André Filipe Afonso de Sousa Fonseca

From data to discovery: Unraveling new pipelines for analyzing high-throughput antibody data in Malaria and Chronic Fatigue Syndrome

2024

André Filipe Afonso de Sousa Fonseca

From data to discovery: Unraveling new pipelines for analyzing high-throughput antibody data in Malaria and Chronic Fatigue Syndrome.

Doutoramento em Ciências Biotecnológicas
(Especialidade em Biotecnologia Molecular)

Trabalho efetuado sob a orientação de:
Professor Doutor Nuno Sepúlveda
Professora Doutora Clara Cordeiro

2024

From data to discovery: Unraveling new pipelines for analyzing high-throughput antibody data in Malaria and Chronic Fatigue Syndrome

Authorship declaration of the work

I declare to be the author of this work, which is original and unpublished. Authors and works consulted are duly cited in the text and included in the reference list. This publication may not be reproduced in whole or in part, nor may it be copied in any way, including electronically, mechanically, photocopies, recorded, or digitized, without the written authorization of the author.

"Today we fight. Tomorrow we fight. The day after, we fight. And if
this disease plans on whipping us, it better bring a lunch,
'cause it's gonna have a long day doing it "

Jim Beaver

Acknowledgments

First I would like to thank FCT - Fundação para a Ciência e Tecnologia which made this work possible with the financial support provided through a PhD fellowship grant (grant ref. SFRH/BD/147629/2019) and to the Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa (CEAUL) for co-funding the publication of my articles and help with other expenses.

Then I would like to thank the University of Algarve for the opportunity to conduct this work and for the unceasing help of all the colaborators, specially professor Deborah Power and miss Sara Ribeiro for always getting back to me as fast as soon as possible whenever I had any doubt concerning any bureaucratic aspect of my PhD.

To Professor Nuno Sepúlveda, I am truly grateful for being able to work under your supervision. Thank you for all the guidance and the companionship, patience, transparency, understanding, and honesty. Thank you for both our scientific and non-scientific conversations. I am very grateful for the all those times when you shared your vast scientific knowledge with me which have allowed me to learn the art of science "up close". Doing science research is an art, and I am amazed to see your mastery of it. I wish you the greatest success. I will never forget all that we went through and I have the upmost respect for you.

To Professor Clara Cordeiro, I thank you for all the support and guidance throughout my doctoral studies, that like a caring mother was always making sure I didn't forget anything. Thank you for the scientific comments and suggestions that greatly enriched the articles and this manuscript. Thank you for always being up to date with the bureaucratic part that assisted and guided me over these four years. Even though you had so much to do, you always found a time to help me when I needed, so you have my upmost gratitude.

To my dear friend João Malato, I am grateful to have shared this journey with you. Thanks fo always helping me specially when I needed to clarify any doubt regarding R and Latex. Your programming skills and knowledge have made me a more efficient researcher. I am very thankful for your constant concern for me and my family, I am very grateful for our friendship and for you to be a part of my academic journey. I wish you all the success in the world with the recommendation to stop being so self-critical and to start trusting more in the tremendous potential you have.

To the entire ImmunoStats group, for the interesting scientific meetings we had, which provided me with a lot of knowledge and laughter.

Finally, to Professor Przemyslaw Biecek, I thank you for receiving me in your laboratory at the Faculty of Mathematics & Information Science, Warsaw University of Technology, Warsaw, Poland, and to all members of the MI^2 group for the fascinating seminars on various topics surrounding the machine learning theme.

Agora em português, para aqueles que me são mais próximos consigam claramente entender a gratidão que tenho por eles:

Aos meus pais por sempre me apoiarem no meu percurso académico. Obrigado por utilizarem os vossos recursos para me darem uma boa formação académica e tudo o que necessitei ao longo de toda a minha vida. Obrigado por acreditarem em mim e me motivarem sempre a terminar o que iniciei, amo-vos.

À minha esposa, a mulher da minha vida, que cuidou da nossa maior benção e fez dela uma menina linda, super educada e inteligente. Peço perdão pelas horas a mais em frente ao computador e por não conseguir estar presente tanto quanto gostarias. Sou eternamente grato por tudo o que tens feito pela nossa família e por toda a dedicação para comigo e com a nossa filhinha, tenho muito orgulho de ti, amo-te.

À minha linda filha a quem devo todo o tempo que não brinquei com ela, todos os momentos em que não pude acompanhar o seu desenvolvimento e todos os momentos que deixei de viver com ela. És o maior presente da minha vida, obrigado por existires e por deixares a minha vida e a vida da mamã mais colorida. És a luz dos nosso olhos, amamos-te.

A Jesus Cristo meu Senhor e salvador que tem-me guiado e protegido ao longo de toda a minha vida. Devo a ele tudo o que tenho e tudo o que sou. Sou extremamente grato pela oportunidade que tive de desencvolver este doutoramento e por todo o trajeto percorrido ao longo destes quatro anos. Que tudo o que consegui alcançar seja para a sua glória, louvado seja.

Contents

Acknowledgments vii

Abstract xvii

Resumo xviii

List of publications included in the thesis xix

List of works not included in the thesis xx

List of abbreviations xxi

Chapter 1 - General introduction 2

1 General Introduction 2

1.1 Introduction to the Immune System 4

1.1.1 Innate immunity 4

1.1.2 Adaptive immunity 4

1.1.2.1 Antibodies 6

1.1.2.2 Antibody structure 6

1.1.2.3 Antibody Diversity 6

1.1.2.4 Antibody Isotypes 9

1.1.2.5 Antibody mediated responses 11

1.2 Immunoassays 12

1.2.1 Enzyme-Linked Immunosorbent Assay (ELISA) 14

1.2.2 Antibody microarrays 16

1.3 Machine Learning analysis on high-throughput antibody data 20

1.4 Objectives and outline of the thesis 22

Chapter 2 - Background knowledge on Machine Learning 24

2.1 Feature selection strategies 24

2.1.1 Filters 24

Finite mixture models 25

2.1.2 Wrappers 27

2.1.2.1 Forward selection 27

2.1.2.2 Backward selection 28

2.1.2.3 Hybrid approaches 29

2.1.3 Embedded 29

2.2 Predictive analysis 30

2.2.1 Logistic Regression . . . 30
2.2.1.1 Maximum likelihood estimation (MLE) . . . 32
2.2.1.2 Gradient Descent . . . 34
2.2.2 Ridge Logistic Regression . . . 34
2.2.3 LASSO Logistic Regression . . . 36
2.2.4 Elastic-Net Logistic Regression . . . 37
2.2.5 Linear Discriminant analysis . . . 37
2.2.6 Quadratic Discriminant analysis . . . 39
2.2.7 Random Forest . . . 39
Gini index . . . 43
2.2.8 Extreme Gradient Boosting . . . 43
2.2.9 Super Learner . . . 45
2.3 Assessing model performance . . . 46
2.3.1 The validation set approach . . . 47
2.3.2 K-fold Cross Validation . . . 48

Chapter 3 - Background knowledge on the studied diseases 51
3.1 Malaria . . . 51
3.1.1 Clinical manifestation . . . 53
3.1.2 Acquired immunity to malaria . . . 54
3.1.3 Diagnosis . . . 55
3.1.4 Treatment and Prevention . . . 58
3.2 Myalgic encephalomyelitis/chronic fatigue syndrome . . . 60
3.2.1 Clinical Manifestation . . . 61
3.2.2 Diagnosis . . . 62
3.2.3 Treatment . . . 62
Bibliography . . . 63
Supplementary Matherials . . . 97

Chapter 4 - Development of antibody selection strategies for multi-sera data 99
4.1 Abstract . . . 99
4.2 Introduction . . . 99
4.3 Materials and Methods . . . 100
4.3.1 Data . . . 100
4.3.2 Measuring association . . . 101
4.3.3 Predictive methodologies . . . 101
4.3.3.1 Multiple logistic and probit regression . . . 101
4.3.3.2 Regularization strategies . . . 101
4.3.3.3 Random forest . . . 101

4.3.4 Predictive accuracy . . . 102
4.3.5 Pipeline . . . 102
4.4 Results . . . 102
4.5 Discussion . . . 107
4.6 Concluding remarks and future work . . . 109
Bibliography . . . 110

Chapter 5 - Antibody selection strategies and their impact in predicting clinical malaria based on multi-sera data 115
5.1 Abstract . . . 115
5.2 Introduction . . . 116
5.3 Materials and Methods . . . 117
5.3.1 Data under analysis . . . 117
5.3.2 Preliminary antibody feature selection using Random Forest . . . 118
5.3.3 Antibody selection based on non-parametric testing . . . 118
5.3.4 Antibody selection based on optimal data dichotomization . . . 118
5.3.5 Antibody selection based on hybrid parametric/non-parametric approach . . . 119
5.3.6 Correction for multiple testing . . . 121
5.3.7 Predictive Stage . . . 121
5.3.8 Statistical Software . . . 123
5.4 Results . . . 123
5.4.1 Preliminary analysis based on the Random Forest approach . . . 123
5.4.2 Analysis based on the simple antibody selection approach . . . 123
5.4.3 Analysis based on the data dichotomization approach . . . 125
5.4.4 Analysis based on the hybrid parametric/non-parametric approach . . . 127
5.5 Discussion . . . 130
5.6 Conclusion . . . 137
Bibliography . . . 138

Chapter 6 - The SARS-CoV-2 receptor ACE2 in ME/CFS: A meta-analysis of public DNA methylation and gene expression data 146
6.1 Abstract . . . 147
6.2 Introduction . . . 147
6.3 Materials and methods . . . 148
6.3.1 Eligible diagnostic criteria of ME/CFS . . . 148
6.3.2 Analysis of published DNA methylation association studies . . . 148
6.3.3 Analysis of gene expression studies . . . 152
6.3.4 Analysis of new RNA data on the *ACE/ACE2* gene expression in ME/CFS153

6.3.4.1 Study participants . . . 153
6.3.4.2 Experimental procedure for RNA isolation and expression . . 154
6.3.4.3 Statistical analysis . . . 154
6.3.4.4 Ethical approval . . . 154
6.3.5 Statistical software . . . 155
6.4 Results . . . 155
6.4.1 Meta-analysis of ACE/ACE2 DNA methylation in ME/CFS patients . . . 155
6.4.2 Meta-analysis of ACE/ACE2 gene expression in ME/CFS patients . . . 157
6.4.3 Analysis of ACE/ACE2 gene expression from a new female cohort . . . 158
6.5 Discussion . . . 158
6.6 Conclusions . . . 162
Bibliography . . . 163
Supplementary Materials . . . 174

Chapter 7 - Revisiting IgG antibody reactivity to EBV in ME/CFS and its potential application to disease diagnosis 177
7.1 Abstract . . . 177
7.2 Introduction . . . 178
7.3 Materials and methods . . . 180
7.3.1 Study participants . . . 180
7.3.2 Peptide array . . . 180
7.3.3 Statistical analysis . . . 181
7.3.4 Statistical software . . . 182
7.4 Results . . . 182
7.4.1 Principal component and linear discriminant analyses . . . 182
7.4.2 Antibody-wide association analysis . . . 184
7.4.3 Analysis of candidate antigens for classifying ME/CFS patients with infectious trigger . . . 185
7.5 Discussion . . . 188
Bibliography . . . 192
Supplementary Materials . . . 199

Chapter 8 - IgG Antibody Responses to Epstein-Barr Virus in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome: Their Effective Potential for Disease Diagnosis and Pathological Antigenic Mimicry 203
8.1 Abstract . . . 203
8.2 Introduction . . . 204
8.3 Materials and methods . . . 205
8.3.1 Study participants . . . 205

8.3.2 Basic Description of Serological Data . . . 205
8.3.3 Statistical Analysis for Predicting the Disease Status . . . 206
8.3.3.1 Dividing the Dataset into Train and Test Sets . . . 206
8.3.3.2 Ranking Antibodies by Their Importance for Predicting the Disease Status . . . 206
8.3.3.3 Individual Statistical and Machine Learning Methods for Predicting the Disease Status from the Anti-EBV Antibodies . . . 206
8.3.3.4 Construction of Final Models for Predicting the Disease Status by Assembling Predictions from Individual Models . . . 207
8.3.4 Bioinformatic Analysis to Test the Importance of Antigen Mimicry in Predicting the Disease Status . . . 207
8.3.5 Statistical Software . . . 208
8.4 Results . . . 208
8.4.1 Construction of a Predictive Model to Distinguish All ME/CFS Patients from HCs . . . 208
8.4.2 Construction of a Predictive Model to Distinguish ME/CFS Patients with Non-Infectious or Unknown Disease Triggers from HCs . . . 210
8.4.3 Construction of a Predictive Model to Distinguish ME/CFS Patients with a Putative Infectious Disease Trigger from HCs . . . 210
8.4.4 Testing the Importance of Antigen Mimicry on Disease Prediction Using a Bioinformatic Approach . . . 211
8.5 Discussion . . . 213
8.5.1 General Comments . . . 213
8.5.2 Clinical and Diagnostic Implications . . . 214
8.5.3 EBV Antigenic Mimicry and Its Putative Role in ME/CFS Pathogenesis 216
8.5.3.1 Replication of Previous Finding on EBNA6_0070 Peptide . . . 216
8.5.3.2 EBNA6_0488 Peptide and the Antigenic Mimicry with CTCF and AEBP1 . . . 217
8.5.4 Interpretation of the Findings under the Lens of the Danger Theory . . 219
8.5.5 Potential Danger Signals in ME/CFS Pathogenesis . . . 219
8.6 Conclusions . . . 220

Bibliography **221**
Supplementary Materials . . . 235

Chapter 9 - Assessing model reliability: The impact of train-test split proportions on the accuracy and reliability of biomarkers against clinical Malaria **238**
9.1 Abstract . . . 238
9.2 Introduction . . . 239

9.3 Materials and methods . . . 240
9.3.1 Data . . . 240
9.3.2 Data Partitioning / Split Ratio: . . . 241
9.3.3 Predictive analysis . . . 241
9.3.4 Power Analysis . . . 241
9.3.4.1 Statistical Software . . . 242
9.4 Results . . . 243
9.4.1 Data partition's effect on variable selection . . . 243
9.4.2 Data splitting's effect on predictive accuracy . . . 244
9.4.3 Data partition's effect on power . . . 245
9.5 Discussion . . . 248
9.6 Conclusion . . . 254
Bibliography . . . 255

Chapter 10 - General Discussion **264**
10.1 Data dichotomizarion: issues and concerns . . . 264
10.2 Predictive analysis: the logic behind using multiple algorithms . . . 267
10.3 Hybrid approach : a comprehensive analysis of the data . . . 270
10.4 Random Forest: a sturdy algortihm . . . 272
10.5 Machine Leaning: powerful tools on large datasets . . . 273
10.6 Antibodies: Unraveling potential new biomarkers against disease . . . 274
10.7 Concluding remarks . . . 275

Abstract

Current serological studies, where thousands of antibodies can now be simultaneously screened, has allowed to enhance our understanding of the immune responses to various pathogens and to support the development of better diagnostic tools and treatment strategies. Nonetheless, the complexity of such data has brought new hurdles regarding the capability of traditional statistical methods to cope with such data. Although Machine Learning (ML) techniques have offered enhanced capabilities to unravel antibody biomarkers, the exact identity of antibody biomarkers against certain diseases remains a struggle. This challenge underscores the pressing need for innovative methodologies to enhance the accuracy in biomarker identification, facilitating more effective diagnostics and targeted therapeutics.

In this thesis I developed analytical pipelines for the analysis of high-throughput antibody data. To illustrate the potential of these pipelines, I focused on antibody data on Malaria and Chronic Fatigue Syndrome/Myalgic Encephalomyelitis. In general, these pipelines were based on an initial variable selection step to identify the most relevant and informative variables, followed by a predictive step where distinct classifiers would be constructed using ML-based approaches. At first, distinct approaches for the analysis of a relative low number of antibodies under analysis to test the suitability on the analysis of such data. We then proceeded to analyze data containing thousands of antibodies. This more-challenging situation motivated me to fine-tune the initial pipelines to better cope with the high dimensionality of the data. Each pipeline leveraged different statistical assumptions and yielded benefits and drawbacks, providing predictive accuracies that ranged from close to 72% up to 90% when implemented on different datasets, surpassing previous published analyzes on he same data.

In conclusion, these new pipelines generated a good predictive performance in the case studies evaluated. Given that they are based on general principles of data analysis, they have the potential to increase the robustness and reproducibility of the analysis of high-dimensional antibody data.

Keywords: Antibody; biomarkers; pipelines; Machine Learning; classifiers

Resumo

Actualmente, estudos serológicos permitem que milhares de anticorpos sejam rastreados simultaneamente, o que tem contribuído significativamente para o aprimoramento de nossa compreensão da resposta imunitária a diversos agentes patogénicos, bem como para o avanço no desenvolvimento de ferramentas diagnósticas e estratgias terapêuticas mais eficazes. Contudo, a complexidade desses dados tem imposto novos desafios no tocante à habilidade de analisar tais dados recorrendo a métodos estatísticos tradicionais. Embora as técnicas de aprendizagem de máquina (ML) tenham melhorado a nossa capacidade de identificar biomarcadores immunes, a precisão na identificação de biomarcadores específicos para certas doenças ainda é um obstáculo a ser superado. Isso ressalta a necessidade urgente do desenvolvimento de novas metodologias que possam melhorar o rigor na identificação de biomarcadores, possibilitando diagnósticos mais efetivos e terapias mais direcionadas.

Neste trabalho, desenvolvi pipelines analíticos para a análise de dados de anticorpos em larga escala. Para demonstrar o potencial desses pipelines, concentrei-me nos dados relacionados a malária e à síndrome de fadiga crônica/encefalomielite miálgica. Esses pipelines basearam-se em uma etapa inicial de seleção de variáveis para identificar aquelas mais relevantes e informativas, seguida por uma etapa preditiva em que diferentes classificadores foram construídos utilizando abordagens baseadas em ML. Inicialmente, explorei diferentes abordagens para analisar um número relativamente pequeno de anticorpos, a fim de avaliar sua adequação na análise desses dados. Posteriormente, expandi a análise para incluir dados contendo milhares de anticorpos. Esta situação mais desafiadora motivou-me a ajustar os pipelines iniciais ou a lidar melhor com a alta dimensionalidade dos dados. Cada pipeline aproveitou diferentes suposições estatísticas e apresentou vantagens e desvantagens, resultando em precisões preditivas variando de aproximadamente 72% a 90% quando aplicadas a conjuntos de dados diferentes, superando análises anteriores publicadas referentes aos mesmos dados

Em conclusão, esses novos pipelines demonstraram um bom desempenho preditivo na avaliação dos estudos de caso. Dado que são fundamentados em princípios gerais de análise de dados, têm o potencial de aumentar a robustez e a reprodutibilidade da análise de dados de anticorpos em larga escala.

Palavras chave: Anticorpos: biomarcadores; pipelines; Aprendizagem de Máquina; classificadores

List of publications included in the thesis

This thesis is based on the following papers:

Chapter 4
André Fonseca, Clara Cordeiro, and Nuno Sepúlveda. Identification of antibody responses predictive of protection against clinical malaria. In Regina Bispo, Lígia Henriques-Rodrigues, Russell Alpizar-Jara, and Miguel de Carvalho, editors, Recent Developments in Statistics and Data Science, pages 227–239, Cham, 2022. Springer International Publishing

Chapter 5
André Fonseca, Mikolaj Spytek, Przemyslaw Biecek, Clara Cordeiro, and Nuno Sepúlveda. Antibody selection strategies and their impact in predicting clinical malaria based on multi-sera data. BioData Mining, 17, 1 2024

Chapter 6
João Malato, Franziska Sotzny, Sandra Bauer, Helma Freitag, André Fonseca, Anna D Grabo wska, Luís Graça, Clara Cordeiro, Luís Nacul, Eliana M Lacerda, et al. The sars-cov-2 receptor angiotensin-converting enzyme 2 (ace2) in myalgic encephalomyelitis/chronic fatigue syndrome: A meta analysis of public dna methylation and gene expression data. Heliyon, 7(8),2021

Chapter 7
Nuno Sepúlveda, João Malato, Franziska Sotzny, Anna D Grabowska, André Fonseca, Clara Cordeiro, Luís Graça, Przemyslaw Biecek, Uta Behrends, Josef Mautner, et al. Revisiting igg antibody reactivity to epstein-barr virus in myalgic encephalomyelitis/chronic fatigue syndrome and its potential application to disease diagnosis. Frontiers in Medicine, 9:921101, 2022

Chapter 8
André Fonseca, Mateusz Szysz, Hoang Thien Ly, Clara Cordeiro, and Nuno Sepúlveda. Igg antibody responses to epstein-barr virus in myalgic encephalomyelitis/chronic fatigue syndrome: Their effective potential for disease diagnosis and pathological antigenic mimicry. Medicina, 60(1):161, 2024.

List of works not included in the thesis

Translation of the book "*Przemysław Biecek, Anna Kozak, Aleksander Zawada. BetaBit - The Hitchhiker's Guide to Responsible Machine Learning. Fundacja Naukowa SmarterPoland.pl. 2022*" to portuguese.

List of abbreviations

A - Adenin
ACE - Angeotensin-converting enzymme
ADAM17 - A disintegrin and metallopeptidase domain 17
ADCC - Antibody-dependent cellular cytotoxicity
ADRA1B - Adrenoceptor Alpha 1B
AEBP - Adipocyte enhancer binding protein
AIC - Akaike's Information Criterion
ALS - Amyotrophic lateral sclerosis
AMA - Apical membrane antigen 1
Anti–SSA/Ro - Anti–Sjögren's-syndrome-related antigen A autoantibodies
AP - Apurinic/apyrimidinic
ASP - Aspergillus
AUC - Area Under de Curve
BALF - Bronchoalveolar avage fuild
BCG - Bacillus Calmette-Guérin
BCR - B cell receptor
BLLF - Ba lotus leaf
BSA - Bovine Serum Albumin
EBA - Erythrocyte-binding antigen
EBNA - Epstein-Barr virus nuclear antigen
CA19-9 - Carbohydrate antigen 19-9
CCC - Canadian Consensus Criteria
CD - Cluster of differentiation
CDC - Center for Disease Control and
cDNA - Complementar deoxyribonucleic acid Prevention
CI - Confidence interval
CMV - Cytomegaloviru
CNDP1 - Carnosine dipeptidase 1
COVID-19 - Corona virus disease 19
CpG - Cytosine and guanine base sequence
CSF - Cerebrospinal fluid
CSP - Circumsporozoite protein
CSR - Class swithcing recombination
CT - Cycle threshold
CTCF - CCCTC-binding factor
D - Diversity
DNA - Deoxyribonucleic acid

DPP-4 - Dipeptidil peptidase-4
DSB - Double-strand breaks
dUTPase - deoxyuridine triphosphatase enzime
EBNA - Epstein Barr Virus Nuclear Antigen
EBV - Epstein-Barr virus
EDS - Ehlers-Danlos syndrome
ELISA - Enzyme-linked immunosorbent assay
ERBB2 - V-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2
ESR1 - Estrogen receptor alpha
FDR - False discovery rate
FOXP3 - Forkhead box P3
G - Guanin
GAMA - Glycosylphosphatidylinositol-anchored micronemal antigen
GAP43 - Growth-associated protein 43
GC - Guanine-cytosine
GEO - Gene Expression Omnibus
GLMNet - Generalized Linear Models Elastic-net regression
GOF - Goodness of fit
HBsAg - Hepatitis B surface antigen
HC - Healthy Control
hCG - Human chorionic gonadotropin
HL - Hosmer-Lemeshow
HLA-DR15 - Humna leukocyte antigen DR15
HMG1A - High mobility group AT-hook 1
HOXA - Homeobox A cluster
HPRT1 - Hypoxanthine phosphoribosyltransferase 1
HRP-II - Histidine-rich protein II
HSP - Heat shock protein
HHV - Human herpesvirus
IFA - Immunofluorescence antibody Ig - Immunoglobulin
IL - Interleukin
KEN - Kenya
Kg/m^2 - Kilogram per square meter
LASSO - Least Absolute Shrinkage and Selection Operator
LDA - Linear discriminant analysis
LDH - Lactate dehydrogenase
LMP1 - Latent membrane protein 1
LRM - Logistic regression model
ME/CFS - Myalgic encephalomyelitis/chronic fatigue syndrome

MERS CoV - Middle East respiratory syndrome coronavirus
MIAME - Minimum Information about a Microarray Experiment
ML - Machine Learning
mRNA - Messanger Ribonucleic acid
MS - Multiple sclerosis
MSP - Merozoite Surface Protein
MSRP- MSP7-related proteins
MTRAP - Merozoite ThromMRospondin-related adhesive protein
N - Total number of samples
NCBI - National Center for Biotechnology Information
NEFM - Neurofilament medium polypeptide
n_{prt} - Number of protected individuals
nr - Non-redundant
NRGN - Neurogranin
n_{sus} - Number of susceptible individuals
NHEJ - Nonhomologous end joining
P - Plasmodium
PAMPs - Pathogen-associated molecular patterns
PBF - Phosphate-buffered saline
PBMC - Peripheral Blood Mononuclear Cell
PCA - Principal component analysis PCR - Polymerase chain reaction
PDA - Pancreatic ductal adenocarcinoma
PDF - Probability density function
Pf - Plasmodium *falciparum*
PRR - Pattern recognition receptor
Prt - Protected
PSA - Prostate-specific antigen
QDA - Quadratic discriminant analysis
R_s - Spearman correlation coefficient
R21 - R21/Matrix-M
RAAS - Renin-angiotensin-aldosterone system
RAG1 - Recombination activating gene 1
RBC - Red blood cells
RDTs - Rapid diagnostic tests
RefSeq - Reference sequence
RF - Random forest
RGS18 - regulator of G-protein signalling 18
Rh - Rhesus
RNAseq - Ribonucleic acid sequencing

