populations. There were 249 metabolite measures (comprised of 168 metabolites and 81 ratios between combinations of metabolites) and 1,468 protein levels considered as potentially-informative features. Performance was evaluated for 10-year onset of type 2 diabetes and death in the test sample, modelling scores individually, concurrently and benchmarking them against the maximal set of 26 possible covariates (see Extended Data Fig. 7).
10.7 Conclusion

In this study, the proteomic profiles of 23 incident diseases and death have been profiled. ProteinScores for the 10-year onset of six incident outcomes added value beyond an extensive set of 24 demographic, lifestyle and clinical covariates. These ProteinScores were for Alzheimer’s Dementia, Parkinson’s disease, COPD, death, type 2 diabetes and ischaemic heart disease. My work suggests that the ProteinScores for these diseases can absorb much of the predictive signal offered by an extensive set of classical covariates, which may offer a route to streamline risk stratification in future. The 3,209 age, sex and lifestyle-adjusted individual protein associations with incident diseases that I identify highlight proteins that may represent causal mediators of disease. These relationships should be examined further in future using pQTLs for the proteins as instruments in Mendelian randomisation analyses. Replication of individual biomarker associations and ProteinScore signals should be attempted in more diverse population, which varied disease profiles, lifestyle, ancestries and age distributions. The Shiny app that I developed allows users to examine temporal variability for each of the 35,232 protein-disease associations tested, in addition to visualising proteins that associate with multiple incident outcomes in an interactive network.

In the next chapter, I discuss the findings from Chapters 6-10. I begin by focusing on the protein EpiScores that I have developed across studies included as Chapters 8 and 9. I then focus on the ProteinScores developed in Chapter 10. Finally, I discuss the molecular mapping studies presented in Chapter 6-7. I also discuss limitations and strengths regarding the cohorts, molecular samples and methodologies used.
11 Discussion

In this thesis, I integrated omics layers from the blood of many thousands of individuals from five cohort studies. I first probed the molecular signatures of brain health. I did this by carrying out MWAS and/or GWAS studies of circulating proteins linked to brain health (Chapters 6-7). I then created protein EpiScores, linking information from the methylome with proteomic outcomes and demonstrating that these scores can be novel tools for quantifying disease risk and brain health (Chapters 8-9). Finally, I mapped the early proteomic signatures of 24 incident outcomes, enriched for neurological diseases, and created ProteinScores to stratify 10-year onset of these outcomes (Chapter 10). In this chapter, I discuss the findings from Chapters 6-10. I start with a focus on the scoring approaches presented in Chapters 8-10, finishing with the studies mapping omics signatures of proteins associated with brain health in Chapters 6-7. I then provide a critical evaluation of limitations associated with the cohort populations, assays and statistical methods used in this thesis. I finish by recommending directions for future research.

11.1 Protein EpiScores

I developed EpiScores for 109 plasma proteins (Chapter 8) and two additional protein EpiScores for serum levels of GDF15 and NT-proBNP (Chapter 9). In my initial study presented in Chapter 8, I considered the levels of 953 possible plasma proteins and used elastic net penalised regression to train EpiScores for them. When tested in independent populations, 109 of the scores were retained (Pearson $r > 0.1$ and $P < 0.05$ when tested against the measured protein). When projected into DNAm at Generation Scotland baseline ($n=9,537$) and modelled in Cox PH, 70 EpiScores were associated with eleven disease outcomes in 130 associations. These results were observed after adjustments for age, sex and lifestyle factors (alcohol consumption, educational attainment, a DNAm-derived score for smoking, social deprivation and BMI). GrimAge outperforms other DNAm scores for ageing in associations with age-related disease outcomes and mortality. Additionally, separating cell-type mediated effects from those that are not is of importance. Therefore, I performed sensitivity analyses on the Cox PH associations, assessing attenuation in the EpiScores-disease associations due to these variables. I saw that of 130 associations,
99 and 78 remained statistically significant (FDR P < 0.05) after further adjustment for estimated immune cell proportions and GrimAge acceleration, respectively. This suggests that the collection of protein EpiScores I have generated provide additional, disease-relevant signals beyond these factors. Another aspect of this study that further validates the potential of EpiScores to capture protein-based signals for disease risk stratification is the replication assessment that I performed in the context of type 2 diabetes. I found that of the 34 SomaScan-derived protein EpiScores associated with incident type 2 diabetes, 28 were consistent with those reported by previous large-scale proteomic studies of prevalent or incident type 2 diabetes in the literature^{327,367}. The WFIKKN2 protein has been reported to be putatively causal for type 2 diabetes in a Mendelian randomisation study and was highlighted by our EpiScore for WFIKKN2 as a marker of incident type 2 diabetes. The individual protein EpiScores I have generated can inform on the health state of individuals, tailored to a specific protein.

I have shown that EpiScores for the 109 protein levels may be of value to early stratification of disease risk. In populations like GS that have excellent phenotypic depth and incident disease mapping but do not have protein data, the EpiScores provide a means to proxy for the protein measures. As GS did not have the relevant proteomics measures available, I could not directly compare protein EpiScores with their equivalent protein measures in Cox PH models employed in Chapter 8. Quantifying whether EpiScores can replicate protein-disease signals in comparisons within the sample is of value when identifying the optimal biomarkers to use in future applications. In Chapter 9, I designed a study that could facilitate the direct comparison of protein EpiScores with equivalent proteins in Cox PH approaches to address this question. I developed EpiScores for serum GDF15 and NT-proBNP, initially training and testing in subsets of Generation Scotland to enable direct comparison of the EpiScores and measured proteins for Cox PH against four incident diseases in the test sample. I found that the GDF15 EpiScore replicated the association observed with the measured protein and incident dementia, stroke and type 2 diabetes. NT-proBNP replicated the inverse association between NT-proBNP and type 2 diabetes, but did not replicate the association between the measured protein and incident stroke. As neither GDF15 nor NT-proBNP were in the set of 953 proteins originally assessed in Chapter 8, these represented novel EpiScores. A previous EpiScore for GDF15 was one of the seven protein EpiScores used to derive
GrimAge. After retraining my EpiScores in the maximal Generation Scotland sample available, I directly compared the performance of my GDF15 EpiScore with the GDF15 EpiScore generated through GrimAge in the external LBC1936 test population (n=322). My EpiScore outperformed the GrimAge GDF15 EpiScore (an increase in $R^2$ of 3.3%), which is likely due to the larger training sample available (n=17,489 in my study, versus n=2,895 in the GrimAge study). In Chapter 10, I identify GDF15 as the top candidate for incident multimorbidity (of 1,468 proteins and 23 diseases tested in 47,600 individuals), which has been supported by previous work.