ROC - Receiver Operating Characteristic
Ron - Recepteur d' origine nantais
ROS - Reactive oxygen species
RSS - Recombinant signal sequence
SARS-CoV2 - Severe acute respiratory syndrome Corona virus 2
scFv - Single-chain fragment variables
Se - Sensitivity
SeroTAT - Serological testing and treatment
SL - Super learner
SLE - Systemic lupus erythromatosus
SLC25A20 - Solute carrier family 25
SNP - Single nucleotide polimorphism
Sp - Specificity
sPLS-DA - Sparse partial least squares discriminant analysis
SSc - Systemic sclerosis
Sus - Susceptible
SVM - Support vector machine
SW - Shapiro-Wilk
TdT - Desoxinucleotidil-transferase terminal
Th1 - Type 1 helper
TMPRSS2 - Transmembrane serine protease 2
TNC - Tenascin C
TNFα - Tumour necrosis factor alpha
TSS - Transcription start site
VIP - Vasoactive intestinal polupeptide
VPAC2 - Vasoactive intestinal peptide receptor type 2
XGB - Extreme gradient boosting
WHO - World Health Organization
μl - microlitter
χ^2 - Chi-square

List of Figures

1 Two-dimensional model of an IgG molecule . . . 7
2 VDJ recombination in the immune system . . . 8
3 Immunoglobulin's class switching recombination . . . 12
4 Antibody mediated responses . . . 13
5 ELISA types . . . 16
6 Antibody microarrays . . . 17
7 Sigmoid curve . . . 32
8 Gradient descent . . . 34
9 Schematic representation of the Random Forest algorithm . . . 42
10 Flow Diagram for Super Learner . . . 47
11 Malaria parasite life cycle . . . 53
12 Natural acquired immunity to the malaria parasite . . . 56
13 Pipeline . . . 103
14 Predictive performance of the best antibody signature . . . 106
15 Ridge Regression regularization strategy results . . . 107
16 Random forest results . . . 107
17 Optimal data dichotomization for antibody selection . . . 120
18 Parametric antibody selection . . . 122
19 Analysis of an RF using all the 36 antibodies as features . . . 124
20 Simple antibody selection results . . . 126
21 Optimal data dichotomizations antibody selection results . . . 128
22 Hybrid antibody selection results . . . 131
23 DNA methylation analysis of 19 and 8 CpG probes located in the ACE and ACE2 genes, respectively . . . 150
24 Boxplots per study, group and gender of the M-values referring to probes identified in Figure 23C and Figure 23D . . . 156
25 Analysis of *ACE*/*ACE2*-related data from eligible microarray-based gene expression studies . . . 157
26 Analysis of *ACE* and *ACE2* expression levels from the German study. Analysis of *ACE* and *ACE2* expression levels from the German study . . . 160
27 Preliminary multivariate analysis of the data . . . 183
28 Antibody-wide association analyses . . . 186
29 Statistical analysis of the antibody levels related to EBNA4_0529, EBNA6_0066, and EBNA6_0070 . . . 187
30 Analysis of the final classification model for predicting ME/CFS patients with an infectious trigger when compared to healthy controls . . . 189
31 Analysis of all ME/CFS patients versus HCs . . . 209

32 Analysis of ME/CFS patients with non-infectious or unknown disease triggers against HCs 211
33 Analysis of ME/CFS patients with an infectious disease trigger against HCs . . 212
34 Bioinformatic analysis of the EBV peptides associated with the 26 antibodies for predicting ME/CFS patients with an infectious disease trigger 215
35 Split ratio impact on feature selection 245
36 Split ratio impact on feature selection 246
37 Power Analysis 249

List of Supplementary Figures

1 Spearman's correlation coefficient 199
2 Pipeline for data analysis where different steps are shown in the flowchart using distinct coloured shapes 235
3 Data concerning to the IgG antibody against EBNA1_0430 236
4 Bioinformatic analysis of the EBV peptides associated with the 26 antibodies for predicting ME/CFS patients with an infectious disease trigger 237

List of Tables

1 Patient Seroprevalence . . . 105
2 Results from the 28 antibodies deemed significant by the data dichotomization approach . . . 129
3 Mixture Model results . . . 132
4 DNA methylation studies . . . 149
5 Microarray-based gene expression . . . 152
6 Summary statistics for the gene expression of *ACE* and *ACE2* from the German female study participants . . . 158
7 Analysis of the linear regression models for the Box-Cox-transformed *ACE* and *ACE2* mRNA levels . . . 159
8 Basic characteristics of ME/CFS patients and healthy controls . . . 180
9 Complmentary log-log estimates . . . 188

List of Supplementary Tables

1 *ACE* and *ACE2* CpG probes 174
2 CpG probes inclduign SNP or coincided with a polymorphic SNP 175
3 Linear models estimates 176
4 EBV peptides 200
5 Null models' results 201
6 Most significant antibodies 202

Part I
General Introduction

Chapter 1

1 General Introduction

The earliest reference to antibodies can be traced back to Emil Von Behring and Kitasato Shibasaburo, who in 1890, published a landmark finding decreeing that transferring serum from animals immunized against diphtheria was able to cure animals that were ill with the disease [1, 2]. In the following year the reference to "Antikörper" or antibody, was first made to describe the neutralizing agent in the blood that could neutralize the diphtheria toxin [3]. This discovery laid the groundwork for understanding the immune system's ability to neutralize harmful agents. As the 20^{th} century unfolded, the significance of antibodies in vaccine development became apparent [4]. In 1920, Albert Calmette and Camille Guérin's pioneering efforts led to the creation of the Bacillus Calmette-Guérin (BCG) vaccine for tuberculosis, showcasing the potential of antibodies in preventing infectious diseases [5]. Moving forward to 1975, Köhler and Milstein published a landmark paper that described the fusion between an antibody-producing plasma cell and a myeloma cell, the latter, which, due to its transformed nature, could propagate indefinitely in culture. This technique enabled the production of unlimited amounts of antibodies in *vitro* and thus, the "hybridoma" was born with its promise of producing unlimited quantities of monospecific antibodies, an innovation that changed the field of immunology forever [6]. Fast forward to today, antibodies are essential for the development of diagnostic and treatment strategies agonist a wide range of diseases [7, 8, 9, 10, 11, 12] and were indispensable in combating the global COVID-19 pandemic. In fact monoclonal antibodies, engineered with high precision and specificity played a crucial role in therapeutic and prophylactic interventions against COVID-19, providing passive immunity and preventing infection in certain high-risk populations [13, 14, 15, 16, 17]. This contemporary application underscores the versatility of antibodies, showcasing their ongoing relevance in addressing evolving health challenges. As I navigate the complexities of infectious diseases in the 21^{th} century, the story of antibodies continues to unfold, from novel discoveries to innovative applications in diagnostics, therapeutics, and disease prevention through vaccination, antibodies have become a cornerstone of modern medicine.

The instrumental role of antibodies however, has been only possible to unveil due to breakthroughs in immunoassays. Until recently, the enzyme-linked immunoassay (ELISA) and other related tests were at the core of the research made in antibody data [18, 19, 20, 21, 22, 23]. These tests, however, were only able to detect or quantify antibodies against a single antigen, lacking the ability to provide a comprehensive overview of the immune systems response to infections. With the recent development of high-throughput technologies, antibody quantification is currently shifting to more advance assays, such as mi-

croarray [24, 25], luminex [26, 27], or cytometry bead assays [28], where a large number of different antibodies can be simultaneously screen in the same biological sample [29, 30, 31]. These techniques have transformed our understanding of the immune system by providing a more holistic view on the intricacies of the immune system's responses [32]. Nonetheless, the large quantity of data generated by these tools has created a demand for more advanced statistical approaches to effectively analyze and extract meaningful insights from the data [33, 34, 35, 36]. Machine Learning (ML) approaches, specifically tailored to handle high-dimensional data, have therefore stablihsed a foothold in the serologic field [37, 38, 39, 40, 41, 42, 43]. However, the implementation of these techniques is still in infancy and innovative strategies are critical for the identification of antibody biomarkers to fight infectious diseases.

Therefore this thesis aims to introduce novel pipelines for the analysis of high-throughput IgG antibody data with the intent of identifying diagnostic biomarkers against disease. Given that the vast majority of features (variables) measured by high-throughput methodologies hold no association with the outcome of interest and thus are redundant or irrelevant for the analysis [44, 45], removal of such features brings several advantages, among which increased accuracy and interpretability of the final models [46]. For this reason, the analysis of this type of data is often divided into a two-step approach. First, a feature selection or variable reduction step which removes redundant antibodies is implemented, followed by a predictive analysis where classifiers are built upon the reminiscent features and their performances assessed [47]. Here several feature selection strategies are proposed which led to the construct of different classifiers with diagnostic potential for the diseases under study. Although initially developed to handle malaria (antibody) data, these strategies were then extended to deal with Myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) data as well. This disease, which was being researched by our group, displayed symptoms remarkably similar to those of patients with long COVID, a disease caused by the SARS-CoV-2 outbreak that struck the wold in 2020 [48]. In this sense, I decided to extend the research effort to the analysis of antibody data in ME/CFS patients in order to accelerate our knowledge of this disease.

Before entering in the depths of this thesis, however, it is necessary to introduce some basic background knowledge on the immune system with a special focus on antibodies. An introduction to instruments currently used to gather antibody data information will follow. I will broadly review the current analysis targeting the identification of antibodies. Finally, I will provide a general overview of the diverse algorithms currently employed for such analysis and end up with a brief outline of the diseases here addressed.

1.1 Introduction to the Immune System

The immune system is a sophisticated network of biological components that comprises various organs, cells, and proteins [49]. This intricate system serves as a defense mechanism against foreign pathogens such as bacteria, viruses, parasites, fungi and other microbes that enter the body, as well as changes in the body's cells, such as cancer cells [49, 50]. The immune system's primary role is to recognize, neutralize and remove harmful substances from the body, thus protecting against illness-inducing agents [49, 51]. Broadly speaking, the immune system is devided into two main branches, the innate immune system and the adaptive immune system which communicate with each other via different molecules and cellular components [49, 51].

1.1.1 Innate immunity

The innate immune system represents the body's first line of defense against pathogens [49]. The innate immune system provides a general defense response, responding in the same way to all foreign substances. For this reason, the innate immune response is often referred to as the non-specific immune system [49]. Upon infection, the innate immune system provides an immediate, quick-acting response that clears out or contains the invading microbes within the first few hours or days after infection, before the adaptive immune responses have developed [49, 51]. The innate immune system is comprised of various cell types, such as macrophages, neutrophils, eosinophiles, basophils, dendritic cells, mast cells and natural killer cells [49, 51, 52]. Innate immune cells express germ-line encoded pattern recognition receptors (PRRs) that recognize and respond to conserved microbial structures, known as pathogen associated molecular patterns called pathogen-associated molecular patterns (PAMPs), like lipopolysaccharides and peptideoglycans commonly expressed in some type of pathogens (i.e, bacteria) [52]. Furthermore, the innate immune system is also composed by the complement, a system of plasma proteins that upon activation leads to a massive amplification of a series of proteins able to clear pathogens via opsonization and cell lysis [52].

1.1.2 Adaptive immunity

Despite being highly efficient in controlling the initial phases of infections, the cells of the innate immune system aren't always capable of eliminating the infectious organisms [49]. Not only that, but the innate immune system is not sufficient to face all different kinds of infections that might occur throughout lifetime of an individual, specially, when we take in consideration the fast evolving parasite world. In such cases, the adaptive immune system takes over [49, 52]. The adaptive immune response is mediated by the B and T lymphocyte cells and their products [49, 51, 50]. Lymphocytes are white blood cells that express

highly diverse receptors capable of recognizing and reacting to a large number of molecular substances (i, peptides derived from foreign proteins) on the surface of pathogens called antigens [49, 51]. Each naive[1] lymphocyte entering the bloodstream, however, bears antigen receptors of a single specificity [51]. The specificity of these receptors is determined by a unique genetic mechanism that operates during lymphocyte development in the bone marrow and thymus (for T lymphocytes) to generate millions of different variants of the genes encoding the receptor molecules [49, 51, 3]. This ensures that distinct lymphocytes in the body collectively carry different antigen receptors with different specificities. Lymphocytes, then undergo a process of clonal selection to select only lymphocytes that carry receptors of a single specificity [49, 51]. This antigen's specificity allows to generate responses that are tailored to specific pathocytes that encounter an antigen to which their receptor binds. This will lead phagocytes to become activate, proliferating and differentiating into either effector or memory cells [53]. Memory cells of the adaptive immunity create an immunological memory after a pathogen has been eliminated, which allows it to provide a faster and more effective response to subsequent reinfections with the same pathogen [49]. This process of acquired immunity is the basis of vaccination and also explains why some illnesses arise only once in a lifetime. Ultimately, depending on the population of lymphocytes used to mediate the immune response, adaptive immunity can be divided into cellular or humoral immunity [49]. The cellular immunity is controlled by T-cells which besides being able to induce the death of infected host cells or cells also play an important role in the activation of B cells of the adaptive humoral immune response [49, 52].

The humoral immunity is mediated by molecules called immunoglobulin (Ig), also known as antibodies, which are produced by B lymphocytes upon antigen recognition [49] . These cells are produced in the bone marrow where they mature to become specialized immune system cells which later travel to secondary lymphoid organs [49]. Each B lymphocyte expresses cell surface antigen receptors, called B-cell receptors (BCRs), which are membrane-bound immunoglobulins with a unique antigen specificity [54, 53]. Once a BCR binds to an antigen, it can become active. Nonetheless, as a general rule naive antigen-specific lymphocytes are difficult to activate by an antigen alone and often require accessory signals that can come from an armed helper T cell [51]. In this case, upon binding to a pathogen, the BCR-antigen complex is taken up by the B cell and processed by proteolysis into peptides. The B cell then displays these antigenic peptides on its surface through specific molecules that attract a matching helper T cell (that recognizes the same antigen as the B-cell), which releases lymphokines and activates the B cell [51]. Besides helping in the activation of B cells, these helper T cells also control isotype switching and have a role in

[1] Naive refers to immune cells that have not encountered an antigen and thus have not been activated. Naive immune cells are cells in a resting state that have not undergone differentiation into effector or memory cells.

initiating somatic hypermutation and affinity maturation [51]. The activated B cell then begins to proliferate and differentiate into progenitor antibody-secreting plasma cells which release a large number of antibody copies of the immunoglobulin that first recognized the antigen [49]. B cells produce different classes of such antibodies, which are highly specific for the immunogen that stimulated the B cell. These different antibody classes serve distinct functions in the adaptive immune system which will be addressed in the next chapter. Some of the activated B cells also transform into memory cells, which are differentiated B cells that can quickly mature to plasma cells, becoming part of the memory of the adaptive immune system [53].

1.1.2.1 Antibodies

Earlier I introduced immunoglobulins, which serve as either cell-surface receptors for antigens allowing for cell signaling and cell activation (BCR) or soluble effector molecules (antibodies) that can individually bind and neutralize antigens at a distance [3]. In this section, I will describe the antibodies structure, molecular genetics underlying their diversity and their functions.

1.1.2.2 Antibody structure

Antibodies are Y-shaped molecules consisting of two identical heavy (H) chains and two light (L) chains bound together in a light heavy-heavy-light arrangement connected by disulfide bounds (Figure 1) [3, 51]. There are five H chain types which define the antibodies isotypes, whereas the L chains consist of either a k or a λ chain [3, 51]. Both H and L chains contain one NH2-terminal region called variable (V) region and one or more COOH-terminal regions called constant (C) region. While only one of such regions is found in L chains (C_L), H chains contain either three or four of such regions (C_Hn). The two V domains of the H (V_H) and L chain (V_L), which are identical in any one antibody molecule, form the antigen binding site (paratope), which determines the antigen-binding specificity of the antibody [3]. The antibody-antigen interactions typically take place between the Ig's paratope, and the antigens epitope, a specific molecular structure on the surface of an antigen [3].

1.1.2.3 Antibody Diversity

As earlier mentioned, each lymphocyte carries antigen receptors of a single specificity, meaning each lymphocyte can only recognize a single antigen [49, 51]. Nevertheless, combined, all lympocytes within the human body can virtually recognize any pathogen. This antibody diversity is ensured by the variable region of the antibodies [3]. The exons that

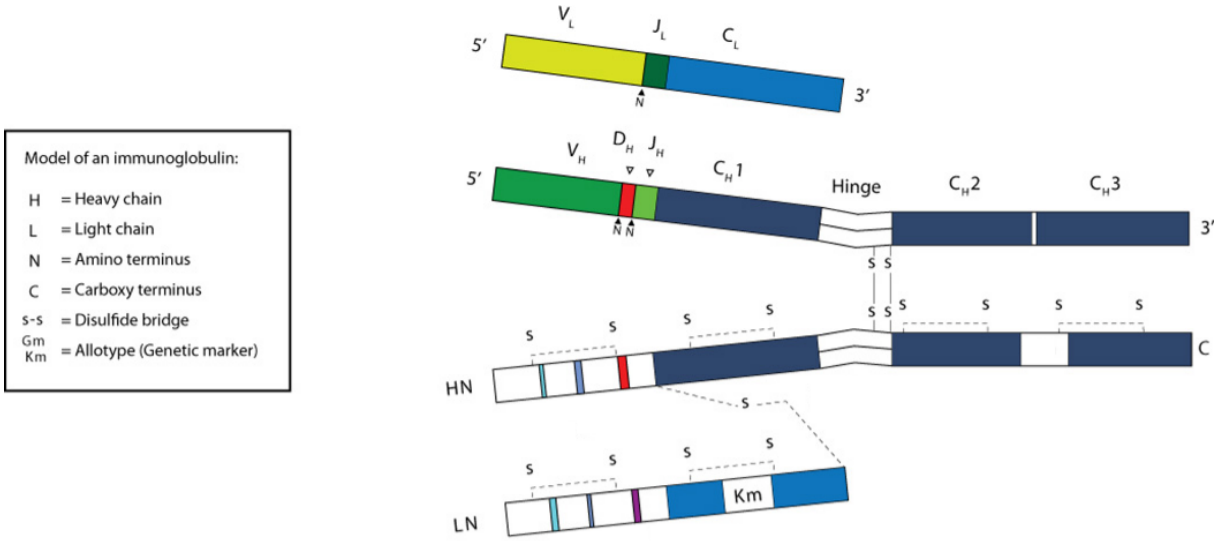


Figure 1: **Two-dimensional model of an IgG molecule.** The H and L chains at the top deconstruct the antibody at a nucleotide level. The chains at the bottom deconstruct the protein sequence. See main text for further details. Source: [3].

encode the antigen binding domains for the V region are assembled through V(D)J recombination of the V(variable), D(diversity) and J(joining) genes on the H chain and the V and J segments, in the L chain (Figure 2) [55, 56]. These genes are assembled in the developing lymphocyte to form a complete immunoglobulin. This recombination is initiated by double-strand breaks (DSB) in the DNA by the Recombination-activating genes 1 an 2 (*RAG1–RAG2*) together with non-lymphoid-specific DNA bending factors, *HMG1A* or *HMG1B* at specific recombination signal sequences (RSS) located adjacent to each V, D, and J coding segment [3, 55, 56, 57]. During this process, DNA rearrangement (inversion or deletion) occurs in a two-step ordered fashion for the H chain. A D and a J segment are initially chosen from among several possibilities and are brought together to form a D-J rearrangement [3, 55, 56]. Then a V region is selected and joined with the D-J arrangement to form a complete VDJ exon (Figure 2) [3, 55, 56]. Immunoglobulin light chain genes however are rearranged in a single step, involving V-J recombination, as D segments are absent from these loci [3]. These different segments are joined together by ubiquitously expressed components of a DNA repair process, known as nonhomologous end-joining (NHEJ) which repair DNA breaks [3, 55, 56]. The NHEJ process creates precise joins between the RSS ends, and imprecise joins of the coding ends. Terminal deoxynucleotidyl transferse (TdT), which is expressed only in lymphocytes, can variably add non-germline encoded nucleotides (N nucleotides) to the coding ends of the recombination product [3]. Overall, this process of combinatorial assembly, where a segment of each V, D and J gene is chosen from several possibilities is the fundamental mechanism driving antigen receptor diversity [3]. As an illustration, if we consider a random assortment of one of the 49 active V_H and one of 27

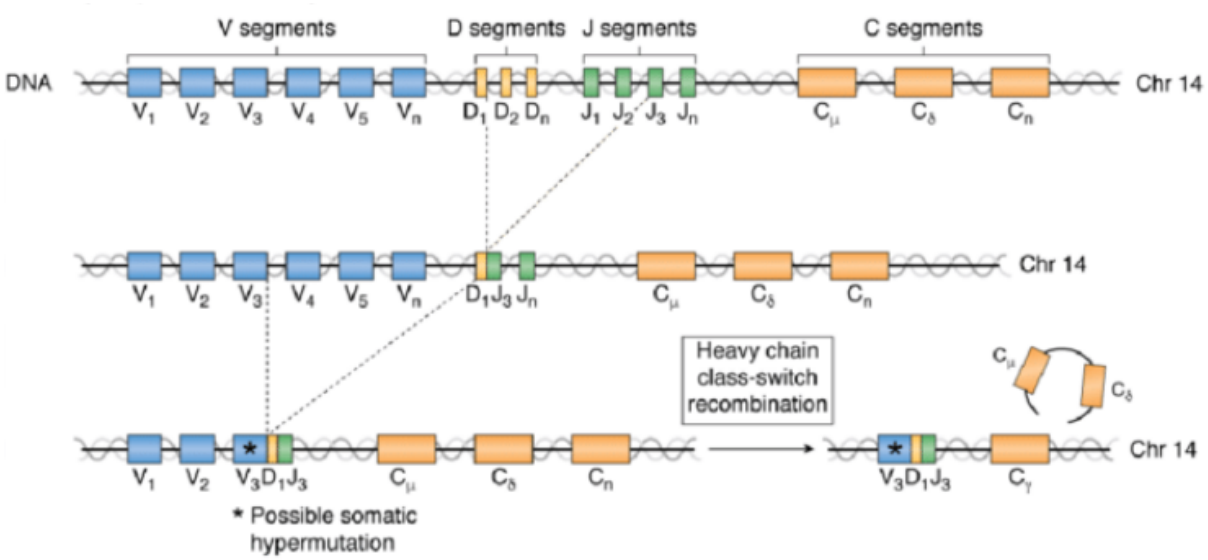


Figure 2: **VDJ recombination in the immune system.** BCR chains are generated through the recombination of a variable (V), diversity (D), and joining (J) segments present in germline loci. This process requires the RAG complex and NHEJ enzymes to break and recombine the native genomic locus in a multi-step process. Somatic hypermutation may occur during this process. Additional diversity is obtained by addition and deletion of nucleotides (N) at the junctions between recombined gene segments. The junction of the J segment to the constant (c) region is made by post-transcriptional splicing. Source: adapted from [57].

D_H with one of the 6 J_H gene segments can generate more than 10^4 different VDJ combinations [3]. While combinatorial joining of individual V, D, and J gene segments maximizes germline encoded diversity, the junctional diversity created by VDJ joining adds another layer of variation in the pre-immune repertoire [3, 55, 56]. D gene segments can rearrange by either inversion or deletion, and each D gene can be spliced and translated in each of the three potential reading frames [3]. This gives each D gene segment the potential to encode six different peptide fragments. Also, during this process, recombinational inaccuracies may arise and nucleotides can be added to the original germline further increasing the potential pre-immune antibody repertoire [3]. Adding to these, the combinatorial association between different L and the 2 possible Variable L chains k and λ consisting of an array of about 40 V and 5 J and about 30 V and 4 J genes, respectively escalates the number of different immunoglobulins to be greater than 10^{16} [3]. This variability of Ig molecules allows each antibody to bind a different specific antigen, and the total repertoire of antibodies to be large enough to ensure that virtually any structure can be recognized [3, 55, 56]. Upon exposure to an antigen, a final mechanism of immunoglobulin diversity is engaged. Aided by Helper T cells, the variable domain genes of lymphocytes undergo somatic hypermutation (SHM) in which mutations are introduced into the V region of the H and L chain, enhancing the affinity of the immunoglobulin for its antigen [3, 51].

1.1.2.4 Antibody Isotypes

While the V region confers antigenic specificity, the effector activity of the antibody is determined by the class or isotype of the heavy-chain C region, which ensures that each antibody generates an appropriate immune response for a given antigen [3]. There are five main classes of H chains denoted by Greek letters: α, σ, ϵ, γ and μ which are found in the IgM, IgG, IgA, IgD, and IgE isotypes respectively [49, 3]. IgG isotype can be split into four subclasses, IgG1, IgG2, IgG3, and IgG4, each with its biological properties and IgA can similarly be split into IgA1 and IgA2 [3, 58]. Early in B cell development, the primary immunoglobulin repertoire after the V(D)J recombination is initially composed of dominated by IgM [53]. In this sense, naive mature B cells express plasma membrane-bound IgM (Fgure 3). However upon activation by the the antigen and a cocktails of cytokine signals transmitted by helper T cells, B cells start to proliferate and undergo immunoglobulin class switching to IgA, IgE and mostly IgG antibody producing cells (Fgure 3) [53]. This process is known as isotype or class switching recombination (CSR) and allows the same VDJ heavy chain variable domain to be juxtaposed to any of the H chain classes [58]. As such, the constant region of the immunoglobulin heavy chain changes but the V domain remains the same, maintaining specificity. This allows different daughter cells from the same activated B cell to produce antibodies of different isotypes with equal specificity enabling the B cell to tailor both the receptor and the effector ends of the antibody molecule to meet a specific need.

The CSR process, like V(D)J recombination starts with double-stranded breaks generated in DNA at conserved nucleotide motifs, called switch (S) regions. S regions reside upstream from gene segments that encode the constant regions of antibody heavy chains, located adjacent to all heavy chain constant region genes with the exception of the Cδ-chain (Figure 3) [59]. DNA is nicked and broken at two selected S-regions by the activity of a series of enzymes, including activation-induced (cytidine) deaminase (AID), uracil DNA glycosylase and apurinic/apyrimidinic (AP)-endonucleases [60, 61]. AID begins the process of class switching by deaminating (removing an amino group from) cytosines within the S regions, converting the original C bases into deoxyuridine and allowing the uracil glycosylase to excise the base. This allows AP-endonucleases to cut the newly-formed abasic site, creating the initial single strand breaks that are converted to double strand breaks through mismatch repair proteins. [59] The intervening DNA between the S-regions is subsequently deleted from the chromosome, removing unwanted μ or δ heavy chain constant region exons and allowing substitution by a γ, α or ϵ constant region gene segment. The free ends of the DNA are rejoined by NHEJ to link the variable domain exon to the desired downstream constant domain exon of the antibody heavy chain [62]. In the absence of non-homologous end joining, free ends of DNA may be rejoined by an alternative pathway biased toward microhomology joins [63]. Finally, even thought each Ig isotye receptor

on the surface of B cells is constituted by a immunoglobulin (Ig) monomer (containing only one Ig unit), the secreted antibodies can be dimeric, with two Ig units as with IgA, or tetrameric with four Ig units like IgM. I will now address each Ig isotype individually:

IgM - IgM is the first immunoglobulin expressed during B cell development, with its monomeric form being expressed in Naive B cells as BCR [53, 64]. As such, IgM antibodies are generally associated with a primary immune response to a microbial infection and are frequently used to diagnose acute exposure to a pathogen [3]. Upon maturation and antigenic stimulation, multimeric, usually pentameric IgM forms are secreted [3]. These pentameric forms work by opsonizing (coating) antigens for destruction and activating the complement system. The pentameric nature of the antibody renders it to be very efficient in this process [3]. Blood circulating IgM account sfor about 10% of human immunoglobulins. IgM also serves as a secretory immunoglobulin at mucosal surfaces and is secreted into breast milk as well [58].

IgD - IgM and IgD, both membrane-bound, are transcribed concomitantly from a shared primary RNA message through differential splicing. IgD is found at very low levels in the serum, constituting less than 0.5% of human immunoglobulins in the serum [3]. While IgD's exact function remains unknown, it is regarded as a BCR involved in the induction of antibody production in B cells [58].

IgG - IgG is the predominant immunoglobulin in blood and tissue fluids, accounting for 70-75% of human immunoglobulins [65]. High affinity IgG antibodies are usually produced during the secondary immune response to the same pathogen (Figure 3). Based on structural, antigenic and functional differences in the C region of the H chain, four IgG subclasses (IgG1, IgG2, IgG3 and IgG4) were identified, which exhibit different functional activities [3]. These subclasses activate the complement cascade, an important means of clearance of opsonized pathogens, at varying degrees, with the IgG4 being the only subclass that fails to do so [3]. Additionally, much of the biological effect of IgG antibodies is exerted via Fc receptors located on surface of certain cells. Macrophages, polymorphonuclear cells and lymphocytes express Fc receptors which triggers many functional effects including phagocytosis, antibody dependent cell mediated cytotoxicity and modulation of lymphocyte function [58]. Unlike other immunoglobulins, during pregnancy, maternal IgG, can cross through the placenta, being transported from mother to baby directly. As such, babies have high levels of antibodies even at birth, with the same range of antigen specificities as their mother [65]. Breast milk or colostrum also contain IgA antibodies that are transferred to the gut of the infant and protect against infections until the newborn can synthesize its own antibodies [3]. Among the different Ig classes, the IgG is the one with the longest lifetime in the plasma. Whereas other classes of antibodies have half-lives of just a few days, most IgG antibodies have half-lives in the circulation of approximately 3 weeks

[65].

IgA - IgA is at mucosal surfaces and secretions such as nasal mucus, saliva, breast milk, and intestinal fluid [3, 64]. While generally a monomer in the serum, IgA at the mucosa, is a dimer (two IgA monomers joined together) [64]. IgA is critical at protecting mucosal surfaces from toxins, virus and bacteria by direct neutralization or by prevention of binding to the mucosal surface [3, 64]. While complement activation doesn't seem to be be a major effector mechanism at the mucosal surface, the IgA receptor is expressed on neutrophils which may be activated to mediate antibody-dependent cellular cytotoxicity (ADCC) [3]. Finally, it has been proposed that IgA may also act as a potentiator of the immune response in intestinal tissue by uptake of antigen to dendritic cells [66]. There are two subclasses of IgA, IgA1 and IgA2, whose structures differ mainly in their hinge regions [3]. IgA1 has a longer hinge region with a duplicated stretch of amino acids that is lacking in IgA2. This elongated hinge region increases the sensitivity of IgA1 to bacterial proteases in spite of partial protection by glycan [3]s. Such increased protection against protease digestion may explain why IgA2 predominates in the many mucosal secretions, such as the genital tract, whereas more than 90% of serum IgA is in the form of IgA1 [3].

IgE - The serum concentration of IgE is the lowest of all immunoglobulins [3, 58]. Nonetheless, IgE is a very potent immunoglobulin associated with hypersensitivity and allergic reactions as well as the response to parasitic worm infections [58]. IgE is primarily found in the lungs and skin [58]. IgE binds with extremely high affinity to Fc receptors expressed on mast cells, basophils, Langerhans cells and eosinophils, which in turn act as potent inducers of inflammatory reactions [3].

1.1.2.5 Antibody mediated responses

Overall, antibodies provide different ways of combating microbes. Here, I briefly describe the best known mechanisms. One way is by coating the pathogens preventing them from invading the host's cells thus inhibiting the toxic effects or infectivity [52, 49]. To invade the host's cells, pathogens may often bind to specific molecules on the target cell surface. By directly attaching to the cell surfaces of pathogens, antibodies prevent latching onto the regular cells of the body and infecting them through a process called pathogen neutralization (Figure 4 bottom left panel) [49]. Another way antibodies can protect against pathogens is by facilitating their uptake via phagocytic cells that are specialized to ingest and kill microbes [52]. This can be reached in either of two ways. In the first, antibodies bound to the pathogen's surface are recognized by phagocytic cell receptors, initiating opsonization, a process where phagocytosis is enhanced by antibodies coating the pathogens (Figure 4 bottom centre panel) [49]. Alternatively, antibodies binding to the surface of a pathogen can also activate the proteins of the complement system, which results in complement proteins being bound to the pathogen's surface [52, 58]. Complement proteins can then

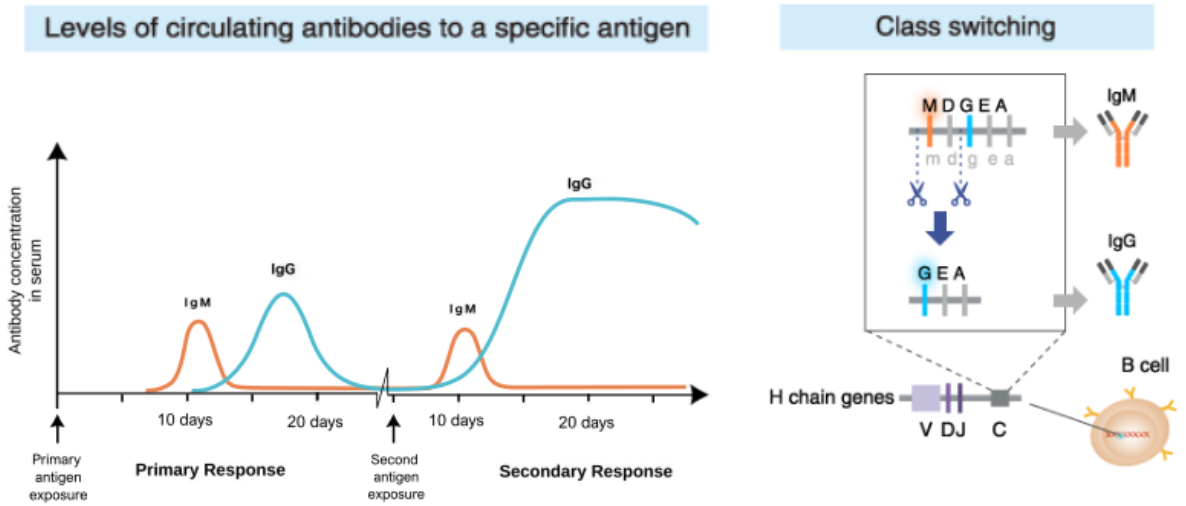


Figure 3: **Immunoglobulin's class switching recombination.** Class switching recombination (CSR) allows for antibodies to maintain their specificity to a certain pathogen while activating different effector mechanisms to combat the invading organism. Source: adapted from https://resources.mblintl.com/scientific-resources/fundamentals-for-planning-research/types-of-antibodies.

strongly enhance opsonization through the recruitment of phagocytes or can directly kill microbial cells by forming pores in their cellular membranes (Figure 4 bottom right panel) [49]. The effector mechanisms engaged in a particular response are determined by the isotype or class of the antibodies produced which as I have explored in the previous section [52].

1.2 Immunoassays

The key role of antibodies in today's medicine is indisputable, serving many purposes such as disease diagnostic markers and vaccine components. Yet, unraveling the potential role of antibodies in a specific disease crucially hinges on their identification, a process usually carried out through immunoassays in a laboratory [67, 68]. Immunoassays are techniques that rely on the specific interaction between antibodies and antigens for detection and quantification of substances such as peptides, proteins, antibodies, and hormones [67, 68]. Until recently, ELISA was the favored immunoassay for the analysis of antibody data. Nonetheless, a major limitation of ELISA was the fact that is could only analyze antibodies against a single antigen at a time [69]. Technological advances however have allowed for the development of high-throughput technologies able of measuring antibodies against thousands of antigens such as Luminex, multiplex bead assays and antibody microarrays. Among theses, antibody microarrays have gained particular prominence as most versatile approaches within multiplexed immunoassay technologies [70].

Here I will provide a brief introduction to both the ELISA and Antibody microarrays,

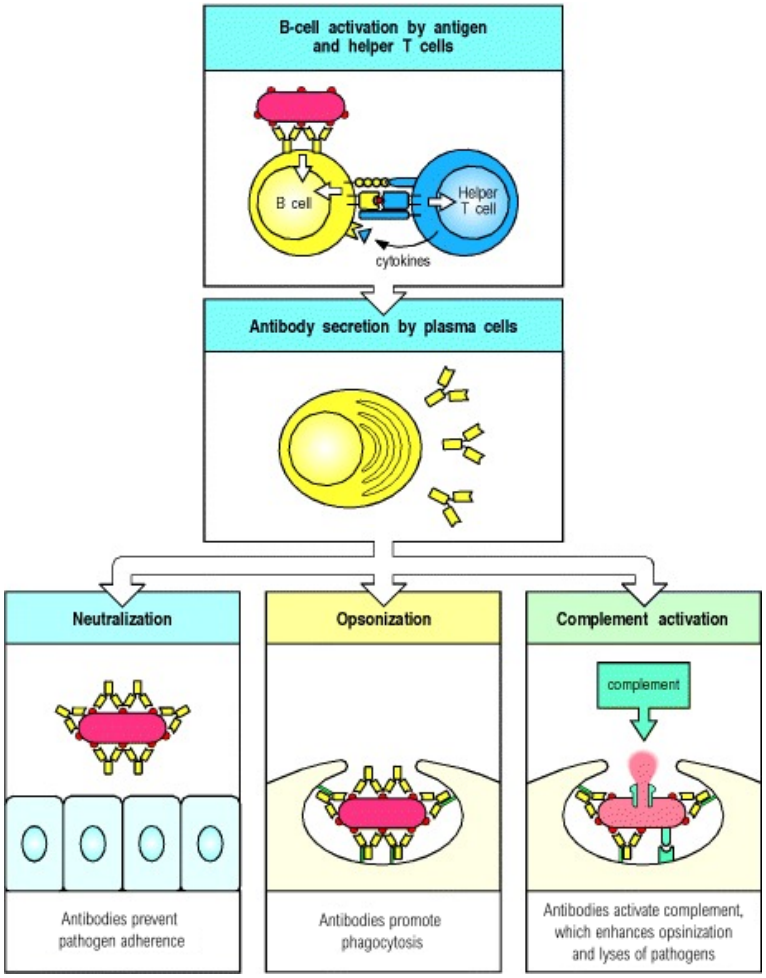


Figure 4: **Antibody mediated responses.** Antibodies protect the host from infection in different ways. They can inhibit the toxic effects or infectivity of pathogens by binding to them: this is termed neutralization (bottom left panel). By coating the pathogens, they can enable accessory cells that recognize the Fc portions of arrays of antibodies to ingest and kill the pathogen, a process called opsonization (bottom center panel). Antibodies can also trigger activation of the complement system. Complement proteins can strongly enhance opsonization, and can directly kill some bacterial cells (bottom right panel). Source: [51].

which concomitantly correspond to the techniques used to obtain the data used in our work [67].

1.2.1 Enzyme-Linked Immunosorbent Assay (ELISA)

Enzyme-Linked Immunosorbent Assay (ELISA) stands out as a cornerstone method for the detection and measurement of antibodies. Developed in 1971, the ELISA is based on the radioimmunoassay procedure originally described by Rosalyn Yalow and Solomon Berson in 1960, that made use of a radioactive isotope to label an antigen or antibody [71]. However, instead of using radioactive isotopes ELISA makes use of enzymes to tag (label) either antigens or antibodies [71]. In an ELISA examination procedure, a target molecule (either antibody or antigen) is first immobilized on a solid polystyrene surface (microplate), topically containing 96 wells [72, 73]. This process is called coating, as the surface of polystyrene microplate wells is coated by these molecules [73, 74]. Following coating, the microwell is washed with a buffer, such as phosphate-buffered saline (PBS) to remove unbound molecules. Following washing, unbound sites on the microplate are blocked with a blocking agent, such as bovine serum albumin (BSA), to prevent nonspecific binding of other molecules [73, 75]. This step, called blocking, helps reduce background noise and ensures that only specific interactions are detected in the later steps [19]. A second washing step is done to remove excess of buffer. The sample, which may contain the target antigen or antibody, is then added to the microplate wells in a process called incubation. If the target molecule is present, it binds to the immobilized molecule on the plate [67]. After incubation, the plate suffers a third wash to remove any unbound substance. These antinbody-antigen complexes are finally detected by adding a substrate (enzyme) which catalyzes a reaction that produces a measurable color change [19]. There are many substrates available for use in ELISA detection. However, the substrates most commonly used are horseradish peroxidase and alkaline phosphatase [74]. The intensity of the color is then measured in light intensity, also known as optical density, via an ELISA reader, which can be converted into a concentration or a titre using a calibration curve of known antibody concentrations [73]. In an usual ELISA protocol, a serial dilution of concentrations is placed in the wells of the plate and after the results are measured, a standard curve from the serial dilutions data is plotted with a concentration on the x-axis using a log scale and absorbance on the y-axis using a linear scale, which gives a sigmoidal curve [76]. Known concentrations give the graph's standard curve, and measurement of unknown substances can be determined when sample values are compared to the linear portion of the graphed standard curve. Overall, there are several formats used for ELISA. These fall into either direct, indirect, sandwich and competitive methods.

Direct ELISA: The direct ELISA relies on an initial coating of the microplate with antigens [77]. Then a detection antibody, which is bound to a enzyme, binds directly to the antigens

attached to the plate's wells leading to its detection (Figure 5 left) [73, 78]. Direct ELISA has the advantage of being faster than the other methods and as the elimination of cross-reaction with secondary antibodies [73]. Its disadvantages include its low sensitivity compared to the other types of ELISA and its high cost of reaction [74].

Indirect ELISA: Similarly to direct ELISA, indirect ELISA begins with the coating of antigens to the ELISA plates [79]. However, indirect ELISA, requires two antibodies, a primary detection antibody that sticks to the antigen of interest and a secondary enzyme-linked antibody complementary to the primary antibody [76]. The primary antibody is added first, followed by a wash step, and then the enzyme-conjugated secondary antibody is added (Figure 5 center left). The indirect ELISA has a higher sensitivity when compared to the direct ELISA [80]. It is also less expensive and more flexible due to the many possible primary antibodies that can be used. The only disadvantages of this type of ELISA are the risk of cross-reactivity between the secondary detection antibodies and longer time to perform than direct ELISA [74].

Sandwich ELISA: Unlike direct and indirect ELISA, the sandwich ELISA begins with a capture antibody coated onto the wells of the plate [76]. Then an antigen is added, followed by incorporation of a second enzyme-conjugated antibody (Figure 5 center right) [73]. Because the antigens are sandwiched between two layers of antibodies (capture and detection antibodies), this method is termed "sandwich" ELISA [72]. The sandwich ELISA has the highest sensitivity among all the ELISA types [81]. However, the major disadvantages of this type of ELISA are the time and expense to conduct [74].

Competitive ELISA: The competitive ELISA tests for the presence of a specific antibody in the serum against antigens bound to the plate well [18]. This type of ELISA method utilizes two specific antibodies, an enzyme-conjugated antibody and another antibody present in the serum (Figure 5 right). Both antibodies are added into the wells which will lead to a competition for binding to antigens. The presence of a color change means that the test is negative because the enzyme-conjugated antibody bound the antigens (not the antibodies of the test serum). The absence of color indicates a positive test and the presence of antibodies in the test serum [74]. The competitive ELISA has a low specificity and cannot be used in dilute samples. However, it can measure a large range of antigens in a given sample, it can be used for small antigens, and it has low variability [82].

Overall, the data obtained from ELISA can be either quantitative, qualitative, or semi-quantitative [83]. The quantitative concentration results are plotted and compared to a standard curve as previously mentioned. The qualitative results confirm or deny the presence of a particular antigen/antibody in a sample [72]. Lastly, ELISA assays are easily standardized, easy to use, very sensitive, relatively cheap and widely available. Such advantages make ELISA particularly suitable for a wide range of uses including rapid antibody

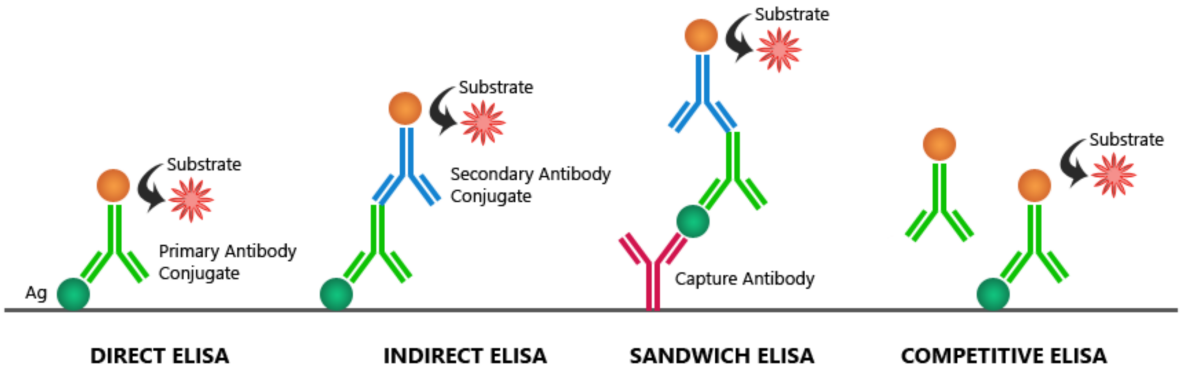


Figure 5: **ELISA types** The direct (left), indirect (center left), sandwich (center right) and competitive (right) ELISA methods are shown. Source: [73].

screening tests for detecting viruses, bacteria, fungi, autoimmune diseases, food allergens, blood typing, presence of the pregnancy hormone hCG and is also used in clinical research, forensic toxicology, and many diagnostic settings [84, 85, 86, 87, 88]. Nonetheless, one of the major drawbacks of ELISA is its inability to measure antibodies against more than a single antigen simultaneously.

1.2.2 Antibody microarrays

Recent advances in high-throughput technologies, however, have made it possible to circumvent this issue, allowing for the simultaneous measurement of antibodies against multiple antigens in the same biological sample at a reasonable cost [69, 89]. Therefore, nowadays the use of high-throughput technologies, are becoming ever more prominent. One of the most commonly used approach are antibody microarrays. Antibody microarrays are a specific form of protein microarrays, the latter referring to miniaturized and parallelized assay systems that allow for the simultaneous characterization of hundreds of thousand of proteins from small amounts of samples, within a single experiment, in a highly high-throughput manner [90, 30]. Since their introduction, these type of arrays have increasingly become an attractive tool for the exploratory detection and study of protein abundance, interaction, function, pathways, drug targets and immune responses [90]. Among the distinct protein arrays available, antibody microarrays have become one of the most powerful multiplexed[2] detection technologies [90]. In this technology, a collection of antibodies are spotted and fixed on a solid surface such as glass, plastic, membrane, or silicon chip using either a standard contact spotter [91, 92] or non contact microarrayer [93, 94]. Then, a sample containing labeled soluble proteins of interest, (i.e., cell lysate or patient body fluid sample) is incubated on the array, and the targeted proteins (antigens) from the sample bind to the immobilized antibodies [30]. The resulting binding events are

[2]Multiplexed arrays refer to technologies that enable the simultaneous detection and measurement of multiple analytes (such as proteins, nucleic acids, or other molecules) within a single sample

then measured by detecting the intensity of fluorescent, colorimetric, chemiluminescent or radioactive indicators [90, 30]. In this format, proteins are detected after antibody capture using direct protein labeling [30]. This method, however lacks specificity in protein target labeling (Figure 6 left) [30]. Alternatively, another antibody array model provides higher sensitivity using the "sandwich" assay format. This format employs two different antibodies to detect the targeted protein [95, 96] (Figure 6 right). One antibody, called the capture antibody, immobilizes the targeted protein on the solid phase, while the other antibody, called the reporter or detection antibody that recognizes a different part of the antigen, generates a signal for the detection system. Using two antibodies significantly increases the specificity and sensitivity of the antibody microarray [90].

Overall, the antibody microarrays have demonstrated a number of advantages compared to traditional, single analyte methods of protein analysis, such as, enzyme-linked immunosorbent assays (ELISA) [90]. Their high throughput and sensitivity, small sample volumes required and ease of standardization have rendered this to be widely adopted [90].

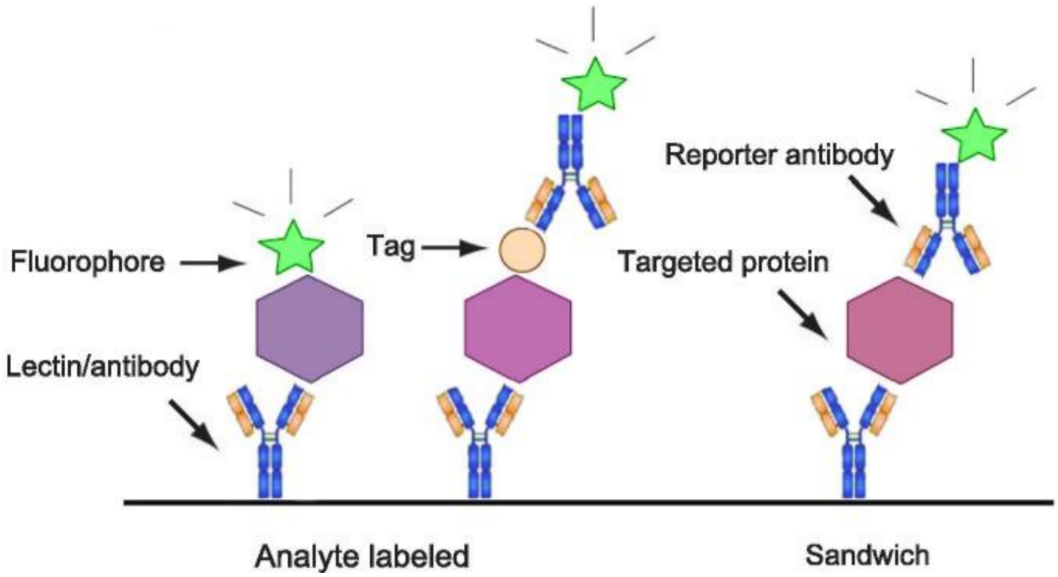


Figure 6: **Antibody microarray.** In this class of arrays, targeted proteins can be detected either by direct labeling or using a reporter antibody in sandwich assay format. Source: [89].

Serological studies using microarrays

The popularity of microarrays in the scientific community its becoming noteworthy. Here I open the floor to discuss the applicability of this technique which have gained a foothold on several research areas. On the topic of autoimmune diseases, Lin et al. [97] have performed a DotScan^{TM} antibody microarray screening of peripheral blood mononuclear cells from 60 systemic lupus erythematosus (SLE) patients of varying disease activity, 25 rheumatoid arthritis patients, 28 other autoimmune disease samples, and 24 healthy controls. The antibody microarray profiles were able to distinguish active SLE patients from healthy controls.