In the analyses included in Chapter 9, my GDF15 EpiScore emerged as a marker of cross-sectional cognitive ability and structural MRI traits in the LBC1936 population. Protein measurements of GDF15 were not available across the LBC1936 waves to run direct comparisons between the GDF15 protein and the GDF15 EpiScore in associations with brain health traits. However, a previous study observed 6.4-fold stronger associations with structural MRI measures using an EpiScore for CRP versus measured CRP. I hypothesise that this would also likely be the case for the GDF15 EpiScore, as GDF15 has a well-documented role in chronic inflammation. These direct comparisons should also be tested for the 84 SomaScan-derived EpiScores developed in Chapter 8, as these were trained in the KORA population and could therefore be evaluated in both STRADL and LBC1936 cohorts (that have structural brain imaging and cognitive test data available). Although multiple time point measures of proteins and their equivalent EpiScores were not available in the cohort populations used in Chapters 8-9, I would recommend that studies integrate data across multiple time points in future to assess biomarker stability. DNAm may reflect a more stable record of the body’s response to exposures such as chronic inflammation than protein levels, that are known to be variable (especially acute-phase and inflammatory proteins). A previously-developed CRP EpiScore had a higher test-retest reliability as compared to measured CRP, indicating that its EpiScore could offer a more reliable measure of chronic inflammation than direct protein measurement. This should also be quantified for all protein EpiScores that I developed in Chapters 8-9, when sufficient data become available.
The EpiScores generated in Chapters 8-9 represent a linear combination of weighted CpG sites selected as features. The individual sites are not necessarily the most biologically meaningful in relation to the protein. While EpiScores can be modelled in place of proteins for disease prediction, caution must be taken when modelling EpiScores in Mendelian randomisation (MR) approaches to assess evidence of causality. A GWAS by McCartney et al. (2021) identified 137 genome-wide significant genetic variants that associated with EpiScores for age measures. These variants were used as instruments for causal association testing in MR, identifying putatively causal relationships with GrimAge (adiposity traits, smoking cessation and educational attainment). In subsequent work, an MR association was also observed between GrimAge acceleration and colorectal cancer (OR = 1, with 95% CI 1.04–1.20 and P = 0.002). However, McCartney et al. advise caution when interpreting the MR results due to weak instruments and potential horizontal pleiotropy. MR relies on the assumptions that: 1) instruments are associated with the exposure, 2) instruments are associated with the outcome only through the exposure and 3) instruments are independent of unobserved confounders influencing the exposure and the outcome after conditioning on observed confounders. Type I pleiotropy involves the instrument influencing many other phenotypes and this can be minimised by restricting instruments to those with plausibility to act directly on the exposure (e.g. a *cis* pQTL acting on a protein as the exposure). As such, modelling complex scores such as an EpiScore (reliant on multiple CpG sites) as an exposure may lead to higher likelihood of pleiotropy and violations in the assumptions of MR. For this reason, I chose not to prioritise running MR involving protein EpiScores in the present work. However, this could be explored in future with careful assessment of these considerations.

11.2 ProteinScores

In Chapter 10, I profiled individual Cox PH associations between 1,468 plasma proteins and 23 incident diseases and mortality (n=47,600). I then developed ProteinScores for the onset of these time-to-event outcomes through elastic net Cox PH regression. Of 19 outcomes that had more than 150 cases, ProteinScores for six outcomes added value in 10-year onset assessments beyond a comprehensive set of demographic covariates, six lifestyle factors and 18 clinically-relevant biomarkers and physical measures. ProteinScores for eight outcomes were also able to absorb much
of the signal offered by the set of 26 possible covariates. By choosing a staged approach to the ProteinScore assessment (informed by a previous metabolomics scoring study \(^{227}\)), I was able to observe augmentation in the AUC difference offered by ProteinScores beyond an increasing number of covariates. For example, a large improvement was observed for the ProteinScore for lung cancer beyond a minimally-adjusted model accounting for age and sex (AUC increased from 0.71 to 0.85 due to the addition of the ProteinScore). When lifestyle factors including smoking (known to account for 70-90\% of the risk of lung cancer \(^{372}\)) were introduced to the minimally-adjusted model, the AUC rose to 0.85 with the further addition of the ProteinScore yielding an AUC of 0.88. However, the ROC model comparison with/out the ProteinScore was no longer statistically significant (P = 0.01, which was above the Bonferroni adjusted threshold of 0.0026). Of note, was the resilience of the ProteinScore signal for Alzheimer’s dementia, which had performance that was largely unaffected by the staged addition of covariates into the incremental Cox model in the test set. An increase in AUC of 0.04 was observed due to the ProteinScore, even when adjusting for the maximal set of 26 covariates. The Parkinson’s disease ProteinScore performance also remained largely unaffected by the adjustment for further risk factors, with a robust addition of 0.04 in AUC added across each of the three staged models tested. As Alzheimer’s dementia and Parkinson’s disease manifest early in the brain without overt symptomatic presentation that is easily detectable \(^{348,373}\), it is imperative to identify high risk individuals many years prior to symptom onset. These individuals can be recruited to clinical trials more effectively, while also potentially benefitting from earlier preventative action.