When combined with the standard laboratory tests, the microarrays provide significantly increased discriminative ability than conventional SLE diagnostics, which will assist better disease management [70]. In a different study, Carlsson et al. [98] constructed 135 human recombinant single-chain fragment variables (scFv) targeting immune proteins to build an in-house antibody microarray. In this study, they examined patients with systemic sclerosis (SSc), SLE and 15 healthy volunteers. This array identified 40 differential expressed proteins creating a candidate proteome signature to delineate SLE and its severity from SSc. This protein signature was a better disease classification than single or even combinations of conventional clinical parameters, illustrating the potential use for antibody microarrays to create new disease-signatures which will add clinical value to disease management.

Antibody microarrays have also been applied for cancer research, mainly to study cancer progression and candidate proteins that may serve as diagnostic biomarkers. Using an in-house developed recombinant antibody microarray platform, Wingren et al. [99] screened sera from 148 patients with pancreatic cancer, chronic pancreatitis, autoimmune pancreatitis (AIP), and healthy controls. In this study the authors were able to identify a panel of 25 protein targets, which contributed to distinguishing pancreatic cancer from healthy control. This 25 protein signature exhibited a high diagnostic potential (AUC of 0.88). Another study developed a customized antibody microarray platform containing 4096 features to interrogate plasma samples spanning pre-invasive and invasive diseases from a mouse model of pancreatic ductal adenocarcinoma (PDA) [100]. They found a protein signature, comprised of differential expression of three proteins, v-erb-b2 avian erythroblastic leukemia viral oncogene(ERBB2), Tenascin C (TNC) and Estrogen receptor alpha (ESR1), which could be used to improve the AUC from 0.86 (95% confidence interval [CI] 0.76–0.96) to 0.97 (95% CI 0.92–1.0) when the PDA marker Carbohydrate antigen 19-9 (CA19-9) was included. In prostate cancer, Schwenk et al. [101] used antibody arrays on suspension bead arrays to compare plasma levels of proteins between different groups in order to find additional biomarkers alongside prostate-specific antigen (PSA). Besides classifying the patients based on PSA, they identified decreased plasma Carnosine dipeptidase 1 (CNDP1) levels which was shown to be associated with more aggressive forms of the disease in subsequent larger studies. In an different study on small intestine neuroendocrine tumours, Darmanis et al. [102] investigated the plasma levels of two independent study sets (77 and 132 samples) and a targeted bead array for 124 unique proteins. Here, the authors were able to achieve a classification accuracy of up to 85% using a panel of proteins.

Within the field of neurodegenerative diseases Haggmark et al.[103] has used antibody suspension bead arrays for profiling cerebrospinal fluid (CSF) from patients with Multiple Sclerosis (MS). In this study, they found Growth-associated protein 43 (GAP43), a cytoplasmic protein involved in the formation, and regeneration of neurons, hence a promising biomarker of diseases of the brain. More recently, Remnestal et al. [104] compared protein levels in 441 CSF samples collected from different neurodegenerative disease sample

sets as well as CSF collected post-mortem. Among the 376 antibodies, the synaptic proteins Growth-associated protein 43 (GAP43) and Neurogranin (NRGN) were found to be associated to Alzheimer's disease patients compared with controls. In a different study, a large-scale screening was conducted utilizing 4500 antibodies on bead arrays using plasma samples, CSF and brain tissue from patients suffering from multiple sclerosis (MS) [105]. In such study a set of proteins was found to be associated with MS subtypes in CSF and plasma. In an different study bead arrays were used to profile plasma from patients suffering from amyotrophic lateral sclerosis (ALS) [106], where 367 ALS patients and 101 controls were analyzed for 278 proteins. The study concluded that neurofilament medium polypeptide (NEFM), solute carrier family 25 (SLC25A20), and regulator of G-protein signalling 18 (RGS18) were valuable proteins related to disease pathophysiology.

Antibody microarrays have also emerged as key tools for the investigation of infectious diseases. Dent et. al. [107] have probed protein microarrays covering 824 P. *falciparum* protein features with plasma from 88 children and 86 adults living in Kenya. In this study, the authors demonstrated that children gradually acquired antibodies to the full repertoire of antigens recognized by adults. Additionally, antibody levels to 106 specific antigens were significantly higher in children who were protected from symptomatic malaria compared with those who were not, highlighting the usefulness of microarrays in he search for malaria antigens associated with protective immunity. [108] also compared antibody responses from residents of two endemic areas in the western Kenyan highlands against 854 polypeptides of P. *falciparum* using high-throughput proteomic microarray technology. Here, they were able to identify 107 proteins as serum antibody targets, which were then characterized for their gene ontology biological process and cellular component of the parasite, showing significant enrichment for categories related to immune evasion, parthenogenesis and expression on the host's cell and parasite's surface. In another study differential protein profiles in plasma in children suffering from malaria and malaria-related complications were profiled [109]. From 1000 proteins screened, 41 proteins had differential expression between malaria infected children and community controls. Furthermore, thirteen further proteins were linked to malaria-disease severity. In an epidemiological study conducted by Kobayashi and colleagues [110], antibody responses in children and adults living in Zambia and Zimbabwe were profiled using a P. *falciparum* protein microarray. While there was little correlation between transmission intensity and antibody signals (magnitude or breadth) in adults, there was a clear correlation in children younger than 5 years of age. Furthermore, antibodies in adults' responses appeared to be durable even in the absence of significant recent transmission, whereas antibodies in children provided a more accurate picture of recent levels of transmission intensities.

In the context of Chronic Fatigue Syndrome, Loebel et. al [25] has performed a comprehensive mapping of the IgG response against teh Epstein-Barr virus (EBV) comparing 50 healthy controls with 92 Chronic Fatigue Syndrome (CFS) patients using a microarray plat-

form. Patients with MS, systemic lupus SLE and cancer-related fatigue served as controls. The authors found significantly enhanced IgG responses to several Epstein Barr Virus Nuclear Antigen 6 (EBNA-6)peptides containing a repeat sequence in CFS patients compared to controls. Moreover, EBNA-6 peptide IgG responses correlated well with EBNA-6 protein responses. The EBNA-6 repeat region showed sequence homologies to various human proteins. In a separate study [111], microarrays used to compare sera from ME/CFS and controls resulted in the identification of a 256-peptide immune signature with the ability to separate ME/CFS cases from controls. More recently, Assis et al. [112] described the validation of a coronavirus antigen microarray containing immunologically significant antigens from Severe acute respiratory syndrome coronavirus SARS-CoV, SARS-CoV-2, Middle East respiratory syndrome coronavirus (MERS-CoV), common human coronavirus strains and other common respiratory viruses. A comparison of antibody profiles detected on the array from control sera collected prior to the SARS-CoV-2 pandemic versus convalescent blood specimens from virologically confirmed COVID-19 cases demonstrated near complete discrimination of these two groups, with improved performance from the use of antigen combinations that include both spike protein and nucleoprotein. In such study the authors concluded that this array can be used as a diagnostic tool, as an epidemiologic tool to more accurately estimate the disease burden of COVID-19, and as a research tool to correlate antibody responses with clinical outcomes.

All together, these studies illustrate the growing interest in high-throughput technologies in serology. Although they have allowed for a broader understating of the immune system's responses to pathogens, in order to handle the data generated by this technologies, more advanced statistical algorithms have become indispensable. The next section will explore the algorithms being currently leveraged to analyze such data.

1.3 Machine Learning analysis on high-throughput antibody data

Data generated by high-throughput technologies has led to the necessity of implementing more capable statistical methodologies, suitable to handle the high dimensionality of the data. Machine Learning techniques which strive in this context have thus been adopted for the analysis of such data. In their paper, Valletta and Recker [40] have used the Random Forest algorithm to analyze antiboy data from malaria patients obtained via luminex and microarrays in an attempt to identify immune signatures predictive of clinical malaria. Bérubé and colleagues [113] have also resorted to the Random Forest to analyzed microarray antibody data in an attempt to reveal candidate biomarkers of the intensity and timing of past exposure to Plasmodium *falciparum*. In a distinct study, the authors have opted to utilize the Random Forest to analyze microarray data on antibody responses to the P. *falciparum* apical membrane antigen 1 (AMA1)variants following natural infection and vaccination [114]. Krumplamp and colleagues also made use of the Random Forest to analyze

microarray data from 35 children with P. *falciparum* malaria with the aim of identifying biomarkers for the disease. In a distinct paper, Mazhari et al. [115] have used two methods to assess potential associations between total IgG responses to individual P. *vivax* proteins and protection against clinical to the Plasmodium. In this study regression-based analysis and Random Forests were used to analyze IgG antibody responses to purified P. *vivax* proteins measured in a multiplexed Luminex assay. Longley and colleagues [116] also resorted to the use of different classifiers to analyze multiplex Luminex assay data. In such paper the authors made use of the Logistic Regression, Quadratic Discriminant Analysis (QDA), Decision Trees, and Random Forests algorithms for the identification o P. *vivax* serological markers. In an different study, the Logistic regression was used to analyze antibodies against P. *vivax* in an attempt to identify protective signatures in populations living in low malaria transmission regions of Brazil and Thailand [117]. In the study conducted by Assefa and colleagues, the Logistic Regression was used on multiplex serological data to analyze the prevalence and spacial distribution of malaria in Ethipoia [118]. Meanwhile, Propriety et. al. [119] established a predictive modeling framework to analyze IgG antibody responses against a large panel of P. *falciparum*-specific antigens obtained using microarrays consisting on distinct algorithms. Here the Logistic Regression, Support Vector Machine (SVM) and Partial-Least Squares Disciminant Analysis(PLS-DA) algorithms were integrated in an attempt to identify immune signature against P. *falciparum* antigens predictive of clinical immunity in distinct malaria endemic communities. Helb and colleagues [24], on the other, have resorted to the use the Random Forest and LASSO Regression embedded in a Super Learner for the analysis of 856 P. *falciparum* antigens obtain by protein microarray in 186 Ugandan children, with the aim of identifying novel serologic biomarkers of malaria exposure.

Foulquier et. al. [114] have also analyzed high-throughput mutiplexed data using the Random Forest algorithm in an attempt to identify signatures for anti–Sjögren's-syndrome-related antigen A autoantibodies (Anti–SSA/Ro), from 60 antibodies on distinct autoimmune diseases. In another study, different Machine Learning algorithms have been used to analyze antibody data obtained from Illumina NextSeq [42]. Here the Random Forest, Support Vector Machine and a Neural Network (Multilayer Perceptron) were implemented for the analysis of dengue antibody responses. In a recent study, Rosado and colleagues [120] used the Random Forest algorithm to analyze multiplex assay data with the aim of the identification serological signatures of SARS-CoV-2 infection.

Given that the objective here is not to exhaust the literature's examples on the ML-based techniques for the analysis of high-throughput antibody data, but to highlight their growing to analyze such data, I redirect the attention of the more interested reader to [121] where several others studies leveraging these algorithms on high-throughput serologic data are documented. Nonetheless, these studies highlight the ongoing implementation of Machine Learning algorithms on high-throughput serological data. In the next chapter, I will

explore in detail the algorithms here introduced, and a few others which were used through the development of this work.

1.4 Objectives and outline of the thesis

The overall aim of the thesis was the development of statistical pipelines for the identification of antibody biomarkers for outcomes related to Malaria and CFS. Although my work focused mostly on the analysis of antibody data, I also dealt with the analysis of DNA methylation and gene expression data due to the emergency of the COVID-19 pandemic.

In part II, I aimed at constructing the backbone of statistical pipelines using low-dimensional antibody data related to susceptibility to malaria. In particular, I developed and assessed the performance of these initial pipelines to understand their susceptibility to such data while elucidating potential disease biomarkers. Initially, a non-parametric statistical pipeline for the identification of antibody responses associated with protection against clinical malaria (Chapter 4) was proposed. Such pipeline relied on the identification of cut-off values in the antibody distributions that maximized the distinction between susceptible and protected individuals,followed by the implementation of distinct Machine Learning methods for the construction of classifiers for clinical malaria.

We then extend the analysis to the same data where the previous pipeline was improved upon and two new pipelines for the analysis of antibody data were developed (Chapter 5). The first relied on a simple selecting strategy based on the use of the non-parametric Mann–Whitney-Wilcoxon test. The second approach, on the other hand, was much more statistically complex, relying on an hybrid parametric/non-parametric analysis that integrated Box-Cox transformation followed by a t-test, together with the use of finite mixture models and the Mann–Whitney-Wilcoxon test as a last resort. Upon selection of the relevant antibodies in each pipeline, a predictive analysis was conducted using an ensemble method called Super Learner, which combined the predictions from distinct classifiers together to provide more accurate result. Finally, data splitting into a train and test subset and correction for multiple testing were elements considered.

In part III, my goal was to introduce new pipelines or to adapt the previous ones for the analysis of high-dimensional data. Moved by the SARS-CoV-2 breakout which resulted in innumerable patients struggling with long-COVID symptoms akin to those of ME/CFS, we set to study the latter. Initially, we set to answer whether the expression of the human angiotensin-converting enzyme 2 (ACE2), the major cell entry receptor for SARS-CoV-2, could be also altered in ME/CFS patients and thus understand if such patients had an increased susceptibility to COVID-19 (Chapter 6). For this, a meta-analysis of public CpG

DNA methylation and gene expression data for ACE2 and its homologous ACE protein was performed, relying mostly on the use of the Logistic Regression model.

Prompted by this initial analysis to ME/CFS data, we then moved to the implementation of two distinct statistical strategies to analyze a dataset of more than 3,000 antibodies with the aim of identifying antibody biomarkers against the EBV in ME/CFS patients using public data (Chapter 7 and 8). While the initial analysis relied on the use of linear models, the second rested on the use of more flexible Machine Learning models.

Finally, in Chapter 9, I investigated the influence of sample size on the performance and the generalization ability of the ML-based approaches. Here, I addressed the research questions associated with using these approaches on datasets with limited sample sizes, which is often done in high-throughput antibody studies. For this study we have analyzed IgG antibody data against more than 2000 P. *falciparum* antigens.

Chapter 2

Background knowledge on Machine Learning

Machine Learning (ML) is a branch of Artificial Intelligence (AI) that uses various statistical and mathematical algorithms to learn from the data in order to make informed predictions or decisions [122]. ML has established its footprint in every day life situations, from personalized recommendations on streaming platforms to self-driving cars and fraud detection in financial institutions, ML is becoming an indispensable tool in today's society. In the healthcare, the story is no different. ML algorithms have made it possible to analyze vast amounts of patient data assisting in early disease diagnosis [123, 124, 125, 126] and the formulation of treatment strategies, that have improved patient outcome [127, 128]. Nonetheless, given that many of these techniques were originally not designed to cope with large amounts of irrelevant features, combining them with feature selection techniques has become a necessity in many applications [129, 46, 130, 44]. Therefore ML-based analysis tend to be conducted in a two-step approach, where an initial feature selection is conducted, followed by a predictive analysis were classifiers are built upon the remaining features [130, 40, 131, 132, 133]. This section is devoted to introducing the different feature selection commonly used and the Machine Learning algorithms traditionally implemented in predictive analysis.

2.1 Feature selection strategies

Feature selection works by removing features that are irrelevant or redundant to explain the outcome of interest [44]. As earlier mentioned, this process not only ensures a more accurate and interpretable model overall, but also allows for faster and more cost-effective models to be implemented [130]. Collectively, feature selection algorithms can be separated into three categories: Filters, Wrappers and Embedded. Since no single algorithm can be specialized to be optimal for all problem settings, here I have implemented models from each category [134].

2.1.1 Filters

Filter methods work without taking the classifier into consideration [44, 135]. In other words, these methods are performed independently of the classification analysis, rendering them to be very computationally efficient and highly scalable [136]. Filter methods, often rely on intrinsic data characteristics such as statistical measures (i.e., p-value, correlation coefficients) and can be divided into univariate or multivariate methods [44, 135]. Univariate methods consider each feature separately, while multivariate are able to find re-

lationships among the features [44]. Although the latter provide obvious advantages over the first, often they are slower and less scalable than univariate techniques [136]. Due to this fact, here I relied on the use of univariate methods such as the t-test, χ^2 and the Mann-Whitney (Wilcoxon rank-sum tests) which are well described in the literature [130, 137, 138, 139, 140]. These methods will not be here addressed, for a more detailed description, I redirect the reader to [141, 142, 143]. Finally, mixture modeling was also used as a filter method in our analysis as described in [44]. Given that mixture models are not of common knowledge, here I will spend a few lines to introduce them.

Finite mixture models

Finite mixture models have been widely used in antibody (or serological) data analysis in order to help classifying individuals into either antibody-positive or antibody-negative [88, 144]. The basic assumption behind these models is that antibody distribution consists of different latent populations, each one representing a distinct antibody state or different degrees of exposure to a given antigen [88]. As such, they can be decomposed in a mixture of two latent seronegative (lower degree of exposure) and a seropositive (higher degree of exposure) populations. Due to their conceptual simplicity and easy interpretation, the most popular finite mixture models invoke the existence of two components related to hypothetical seronegative and seropositive individuals or, equivalently, antibody-negative and antibody-positive individuals [144, 145, 146, 140]. The most popular are the Log normal distribution in the original scale of the measurements or, equivalently, the Normal distribution after logarithmic transformation of the data [88]. Two component mixture Gaussian models invoke a Gaussian distribution with mean value μ_0 and standard deviation σ_0 for the seronegative population and another one with mean value μ_1 and standard σ_1 for the seropositive population. For independent and identically distributed random sample of n individuals, the corresponding sampling distribution is described by the following equation:

$$f(\{x_i\}|\mu_0,\mu_1,\sigma_0,\sigma_1,\pi)=\prod_{i=1}^{n}\Big((1-\pi)f_{N(\mu_0,\sigma_0)}(x_i)+\pi f_{N(\mu_1,\sigma_1)}(x_i)\Big), \tag{1}$$

where x_i is the antibody level of the i^{th} individual in the sample, $f_{N(\mu_0,\sigma_0)}(x_i)$ and $f_{N(\mu_1,\sigma_1)}(x_i)$ are the probability density functions of the Gaussian distributions associated with seronegative and seropositive populations, respectively, and π is the probability of sampling a seropositive individual from the population [144]. The parameter estimation can then be performed by maximum likelihood estimation [144]. Although the normal distribution is most popular choice for the mixing distribution in serological analysis often, the distributions of this data shows long tails and skewed to the right in each latent population, even after applying log-transformation [140]. In such cases, one can use instead the Generalized

Student's t-distribution as the mixing distribution, because it has heavier tails than the Normal distribution [88]. Therefore, to better accommodate this type of data I also employed mixture models with the Generalized Student's t-distribution.

Apart from these, I also implemented less trivial mixture models, namely Skew-Normal distributions which exhibit higher flexibility [88]. The flexibility of this family is attributed to four parameters that control the location, the scale, the skewness and the flatness of the resulting distribution [88]. Note that Skew-Normal distributions include both the normal distribution, and the Generalized Student's t-distribution. To incorporate asymmetry in the modeling, one can use the Skew-Normal as the mixing distribution in which a random variable W_k has a Skew-Normal distribution with location parameter μ_k, scale parameter σ_k^2 and skewnwess parameter α_k (denoted as $W_k \frown SN(\mu_k,\sigma_k^2,\alpha_k)$) [147]. In this case the probability density function (PDF) can be written as

$$f_{W_k}(w)=2\frac{1}{\sqrt{2\pi\sigma_k}}e^{-\frac{(w-\mu_k)^2}{2\sigma_k^2}}\int_{-\infty}^{\alpha_k\frac{(w-\mu_k)}{\sigma_k}}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx=$$

$$=2\phi(w;\mu_k,\sigma_k^2)\Phi\left(\frac{\alpha_k(2\ w-\mu_k)}{\sigma_k}\right), w,\mu_k,\alpha_k\in\mathbb{R},\sigma_k\in\mathbb{R}^+$$

which can be plugged into equation (1), where $\phi(w;\mu_k,\sigma_k^2)$ denotes the PDF of the normal distribution with mean μ_k and variance σ_k^2; $\Phi(.)$ denotes the the cumulative distribution function of the standard Normal distribution [148, 149, 150]. The mean and variance of the Skew-Normal distribution are respectively given by,

$$E(W_k)=\mu_k+\sigma_k\sqrt{\frac{2}{\pi}}\frac{\alpha_k}{\sqrt{1+\alpha_k^2}},\quad V(W_k)=\left(1-\left(\frac{2}{\pi}\frac{\alpha_k}{\sqrt{1+\alpha_k^2}}\right)^2\right)\sigma_k^2.$$

Concerning the Skew-t and Student's t-distributions these were obtained as follows, where if $Z_k \frown ST(\mu_k,\sigma_k^2,\alpha_k,\nu_k)$, then its PDF is given by

$$f_{Z_k}(z)=2t(z;\mu_k,\sigma_k,\nu_k+1)T\left(A\sqrt{\frac{\nu_k+1}{d+\nu_k}};\nu k+1\right),$$

with A $=\frac{\alpha_k(z-\mu_k)}{\sigma_k}$ and $d=\left(\frac{x-\mu_k}{\sigma_k}\right)$, where $t(z;\mu_k,\sigma_k,\nu_k+1)$ denotes the probability density function of a Student-t distribution with location parameter with μ_k, scale parameter σ_k and ν_k+1 degrees of freedom, $T(.;\nu_k+1)$ represents the cumulative distribution function of a standard Student-t distribution with ν_k+1 degrees of freedom [88].When the skewness parameter is equal to zero, i.e., $\alpha_k=0$, the quantity $A=\frac{\alpha_k(z-\mu_k)}{\sigma_k}=0$, and the above expression takes the form

$$f_{Z_k}(z)=t(z;\mu_k,\sigma_k,\nu_k+1)$$

which corresponds to the probability density function of a Generalized Student-t distribution with location parameter μ_k, scale parameter σ_k and $\upsilon_k + 1$ degrees of freedom [88].The mean and variance of the Skew-t distribution are respectively given by,

$$E(Z_k) = \mu_k + \sigma_k b_{\upsilon k} \delta_k \ \ if \ \ \upsilon_k > 1, \quad V(Z_k) = \sigma_k^2 [\frac{\upsilon_k}{\upsilon_k - 2} - (\upsilon_{\upsilon k} \delta_k)^2] \ \ if \ \ \upsilon_k > 2,$$

where $b_{\upsilon_k} = \frac{\sqrt{\upsilon_k}\Gamma(\frac{1}{2}(\upsilon_k - 1))}{\sqrt{\pi}\Gamma(\frac{1}{2}\upsilon_k)}$ and $\delta_k = \frac{\alpha_k}{\sqrt{1+\alpha_k^2}}$ are parameters involved in calculating the variance of the distribution and Γ refers to the Gamma function that generalizes the factorial function to complex and real numbers [88]. These equations can then be plugged into equation 1 to obtain a mixture model of the respective distributions.

2.1.2 Wrappers

A significant disadvantage of filter methods is the inability to detect complex relations between multiple features and the outcome of interest, which generally translates into poorer results in the predictive phase [44]. Wrappers on the other hand detect feature dependencies to the outcome variable [130]. Thus, wrapper methods identify the best-performing set of features for the chosen classification algorithm. This is achieved through an iterative search process that uses the performance of the classifer at each iteration to guide the search process [140]. Therefore, generally, wrappers achieve a better performance than filter methods [130]. However, these comes at a certain cost. Wrappers work by constructing models from scratch for each generated subset, then using prediction performance as a criterion function to select the best subset [151]. To be able to determine and compare the relative output of each generated subset, wrappers need to used learning algorithms which renders them to be complex and expensive in terms of execution time. This becomes a critical aspect when thousands of features are considered and therefore wrappers have been largely avoided when dealing with high-dimensional data [135]. Notwithstanding, diverse wrapper methods have been proposed in the literature [152, 153, 154, 136, 155]. Here I have made use of three distinct Wrapper which follow under the Subset Selection models category: *Forward Selection*, *Backward Selection* and *Stepwise Selection* [137]. I will now briefly discuss this models.

2.1.2.1 Forward selection

Forward selection works by fitting an empty model to the data and adding a feature (predictor) until there are no more predictors to add [156, 135]. In this sense, a null model, containing no features, is initially fitted. Then, the predictor that contributes the most to an increase in the model's performance is added to the model, and the model is refitted [137]. This process of adding the highest contributing predictor and refitting a new model with the added variable is repeated until all the predictors have been added [137]. Finally,

from all the models generated, an overall single best model is selected using the prediction error, such as the classification error rate[3] or the deviance [156, 154]. Alternatively, cross-validation or information criteria measures such as the Akaike Information Criterion (AIC)[4] [157], or Bayesian Information Criterion (BIC)[5] can also be used [158]. Instead, predictors can be added until a predetermined stopping criterion is met. This criterion could be to reach a specific number of predictors or to achieve a certain level of model performance [156]. As an example, one might decide to stop when adding more predictors doesn't significantly improve the model's performance

In general, forward selection can be applied even in the high-dimensional setting in which $n < p$, where p represents the number of variables and n the number of samples [156]. Nonetheless, given that forward selection doesn't perform an exhaustive search across all possible combination of predictors, it may not always yield the optimal subset of predictors [156]. Also, the order in which the predictors are added affects this approach. To better understand why this is the case I will explore a simple example: Consider a given dataset with $p = 3$ predictors, in which the best possible one-variable model contains the predictor X_1, and the best possible two-predictor model instead contains X_2 and X_3. In such case, Forward stepwise selection would not be able to select the best two-predictor model because, since the model with one predictor will contain X_1, the model with two predictors must also contain X_1 together with an additional variable [156].

2.1.2.2 Backward selection

Similarly to forward selection, backward selection provides an efficient way to select relevant features. However, in opposition to forward selection, backward selection begins with a full model containing all p predictors [137, 154]. In each iteration, the predictor that contributes the least to the model's performance is removed and the model is refitted. This process continues until a predetermined stopping criterion is met or until there is just a single predictor remaining [156]. Finally, the best overall model is obtained as previously described. A major disadvantage of backward selection is that it requires the sample size

[3] The classification error rate is a common approach to determine the performance of a model that denotes the number of wrongly label observation predicted by a model, given by: $Err = \frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$, where y_i is the true label (class) for the i^{th} observation, $\hat{y}_i$ the predicted label for such observation and I is the indicator function, which returns 1 if the condition inside is true and 0 otherwise.

[4] AIC is a statistical criterion used for model comparison and selection. The basic idea behind AIC is to balance the goodness of fit of a model with its complexity or number of parameters. The AIC score is calculated as follows: $AIC = -2 * \text{log-likelihood} + 2 * \text{number of parameters}$, where the log-likelihood measures how well the model explains the observed data, and the penalty term ($2 *$ number of parameters) discourages overly complex models. The lower the AIC value, the better the model is considered, as it indicates a good trade-off between goodness of fit and complexity.

[5] BIC is also a statistical criterion used to compare and select models. Like AIC, BIC aims to strike a balance between model fit and model complexity. BIC penalizes complex models more heavily than AIC. The formula for BIC is given by: $BIC = -2 * \text{likelihood} + \log(n) * \text{number of parameters}$, where n is the sample size. The penalty term in BIC ($log(n)$* number of parameters) includes a factor related to the sample size, making BIC tend to prefer simpler models, especially when the sample size is large.

n to be larger than the number of predictors p, so that the full model can be fitted [156]. In contrast, forward stepwise selection can be used even when $n < p$, and so in cases it represents a more viable option [156].

2.1.2.3 Hybrid approaches

As an alternative to using either one of these selection approaches, one can implement hybrid versions of forward and backward selection. In these approaches an empty mode is fitted and predictors are iteratively added, however after a new predictor has been added, such as in forward selection, the method may also remove any variable that no longer provides an improvement in the model fit, such as in backward selection [159, 156]. The overall idea is to iteratively refine the model by adding variables that improve its fit and removing variables that do not contribute significantly. Such an approach increases the chance of identifying the optimal subset of the p predictors, while retaining the computational complexity of forward and backward stepwise selection [156]. Such hybrid approaches are also referred to as bidirectional or stepwise selection.

2.1.3 Embedded

Embedded methods have emerged as an alternative to reduce the computation time taken by wrapper methods that require the reclassification of the different data subsets [160]. The main advantage of embedded methods is that feature selection is integrated or built into the classification algorithm [130]. These classifiers adjust their internal parameters and determine the appropriate weights/importance given for each feature to produce the best classification accuracy [130]. Therefore, the search for the optimal feature subset is embedded in the model construction which are combined in a single step. These render embedded methods to be more efficient and perform better than wrappers [135, 140]. All around, embedded methods can bee seen as an intermediate solution between filter and wrapper methods, in the sense that they combine the low computational cost of implementation (filters) while retaining the ability to detect the feature interactions with the classifier (wrappers) [130]. Some examples of embedded methods include tree-based algorithms (i.e., Decision Tree, Random Forest, Gradient Boosting), and regularization models (i.e., LASSO or Elastic-Net) which usually work with linear classifiers, such as Logistic Regression, by penalizing/shrinking the coefficient of features that do not contribute to the model in a meaningful way [130, 160, 135, 137]. Given that these models are used fo classification, they will be addressed in the following section where I explore the classification algorithms commonly used during predictive analysis.

2.2 Predictive analysis

Predictive analysis refers to a data-driven approach that uses statistical algorithms and Machine Learning methods to exploit current data as a means to make informed predictions about future events or trends [161]. Within this framework, models are typically implemented on a set of data to learn its patterns and relationships subsequently providing an accurate prediction on new unseen data. Predictive analysis belongs to the realm of supervised learning, in which the presence of the outcome variable guides or "supervises" the learning process [161, 156]. Contrarily, in instances where one observes only the features and have no measurements of the outcome, the analysis will belong to the unsupervised learning field [156]. The latter is very helpful in detecting patterns and similarities between the variables within a data set, however, I will not address it here [156, 162]. When predictive analysis explores the relationship between two or more inputs (predictors or independent variables) and a single outcome (dependent variable) an analysis is said to be multivariable [163].

Multivariable analyses can be divided into regression or classification analyses depending on the type of data of the outcome [161]. While the first refers to cases where the outcome variable is numeric, the second alludes to situations where the outcome variable is categorical, this is, the outcome is divided into distinct non-overlapping categories or classes [156]. Classification analyses thus rely on classifying or assigning the observation to specific categories, or classes [156, 161]. Nonetheless, instead of directly predicting a class, often the methods used for classification first predict the probability that the observation belongs to each of the categories of a qualitative variable, as the basis for making the classification [156]. Classification problems, especially binary classification problems, in which there are only two available classes, are one of the most recurring and relevant tasks [164]. Currently, aided by the advents in ML, hundreds of models for classification problems are becoming available each year [156]. In this chapter, I will explore the ML algorithms here implemented, which represent some of the most commonly used for classification purposes.

2.2.1 Logistic Regression

Logistic Regression is one of the most commonly used algorithms for classification purposes [165]. This approach is well suited for describing relationships between one or more categorical or numerical predictors and a categorical outcome. Usually the Logistic Regression is used in cases where the outcome has only two categories (i.e., it is binary/dichotomous) [156]. In such cases, the underlying distribution is Binomial, where each observation represents a Bernoulli distribution [166]. The Bernoulli distribution is a discrete probability distribution that can only have two outcomes, 1 referring to the probability p of the event of interest occurring or 0 the probability $1-p$ of the event not occurring [166]. Hence, I can

define the variable (Y) as

$$Y = \begin{cases} 1 & \text{event occurring} \\ 0 & \text{event not occurring} \end{cases}$$

Then to assign a class k^{th} to which Y belongs, a model must predict the probability of Y belonging to each of these two classes [156]. Given that probabilities range from 0 to 1, the predicted probability by the model must use a function that also gives outputs limited to such interval. This can be achieved using the Logistic function:

$$p(x) = P(Y = k|X = x) = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}} = \frac{1}{1+e^{-(\alpha+\beta X)}} \tag{2}$$

where x is a specific value of the variable X, p is the probability of the event of interest, α is the Y-intercept, β is the regression coefficient [167]. The Logistic Regression will always produce a sigmoidal or S-shaped curve, which will constrain the predicted values to lie between 0 and 1 [156, 167] (Figure 7). However, these S-shaped curves make it difficult to estimate the probability of Y via a linear equation. This is because the errors are neither normally distributed nor constant across the entire range of data [156]. The Logistic Regression solves this problem by applying a link function, the logit transformation to the dependent Y variable [156, 167]. So, after manipulating equation (2), one reaches to the Logistic Regression model, which estimates the probability of an event occurring by having the logit (log odds) of the outcome being a linear combination of the independent variables [156, 167]:

$$log(\frac{p(x)}{1-p(x)}) = \alpha + \beta X$$

The left side of the equation is called log odds because it takes the natural logarithm (log) of the odds of Y, given by $\frac{p}{(1-p)}$, where p is the probability of an event occurring and $1-p$ the probability of the event not occurring. Thus, the Logistic Regression model can also be denoted as:

$$logit(Y) = natural\ log(odds) = log(\frac{p}{1-p}) = \alpha + \beta X \tag{3}$$

Therefore, the logit function enables the probabilities of the categorical response variable to be transformed to a continuous scale allowing the relationship between the predictors and the response variable to be modeled using a linear function [165]. For this reason, it is also called a link function as it links the response variable to the predictors via a linear relationship [165, 167]. Extending the logic of the simple Logistic Regression (3) to multiple predictors, one can construct a complex Logistic Regression for Y as follows:

$$logit(Y) = log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \tag{4}$$

where $X=(X_1,...,X_p)$ are p predictors. Equation (4) thus be rewritten as,

$$p(x)=\frac{e^{(\alpha+\beta_1 X_1+\beta_2 X_2+...+\beta_n X_n)}}{1+e^{(\alpha+\beta_1 X_1+\beta_2 X_2+...+\beta_n X_n)}}.$$

However, as earlier stated the Logistic Regression only estimates the probability $p(X)$ that an observation belongs to each of the categories of Y, it does not directly classify into one of both classes. However, based on the estimated probabilities given by the Logistic Regression one can attribute the output to one of the classes by choosing a cutoff value and assigning the observations with a probability greater than the cutoff as one class and below as another [156]. Usually, the 0.5 threshold is the most commonly used for such classification, where a prediction assigns the observation to class 1 if $P(Y=k \mid X=x)>0.5$ and 0 otherwise. According to this method, the predicted class $\hat{y}$ can be denoted as:

$$\hat{y}=\begin{cases}1 & if\ p\geq 0.5\\ 0 & if\ p<0.5\end{cases}$$

Nevertheless, the selection of the probability cut-off can be estimated using other thresholds. Finally, here I have introduced the Logistic Regression which makes use of the logit link function to provide a relationship between the linear predictors and the expectation of the response variable [167]. However, different models that rely on the use of other link functions can be implemented to model such relationships. The probit and the complementary log-log link functions are some of the most commonly used alternatives, however, the logit link function is preferred because of the easier interpretation of the coefficients [168].

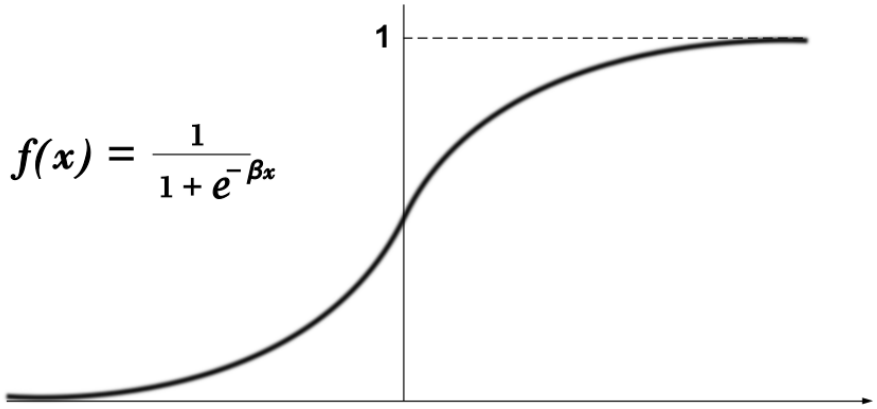


Figure 7: **Sigmoid curve.** The sigmoid function yields predictions that lie within the 0 to 1 probability interval. Source: [169].

2.2.1.1 Maximum likelihood estimation (MLE)

To fit the Logistic model, I use the Maximum Likelihood Estimation (MLE) [165, 167, 170].

MLE is utilized for estimating the α and β parameters of a statistical model. When fitting a Logistic Regression model, these coefficients are unknown and thus must be estimated based on the available training data [156]. Estimation of such coefficients is achieved through the maximization of the likelihood function, which measures how well the model predicts the observed labels in the training data [156]. The basic intuition behind maximum likelihood is to estimate the Logistic Regression coefficients α and β, such that the predicted probability $p(X)$ for an observation corresponds as closely as possible to its true class [167]. Put it differently, one seeks the model coefficients, that maximize the probability (or likelihood) of observing the given data used to train the model. In this sense, the likelihood function can be formalized by the following mathematical equation:

$$
\begin{aligned}
L(\alpha,\beta) &= \prod_{i=1}^{N} P(Y_i = y_i \mid X_i) \\
&= \prod_{i=1}^{N} p(x_i)^{y_i} p(1-x_i)^{1-y_i} \\
&= p(x_i)^{\sum_{i=1}^{N} y_i} p(1-x_i)^{\sum_{i=1}^{N} 1-y_i}
\end{aligned}
$$

where the product of the probability of an event occurring $p(x)$ by the probability of an event not occurring $p(1-x_i)$ is obtained across all the samples (i^{th}). However, to simplify the expression one can take the log of the likelihood, which allows us to use the property $log(x*y) = log(x) + log(y)$ to transform the product of a logarithm into a sum of logarithms:

$$
\begin{aligned}
log(L(\alpha\beta) &= log\left(p(x_i)^{\sum_{i=1}^{N} y_i} p(1-x_i)^{\sum_{i=1}^{N} 1-y_i} \right) \\
&= \sum_{i=1}^{N} y_i.log(p(x_i)) + \sum_{i=1}^{N} (1-y_i).log(p-p(x_i)) \\
&= \sum_{i=1}^{N} [y_i.log(p(x_i)) + (1-y_i).log(p-p(x_i))]
\end{aligned}
$$

Finally, estimation of the parameters α and β is done by calculating the partial derivatives of log-likelihood with respect to each coefficient/parameter and setting them equal to zero:

$$
\frac{\delta}{\delta\alpha} Log(L(\alpha,\beta)) = 0 \quad and \quad \frac{\delta}{\delta\beta} Log(L(\alpha,\beta)) = 0
$$

These equations are then typically solved using numerical optimization methods, such as Gradient Ascent, which is used to find the values of α and β that maximize the likelihood function [170]. In this sense, while the mathematical expressions for the MLE set up a system of equations, numerical methods are used to solve for the coefficient values that maximize the likelihood function [170]. The Maximum Likelihood is a very general approach used to fit not only the Logistic regression but many other models that I will introduce throughout the following sections [156].

2.2.1.2 Gradient Descent

While optimization of the parameters in the logistic regression can be achieved using gradient ascent, which is used for maximization of a function (i.e., MLE), it is more common to use gradient descent to minimize a cost function (i.e., negative log-likelihood) [170]. Therefore, the standard practice is to use gradient descent to update the parameters in the direction that minimizes the loss function. In other words, gradient descent aims to find the minimum of a function by adjusting parameters in the direction of the steepest decrease of such function [170, 171]. Initially, the Gradient Descent sets an arbitrary value for the coefficients, usually 0, and the negative log-likelihood objective function is computed, leading to a variation of coefficients [171]. This process is iteratively repeated until convergence[6] or the specified number of iterations is completed. Once convergence is achieved, the coefficients α and β at the end of the iteration represent the maximum likelihood estimates for the logistic regression model [170, 171].

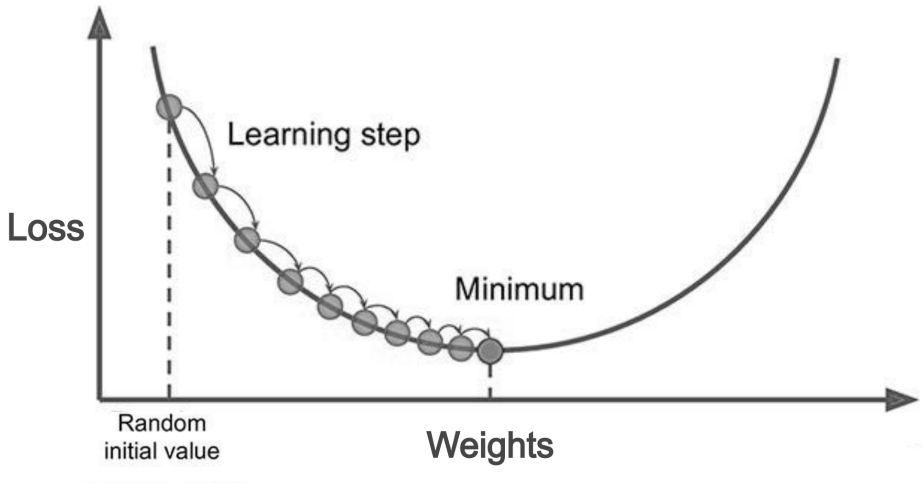


Figure 8: **Gradient descent.** The gradient descent adjusts the parameters in the direction of the steepest decrease of a function. Source: [172].

2.2.2 Ridge Logistic Regression

The Ridge Logistic Regression is a well-known shrinkage technique [135]. Shrinkage approaches involve fitting a model with all predictors and then applying a technique to constrain or regularize the coefficient estimates, shrinking them towards zero [156]. This is achieved by modifying the loss function with a penalty term which effectively shrinks the estimates of the coefficients [156, 173]. For this reason, these types of methods are also

[6]Models are said to converge when the classification error (or loss) stops decreasing or reaches a minimum error level. It refers to the state where the model has reached an optimal solution. This means that the model's parameters have been adjusted to the point where further training will not significantly improve its performance. Convergence is typically achieved through iterative optimization algorithms, such as gradient descent, which update the model's parameters to minimize a specified loss function. When a model converges, it has effectively learned the underlying patterns in the training data.

called “shrinkage” or “regularization" methods [156]. The underlying logic for the use of such constraint lies on the fact that by shrinking the coefficient estimates one can significantly improve the fit of the model. The best-known techniques for shrinkage are the Ridge, LASSO and Elastic-Net regression. However, as their names imply, these techniques are only suitable for regression analysis [156]. When applied to the classification setting, regularization is often performed using Logistic Regression or other classification algorithms [130]. In the case of the Logistic Regression, a shrinkage penalty is applied to the model, and depending on the penalty used one obtains the Ridge, LASSO and Elastic-Net Logistic Regressions. The Ridge Logistic Regression is given by:

$$\sum_{i=1}^{N}[y_i.log(p_i)+(1-y_i).log(1-p_i)]-\lambda\sum_{j=1}^{p}{\beta_j}^2$$

where λ is a regularization or tunning parameter and $\sum_{j=1}^{p}{\beta_j}^2$ is the penalty term also known as the L2 norm or penalty, calculated as the square root of the sum of the coefficients vector values [135, 174]. Similar to the Logistic Regression model, the overall goal of ridge classification regression is to obtain the coefficient estimates that maximize this slightly different log-likelihood function where the L2 ridge penalty is added [156]. Thus, the objective function for Ridge Logistic Regression can be written as:

$$\hat{\beta}^R = \underset{\beta}{maximize}\{\sum_{i=1}^{N}[y_i.log(p_i)+(1-y_i).log(1-p_i)]-\lambda\sum_{j=1}^{p}{\beta_j}^2\} \tag{5}$$

Here, the λ tuning parameter controls the relative impact of the penalty term on the regression coefficient estimates (the amount of shrinkage) and is calculated separately [174]. When $\lambda = 0$, the penalty term has no effect, and Ridge Logistic Regression will produce the Logistic Regression estimates[174]. However, as λ increases, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero. Nonetheless, none of the coefficients will be exactly zero [174, 156]. Although this may not be a problem for prediction accuracy, it can create a challenge in model interpretation, especially in settings in which the number of variables is quite large.

Overall, the penalty term penalizes large coefficients helping to prevent overfitting[7] and stabilising the estimates in the presence of multicollinearity[8] which in turn improves the generalization[9] of the model [156, 174]. Finally, for a set of random predictor $X = X_1,...,X_j$,

[7] Overfitting is a phenomenon in Machine Learning where models learn the training data too well, capturing noise or random fluctuations, consequently, performing poorly on new, unseen data

[8] Multicollinearity refers to a high degree of correlation among independent variables in a regression model, which can lead to difficulties in estimating individual variable effects and results in an unstable and unreliable coefficients.

[9] A model's generalization refers to its ability to accurately predict outcomes on new, unseen data based on what the model learned on the training data.

the estimation of each coefficient $X_j\hat{\beta,\lambda}_j^R$ depends not only on the value of λ but also on the scaling of the j^{th} predictor itself. Therefore, it is best to apply ridge regression after standardizing[10] the predictors. As a result, the final fit will not depend on the scale on which the predictors are measured [173]. Furthermore, normalizing the data provides a way to rank predictor variables based on the regression coefficients. This however is only possible if the data is normalized, as one can't compare the coefficients of different scaled variables [173].

2.2.3 LASSO Logistic Regression

As I have just described, in Ridge Logistic Regression, increasing the value of λ will tend to reduce the magnitudes of the coefficients, but will not result in the exclusion of any of the variables. Therefore, Ridge Logistic Regression will always generate a model involving all predictors. A recent alternative to ridge regression called the Least Absolute Shrinkage and Selection Operation (LASSO), overcomes this disadvantage [175]. In LASSO Logistic Regression, the penalized version of the log-likelihood function to be maximized takes now the form:

$$\hat{\beta}^L = \underset{\beta}{maximize}\{\sum_{i=1}^{N}[y_i.log(p_i)+(1-y_i).log(1-p_i)]-\lambda\sum_{j=1}^{p}|\beta_j|\}. \quad (6)$$

Comparing equation (5) to (6), one can see that the LASSO and Ridge Regression have similar formulations. The only difference is that the β_j^2 term in the ridge regression has been replaced by $|\beta_j|$ in the LASSO. This penalty term is called the L1 penalty term [175]. As with ridge regression, the lasso shrinks the coefficient estimates towards zero. However, in the case of the lasso, the L1 penalty has the ability to force some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large [175]. Hence, depending on the value of λ, the LASSO can produce a model involving any number of variables [135]. Hence, LASSO performs variable selection. As a result, models generated from the LASSO are generally much easier to interpret than those produced by ridge regression.

Both, Ridge and LASSO Logistic Regression produce a set of coefficient estimates whose values depend on the different values of λ. Therefore, choosing a good value of λ is a critical step for both methods [135]. To accomplish this task, different approaches are described in the literature. Among these, cross-validation , which I will explore latter, provides a simple way to tackle this problem [175]. Briefly, a grid of λ values is initially chosen and the cross-validation error for each value is computed. The tuning parameter value for which the cross-validation error is smallest (or the log-likelihood maximized) is then selected and finally, the model is refitted using all of the available observations and the selected value of

[10] Standardization is the process of transforming data by rescaling it to have a mean of 0 and a standard deviation of 1, facilitating comparisons between variables with different units and scales

the tuning parameter [174].

2.2.4 Elastic-Net Logistic Regression

More recently, another regularization method, called Elastic-Net Logistic Regression, was developed. This combines the L1 (Lasso) and L2 (Ridge) regularization penalties, however further includes a tuning parameter $\alpha \geq 0$:

$$\sum_{i=1}^{N}[y_i.log(p_i)+(1-y_i).log(1-p_i)]-\alpha\sum_{j=1}^{p}\beta_j^2+(1-\alpha)\sum_{j=1}^{p}|\beta_j|$$

Thus the Elastic-Net Logistic Regression combines the Ridge and Lasso regularizations, providing a way to control the balance between the L1 and L2 penalties offering a more versatile approach [162, 176].

2.2.5 Linear Discriminant analysis

Linear Discriminant Analysis (LDA) is a classification method similar to Logistic Regression with a different way to model $P(Y=k|X=x)$. Here, rather than calculating this probability, LDA models the distribution of the predictors X separately for each of the response classes and then use the Bayes theorem[11] to flip these around into estimates for $P(Y=k|X=x)$ [156, 177]. LDA classifies an observation into one of K possible classes by estimating the prior probability that a randomly chosen observation belonging to the K^{th} class, π_k and the density function of X for an observation that comes from the K^{th} class, $fk(x)=Pr(X|Y=k)$, and plugging them into the Bayes theorem[156]:

$$P(Y=k|X=x)=\frac{\pi_k f_k(x)}{\sum_{k=1}^{K}\pi_k f_k(x)}. \tag{7}$$

This provides the posterior probability of an observation $X=x$ to belonging to the k^{th} class, that is, the probability that the observation belongs to the k^{th} class, given the predictor value for that observation [156]. The prior probability, π_k, helps accounting for imbalances in class sizes [156]. If one class has significantly more samples than another, it might have a greater influence on the classification. By incorporating prior probabilities, LDA can balance the impact of different classes. The density function, on the other hand, represents the distribution of the predictor variables for observations belonging to the k^{th} class and thus helps to quantify how likely a particular set of predictor values is for the k^{th} class [177]. The larger $fk(x)$ the higher the probability that a data point belongs to the k^{th}

[11] The Bayes classifier is a simple classifier that, assigns each observation to the most likely class. Put differently, Bayes assigns an observation to the class for which: $P(Y=k|X;=x)$ is largest, where k is one of distinct classes. In a two-class problem, where there are only two possible response values, the Bayes classifier predicts a given class if $P(Y=k|X=x)>0.5$ and another otherwise

class. The estimation of π_k is given by the fraction of the training observations that belong to the k^{th} class:

$$\hat{\pi}_k = \frac{n_k}{n}.$$

The estimating of the density function $f_k(x)$ on the other hand can be done by approximation to the Bayes classifier. LDA does this based on the assumption that $f_k(X)$ is a multivariate Gaussian distribution in which each individual predictor follows a one-dimensional normal distribution with some correlation between each pair of predictors [156]. Formally, the multivariate Gaussian density is defined as:

$$f(x) = \frac{1}{2\pi^{\frac{p}{2}} |\sum|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T \sum^{-1} (x-\mu)\right),$$

where μ is the mean value of X and $\sum$ is the covariance matrix of X. Furthermore, LDA assumes that the observations in the k^{th} class are drawn from a multivariate Gaussian distribution $N(\mu_k, \sum)$, where μ_k is a class-specific mean vector, and $\sum$ is a covariance matrix that is common to all K classes [156, 177]. Plugging the density function for the k^{th} class, $f_k(X = x)$, into equation (7) and performing a little bit of algebra reveals that the Bayes classifier assigns an observation $X = x$ to the class for which

$$\delta_k(x) = -\frac{1}{2}(x-\mu_k)^T \sum^{-1} (x-\mu_k) + log\pi_k \tag{8}$$

is largest, where $\delta_k(x)$ is the discriminant function for the k^{th} component of a Gaussian mixture model. For a detailed explanation of how the above equation was derived, please refer to the supplementary materials [156, 177]. However, even if one is certain that X is drawn from a Gaussian distribution within each class, to apply the Bayes classifier one still must estimate the parameters μ_1,..., μ_k,π_1,...π_k and σ^2. These are estimated using:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\sigma^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

where n is the total number of training observations, and n_k is the number of training observations in the k^{th} class. The estimate for μ_k is simply the average of all the training observations from the k^{th} class, while σ^2 represents the weighted average of the sample variances for each of the K classes [156]. To assign a new observation $X = x$, LDA plugs these estimates int equation (8) to obtain quantities $\hat{\delta}_k(x)$ and classifies the estimates for which $\hat{\delta}_k(x)$ is largest. The fact that the discriminant function $\delta_k(x)$ is a linear function of x, that is, that the LDA decision rule depends on X only through a linear combination of

its elements, explains why this model is called Linear Discriminant Analysis.