The Shiny app (https://protein-disease-ukb.optima-health.technology, Username: ukb_diseases, Password: UKBshinyapp) that I chose to develop in Chapter 10 visualises the 3,209 fully-adjusted, individual Cox PH associations. Individuals can search for top biomarkers and view proteins in an interactive network showing multi-protein and multi-disease view. This feature allows for exploration of the proteins I identified as markers for multiple incident morbidities. It also allows for evaluation of protein-disease associations by successive yearly intervals of case follow-up, to understand the stability of marker associations when considering near-term versus longer-term diagnoses. The top markers identified in the Cox PH associations hosted in this Shiny app are likely to be more biologically-relevant to a given disease than the
collection of proteins selected for ProteinScores. For example, of the proteins associated with incident Alzheimer’s dementia (NEFL, GFAP, BRK1, BCAN, NPTXR and GDF15), the majority are linked to neurological or synaptic function: NEFL (axon al cytoskeleton and neuronal transport) \(^{231}\), GFAP (glial cells and blood brain barrier function) \(^{374}\), BRK1 (actin microtubule organization) \(^{375}\), BCAN (regulation of brain extracellular matrix) \(^{376}\) and NPTXR (neuroplasticity and synaptogenesis) \(^{377}\). In contrast, GDF15 is a marker of multiple morbidities and I hypothesise that the association with Alzheimer’s dementia for this protein may be picking up morbidities such as type 2 diabetes and cardiovascular disease, which are risk factors for dementia onset that can occur many years prior. Increased levels of NEFL were associated with higher incidence of multiple neurological traits (Alzheimer’s dementia, Parkinson’s disease, ALS, multiple sclerosis and ischaemic stroke). These diseases are hallmarked by neuron degradation, which may indicate that NEFL may be a consequential marker released into circulation when synapses breakdown \(^{232,355}\). To delineate proteins that are likely mediators of disease, I encourage further exploration through techniques such as Mendelian randomisation and colocalisation \(^{233,239}\).

Across Chapters 8-10, incident type 2 diabetes emerged as a trait that is particularly well-captured by the epigenetic and proteomic scores I developed. My preliminary assessment comparing the type 2 diabetes ProteinScore to a metabolomic score and a joint (metabolomic + protein) score suggests that while diabetes is typically defined as a metabolic disease, the ProteinScore absorbs much of the predictive signal offered by metabolomics. However, an additive signal when modelling both omics types was also observed. Having type 2 diabetes is one of the leading risk factors for subsequent dementia onset \(^{12}\). Brain ageing has also been shown to accelerate by around 26% in individuals with progressive type 2 diabetes compared to those without diabetes \(^{44}\). Preventing type 2 diabetes onset should be a goal for the preservation of brain health. The EpiScores and ProteinScore for type 2 diabetes that I contribute to the field may be useful in these efforts. While modest, the type 2 diabetes ProteinScore did add value beyond HbA1c, the well-validated, clinically-used marker \(^{344,345}\).

11.3 Molecular and phenotypic signatures of protein levels

In Chapters 6 and 7, I performed molecular association studies of blood-based protein markers relevant to brain health. The GWAS of S100β in Chapter 6 led to the
identification of one sentinel SNP (rs8128872, \(P = 5.0 \times 10^{-17}\)). When modelled in MR analyses, that there was no evidence of a putatively causal relationship between S100\(\beta\) and Alzheimer’s dementia. My results suggest that despite the links made between S100\(\beta\) and Alzheimer’s dementia in the literature \(276,277,378–380\), the former may be an indicator of neuroinflammation more widely. Of interest, is a study that utilised the GWAS summary statistics I generated in LBC1936 (n=769) to run MR for five neuropsychiatric diseases \(280\). The authors replicated the null result I identified between S100\(\beta\) and Alzheimer’s dementia, but found an association between S100\(\beta\) and bipolar disorder (OR = 1.07, 95% CI 1.03–1.13, \(P = 0.002\)).

My MWAS findings in Chapter 6 did not identify any CpG site that was associated with S100\(\beta\) levels with epigenome-wide significance. This null result highlights an important point, which is that not every protein has a detectable DNAm signature. This may be due to sample size restrictions precluding the detection of CpG-protein signatures, or it may be that there are no probes that biologically associate with the protein. In Chapter 7, I applied a stringent Bonferroni-adjusted threshold, reporting 2,928 pQTMs that were detectable for 191 of a possible 4,058 proteins tested. Two of the proteins (PAPPA and PRG3) accounted for 987 and 1,116 associations, respectively, leaving 825 pQTM associations involving 189 unique proteins. Of the 2,928 pQTMs identified, 2,892 were previously unreported. Of 81 comparable associations reported by Zaghlool et al. (2020) \(177\), 26 replicated at our significance threshold \((P < 4.5 \times 10^{-10})\) with the same direction of effect; a further 16 replicated at the epigenome-wide significance threshold \((P < 3.6 \times 10^{-6})\) and a further 39 replicated at nominal \(P < 0.05\). Taken together, the MWAS results from Chapters 6-7 suggest that some proteins will be more viable candidates for EpiScore development than others. This is something I observed in my work in Chapter 8, as only 109 of 953 tested proteins met the threshold for retention as EpiScores. Previous work has demonstrated that there is variability in the variance explained in protein levels by either genetic or epigenetic omics types. Hillary et al. (2020) showed that for the inflammatory panel of proteins I used to train EpiScores in Chapter 8, 46% of variance in the levels of CXCL10 could be explained by DNAm \(314\). The EpiScore for CXCL10 was one of the 109 EpiScores selected in my analyses (Pearson’s \(r = 0.23\) and \(P = 0.0049\) with measured CXCL10). I observe an enrichment for inflammatory-associated proteins in the set of 109 EpiScores I developed from the 953 proteins tested. The close interplay between DNAm and
proteins linked to chronic inflammation was also observed in my MWAS results in Chapter 7, as there was enrichment for inflammatory pathways in the 191 proteins that had pQTMs identified. This is also supported by the previous MWAS of protein levels performed by Zaghloul et al. (2020), which identified DNAm at the NLRC5 gene as a potential regulator of the inflammasome that I also replicated in my findings.

Protein levels used in Chapter 7 to derive pQTMs and Chapter 8 to train EpiScores were pre-adjusted for the effects of known pQTLs. This was done to ensure that the statistical signals were predominantly capturing non-genetic influences on DNAm. In Chapter 9, the protein EpiScores for GDF15 and NT-proBNP were trained without pre-adjusting for pQTLs and a polygenic risk score for each protein was assessed in incremental modelling in the test set to understand the proportion of the EpiScore performance that was potentially capturing genetic influences. Although these analyses suggested there was some overlap in the signals that indicated leakage of genetic effects in the EpiScores, the EpiScores contributed a distinct, additive performance beyond the PRS alone. If the goal of EpiScore use is to predict and stratify disease risk, it may not matter that some level of leakage from genetics is captured. However, if the goal is to examine non-genetic influences on protein levels, adjusting for pQTLs would be advisable. Even with adjustment for pQTLs it is unlikely that I was correcting for all possible genetic influences. For example, recent GWAS of protein levels have been released with ever-expanding, novel pQTLs discovered in larger sample sizes.

Finally, as with the individual protein associations I report in Chapter 10, 191 proteins associated with brain health traits that I identify in Chapter 7 should be investigated further using Mendelian randomisation and colocalisation methods. This will identify proteins that are most likely to be causally-implicated in brain morphology and cognitive ability. The full set of 405 protein-trait associations I report are informative for brain health stratification, even if they do not represent causal mediators. For example, an inflammatory cluster of proteins emerged whose elevated levels were associated with poorer brain health. This suggests that there is a detectable signature of neuroinflammation that associates with (and may drive) brain ageing, which is supported by literature in the wider field. Of the 191 proteins that were implicated in associations with brain health traits in Chapter 7, 17 had pQTMs in the
MWAS studies of protein levels within the sample (n=774). One of the most interesting signatures emerged around DNA methylation (DNAm) at the CpG site cg06690548 in the \textit{SLC7A11} gene locus, which was associated with RBP5, ACY1 and SCUBE1 levels. Differential DNA methylation at this CpG site has been identified as a causal candidate for Parkinson’s disease (N > 900 cases and N > 900 controls) and was implicated in the largest MWAS of ALS to date (6763 cases, 2943 controls). Given that ACY1, SCUBE1 and RBP5 were protein markers that I linked to either lower processing speed or higher relative brain age, the CpGs that I have identified in this study (such as cg06690548) could represent plasma markers for mediation of environmental risk on brain health.

11.4 Limitations

I have discussed the limitations relevant to each study in Chapters 6-10. In the following section, I provide some general limitations that apply to the cohorts, methodologies and assays used in this thesis.

11.4.1 Cohorts

In this section, I will provide a brief outline of the strengths and weaknesses relevant to the cohort studies I have used in my empirical chapters. The work performed in this thesis is based on five cohorts: GS (Chapters 7, 8 and 9), LBC1921 (Chapter 8), KORA (Chapter 8), LBC1936 (Chapters 6, 8 and 9) and the UKB (Chapter 10) (see Figure 5-1).

All of the described cohort populations are phenotyped extensively, with a range of lifestyle, health and molecular measures from blood samples taken at baseline clinic visits. One strength of the GS and UKB studies is their extensive electronic health linkage to primary and secondary care records, cancer and death registries and prescribing records. GS is the largest published single cohort with DNA methylation data. The LBC1936 study is well-positioned to model change in cognitive ability and brain morphology measures over time, given its rarely-available successive waves of follow-up in these measures. The UKB is the largest published cohort with Olink proteomics data. Unlike the other cohorts, KORA is not based within the UK, strengthening its use as a diverse training set.
The majority of cohorts used in my empirical work only included participants of white European ancestry. The UKB-PPP sample included a small proportion (7%) of individuals from non-European ancestries. The lack of diversity in cohort populations is a major issue in epidemiological and genetic studies as results may not necessarily translate to non-European groups. In a recent study comparing pQTL signatures across European and African ancestries\textsuperscript{383}, 10\% of the sentinel cis-pQTLs identified in the European population were either unidentifiable or rare in the African sample. When training PRS for protein levels and testing them across ancestries, the authors found that scores trained in the European population had worse performance in the African population than the converse, despite the African population being much smaller in sample size\textsuperscript{383}. Whether protein EpiScores translate across ancestries should be evaluated when sufficient data are available. Certain CpG sites that are typically highly correlated with their cis-SNPs have been identified as ancestry-dependent DNAm sites\textsuperscript{384}. There is also a need for multi-ancestry MWAS studies and meta-analyses on diverse groups\textsuperscript{385}. For example, a study published this year showed replication of MWAS findings for five CpG sites associated with insulin resistance in individuals of European and South Asian ancestries\textsuperscript{386}. Additionally, whether the poor translation of PRS that is observed across ancestries\textsuperscript{192} is also reflected in EpiScore analyses should be explored further. A new EpiScore that we have developed for the assessment of biological age acceleration called bAge (based on the protein EpiScore measures generated in Chapter 8) showed significant associations with time-to-mortality across six test sets including individuals from hispanic, white and black ancestry groups\textsuperscript{42}. This suggests that EpiScores may translate across ancestries. The translation is likely to be dependent on whether the CpG sites included in scores are conserved across ancestry groups.

The cohorts included in this thesis are also likely to suffer from selection bias, as studies such as these typically include individuals with better overall health profiles and socioeconomic rankings than the general population. Individuals in Scotland have a well-documented, longstanding high prevalence of unhealthy lifestyle behaviours (such as smoking and alcohol consumption) and obesity as compared to the UK and Europe\textsuperscript{387,388}. Cohort profiling of the Generation Scotland and Lothian Birth Cohorts suggests that these populations tend to be healthier and of higher socioeconomic status than the wider Scottish population\textsuperscript{247,262}. In the Lothian Birth Cohorts,
longitudinal attrition has also introduced sampling bias, as those who participate in
cognitive tests and brain scans are those that are physically and mentally able to do
so\textsuperscript{262}. Furthermore, the data collected in these cohorts for health variables is largely
reliant on the self-reported information from interviews, in addition to questionnaires
or surveys completed at study baseline. This can introduce recall bias\textsuperscript{389}. For
example, if cases for a given disease are more likely to report or remember certain
risk factors than controls, this could skew comparisons between these groups. In
disease contexts, modelling the agreement between self-report data and linkage to
healthcare records is often recommended to improve reliability of disease mapping\textsuperscript{390}.
Even when taking this approach, the rate of correct identification of cases has been
questioned. Recent work has mapped diseases in the UK Biobank through a
combination of self-report and linkage data and compared the classification of cases
to clinical expert adjudication of the full medical records of individuals. The proportion
of cases identified that were true positives was 71.4\% for dementia\textsuperscript{312} and 83\% for
ischaemic stroke\textsuperscript{391}. These studies suggest that electronic health data linkage is
inherently limited as a means to identify cases in large population studies. It is worth
noting that electronic health linkage data was designed to record diagnoses data for
clinical records primarily, rather than specifically for statistical research studies such
as those performed in this thesis. In the UK, primary care physicians may document
certain codes more than others as part of incentives to document diagnoses in
treatment areas\textsuperscript{220,221,392}. The extent to which these biases may over-inflate, or under-
inflate the true rate of certain diseases should be documented further in future.