2.2.6 Quadratic Discriminant analysis

Quadratic Discriminant Analysis (QDA), similarly to LDA, results from assuming that the observations from each class are drawn from a Gaussian distribution and plugging estimates for the parameters into Bayes' theorem in order to perform prediction. However, unlike LDA, which assumes a covariance matrix that is common to all K classes, QDA assumes that each class has its own covariance matrix [156]. That is, it assumes that an observation from the k^{th} class is of the form $X \frown N(\mu_k, \sum_k)$, where $\sum_k$ is a covariance matrix for the k^{th} class. Under this assumption, the Bayes classifier assigns an observation $X = x$ to the class for which

$$\begin{aligned} \delta_k(x) &= -\frac{1}{2}(x-\mu_k)^T \sum_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\Sigma_k| + \log\pi_k \\ &= -\frac{1}{2}x^T \sum_k^{-1} x + x^T \sum_k^{-1}\mu_k - \frac{1}{2}\mu_k^T \sum_k^{-1}\mu_k - \frac{1}{2}\log|\Sigma_k| + \log\pi_k \end{aligned}$$

is largest [162]. Therefore QDA results from plugging the estimates for $\sum_k$, μ_k, and π_k into (9), and then assigning an observation $X = x$ to the class for which this quantity is largest. This times, since x appears as a quadratic function of the decision rule, $\delta_k(x)$ the model is labeled Quadratic Discriminant Analysis. By not assuming linearity QDA becomes much more flexible than LDA, having substantially higher variance[12], which can potentially lead to improved prediction performance [156].

2.2.7 Random Forest

Random Forest (RF) is an ensemble method[13] that relies on decision trees, which are then combined to yield a single consensus prediction [178] (figure 9). Classification Trees or Decision Trees are Machine Learning algorithms that rely on the recursive segmentation of the predictor space into distinct non-overlapping regions which are then used to assign a class label to each observation, based on the majority class of the instances that end up in the same region [156]. To construct such regions, Classification Trees use a top-down, greedy approach known as recursive binary splitting. In such approach, all predictors are

[12] Variance is defined as the amount by which the performed of a model changes by using a slightly different dataset to fit the model. Models with high variance, are very flexible, and can fit the data closely, even to the extent of capturing noise. In this sense, if a method has high variance, small changes in the data can result in large changes in a model's predictions. This leads the model to perform poorly on new data because the model has essentially memorized the training set and does not generalize well to new data leading to an increased classification error.

[13] Ensemble methods are approaches that combine multiple simple *building block* models with lower performances to obtain a single powerful model. The idea is to leverage the strengths of each different *building block* model and mitigate their individual weaknesses, ultimately creating a more accurate and robust model

initially located in a single region (at the top of the tree). Then, the predictors are spitted via two new branches further down on the tree, where in each split the predictor and respective cut point leading to the greatest possible reduction in the classification error rate are selected [156]. In this sense, for each split in a tree, all predictors and all possible values of the cut points for each of the predictors are considered. Then, the predictor and cut point that lead to the tree with the lowest classification error rate are chosen. This process is then successively repeated, where the best predictor and cut point for splitting the data further are searched once again so that the classification error rate is minimized within each of the resulting regions. However, instead of splitting the entire predictor space, this time, the split only considers one of the two previously identified regions. The process continues until a stopping criterion is reached (e.g. until no region contains more than five observations) [156]. Once the different regions have been created, the response for a given test observation can be predicted according to the most commonly occurring class or mode response value for the training observations in the region to which that test observation belongs. Together, these splitting rules used do segment the predictor space can be summarized in a tree and thus this approach is know as decision trees.

Decision trees, however, suffer from high variance [156]. Bootstrap aggregation[14], also known as bagging, is a technique for reducing the variance of an estimated prediction function [156, 162]. While bagging can be used to improve the prediction of any statistical method it seems to work especially well for high-variance procedures, such as trees. To better illustrate the need of Bootstrap aggregation let's use an illustration [156]. Consider a given set of n independent samples Z_1,...,Z_n, each with variance σ^2, where the mean variance of the observations ($\overline{Z}$) is given by $\frac{\sigma^2}{n}$. In this sense, averaging a set of observations reduces variance. Hence a natural way to reduce the variance and increase the test set accuracy of a statistical learning method would be to take many datasets sets from the population, build a separate prediction model using each set, and average the resulting predictions [156]. However, generally, one does not have access to multiple data sets. Instead, one can bootstrap by taking repeated samples from a single data set. Therefore one can generate B different bootstrapped training data sets and then fit the model on the b^{th} bootstrapped set. Finally, for a given observation, one can record the class predicted by each of the B trees, and take a majority vote where the overall prediction is the most commonly occurring majority vote class among the B predictions [156, 162], which can be translated by the following formula:

[14] Bootsrap aggregation is a powerful method used to simulate new data by repeatedly predicting the error on a distinct subsets of a single dataset. The idea is to randomly draw (randomly resample) datasets with replacement from the training data in order to produce bootstrap data sets. Replacement meaning that the same observation can occur more than once in the bootstrap data sets. Usually, this process is repeated B times for some large value of B (say B = 1000), producing B bootstrap datasets with the model being refitted to each of the bootstrap datasets. Each time a model is refitted one can estimate the error on the bootstrap data sets by tracking how well it predicts the original training set.

$$\hat{y}(x) = \text{argmax}_y \left(\sum_{b=1}^{B} (\hat{y}_b(x) = y) \right)$$

where $\hat{y}_b(x)$ is the predicted class of the b^{th} decision tree.

This process is the backbone of Random Forests, nonetheless, Random Forests provide an improvement over banging by introducing a small tweak that decorrelates the trees [162, 178]. As in bagging, many decision trees are built on bootstrapped samples. However, each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors ($m < p$) [162, 178]. Then, from these m predictors, the split is only allowed to use one of them. Therefore a fresh sample of m predictors is taken at each new split. Typically, the number of predictors considered at each split is approximately equal to the square root of the total number of predictors ()$m = \sqrt{p}$) [156]. This, renders Random Forest to be unable to consider a majority of the available predictors at each split in the tree [162, 178]. The rationale behind this principle can be highlighted by the following example. Suppose that in a given dataset there is one very strong predictor along with several other moderately strong predictors. If one used bagging, most if not all of the trees would use this strong predictor in the top split [156]. Consequently, all of the bagged trees will look quite similar to each other. Hence the predictions from the bagged trees would be highly correlated. Averaging many highly correlated quantities does not lead to as large of a reduction in variance as averaging many uncorrelated quantities. In particular, this means that bagging will not lead to a substantial reduction in variance over a single tree in this setting. Random Forests, on the other hand, overcome this problem by forcing each split to consider only a subset of the predictors [178, 156]. Therefore, on average $\frac{p-m}{p}$ of the splits will not even consider the strong predictor, and so other predictors will have more of a chance [156]. This process of decorrelating makes the resulting trees less variable and hence more reliable. In this sense, the main difference between bagging and Random Forests is the choice of predictor subset size m [156].

The error rate can be then determined using the *out-of-bag* (OOB) observations, avoiding the need for any additional resampling method (i.e. cross-validation) [178, 156]. When performing bagging, each tree in the forest is trained on a bootstrap sample, meaning that not all observations from the original training dataset are included in such a bootstrapped set [178]. Given that when drawing a random sample with replacement, each observation has approximately a 1 - $\frac{1}{e}$ probability of being included in the sample, usually, each bagged tree makes use of only around two-thirds of the observations [156]. The remaining one-third of the observations not used to fit a given bagged tree are referred to as the OOB observations, which can then be used to predict the response for the i^{th} observation using each of the trees in which that observation was OOB [156]. In this sense, for each observation the Random Forest averages only those trees corresponding to bootstrap samples in which z_i

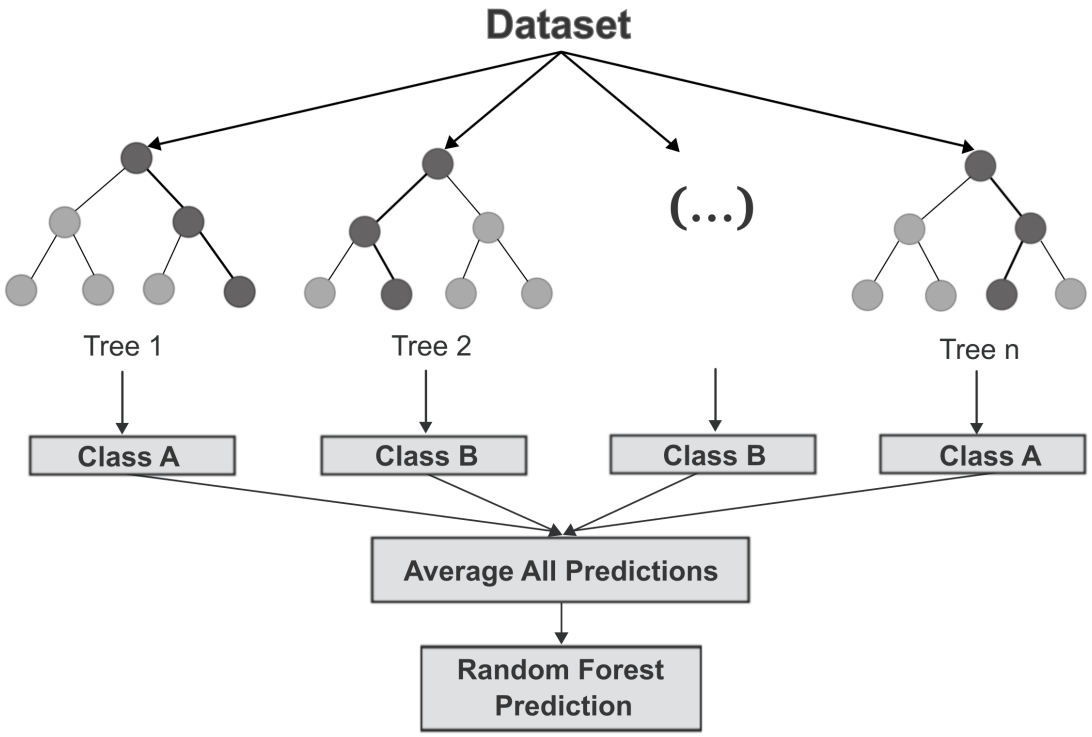


Figure 9: **Schematic representation of the Random Forest algorithm.** Several decision trees are constructed on bootstrapped samples and a class is predicted by each tree. The predictions across all decision trees are averaged and, finally, an overall prediction is made. Source: [157].

did not appear. The single prediction for the i^{th} observation is finally obtained by majority voting (most frequent class), leading to a single OOB prediction [156, 162]. The resulting OOB error is a valid estimate of the test error for the bagged model since the response for each observation is predicted using only the trees that were not fit using that observation. The OOB samples can also be used to determine variable importance. At each b^{th} grown, the OOB samples are passed down the tree, and the prediction accuracy is recorded [156]. Then the values for the j^{th} variable are randomly permuted and the accuracy is once again computed. The decrease in accuracy, as a result of this permuting, is averaged over all B trees and can be used as a measure of the importance for the j^{th} variable in the Random Forest [156]. In the context of bagging Classification Trees, variable importance can also be computed using the mean decrease in the Gini index, averaged over all trees, which I will address ahead.

Earlier I mentioned that the number of predictors to be considered at each split while assembling a Random Forest is approximately $m = \sqrt{p}$. However this value, also known as an hyperparameter[15] can be changed to a value one desires to establish before implementation of the Random Forest model [179]. Besides the number of predictors, several other

[15]A hyperparameter is a configuration setting external to a model that influences its implementation and performance and can be manually tuned by the practitioner

hyperparameters can be manually tuned. Two major ones are the number of trees and the minimal node size. The number of trees determines the total number of decision trees to be built in the ensemble [179, 162]. A higher number often leads to more robust models but may increase computational cost. The minimal node size, on he other hand, sets the minimum number of samples required to be present in a leaf node during the tree-building process [179, 162].

Gini index

Although the classification error rate can be used as a criterion for making the binary splits, such metric is not sufficiently sensitive for tree-growing, and thus other measures such as the Gini index and entropy are preferable [156, 162]. The first, here implemented, is often used instead of the classification error rate and thus I will focus on the same. The Gini index is defined by:

$$G = \sum_{k=1}^{K} (\hat{p}_{mk}(1-\hat{p}_{mk}),$$

where $\hat{p}_{mk}$ represents the proportion of the training observations in the m^{th} region that belongs to the k^{th} class. The Gini index is a measure of total variance across the K classes, taking on small values if all of the $\hat{p}_{mk}$'s are close to 0 or 1 [156, 162]. For this reason the Gini index is also referred to as a measure of node purity, in which a small value indicates that a node contains predominantly observations from a single class [162].

2.2.8 Extreme Gradient Boosting

Boosting, just like banging is an approach that can improve the predictions resulting from several statistical learning such as decision trees [156]. Nevertheless, unlike bagging, where separate decision trees are fitted to each bootstrapped copy, in boosting, trees are grown sequentially, that is, each tree is grown using information from previously grown trees [180, 156, 162]. For this, shallow decision trees[16] are fitted with each observation being initially given equal weights [180, 156]. After each iteration, the weights are adjusted based on the performance of Classification Trees. Misclassified observations receive higher weights, making them more influential in the next iteration. This focuses the attention of the subsequent Classification Trees on the misclassified instances [180, 162]. Thus, unlike in bagging, the construction of each tree depends strongly on the trees that have already been grown [156]. Finally, the predictions from each Classification Tree are combined and all predictions are determined by a weighted majority vote, where worst-performing

[16] A shallow tree refers to a decision tree with a limited depth or number of levels, which tends to capture simpler relationships in the data. Given that, in boosting, the growth of a particular tree takes into account other trees that have already been grown, smaller trees are typically sufficient to contribute to the overall model's predictive power.

classification trees are given a lower weight and better-performing learners receive higher weights.

Some of the most popular boosting algorithms are *Adaptive Boosting* [181] (AdaBoost), *Light Gradient Boosting Machine* [182] (LightGBM) and *Extreme Gradient Boosting* [183] (XGBoost). Here I will focus on the latter, which is an optimized version of *Gradient Boosting*. XGBoost is a powerful and efficient algorithm designed to provide high performance, scalability, and flexibility for a variety of machine-learning tasks, including classification. Besides its speed and effectiveness XGBoost supports parallelization of tree construction, making it highly suitable for large datasets [183]. All these attributes have rendered XGBoost to become a popular tool. This algorithm is based on the gradient boosting framework, which builds a series of sequential decision trees, where each new learner corrects the errors made by the ensemble of existing learners [183]. However, XGBoost introduces regularization terms to the objective function, which helps to control the complexity of the model and prevents overfitting. The way XGBoost works is as follow: if I have a database that has m features and n number of examples, given by $database = \{(x_i, y_i) : i = 1...n, x_i \in \mathbb{R}^m, Y_i \in \mathbb{R}\}$, let $\hat{y}_i$ be the predicted output of an ensemble tree model generated from the following equations:

$$\mathring{A}._i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathscr{F}$$

where $\mathring{A}._i$ is the output of the i^{th} observation, $\phi(x_i)$ is the prediction for the i^{th} observation obtained though the function ϕ applied to the input x_i, $\sum_{k=1}^{K} f_k(x_i)$ is the prediction modeled as the sum K (number of trees) in the model $f_k(x_i)$ each belonging to the function space $\mathscr{F}$ [183]. To solve the above equation, one needs to find the best set of functions by minimizing the loss function and the regularization terms:

$$\mathscr{L} = \sum_i l(y_i, \mathring{A}._i) + \sum_k \Omega(f_k)$$

where l represents loss[17], the loss function which is the difference between the predicted output $\hat{Y}_i$, denoted by $\mathring{A}._i$ in the above formula and the actual outcome y_i. While Ω represents the regularization term that penalizes certain properties of the individual functions f_k, promoting the simplicity of the model and assists in avoiding over-fitting of the model and it is calculated using:

[17] The loss, more commonly known as log loss or cross entropy, just like the classification error rate, is a as a metric to evaluate the performance of a classifier. The term loss represents the discrepancy or error in the model's predictions. When the model's predictions align well with the actual values, the loss is low. The Log Loss is given by: $l = \frac{1}{N}\sum_{i=1}^{N}[y_i.log(p_i)+(1-y_i).log(1-p_i)]$, where N is the number of observations in the dataset, y_i is the true labeled instances (0 or 1) and p_i is the predicted probability that instance i belongs to the positive class (usually denoted as the class = 1) and $p_i = \frac{1}{1+e^{\alpha+x_i\beta}}$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda ||w||^2$$

where T, in the above equation represents the number of leaves of the tree and w is the weight of each leaf [183].

Finally, the term *Gradient* in XGBoost refers to the use of gradient descent optimization during the training process, where the objective is to minimize the loss function that measures the difference between the predicted and actual classes (i.e classification error rate) [180]. In this sense, XGBoost uses gradient descent to find the optimal parameters for the model by iteratively moving towards the minimum of the loss function.

Similar to the Random Forest, XGBoost also provides a measure of feature importance, allowing to understand which features contribute the most to the model's predictions [156]. Furthermore, just like the Random Forest, one can adjust the hyperparameters in XGBoost, such as the number of tress and the number of splits in each tree. Although, additionally, in XGBoost the shrinkage parameter λ, can also be defined by the user, which will control the rate at which the boosting learns.

2.2.9 Super Learner

The Super Learner (SL) is a prediction method designed to find the optimal combination of a collection of candidate prediction algorithms [184]. The Super Learner framework is built on the theory of cross-validation and allows for a set of prediction algorithms to be considered for an ensemble, where the model then attributes weights for each candidate learner in order to find the combination of algorithms that minimize the cross-validated error [185] (Figure 10). In the context of prediction, this learner is itself a prediction algorithm, which applies a set of candidate learners to the observed data, and chooses the optimal learner for a given prediction problem based on cross-validated risk. The Super Learner for perdition works as follows: for a dataset $X_i = (Y_i, W_i), i = 1, ..., n$ where Y is the outcome of interest and W is a p-dimensional set of covariant, the objective is to estimate the function $\psi_0(W) = E(Y|W)$ [184]. Such function can be expressed as the minimizer of the expected loss:

$$\psi_0(W) = arg\min_{\psi} E[L(X, \psi(W))]$$

where the loss function is the log loss, earlier mentioned. For a given problem, a library or set of prediction algorithms that minimizes the loss function ($\mathcal{L}$) can be proposed. Each algorithm is then fitted to the entire data set $X = X_{i:i=1}, ..., n$ to estimate $\hat{\Psi}_k(W), k = 1, ..., K(n)$. The data is then split into a training an validation subsets, according to a V-fold cross-validation, where the n observations are split into V equal size groups [184]. The V^{th} group is then defined as the validation $V(v)$ subset, while the remaining group, defined as $T(v)$ represents the training subset, $v = 1, ..., V$. For each v^{th} fold, each algorithm in $\mathcal{L}$

is fit on the training subset $T(v)$ and the predictions are saved on the corresponding validation data, $\hat{\Psi}_{k,T(v)}(W_i), X_i \in V(v)$ for $v = 1,...,V$. The predictions from each algorithm are then stacked together $Z = \hat{\Psi}_{k,T(v)}(Wi), X_i \in V(v)$ & $k = 1,...,K$. A family of weighted combinations of the candidate estimators indexed by weighted-vector α is proposed:

$$m(z|\alpha) = \sum_{k=1}^{K} \alpha_k \hat{\Psi}_{k,T(v)}(W_i), \alpha_k \geq 0 \forall_k, \sum_{k=1}^{K} \alpha_k = 1.$$

The α that minimizes the cross-validation error of the candidate estimator $\sum_{k=1}^{K} \alpha_k \hat{\Psi}_k$ over all allowed α-combinations is determined:

$$\hat{\alpha} = \min_{\alpha} \sum_{i=1}^{n} Loss(Y_i, P(y_i|z_i, \alpha))$$

where Y_i represents the true class label for the i^{th} instance, $P(y_i|z_i, \alpha))$ is the predicted probability in the i^{th} instance belonging to the y class given the input features $z_i$and parameter values α and $Loss$ is the chosen loss function that quantifies the difference between the true class label and the predicted probability such as the cross-entropy loss [184]. Finally, $\hat{\alpha}$ is combined with $\hat{\Psi}_k(W), k = 1,....,K$ according to the familiy m(z|α) of weighted combinations to create the final Super Learner fit:

$$\hat{\Psi}_{SL}(W) = \sum_{k=1}^{K} \hat{\alpha}_k \hat{\Psi}_k(W)$$

Theoretical, the Super Learner will perform asymptotically as well as, or better than any of the candidate learners.

2.3 Assessing model performance

The SL makes use of a procedure that splits the data into a train and validation subsets in order to evaluate the performance of the candidate models. To understand the underlying idea behind this procedure, on should recall that the central objective when conducting predictive analysis is to predict future events [122]. In this sense, one is not particularly interested in understanding how well a model performed on the data it was trained on, rather the primary focus is to evaluate the model's performance on new previously unseen data. However, as earlier mentioned in section 2.2.7 new data is often not available. This represents a significant problem as a model's performance on the data it is used to trained on, can be quite different from the one when implemented on new data, and in particular the former can dramatically overestimate the latter [156]. Therefore, in the absence of a designated test set on which one can directly estimate the model's performance, alternative

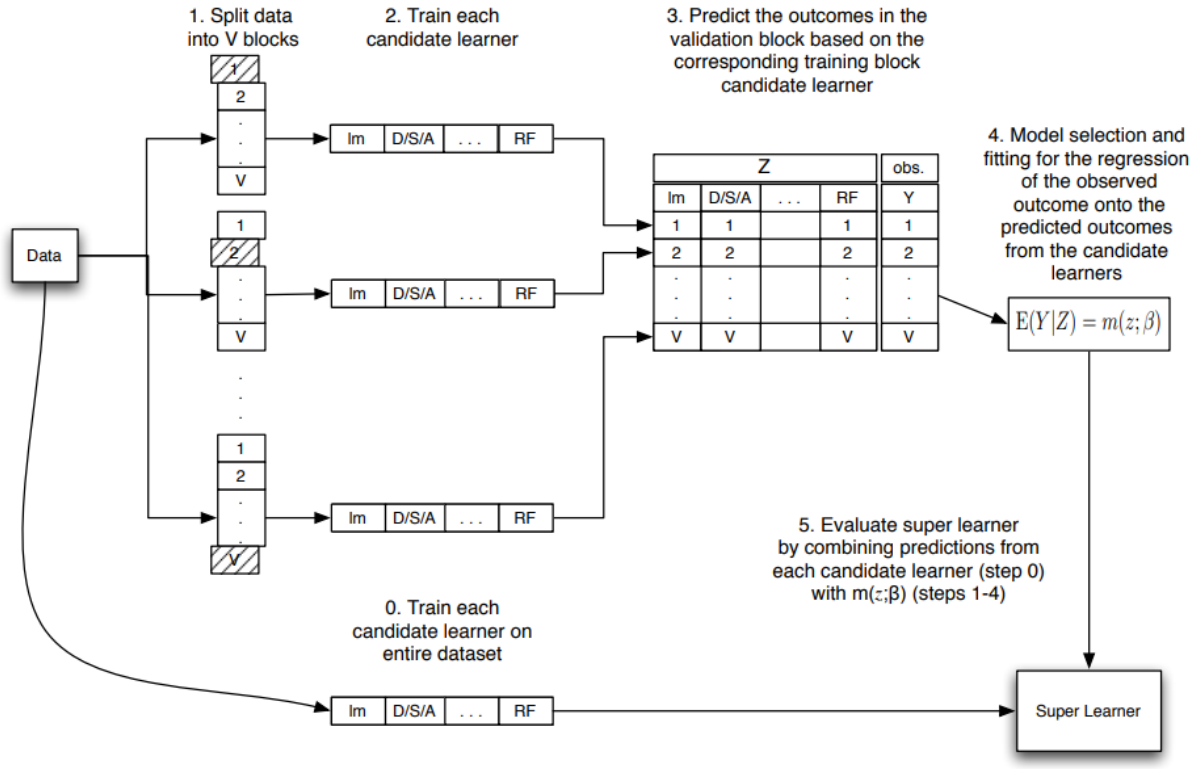


Figure 10: **Flow Diagram for Super Learner.** The different steps of the Super Learner implementation are show: Initially each candidate algorithms are trained on the entire dataset. The dataset is then split into V blocks. Each candidate learner is then fitted to the train subset and the outcomes are predicted on the validation subset. Finally, the predictions from each candidate learner are then combined. Source: [185]

techniques to estimate such metrics using the available training data must be used which guides the choice of the most appropriate learning method or model, and gives us a measure of the quality of the ultimately chosen model [156]. This chapter is devoted to describe such techniques, which allow to simulate the estimation of the error rate on new data.