11.4.2 Assays

Proteomic assays

Multiplexed assays (such as Olink and SomaScan) allow for multiple protein levels to
be simultaneously measured from a blood sample and therefore open more
possibilities for biomarker identification. The Olink assay allows for multiplexed
immunoassays to be performed, while avoiding the cross-reactivity that would typically
arise in ELISA approaches at scale\textsuperscript{393}. SOMAmers have deep coverage across the
circulating proteome, but suffer from non-specific aptamer binding estimated to occur
in roughly 7\% of SOMAmers\textsuperscript{145,392}. Understanding what these protein assays are truly
measuring remains challenging. Olink and SomaScan do not always correlate, as
studies are emerging that compare the same target protein measurements across the two platforms. In a recent analyses comparing 1,097 proteins targeted by both the Olink Explore and SomaScan platforms in 1,000 Icelandic individuals, a median Spearman correlation of 0.46 was observed (with two bimodal distributions at 0 and 0.7) \(^{394}\). In 417 comparisons between SomaScan v4 and Olink Proseek v5003 panels in the Athlerosclerosis Risk in Communities (ARIC) study \((n=427)\), a mean correlation of 0.46 was identified (ranging from -0.21 to 0.97) \(^{395}\). That study also compared 18 of the SomaScan proteins against immunoassays in 110 participants, with results suggesting that six of the assays were highly correlated between technologies \((r > 0.8)\), two had modest correlations \((r = 0.5 \text{ to } 0.8)\) and 10 were poorly correlated \((r < 0.5)\). GDF15 was one of the proteins in the highly correlated group. Of 35 Olink protein measures, eight had correlations above 0.5 in 173 individuals from a German cohort population when compared to mass spectroscopy measures \(^{396}\). In a study that compared 591 common proteins measured on the Olink explore and SomaScan 1.3k assays in 568 individuals from the Jackson Heart Study, \(cis\)-pQTLs were identified for 25\% (368 of 1472) Olink proteins and 16\% (206 of 1301) SomaScan 1.3k proteins \(^{397}\). The higher rate of \(cis\) detection for Olink points towards greater confidence in the specificity of the assay for these protein measures. Variance in traits that could be explained by predictive scores generated using either platform were also examined by the same study. Despite SomaScan having deeper proteomic coverage, predictive performance was similar (e.g. eGFR \(R^2\) Olink = 0.655 and SD 0.058, eGFR \(R^2\) SomaScan = 0.643 and SD 0.053; BMI \(R^2\) Olink = 0.758 and SD 0.026, BMI \(R^2\) SomaScan = 0.718 SD 0.037). Recent work from Pietzner et al. (2020) \(^{398}\) compared the pQTLs identified from GWAS of protein levels measured from both Olink and SomaScan platforms to understand whether differences in colocalisation were observed with disease outcomes. GDF15 emerged as an example of divergence, with lead variants identified on each platform showing opposing effects when tested for colocalisation with BMI \(^{398}\). The authors hypothesised that the two measurement platforms may be capturing different dimer confirmations of the GDF15 protein, which have alternate functional roles. Therefore, in cases whereby protein measures differ across platforms, they may be indicating a true biological difference rather than a technical inaccuracy introduced by cross-reactivity.
In the field of pQTL mapping, there have been multiple attempts to quantify consistency and differences in genetic architectures of proteins across panels. Sun et al. (2018) assessed the replication of pQTLs (n=4,998) and found a high correlation in effect sizes between pQTLs identified using SOMAscan and Olink measures (r = 0.83). A total of 65% pQTLs replicated (81% cis, 52% trans). Similarly, Pietzner et al. (2021) identified 64% of pQTLs that replicated across SomaLogic and Olink measures in 10,708 individuals. However, the authors observed correlations in effect sizes that were lower than those reported by Sun et al. (cis r = 0.41 and trans r = 0.34). To my knowledge, no studies have sought to perform MWAS comparisons between Olink and SomaScan platforms, which should be explored in future work. In Chapter 7, I use 4,235 SomaScan protein measures, however future studies should corroborate signals in populations that have equivalent Olink measures available. The protein EpiScores I produce in Chapter 8 represent the first DNAm scoring of multiplexed protein measures. Interestingly, there were five EpiScores for proteins common to both Olink and SomaScan panels that had variable correlation strength (GZMA r = 0.71, MMP.1 r = 0.46, CXCL10 r = 0.35, NTRK3 r = 0.26, and CXCL11 r = 0.09). This suggests that protein EpiScores may reflect underlying variability in the protein platforms. Protein assay measures are also quantified on a relative scale, which means that absolute values are not available and limits interpretation of protein biomarker effects at the level of individuals.

**DNA methylation assays**

The cohort populations that had DNA methylation measures used in this thesis generated measurements through Illumina microarrays. These measure either ~450,000 or ~850,000 CpG sites in a genome-wide capacity, but cover only around 3-5% of the total number of sites measureable across the human genome. New technologies (such as Twist with ~3 million CpGs, or Nanopore sequencing technologies) that have deeper coverage of genome-wide DNA methylation have been developed, but are currently limited to smaller samples. These could identify new sites for traits. They also may capture distal regulatory elements that may inform on functional relationships between the epigenome and the proteome.
11.4.3 Methodologies

Statistical models

The elastic net penalised regression method used to generate protein EpiScores in Chapters 8-9 and ProteinScores in Chapter 10 is well-suited to reducing the total number of possible features to a manageable subset of features. While the Beta weighting coefficient selected in elastic net provides some indication of relative importance for selected CpG sites, interpretability is still limited. One avenue to assess the importance of features may be to iterate the elastic net model (i.e. 1000 times) and assess the most frequently selected features. Conditional variable importance can also be extracted for random forest approaches. However, random forests are computationally more expensive than elastic net penalised regression owing to the construction of tree-based layers mapping relationships between input features (number of trees and node sizes). In a random forest approach, the model averages outputs from independently trained decision trees, with each tree built using different subsets of variables to prevent overfitting. Therefore, typically when dealing directly with DNAm measures (~500,000 CpG sites), a subset must be selected as part of feature reduction and pre-selection steps prior to running predictive scoring models. In Cheng et al. (2023), only the CpG features from a linear Cox LASSO model were taken forward for random forest modelling, as modelling all CpG sites available would be computationally impractical. The authors note that this approach is likely to limit the potential value that may be added from the non-linear model fitting steps implemented as part of tree-based methods (as elastic net, LASSO and ridge regression are weighted-linear methods). For these reasons, methods such as random forests may be more applicable in modelling proteomic data, as the number of protein features considered in Chapter 10 of this thesis was 1,468. This should be explored in future to compare random forest with elastic net methods for ProteinScore development and extract feature importance rankings.