2.3.1 The validation set approach

The simplest method for estimating the classification error rate is the validation set approach. This approach consists of randomly dividing the available set of observations into two sets, a training set and a validation or hold-out set. The model is fitted on the train set and then, used to predict the responses for the observations in the hold-out validation set [156]. Since the hold-out validation set was not used in the fitting process, it's classification error rate provides an approximately unbiased estimate for the test error [156]. Although conceptually simple and easy to implement this approach carries two main drawbacks. Firstly, the validation estimate of the classification error rate can be highly variable, depending precisely on which observations are included in the training set and which observations are included in the validation set [156]. Put differently, if one repeats the process of

randomly splitting the sample set into two parts, it will get a somewhat different estimate for the test classification error rate. Moreover, statistical methods tend to perform worse when trained on fewer observations. In this regard, since the validation approach model is fitted only in the subset of the observations included in the training set, rather than on the entire set, the validation set error rate will tend to overestimate the test error rate [156]. To address these limitations, a more refined approach, termed cross-validation, was introduced.

2.3.2 K-fold Cross Validation

K-fold cross-validation is the most widely used method for estimating the test error rate [162], representing the one implemented by the Super Learner. The k-fold cross-validation (k-fold CV) approach involves randomly dividing the set of observations into k groups, or folds, of approximately equal size [156, 185]. The first fold is treated as a validation set, while the remaining k - 1 folds are used to fit the model [156, 162]. The number of misclassified observations (classification error rate) is then computed on the observations in the held-out validation fold [185]. This procedure is repeated k times, each time, with a different group of observations being treated as the validation set. This process results in k estimates of the test classification error rate [156, 185]. The k-fold CV estimate is computed by averaging these values:

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} Err_i$$

where $Err_i = I(y_i \neq \hat{y}_i)$. K-fold CV is typically performed using $k = 5$ or $k = 10$ [156, 162]. A special case of k-fold cross-validation is the Leave-one-out cross-validation (LOOCV) in which $k = n$ [156]. In this approach, a single observation is used for the validation set, while the remaining observations make up the training set. Thus, the statistical learning method is fitted on the $n-1$ training observations, and the classification error rate is determined on the excluded observation. This procedure is repeated n times and the classification error rate is obtained based on the average n error rates:

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} Err_i$$

While LOOCV has a couple of major advantages over the validation set it also has a major drawback. The advantage of such approach over the validation set is that it has far less bias[18] [156, 186]. In LOOCV the statistical learning method is repeatedly using training sets

[18] Bias refers to the error introduced by approximating a complex real-world problem by a simplified model, which may not be flexible enough to capture the true complexity of the relationships present in the data. Linear models, which assume a linear relationship between the predictors X and the response variable y are a good example of models with high bias. Given that real-life problems are highly unlikely to have such a

that contain $n-1$ observations, almost as many as are in the entire dataset [186]. This contrasts with the validation set approach, in which the training set is typically around half the size of the original dataset, leading LOOCV not to overestimate the test error rate as much as the validation set approach does [156]. Furthermore, unlike the validation approach which yields different results when repeatedly applied due to randomness in the training/validation set splits, repeat implementation of LOOCV will always yield similar results since there is no randomness in the training/validation set splits [156]. LOOCV, however, can be very expensive[19] to implement [156]. Specially if a model is slow to fit, LOOCV can be very time-consuming. This is where the computational advantage of using $k=5$ or $k=10$ rather than $k=n$ steps in. Given that some statistical learning methods have computationally intensive fitting procedures, performing LOOCV may pose computational problems, especially if n is extremely large. Alternatively, performing 10-fold CV requires fitting the learning procedure only ten times, which may be much more feasible. Apart from the computational advantage, k-fold CV offers another upside, in that it often gives more accurate estimates of the test error rate than LOOCV [162]. This has to do with a bias-variance trade-off [20]. Earlier we saw that the validation set approach tends to overestimate the test error rate because the model statistical learning method is trained on only half the observations of the entire dataset. Applying the same reasoning, it is clear that the LOOCV will give approximately unbiased estimates of the test error, since each training set contains $n-1$ observations, while k-fold CV for, say, $k=5$ or $k=10$ will lead to an intermediate level of bias. So, if one was only to consider bias reduction, LOOCV would be preferred over k-fold CV [156]. However, one needs to consider variance too, and it turns out that LOOCV has a higher variance than does k-fold CV. Given that during the implementation of LOOCV, the models are trained on an almost identical set of observations the outputs provided by the model are highly (positively) correlated with each other [156]. In contrast, when performing k-fold CV, the overlap between the training sets in each model is smaller, thus the average of the outputs of k-fitted models are somewhat less correlated with each other. As the mean of many highly correlated quantities has a higher variance than the mean of many less correlated quantities, the test error estimate resulting from LOOCV has a higher variance than the test error estimate resulting from k-fold CV [156].

simple linear relationship, implementation of such methods will result in some bias. Therefore, highly biased models may lead to underfitting, where the model is too simplistic to capture the underlying patterns in the data leading to an increased error. A low bias model, on the other hand, is more flexible and less constrained by assumptions, thus being able to capture complex relationships in the data

[19] The term *expensive* in the context of Machine Learning, implies that implementing a particular model may require significant investment, be it financial, time, or resource-related.

[20] The bias-variance trade-off concept relies on finding the right balance between the bias and the variance. As I have already mentioned, bias occurs when a model is too simple and fails to capture the underlying patterns in the data. Highly bias models can lead to poor performances, translated by an underfitting of the model. Variance, on the other hand occurs when a model is too complex and captures noise or random fluctuations in the training data. High variance can result in a poor generalization to new, unseen data, translated as overfitting. Thus, the bias-variance trade-off involves finding or adjusting the model to minimize both bias and variance which is crucial for creating a model that generalizes well to new data.

These strengths often render k-fold CV to be preferred over LOOCV.

Chapter 3

Background knowledge on the studied diseases

3.1 Malaria

Malaria is a parasitic disease that has been afflicting the human population over the millennium, and continues to be a major health problem [187]. Despite continuous worldwide efforts and investments, malaria is still the principal cause of disease and death caused by a parasite [188]. According to the World Health Organization's (WHO) latest report, in the year of 2022, there were an estimated 246 million malaria cases worldwide and a total of 608 000 deaths estimated globally [187]. Although relatively uncommon in developed countries, malaria still represents a significant social and economic burden in developing tropical and sub-tropical, countries where it is endemic [189]. Africa is the most affected region, where the poor show the highest morbidity and mortality rates [187]. The term malaria originates from the 18^{th} century Italian expression mala aria, meaning "bad air", referring to the foul air evaporating from stagnant waters of marshes that used to be thought as the origin of the disease [189]. It wasn't until 1897 that Ronald Ross identified the real causative agent of malaria in a mosquito that had previously fed on an infected patient [190]. This discovery provided the first step in understanding of the parasites' life cycle of development and transmission and laid the foundation for the development or more specialized methods for malaria treatment and control.

Malaria is caused by infections with protozoan parasites of the genus *Plasmodium*, unicellular eukaryotic microorganisms that invade and live in a host, from which they derive nourishment and shelter [189, 191]. These microorganisms have a 23-megabase nuclear genome consisting of 14 chromosomes, which encodes about 5300 genes from which a large proportion are devoted to immune evasion and host–parasite interactions [192]. Currently, from more than 200 *Plasmodium* species known to infect various species of vertebrates, only 5 infect humans: *falciparum*, *vivax*, *ovale*, *malariae* and *knowlesi* [191]. The first four species infect exclusively humans, while *knowlesi* is naturally maintained in maca ques and monkeys and causes zoonotic malaria [21] [191]. Of all, P. *falciparum*is the most prevalent, being the principal cause of malaria morbidity and mortality [187], specially in sub-Saharan Africa.

The main carriers of malaria parasites known to affect humans are some species and subspecies of mosquitoes belonging to the genus *Anopheles*. More precisely female night-biting mosquitoes which feed on blood meals to support their development [191]. Of the more than 400 species of *Anopheles* mosquitoes that have been described, only about 30 or

[21] Zoonotic malaria refers to malaria that can be transmitted between animals and humans.

40 species commonly transmit human malaria [193]. The *Anopheles* mosquitoes, however, are not just the vector for malaria, they are the definitive host, where sexual reproduction of the parasite occurs, an essential step for its life cycle [191].

The malaria parasite's life cycle is highly complex and involves two hosts, a vertebrate and an a invertebrate host [189]. During a blood meal, a malaria-infected female *Anopheles* mosquito inoculates sporozoites into the bloodstream of the vertebrate host (Figure 11(1)). Sporozoites then travel to the liver where they infect hepatocyte [194]. This stage, characterized by the invasion of hepatic cells, termed hepatocytes, is known as the pre-erythrocytic stage (Figure 11(2)). Within the hepatocytes, the sporozoites' period of persistence varies between the *Palsmodium* species. Nevertheless, all species develop into the primary or pre-erythrocytic form named schizont (Figure 11(3)). These pre-erythrocytic forms begin dividing repeatedly, resulting in a large number of exoerythrocytic merozoites and at the end of the pre-erythrocytic phase, thousands of these merozoites are released into the circulation to invade red blood cells (RBC) [189] (Figure 11(4)). Although many merozoites are destroyed by the immune system, some immediately invade RBCs triggering the erythrocyte phase of asexual reproduction called schizogony (Figure 11(5)). For this invasion to occur, the merozoites must recognize specific surface receptors on the RBC membrane. The membranes of the RBC and merozoite fuse, and the parasites become vacuolized inside the RBC where they continue to grow feeding on haemoglobin [195]. The earliest parasite forms in the red cell are ring form trophozoites that will develop into mature schizonts. At this stage serial cycles of asexual replication produce rising parasite numbers. After 48 to 72 hours a single trophozoite will produce about 4 to 36 merozoites and when parasitemia reaches a threshold of roughly 100 parasites per microliter (μL) clinical manifestations of the disease occur [189]. As the nucleus begins to divide, trophozoites become to be called a developing schizont. The rupture of the erythrocytes at the end of schizogony frees the merozoites to invade new red cells and perpetuates the host infection (Figure 11(6)). Finally, a subpopulation of the merozoites switches to sexual development producing female (macrogametocytes) and male (microgametocytes) gametocytes [189] (Figure 11(7)).

These gametoccytes are then ingested by feeding mosquitoes during a new blood meal (Figure 11(8)). These gametocytes then travel through the mosquitoe's bloodstream and once in the mosquitoe's stomach, the male and female gametocytes develop into gametes [196] (Figure 11(9)). There, the female gametocytes undergo a process of maturation called exflagellation and the microgametes divide into a four to eight nuclei, each of which grows a flagella [189]. Their motility allows them to rupture out of the confines of the gametocyte membrane to fertilize the product of the female gametocyte, called macrogamete. A zygote is thus formed by the sexual fusion of the macrogamete and the microgamete. Twelve to twenty-four hours later, the zygote elongates and becomes motile being at this stage termed ookinete (Figure 11(10)) which then travels to the mosquito's midgut outer wall to form an oocyst, dividing into thousands of spindle-shaped sporozoites [196, 189]

(Figure 11(11)). All this process, from which gametocytes mature into sporozoites is called sporogony, and depending on the species of plasmodium, can take from 8 to 35 days. The sporozoites are then released into the mosquito's bloodstream and migrate throughout the mosquitoe's body to the salivary glands (Figure 11(12)). Inoculation of the sporozoites into a new human host perpetuates the malaria life cycle [196].

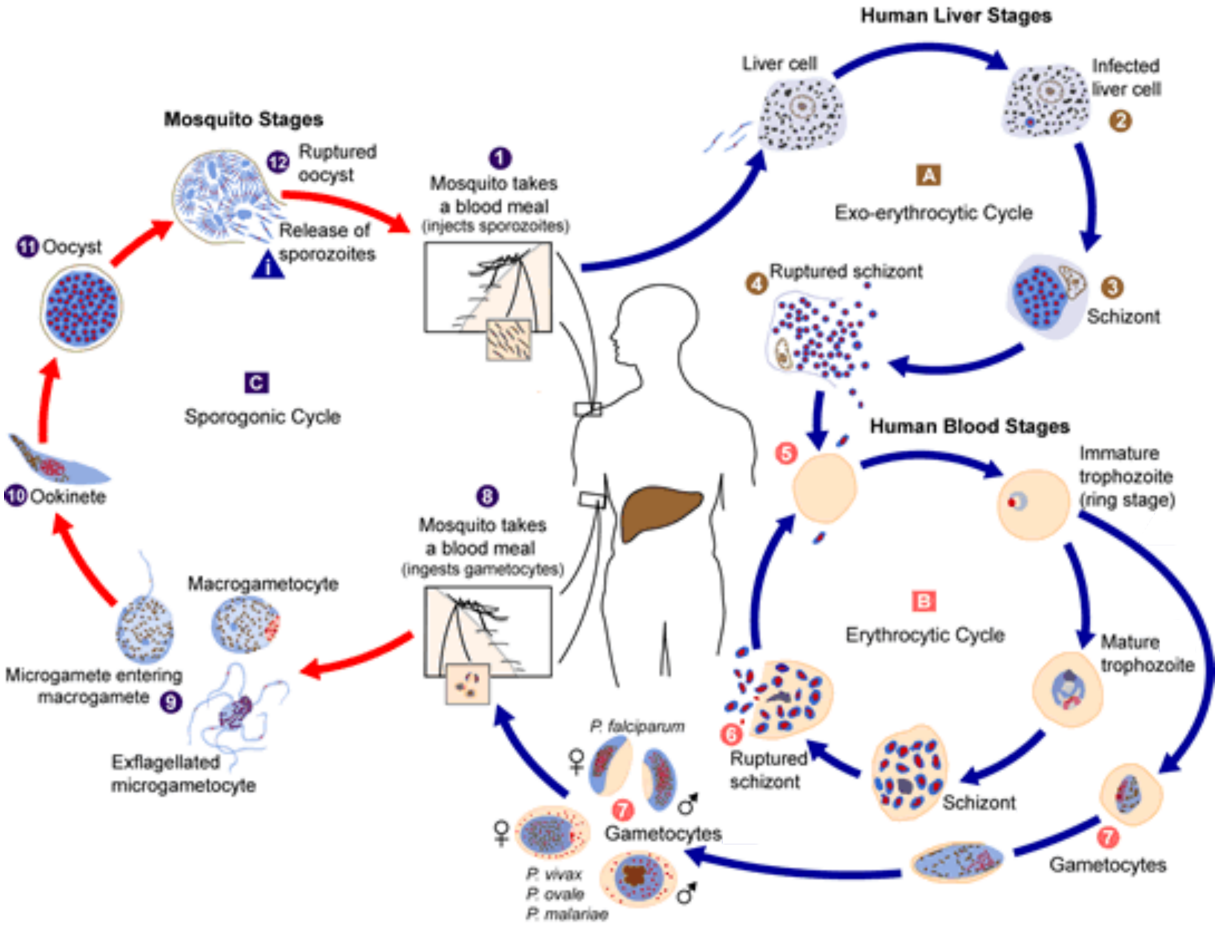


Figure 11: **Malaria parasite life cycle.** The different stages of the malaria parasite are shown with different numbers and letters. Source: adapted form https://www.cdc.gov/malaria/about/biology/index.html.

3.1.1 Clinical manifestation

From the time of the mosquito bite until approximately a week later, patients remain asymptomatic. This because parasites are undergoing multiplication in the liver [197]. It isn't until a later stage, when merozoites within the RBCs reach a high enough parasitemia level and are released to the bloodstream, that symptoms start to be felt [189]. In malaria-endemic regions, the clinical presentation of malaria is mostly seen in children bellow the age of 5 years or in pregnant women in their first gestation [198].

Malaria symptoms are very non-specific, including splenomegaly, hepatomegaly, jaun-

dice (yellowish pigmentation of the skin and sclera) paroxysm, fever[22], body aches, myalgias (muscle pains), nausea, and diarrhea [199, 200]. Another classical symptom of malaria is paroxysms which comprise three successive stages [199]. The first is characterized by a cold stage, where patients experience shivering and a feeling of cold. In the next stage, the hot stage, patients have high fevers, sometimes reaching 41°C, coupled with dry skin, headaches, nausea, and vomiting. Finally, there is the sweating stage during which the fever drops rapidly and patients experience diaphoresis (excessive sweating) [199].

If untreated, malaria patients can experience more sever symptoms, which include cerebral malaria, multi-organ failure, acute lung injuries, acute kidney injury and severe anemia [201, 200]. Cerebral malaria is one of the most serious symptoms of severe malaria and it is characterized by seizures/convulsions, confusion, delirium, hyperpyrexia (very high fever), cerebellar dysfunction, and coma, the most severe manifestation [201, 200].

Finally, it is interesting to highlight that although clinical manifestations are mostly observed in children, infants between 0 and 1 year of age have a very low incidence of disease [202]. After that age, incidence increases until the third year of age, gradually declining past that point. These observations correlate with the year of natural immunity in infancy, followed by a period of relative immune susceptibility prior to the slow development of acquired immunity [203, 204].

3.1.2 Acquired immunity to malaria

Acquired immunity to malaria is a process by which, individuals living in endemic areas, consistently exposed to malaria over long periods of time, became resistant to infections with the malaria parasite, rarely experiencing the disease even with blood infections presumed to be lethal to non-immune, malaria-naive visitors [205, 206] (Figure 12). This immunity, also denoted as clinical immunity (i.e., immunity against the symptoms of malaria) is mostly perceived as a limitation of the peak parasitaemia, a reduction in parasite density or decrease in prevalence of disease [206, 207]. However, this immunization as been shown to be species-specific, not sterilizing in which it fails to eliminate the infection completely, leading to persistent low-grade parasitaemia, and relatively short-lived, as maintenance of clinically protective immune responses requires recurrent exposure to the parasite [208]. Differences in antigen load, variability and the expression of the more than 5,000 parasite gene products during the Plasmodium life cycle explain the difficulty that hosts have in inducing sterilizing immunity to P. *falciparum* [208].

Although the exact mechanisms by which natural immunity to malaria is achieved and maintained is still not fully understood, growing evidence suggests that over time, natural infection elicit a robust immune response against the blood stage of the parasite which

[22] Fever in children is an hallmark symptom of malaria in endemic areas, with almost all children infected with malaria experiencing fever

provide protection against malaria [209]. Indeed, recently, antigen-specific immune responses associated with protection against malaria infection have been proposed. Among such responses, the antibodies against the Circumsporozoite Surface Protein (CSP) have been highlighted to confer protection against clinical malaria [210, 211, 212, 213, 214]. Indeed the only available vaccine until recently targeted the CSP protein [215, 216]. The Merozoite Surface Protein 1 (MSP1) [217, 218, 219, 214, 220] and the Apical Membrane Antigen 1 (AMA1) [217, 221, 222, 223] have been extensively cited as protective agents against malaria. The pfEMP1 protein [224, 225, 226, 227] and the Erythrocyte-Binding Antigen 175 (EBA175) [218, 222, 219, 228, 229] are also common antibody signature found to be associated with protection to clinical malaria within the literature. Apart, from these best described antibodies, many others have been linked with protection to clinical malaria. These can be found in the following [222, 119, 230, 24, 40, 116, 231]. However, ongoing efforts are being made to uncover antibody profiles that could grant immunity against clinical malaria, offering new venues for vaccine research and development [229, 214, 220, 224].

3.1.3 Diagnosis

For non-immune individuals, prompt and accurate malaria diagnosis is critical for an effective management of malaria, with delayed diagnosis and treatment being the leading causes of death in many countries [232]. Disease diagnosis can either be made in the clinic on in the laboratory. Clinical diagnosis is based on the patients' signs, symptoms and on physical findings at examination. Nonetheless, malaria clinical diagnosis is still very challenging because of the non-specific nature of the signs and symptoms, which overlap considerably with other common, as well as potentially life-threatening diseases (i.e, common viral or bacterial infections, and other febrile illnesses) [232].

In the laboratory, malaria diagnosis involves identifying the malaria parasites or the parasite's antigens in the patient's blood [232]. To this date, The reference test method for malaria diagnosis is light microscopy of stained blood smears by Giemsa or Wright's, or Field's stains [188, 233]. Its simplicity, low-cost, ability to identify the presence of the parasite together with the infection species and ability ot assess parasite density have led to the wide acceptance of this technique by laboratories around the world [232]. However, processing and interpretation of malaria smears are labor intensive, time consuming and ill-suited for high-throughput use. Furthermore this technique requires appropriate equipment as well as considerable training of healthcare workers, factors that limit their use in many parts of sub-Saharan Africa where malaria is endemic [234]. Notwithstanding the most important shortcoming of microscopic examination is its relatively low sensitivity, particularly at low parasite levels, which may lead inexperienced microscopist to be unable to detect parasitic infections [232].

Rapid diagnostic tests (RDTs), which do not rely on specific equipment and require min-

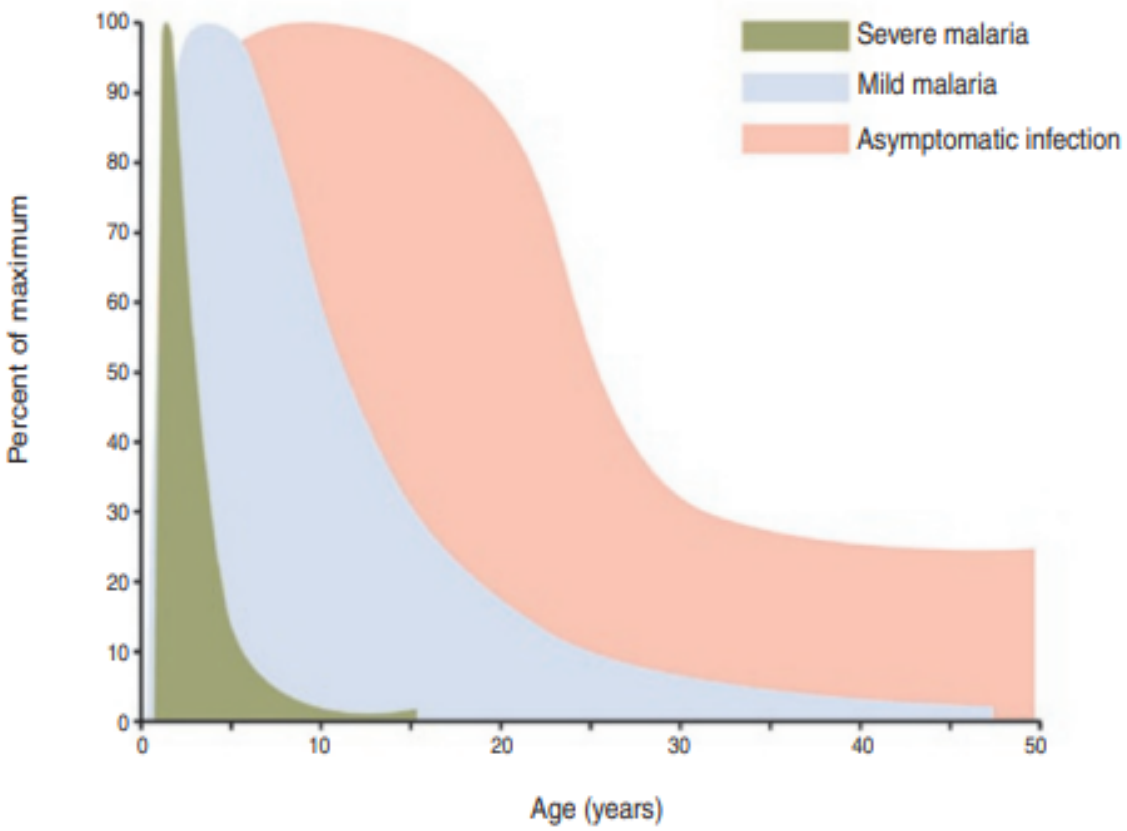


Figure 12: **Natural acquired immunity to the malaria parasite.** Population indices of immunity in an endemic area of P. *falciparum* transmission (adapted from ref. 96). Change over time of various indices of malaria in a population living in an endemic area of P. *falciparum* transmission: asymptomatic infection (pink), mild disease (febrile episodes caused by malaria; blue) and severe or life-threatening disease (green). The data are normalized and are presented as the percent of maximum cases for each population index.

imal skill to perform and interpret, have been developed to overcome the problems of conventional microscopy [232]. Ever since their approval in 2007, RDTs have gained particular prominence in malaria diagnosis, and currently, the WHO recommends their use as the first choice of test all across the world in all malaria-endemic areas [188]. RDTs are small, easy to use devices that detect malaria antigens in a small amount of finger-prick blood samples, by immunochromatographic assay with monoclonal antibodies directed against target parasite antigens impregnated on a test strip [235]. Most available RDTs are manufactured target antigens a P. *falciparum*-specific protein, such as Histidine-Rich Protein II (HRP-II) or Lactate Dehydrogenase (LDH). Although RDTs have been demonstrated to be quite sensitive [236, 237, 238, 239, 240, 241] and thus have represented a highly valuable, rapid malaria-diagnostic tool for healthcare workers, their performance has also been criti-

cized in other publications and occasional false-negative[23] results, specially if parasite density is low, have been evidenced [242, 243, 244, 245, 246, 247]. Thus the use of RDTs should currently be used in conjunction with other methods such as microscopy to confirm the results [232].

Alternatively, to RDTS, the diagnosis of malaria has also been done through serologic testing using either immunofluorescence antibody(IFA) or ELISA testing [232, 248]. Despite being highly sensitive and specific, these test are quite time-consuming, hard to standardize and subjective, requiring qualified technicians [232]. Even thought the literature clearly illustrates their reliability, for malarial serology testing, these assays have been proven to be most useful in epidemiological surveys, for screening potential blood donors, and occasionally for providing evidence of recent infection in non-immune individuals [249, 250, 251, 248, 232].