In Chapters 8-10, the overarching aim was to develop protein EpiScore and ProteinScore measures and evaluate these as early stratifiers of disease across cohort populations. The methods I have used should be extended in future work to model potential thresholds for scores, which may inform clinical translation. Polygenic
risk scores have been found to consistently highlight high risk individuals in the upper and lower percentiles when assessed in a distribution across populations \(404\). However, thresholding based on specific, absolute levels of a score is required to set guidance for clinical decision-making of risk status at the level of an individual patient.

There are several metrics that can be used to breakdown score performance beyond AUC alone, which describes the overall performance metric across all potential classification thresholds. Sensitivity provides the true positive rate, which represents the proportion of individuals with the disease that are identified \(405\). Specificity provides a true negative rate, which indicates the proportion of controls that are correctly identified. While these metrics are useful, it is also important to consider the number of false positives in light of true positives and false negatives in light of true negatives \(405\). Positive predictive values (PPVs) represent the number of individuals with the disease (true positives) that are correctly classified, as a proportion of the true positives and falsely-labelled controls (false positives). Negative predictive values (NPVs) indicate the number of controls (true negatives) that are correctly identified, as a proportion of the true negatives and falsely-labelled controls (false negatives). By simulating all four discussed metrics across a range of potential clinical score thresholds, the number of individuals that are falsely labelled as cases or controls can be identified at each threshold. Understanding how PPV and NPV change across various thresholds is important in clinical contexts, as a decision must be made in terms of how many false positives or false negatives are tolerated \(186\).

Cheng et al, (2023) showed that an EpiScore for 10-year onset of type 2 diabetes increased the true positive and true negative identification as compared to typical risk factors across a range of probability thresholds \(347\). Recent work in the UK Biobank modelled ProteinScores in the context of a population-level clinical screening paradigm, whereby the specificity (also termed detection rate) of scores was calculated across multiple false detection rate acceptability thresholds (ranging from 5% to 40%) \(406\). The false detection rate was calculated as the proportion of false positives over the number of false positives and true negatives. The authors show that there is a tradeoff between maximising detection rate of incident disease cases, versus increasing the false detection of cases that remained controls over the censor period. The decision of what an acceptable rate of false positives is must be made on a
disease-by-disease basis at the point of translation of scores to the clinic. Further refinement will also be required to ensure that scores are well-calibrated, with minimised differences between the estimated true risk of disease \(^{407,408}\). This should be tested both at the level of population cohorts, but also to evaluate how well a score is capturing risk at the level of diverse individuals.

Further to the methods for score thresholding discussed above, there are several factors that currently limit the translation of omics scores to the level of an individual patient in clinic. The EpiScores I have created in Chapters 8-9 can be projected into DNAm samples as part of a wide range of EpiScores for lifestyle, disease and protein biomarkers. This would provide a general picture of the health state of an individual. However, the protein and DNAm measures used to generate the protein EpiScores are normalised to a relative scale. As such, it would be difficult to detangle technical and biological variation in scores when projected into DNAm at the levels of an individual. Recent work suggests that the impacts of technical variability in EpiScores across populations can be minimised by choice of normalisation methods \(^{409}\). Due to relative scale of measurement, EpiScores must still be contextualised within a reference population (i.e. the test set) \(^{210}\). An example of this in practice is the MethylDetectR Shiny app \(^{410}\), which houses the 109 protein EpiScores I have created alongside other EpiScores generated by our research group. This allows the user to view the quantile an individual sits within for a given protein EpiScore marker as compared to the wider population. Users can upload their own DNAm data and generate both the protein EpiScores and the visualisations for their own populations. Future work is required to detangle the reliability of scores projected at the level of a single individual, to ensure biological effects are separated for technical variability. This is limiting factor that could preclude the translation of EpiScores to individualised, personal care applications. ProteinScores may also be impacted by the challenge of separating technical and biological variability, however the Olink data I make use of in Chapter 10 has been previously shown to have very few batch effects \(^{144}\). A more important factor that must be considered with ProteinScores developed using proteins from Olink and SomaScan platforms is that the proteins are measured on relative scales, reducing interpretability in practice. As mass spectrometry measures absolute abundances of proteins, it could represent a more viable avenue to generate clinically-meaningful thresholds based on scores. The Cox PH models employed in Chapters
test for relationships between one marker (proteins or protein EpiScores) and one time-to-event outcome. However, the Cox PH approach is not well-suited to modelling complex proteomic patterns associated with multimorbidity. The overlapping morbidities individuals have, in addition to the years lived with certain morbidities and the order of diagnoses across the life course are important considerations that I do not address in this thesis. One approach to rectify this would be to adjust for the baseline prevalence of certain diseases, observing whether associations are attenuated. However, this does not account for the temporal ordering and length of exposure to certain diseases across the life course. Approaches such as growth curve modelling and latent class analyses can help to cluster individuals longitudinally into multimorbid subgroups. Recent work in the UK Biobank suggests that five main multimorbid endotypes exist, with the principal dimensions that dictated clustering shown to be dependent on paralysis, stroke and dementia within individuals. Despite these limitations, I do highlight a set of proteins (including GDF15) in the UK Biobank that were indicators of multimorbidity in individual associations and in associations with a multimorbidity binary variable.

In Chapter 6, I chose to use the Wald ratio test in MR but this is inherently unreliable given that a single SNP instrument is modelled for the exposure variable. The Wald ratio represents the ratio of the SNP-outcome effect over the SNP-exposure effect. If horizontal pleiotropy exists (the SNP affects the outcome via a pathway other than the exposure), then an association may be detected if separate causal variants for the exposure and outcome are in LD with each other. In my empirical work, no association was identified and colocalisation provided additional confirmation for this.

While the pQTMs identified in Chapter 7 identify potential biological signatures that aid in our understanding of protein markers and brain health, they cannot be used to infer causality. Similarly, the linear mixed models used in Chapter 7 identify individual protein markers of brain health traits, but do not infer causality. Recent work I have been involved with has taken forward protein markers of cognitive ability for causal testing in MR, leading to the identification of putatively causal relationships between higher CA14 and CDCP1 and stroke risk. CA14 was also implicated as a potential causal factor associated with hippocampal volume.
11.5 Recommendations

The work undertaken in Chapters 8-9 of this thesis provides a set of protein EpiScores that can be used as tools for biomarker discovery and disease risk stratification. The protein EpiScores I have created can be used as features in score generation, in approaches analogous to the original generation of the GrimAge score. This has already begun, with an updated biological age acceleration score (bAge) using protein EpiScores I created in Chapter 8 as input features in addition to the EpiScore features from the original GrimAge score \(^\text{212}\). The score was trained using a Cox PH survival elastic net on 18,365 individuals in Generation Scotland (including 1,214 that had died), with 35 features selected that included 28 of the 109 protein EpiScores I had developed in Chapter 8. bAge slightly outperformed GrimAge in the strength of its association to survival (GrimAge HR = 1.47, 95% CI = [1.40, 1.54], P = 1.08 \times 10^{-52}, versus bAge HR = 1.52, 95% CI = [1.44, 1.59], P = 2.20 \times 10^{-60}) \(^\text{212}\). The set of protein EpiScores I developed have also been utilised in two studies assessing incident cardiovascular and type 2 diabetes risk \(^\text{289,347}\). In Cheng et al. (2023), the best-performing approach for 10-year onset of type 2 diabetes was the joint inclusion of both a ‘direct’ score trained on DNAm features, with a ‘composite’ score created using the set of 109 protein EpiScores from Chapter 8 as input features. The full model (risk factors + protein EpiScore composite + direct score) had an AUC of 0.872 and PRAUC of 0.302. Interestingly, tree-ensemble methods did not improve beyond this approach, with AUCs of 0.866 and 0.864 for random survival forest (RFS) \(^\text{414}\) and survival Bayesian additive regression tree (sBART) \(^\text{415}\) approaches, respectively. When modelled singularly, the composite protein EpiScore predictor had an AUC of 0.864 and the direct DNAm predictor had an AUC of 0.870. Given the number of features considered by these models (109 protein EpiScores, versus 200,000 CpG sites chosen as they had the highest variance), it is clear that the protein EpiScores offer a means to achieve similar risk stratification but with much-reduced computational burden. Feature selection and dimensionality reduction is a well-known issue in DNAm-based score generation \(^\text{187}\). The studies I have outlined suggest that the 109 protein EpiScores capture a disease-relevant signal and require relatively few CpG measurements. There may also be additional predictive information conferred through the inclusion of protein data when training the protein EpiScores. Continued
generation and testing of protein EpiScores as both individual biomarkers and features to augment direct DNAm scores and reduce computational burdens is therefore recommended.

**Chapter 10** of this thesis mapped individual, early protein markers of incident diseases and generated a set of ProteinScores for these outcomes. Taken together, the EpiScores and ProteinScores I have created in **Chapters 8-10** may have clinical utility in the future. For example, the ProteinScore work in **Chapter 10** of this thesis suggests that the type 2 diabetes ProteinScore vastly outperformed the PRS for type 2 diabetes alone. The ProteinScore marginally improved beyond HbA1c – the gold-standard clinical marker for type 2 diabetes. An important future consideration when assessing clinical utility of predictive scores will be sensitivity and thresholding to categorise high-risk individuals. Sensitivity of a score refers to the proportion of individuals that are categorised as high risk based on a score that actually develop a disease. PRS typically have limited sensitivities of 10-15%, but have been shown to be more accurate for individuals lying at the extremes of the PRS distribution. Studies comparing whether ProteinScore and EpiScore measures can improve sensitivity beyond PRS should be conducted. The scores I have created further improve performance beyond PRS, which is likely to translate into improved sensitivity for clinical decision-making at the point of care. Given that they capture an environmental component of disease risk that is likely to reflect modifiable lifestyle influences on DNAm and protein levels, the scores I have created are likely to improve identification of high-risk individuals. However, health economic quantification of the likely benefits to health services and patients is required to understand whether the improvements offered by scores should be rolled out in practice. Ethical considerations around choosing the ‘cut-off’ point to delineate who should be categorised as high-risk and whether this information should be communicated to individuals should also be explored, given that being high-risk does not guarantee diagnosis of a disease.

It is important that future work examines how well the performance I report in **Chapters 8-10** translates when scores are applied across diverse populations. This should involve assessment of translation across varied age groups and sexes, given the differences in DNAm profiles and disease risk that can exist across these groups. For example, a study in 1,049 individuals from Avon Longitudinal Study of...
Parents and Children (ALSPAC) cohort found that an alcohol EpiScore comprising 144 CpGs trained in an adult population explained considerably less variance in adolescents (0.8%) versus individuals in midlife (7.6%)\textsuperscript{153}. In contrast, a study of 1,183 children and adolescents from the Texas-Twin project found that the cognitive ability score trained by McCartney \textit{et al.} using DNAm from adults in GS\textsuperscript{175} explained a large proportion of variance (11.1%) in performance on a mathematics test\textsuperscript{418}. Translation should be investigated in future work in studies such as these and is likely to be variable by EpiScore.

The present thesis did not include direct involvement of members of the public or individuals with lived experience of diseases to shape the research. Patient and public involvement (PPI) in research has risen across the past decade, with legislation introduced in the UK enabling diverse perspectives to shape research\textsuperscript{419}. PPI groups are also becoming more frequent as co-collaborators in research, which is important to ensure that ethical and patient-centred concerns are raised regarding therapies or interventions\textsuperscript{420}. I would recommend that researchers open up further avenues for dialogue with these groups to shape research moving forwards. This is pertinent in the field of disease risk scoring, as there are a range of ethical concerns regarding implementation of scores at a clinical level. For example, the field of PRS, there are concerns around data storage and access, thresholding of scores and communication of risk\textsuperscript{186,188,421}. Taking a personalised view of preventative healthcare, I would advise that each individual should decide the extent of their participation in screening and omics risk score testing, in addition to the level of information they would like to receive regarding their risk level. In this thesis, omics data were typically taken from a single baseline blood sample (GS and the UKB are examples of this). In the statistical approaches, I analyse associations and risk profiles of individuals using population-level data. Recent work has focused on deep profiling by building ‘personal omics profiles’ for individuals that include multiple omics measures taken at multiple time points\textsuperscript{422}. This stream of research has highlighted that an abnormal change in a marker that should be a warning signal for an individual may fall within the ‘general healthy range’ for that marker in the wider population\textsuperscript{423}. Individualised profiling is arguably of huge value for monitoring health in a highly personalised way. The nature of the data available also meant that it was not possible to compare omics signatures at multiple time points. If the data were to exist in the future to facilitate both protein
and EpiScore comparisons at multiple time points, I recommend this. For example, modelling both individual CpG sites and the projection of EpiScores at regular time points may help to identify proteins for which EpiScore equivalents may be most useful as stable measures of the body’s health state. Recent work combining the deep molecular profiling with wearable sensor information has shown that an individual’s response to dietary intervention and the omics signatures that associate with their natural daily patterns in measures such as heart rate\textsuperscript{424}. Looking at the change in omics measures around interventions and everyday biological fluctuations may help to understand the role DNAm has in pathways to disease. It may also help to develop omics-based endpoints reflecting health states. Similarly, deep longitudinal profiling of 150 individuals over 4 years revealed that certain seasonal patterns in blood omics measures exist\textsuperscript{425}. In that study, proteins that exhibited seasonal patterns predominantly those with functions in inflammation, immunity and allergies (IL5, SIGLEC15, ILR1AP, C2 and C9)\textsuperscript{425}. DNAm patterns in children have been found to be sensitive to season of birth and latitude (indicating temporal and spatial influences) in an analysis that spanned 21 cohorts and 9,358 individuals\textsuperscript{426}. Understanding the interplay between temporal and geographical factors and omics signatures may help to better-adjust for these variables in analyses, while also potentially informing timings around when therapeutic interventions may be most effective or safe for individuals.

Subtyping individuals to understand heterogeneity within diagnoses should also be explored. For example, within individuals with Alzheimer’s dementia diagnoses codes recorded through data linkage, there may be a subset of individuals with early onset and a greater influence of genetic factors in the disease pathogenesis\textsuperscript{427}. Similarly, there may be clusters that emerge with specific omics profiles that indicate they may respond to specific therapies\textsuperscript{428}. Another class of individuals may emerge that have a strong influence of modifiable, unhealthy lifestyle behaviours and the presence of multimorbidity or polypharmacy. In this thesis, the cohort studies that I worked with had very few prevalent cases and limited incident cases of Alzheimer’s dementia and typically even fewer cases of other neurological diseases. With expanding linkage, further clustering of individuals to identify heterogeneity in underlying pathogenesis of diseases will become more viable as cases grow.
In Chapters 6 and 7, I profiled the epi/genetic architectures of proteins associated with brain health. Future analyses should identify the pQTM findings that differ or replicate across protein panels. Similarly, it is also important to understand the extent that both pQTM associations translate from blood to cerebrospinal fluid and brain. Fewer than 10% of probes have been found to significantly correlate between blood and brain DNAm samples. Due to the high variability and poor correlation in DNAm between blood and brain samples, it is likely that some of the blood-specific DNAm findings I identify here will not translate to brain DNAm. Future MWAS on brain tissues are required to resolve DNAm patterns specific to brain. I previously published findings suggesting that variability in DNAm also exists across brain regions, as EpiScores for lifestyle traits had highly variable performance depending on the brain region DNAm was sourced from. Finally, further development of methods that can model interrelatedness between multiple omics features may help to resolve complex causal pathways between genetic, epigenetic, proteomic and brain health.

11.6 Final summary

Advances in large-scale epigenetic and proteomic datasets, combined with innovations in statistical modelling techniques have facilitated unprecedented access to the molecular signatures of brain health and disease. In this thesis, an integrative, multi-omics approach was employed to map molecular signatures of brain health and stratify neurological disease risk. Incident diseases affecting the brain directly, but also the wider body (which impact the brain) were examined. Protein EpiScores and ProteinScores were shown to be useful tools for biomarker discovery and disease risk stratification. These can be used by a range of researchers for diverse applications to many different health states in future. Proteomic assessments of leading incident diseases and brain health traits were also conducted, which highlighted potential proteomic candidates and pathways that can be explored further as preventative targets for intervention.

The body of work in this thesis provides evidence that a relatively small subset of blood omics measures can be used to track high-risk individuals, often a decade prior to disease onset. ProteinScores were shown to be of value for 10-year onset classification of Alzheimer’s dementia, Parkinson’s disease and ischaemic stroke.
Identification of high-risk individuals is critical for early intervention and prevention strategies. Translation of findings from cohort studies to individual-level data is likely to be one of the key steps required to realise the potential of omics sampling at the point of care. Scores represent modifiable, trackable endpoints for measuring processes such as inflammation in the body and can be used as endpoints to quantify the time-sensitive effects of interventions in the future.
12 References


138. Houseman, E. A. et al. DNA methylation arrays as surrogate measures of cell

139. Nazarenko, T. et al. Technical and biological sources of unreliability of
Infinium type II probes of the Illumina MethylationEPIC BeadChip microarray.

methylation at ANK1, BIN1, RHBDF2 and other loci. *Nat. Neurosci.* **17**, 1156–
1163 (2014).

141. Lambert, E. et al. The Alzheimer susceptibility gene BIN1 induces isoform-
dependent neurotoxicity through early endosome defects. *Acta Neuropathol.

142. Nabais, M. F. et al. An overview of DNA methylation-derived trait score

143. Ferkingstad, E. et al. Large-scale integration of the plasma proteome with

144. Sun, B. B. et al. Plasma proteomic associations with genetics and health in


146. Hillary, R. F. et al. Genome and epigenome wide studies of neurological
(2019).

147. Hillary, R. F. et al. Multi-method genome- And epigenome-wide studies of


Appendix – Publications

First author publications

Published


Submitted


Middle author publications

Published


Submitted


*, indicates joint first authorship