The problems of traditional malaria diagnostic methods mentioned so far have driven the development of molecular diagnostic techniques that display high sensitivity and high specificity, without subjective variation [252]. The most noteworthy technique is the Polimerase Chain Reaction (PCR), which detects the presence of the parasite's nucleic acid (DNA) in a sample [253]. PCR has been proven to be one of the most specific and sensitive diagnostic methods, particularly for malaria cases with low parasitemia [254] surpassing traditional methods [254]. Although overcoming the major problems of malaria diagnosis such as sensitivity and specificity, the utility of PCR is limited by its complex methodology, high cost, and the need for specially trained technicians [232]. PCR, therefore, is not routinely implemented in developing countries because of the complexity of the testing and the lack of resources to perform these tests adequately and routinely [255]. Quality control and equipment maintenance are also essential for the PCR technique, so it may not be suitable for malaria diagnosis in remote rural areas or even in routine clinical diagnostic settings [256]. Therefore, PCR remains largely an investigation tool often used to confirm malaria species infection after diagnosis by microscopy or a RDT in laboratories that might not have microscopic experts, follow-up therapeutic response and identify drug resistance [257].

Despite the advances in diagnosis strategies, diagnosing malaria is far form being a simple task, as diagnostics techniques are subjective to many factors such as the variety of species, the different stages of the parasite life cycle, differences in the immune responses, the variety of signs an symptoms among other factors that hamper disease identification [232]. Thus, new cost-effective, easy to use efficient and reliable tools for malaria diagnosis are in need. To tackle this challenge, recently, great attention has been given to the identification of novel serological biomarkers against malaria [40, 119, 222, 230, 115, 116, 117], with may help to overcome the current limitation of traditional serological techniques.

[23] A false negative occurs when a test incorrectly indicates the absence of a condition that truly exists. Increase values of false negative are directly correlated with lower sensitivities.

3.1.4 Treatment and Prevention

Malaria treatment has largely relied on antimalarial drugs, which have been developed to target different stages of the Plamodium development. Such drugs have largely relied on the quinine, chloroquine, rimaquine and sulfoncs or sulfonamides in combination with the others compounds [258].

Quinine, the active ingredient in Cinchona bark, isolated in the year of 1820, became the first effective antimalarial compound [259]. Since then, quinine has become an important and effective treatment option for malaria, that continues to play a significant role in the management of malaria [259]. Althought the exact anti-malarial mechanism of action remains unknown, quinine seems to acts on the asexual stages of the malaria parasite by inhibiting its heme polymerase enzyme, thereby inhibiting hemozoin formation, an essential process for the survival of the malaria parasite. Quinine remained the mainstay of malaria treatment until the 1920s, when more effective synthetic anti-malarials became available. The most important of these drugs was chloroquine which became extensively used [260]. Similar to quinine, Chloroquine's antimalarial action suppressed the parastite's asexual stages [258]. Nonetheless, due to its heavy use, chloroquine resistance started to develop slowly. Resistance of Plasmodium *falciparum* to chloroquine was seen in parts of Southeast Asia and South America by the late 1950s, and was widespread in almost all areas with *falciparum* malaria by the 1980s. With increasing resistance to chloroquine, quinine in combination with sulfones and tile sulfonamides played a key role, particularly in the treatment of severe malaria [260]. Primaquine, introduced in 1950, was the first available drug against the hypnozoites within the liver (P. *vivax*, P. *ovale*). More specifically, primaquine eradicated hepatic exoerythrocytic parasites preventing long-term relapses and provided sterilization against the sexual plasmodium, particularly of P. *vivax*. [261, 262]. Although not effective as before, primaquine therapy is still used [262].

Apart from these drugs, artemisinin-based combination therapy (ACT) has also been prominently used [263]. Artemisin was first isolated by chinese scientist in 1972 from the *Artemisia annua* plant, an year after the antimalarial activity such plant's extract was experimentally proved in a primate model [264]. Since that period, a number of semi-synthetic derivatives were developed to improve the drug's pharmacological properties and antimalarial potency [263]. These derivates are highly active against asexual forms of the distinct species of Plasmodium that infect humans [263]. They are also active against the sexual form of the parasites (gametocytes) taken up by the mosquito and can therefore reduce transmission rates [263]. Although their exact mechanism remains unknown it is presumed that endoperoxide moiety, essential for antimalarial activity, may cause destructive free-radical generation within the parasite and, through the formation of covalent bonds, alter the function of key parasite proteins, including membrane transporters [265, 266]. In 2001 the WHO endorsed ACT as the standard treatment for all malaria infections in areas

where P. *falciparum* is the predominant infecting species, and today most malaria endemic countries have now adopted ACT treatments as first-line treatment of *falciparum* malaria [263, 267].

However patients compliance remains a problem in several regions [268, 269, 270]. Not only that, but side effects are often a problem with many drugs which further hampers their use [263, 271, 272, 273]. Finally, the increased resistance of malaria parasites to most of the antimalarial drugs (including the most efficacious artemisinin derivatives) and their declining efficacy has become a great hurdle to malaria management highlighting the need for the development of new therapeutic agents/drug combinations to effectively tackle drug resistance [258]. Ideally, an effective vaccine against malaria would strongly reduce the need for drug administration [215]. Despite tremendous efforts towards the identification of a vaccine against malaria, little success has been achieved.

Until recently, the only available vaccine against malaria, approved by the WHO, was the RTS,S/AS01, which targets the circumsporozoite surface protein (CSP) expressed by the P. *falciparum* parasite at the pre-erythrocytic stage [215]. The sames consists on part of the circumsporozoite sequence fused with a viral envelope surface antigen from the hepatitis B virus (HBsAg) to which the chemical adjuvant (AS01) is added to increase the immune system response [215, 274]. Notwithstanding the great benefits in immunizing against the disease, the RTS,S/AS01 vaccine has shown modest efficacy against malaria illness [215, 275, 216, 274, 276]. Although improvements to the vaccine have been attempted, the same have not shown great benefit [277, 211, 278]. In October of last year (2023), the R21/Matrix-M (R21) pre-erythrocytic malaria vaccine became the second vaccine recommended by WHO to prevent malaria in children living in areas of risk [187]. The R21 is a virus-like particle comprising the central repeats of Asn-Ala-Asn-Pro (NANP) and C-terminal sequence of the circumsporozoite protein fused to the hepatitis B surface antigen (HBsAg) administered with a saponin adjuvant, Matrix-M [279]. So far, this vaccine has shown increased performance over RTS,S/AS01 [279, 280, 279, 281]. Such addition of the R21 malaria vaccine to complement the ongoing use of the first malaria vaccine, RTS,S is now expected to result in major benefits for individuals, specially children, living in areas where malaria is still a major public health problem [187]. Until now, vaccine development has been mostly hampered by the tremendous complexity of the parasite, which has several developmental stages expressing unique sets of stage-specific genes and multiple antigens, and the incomplete understanding of the molecular mechanisms that underlie the interactions between the parasite with its hosts [208].

3.2 Myalgic encephalomyelitis/chronic fatigue syndrome

Myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) is a chronic and debilitating systemic neuro-immunological clinical condition characterized by an unexplained and persistent fatigue not alleviated by rest and post-exertional malaise (PEM) [282]. Additionally, patients also experience a plethora of other symptoms related to immunological, autonomic, cognitive, neuroendocrine, or neurological systems dysfunctions which can severely impair patients' ability to conduct the activities of daily living [283, 284, 285, 286]. ME/CFS results from the combination of two diseases, Myalgic encephalomyelitis (ME) and chronic fatigue syndrome (CFS). Despite having different definitions, ME and CFS have been often used interchangeably or collectively to characterize the ME/CFS condition. ME/CFS prevalence is estimate to range between 0.2% and 2.8% of the worldwide population [287, 288, 289, 286]. Although the disease affects all ages, races and socioeconomic groups, studies have shown that women are roughly three to four more times more susceptible to the disease then man [290]. Furthermore it is most commonly reported during adulthood, between the ages of 20 and 45 years [282]. Etiological factors for ME/CFS include genetic predisposition, stress, trauma, exposure to toxins, physical activity and rest ratio, as well as a recent history of infectious disease [291]. Finally, while most ME/CFS cases are sporadic, there are reports of cluster outbreaks [292, 293].

Although thought to be multifactorial, ME/CFS's exact etiology remains unknown. None theless, several triggering factors have been proposed, including immune and inflammatory dysfuctions, chronic neuro-inflammation, cell receptor anomalies, decreased metabolism and mitochondrial dysfunction, among other causes [294, 295, 296, 25, 297, 298, 299, 300, 301]. Genetic pre-desposition has also been linked to ME/CFS. In this context,several paper have found an association between genetic [302, 303, 304] and epigenetic (DNA methylation) alterations [305, 306, 307, 308] and ME/CFS pahtology. The latter refers to a biochemical process where a methyl group (CH3) is added to DNA molecules, often at specific sites, regulating gene expression by influencing the accessibility of DNA to transcription factors and other proteins involved in gene regulation [305]. Such modification can either activate or silence gene expression depending on where it occurs within the gene sequence, thus impacting gene expression [305]. However, given the disease heterogeneity, there is likely more then a single pathological mechanisms leading ot the disease. Nonetheless, growing evidence points towards and immune pathology as disease is often preceded by infection [25, 309, 296]. In fact, viral infections appear to frequently precede ME/CFS onset in a vast subset of ME/CFS patients [296, 310]. The more commonly mentioned viral triggers are the EBV [311, 311], cytomegalovirus (CMV) [312], human herpesvirus (HHV) 6, HHV-7, HHV-8 [313, 314, 315], enteroviruses [316], lentivirus [317], herpes simplex 1 and 2 (HSV1 and HSV2) [318] and varicella-zoster virus (VZV) [319, 320] which are globally distributed and highly prevalent in the the adult population [296]. After infection, such

viruses, remain in the body as mostly latent, persistent infections and may reactivate under various conditions [321, 322]. Immunologic disturbance associated with ME/CFS may be the result of viral infection or may lead to reactivation of latent viruses [296]. Once reactivated, the viruses may contribute to the morbidity of ME/CFS via inflammation and immune dysregulation, especially the herpesviruses EBV and HHV-6, which infect immune cells [323]. Still, the association of ME/CFS with a single infectious agent has not been confirmed, and the role of viral infections in ME/CFS remains obscure [291, 324]. On another note, viral infections have also been proposed to trigger an autoimmune response as well [325]. Finally, molecular mimicry has also been proposed within the viral hypothesis. This mechanisms occurs when foreign and self-peptides share sequence similarities, inducing the cross-activation of autoreactive cellular populations from the adaptive immune system [326]. This can lead to chronically activated immune responses that aim at controlling latent infections, posing a high deleterious autoimmune potential [327, 328, 329]. Indeed, specific EBV antigens have been reported to share sequence homology with certain human peptides [330, 331, 332].

Recently, ME/CFS has gained particular prominence because of its significant overlap with the post-COVID syndrome (long COVID or post-acute sequelae of COVID), with several studies estimating that 50% of patients with post-COVID syndrome fulfill ME/CFS criteria [333, 334]. This has led some researches to estimate that COVID-19 infections could be a trigger for the onset of ME/CFS [335], while others rather point towards a reactivation of EBV by the COVID-19 virus [336].

3.2.1 Clinical Manifestation

Apart form the persistent fatigue and Post-Exertional Malaise (PEM) hallmark symptoms, ME/CFS patients also experience an array of symptoms. These include neurological impairments such as slowed thought, impaired concentration, short-term memory loss, confusion, disorientation, cognitive overload, difficulty with making decisions, slowed speech, acquired or exertional dyslexia [337, 296]. ME/CFS patients also experience a number of autonomic symptoms such as nausea, vertigo, dizziness, drop in core temperature and heart pounding [296]. Brain fog[24] or confusion, trouble concentrating, short-term memory problems, attention deficits, slow thinking,trouble to process and retrieve words are also some of the neurocognitive troubles experience by ME/CFS patients [296]. Neurosensorial and neuromuscular disturbances include inability to focus vision, sensitivity to light, noise, vibration, odor, taste and touch impaired depth perception together with muscle weakness, twitching, poor coordination, feeling unsteady on feet and ataxia [337, 296, 282]. Sleep disturbance, unrefreshing sleep, insomnia and day-time sleepiness or dif-

[24] Brain fog is an impairment in short-term memory or concentration severe enough to cause a reduction in previous levels of personal activities

ficulties falling asleep, have also been reported [337, 296]. Pain such as headaches, muscle and joint pains are also common sympotms reported by patients [296, 282]. Flu like symposium, fever or chills, sore throat and swollen lymph nodes are also some of the immunological sympoms eperienced by ME/CFS patients. For a more detailed review on ME/CFS symptoms I redirect the interested reader to [337, 296, 285].

3.2.2 Diagnosis

The lack of a reliable biological markers has hampered ME/CFS diagnosis, which currently relies on the use of symptom-based case criteria while excluding any other fatigue-inducing illnesses that could explain the symptoms [338, 339, 340]. These criteria are often based on self-report assessments designed to screen for non-specific symptoms that may overlap with those of other clinical conditions [283]. Currently, several diagnostic criteria have been proposed for ME/CFS diagnosis, however most frequent definitions commonly used in research and clinical practice include the following: the Fukuda criteria (FC) [285], the Canadian Consensus Criteria (CCC) [284], and the International Consensus Criteria (ICC) [337]. Although these criteria share similarities, the lack of consistency across case definitions poses a major challenge in ME/CFS research [341, 342]. This, coupled with the high heterogeneity of patient's symptoms and the lack of a reliable diagnostic biomarker, have contributed to delays in the identification of biomarkers and effective treatments and possible misdiagnosis of patients [343]. Nonetheless, research efforts continue seeking biomarkers to aid etiological understanding, clinical selection and treatment options for this condition [344]. On their paper, Maksoud and colleagues [344] provide an overview of the different biomarkers identified to date in the literature. These include genetic/epigenetic, endovascular/circulatory, neurological, physiological, immunological biomarkers and many others. However to this date, no conclusive biomarker for disease diagnosis has been established.

3.2.3 Treatment

Currently, there is no known cure for ME/CFS. However, drug-based treatment has been found to be a safe and effective solution for ME/CFS [333]. Therefore, ME/CFS treatment relies mostly on pharmacological interventions such as pyridostigmine, aripiprazole, and naltrexone which are mostly prescribed for symptoms relief and management [333, 345, 346, 347, 333]. These are often recommended on the basis of the patient's primary symptoms [333]. Alternatively, non-pharmacological interventions and medications may also be prescribed for the treatment and management of symptoms [348]. Medications such as modafinil or methylphenidate may help fatigue and brain fog but risk worsening hyperadrenergic symptoms and PEM, and so should be employed with care [333]. However, ME/CFS patients often exhibit hypersensitivity to standard medications given in the usual

doses and can experience side effects or worsened symptoms [284]. Finally, in addition to a direct action for symptom treatment, lifestyle adjustments may also help ME/CFS patients to manage the disease [333].

References

[1] Emil von Behring. Ueber das zustandekommen der diphtherie-immunität und der tetanus-immunität bei thieren. *Drucke 19. Jh.*, 1890.

[2] Emil von Behring. Untersuchungen ueber das zustandekommen der diphtherie-immunität bei thieren. *Drucke 19. Jh.*, 1890.

[3] Harry W Schroeder Jr and Lisa Cavacini. Structure and function of immunoglobulins. *Journal of allergy and clinical immunology*, 125(2):S41–S52, 2010.

[4] Stefan HE Kaufmann. Immunology's coming of age. *Frontiers in immunology*, 10:684, 2019.

[5] Simona Luca and Traian Mihaescu. History of bcg vaccine. *Maedica*, 8(1):53–58, 2013.

[6] Georges Köhler and Cesar Milstein. Continuous cultures of fused cells secreting antibody of predefined specificity. *nature*, 256(5517):495–497, 1975.

[7] Julie Overbaugh and Lynn Morris. The antibody response against hiv-1. *Cold Spring Harbor perspectives in medicine*, 2(1):a007039, 2012.

[8] Natalia V Voge and Enrique Alvarez. Monoclonal antibodies in multiple sclerosis: present and future. *Biomedicines*, 7(1):20, 2019.

[9] David Zahavi and Louis Weiner. Monoclonal antibodies in cancer therapy. *Antibodies*, 9(3):34, 2020.

[10] Keiichi Mitsuyama, Mikio Niwa, Hidetoshi Takedatsu, Hiroshi Yamasaki, Kotaro Kuwaki, Shinichiro Yoshioka, Ryosuke Yamauchi, Shuhei Fukunaga, and Takuji Torimura. Antibody markers in the diagnosis of inflammatory bowel disease. *World journal of gastroenterology*, 22(3):1304, 2016.

[11] Eunhye Ji and Sahmin Lee. Antibody-based therapeutics for atherosclerosis and cardiovascular diseases. *International Journal of Molecular Sciences*, 22(11):5770, 2021.

[12] Giuseppe Pantaleo, Bruno Correia, Craig Fenwick, Victor S Joo, and Laurent Perez. Antibodies to combat viral infections: development strategies and progress. *Nature Reviews Drug Discovery*, 21(9):676–696, 2022.

[13] A Medina Gamero and M Regalado Chamorro. Monoclonal antibodies as treatment for covid-19. *Neurology Perspectives*, 2(1):47, 2022.

[14] Juthaporn Cowan, Ashley Amson, Anna Christofides, and Zain Chagla. Monoclonal antibodies as covid-19 prophylaxis therapy in immunocompromised patient populations. *International Journal of Infectious Diseases*, 2023.

[15] Ashraf A Tabll, Yasser E Shahein, Mohamed M Omran, Mostafa M Elnakib, Ameera A Ragheb, and Khaled E Amer. A review of monoclonal antibodies in covid-19: Role in immunotherapy, vaccine development and viral detection. *Human antibodies*, 29(3):179–191, 2021.

[16] Michael Chary, Alexander F Barbuto, Sudeh Izadmehr, Marc Tarsillo, Eduardo Fleischer, and Michele M Burns. Covid-19 therapeutics: use, mechanism of action, and toxicity (vaccines, monoclonal antibodies, and immunotherapeutics). *Journal of Medical Toxicology*, 19(2):205–218, 2023.

[17] Lin Ning, Hamza B Abagna, Qianhu Jiang, Siqi Liu, and Jian Huang. Development and application of therapeutic antibodies against covid-19. *International journal of biological sciences*, 17(6):1486, 2021.

[18] Rudolf M Lequin. Enzyme immunoassay (eia)/enzyme-linked immunosorbent assay (elisa). *Clinical chemistry*, 51(12):2415–2418, 2005.

[19] Jefte M Drijvers, Imad M Awan, Cory A Perugino, Ian M Rosenberg, and Shiv Pillai. The enzyme-linked immunosorbent assay: the application of elisa in clinical research. In *Basic science methods for clinical researchers*, pages 119–133. Elsevier, 2017.

[20] Katherine M McKinnon. Flow cytometry: an overview. *Current protocols in immunology*, 120(1):5–1, 2018.

[21] Cecil C Czerkinsky, Lars-Åke Nilsson, Håkan Nygren, Örjan Ouchterlony, and Andrej Tarkowski. A solid-phase enzyme-linked immunospot (elispot) assay for enumeration of specific antibody-secreting cells. *Journal of immunological methods*, 65(1-2):109–121, 1983.

[22] Juan S Bonifacino, Esteban C Dell'Angelica, and Timothy A Springer. Immunoprecipitation. *Current protocols in protein science*, 18(1):9–8, 1999.

[23] Biji T Kurien and R Hal Scofield. Western blotting. *Methods*, 38(4):283–293, 2006.

[24] Danica A Helb, Kevin KA Tetteh, Philip L Felgner, Jeff Skinner, Alan Hubbard, Emmanuel Arinaitwe, Harriet Mayanja-Kizza, Isaac Ssewanyana, Moses R Kamya,

James G Beeson, et al. Novel serologic biomarkers provide accurate estimates of recent plasmodium falciparum exposure for individuals and communities. *Proceedings of the National Academy of Sciences*, 112(32):E4438–E4447, 2015.

[25] Madlen Loebel, Maren Eckey, Franziska Sotzny, Elisabeth Hahn, Sandra Bauer, Patricia Grabowski, Johannes Zerweck, Pavlo Holenya, Leif G Hanitsch, Kirsten Wittke, et al. Serological profiling of the ebv immune response in chronic fatigue syndrome using a peptide microarray. *PloS one*, 12(6):e0179124, 2017.

[26] Patrick J Lammie, Delynn M Moss, E Brook Goodhew, Katy Hamlin, Alejandro Krolewiecki, Sheila K West, and Jeffrey W Priest. Development of a new platform for neglected tropical disease surveillance. *International journal for parasitology*, 42(9):797–800, 2012.

[27] Jonas Blomberg, Muhammad Rizwan, Agnes Böhlin-Wiener, Amal Elfaitouri, Per Julin, Olof Zachrisson, Anders Rosén, and Carl-Gerhard Gottfries. Antibodies to human herpesviruses in myalgic encephalomyelitis/chronic fatigue syndrome patients. *Frontiers in Immunology*, 10:1946, 2019.

[28] Mandy Sowa, Rico Hiemann, Peter Schierack, Dirk Reinhold, Karsten Conrad, and Dirk Roggenbuck. Next-generation autoantibody testing by combination of screening and confirmation—the cytobead® technology. *Clinical Reviews in Allergy & Immunology*, 53:87–104, 2017.

[29] Gavin MacBeath. Protein microarrays and proteomics. *Nature genetics*, 32(4):526–532, 2002.

[30] FX Reymond Sutandy, Jiang Qian, Chien-Sheng Chen, and Heng Zhu. Overview of protein microarrays. *Current protocols in protein science*, 72(1):27–1, 2013.

[31] Mohamed F Elshal and J Philip McCoy. Multiplex bead array assays: performance evaluation and comparison of sensitivity to elisa. *Methods*, 38(4):317–323, 2006.

[32] Jennifer A Maynard, Ryan Myhre, and Benjamin Roy. Microarrays in infection and immunity. *Current opinion in chemical biology*, 11(3):306–315, 2007.

[33] Stephane Robin. Some statistical issues in microarray data analysis. In *Between Data Science and Applied Data Analysis: Proceedings of the 26th Annual Conference of the Gesellschaft für Klassifikation eV, University of Mannheim, July 22–24, 2002*, pages 337–347. Springer, 2003.

[34] Robert Nadon and Jennifer Shoemaker. Statistical issues with microarrays: processing and analysis. *TRENDS in Genetics*, 18(5):265–271, 2002.

[35] Nancy Naichao Wang. *Statistical problems in DNA microarray data analysis*. University of California, Berkeley, 2009.

[36] Kouros Owzar, William T Barry, and Sin-Ho Jung. Statistical considerations for analysis of microarray experiments. *Clinical and translational science*, 4(6):466–477, 2011.

[37] Leonard Wossnig, Norbert Furtmann, Andrew Buchanan, Sandeep Kumar, and Victor Greiff. Best practices for machine learning in antibody discovery and development. *arXiv preprint arXiv:2312.08470*, 2023.

[38] Sven-Kevin Hotop, Susanne Reimering, Aditya Shekhar, Ehsaneddin Asgari, Ulrike Beutling, Christine Dahlke, Anahita Fathi, Fawad Khan, Marc Lütgehetmann, Rico Ballmann, et al. Peptide microarrays coupled to machine learning reveal individual epitopes from human antibody responses with neutralizing capabilities against sars-cov-2. *Emerging microbes & infections*, 11(1):1037–1048, 2022.

[39] Laura Abel, Simone Kutschki, Michael Turewicz, Martin Eisenacher, Jale Stoutjesdijk, Helmut E Meyer, Dirk Woitalla, and Caroline May. Autoimmune profiling with protein microarrays in clinical applications. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1844(5):977–987, 2014.

[40] John Joseph Valletta and Mario Recker. Identification of immune signatures predictive of clinical protection from malaria. *PLoS computational biology*, 13(10):e1005812, 2017.

[41] Christiane Sokollik, Aurélie Pahud de Mortanges, Alexander B Leichtle, Pascal Juillerat, and Michael P Horn. Machine learning in antibody diagnostics for inflammatory bowel disease subtype classification. *Diagnostics*, 13(15):2491, 2023.

[42] Eriberto Noel Natali, Alexander Horst, Patrick Meier, Victor Greiff, Mario Nuvolone, Lmar Marie Babrak, Katja Fink, and Enkelejda Miho. The dengue-specific immune response and antibody identification with machine learning. *npj Vaccines*, 9(1):16, 2024.

[43] Ralf Krumkamp, Nicole Sunaina Struck, Eva Lorenz, Marlow Zimmermann, Kennedy Gyau Boahen, Nimako Sarpong, Ellis Owusu-Dabo, Gi Deok Pak, Hyon Jin Jeon, Florian Marks, et al. Classification of invasive bloodstream infections and plasmodium falciparum malaria using autoantibodies as biomarkers. *Scientific reports*, 10(1):21168, 2020.

[44] Zena M Hira and Duncan F Gillies. A review of feature selection and feature extraction methods applied on microarray data. *Advances in bioinformatics*, 2015, 2015.

[45] Todd R Golub, Donna K Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P Mesirov, Hilary Coller, Mignon L Loh, James R Downing, Mark A Caligiuri, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531–537, 1999.

[46] Nurhawani Ahmad Zamri, Nor Azlina Ab. Aziz, Thangavel Bhuvaneswari, Nor Hidayati Abdul Aziz, and Anith Khairunnisa Ghazali. Feature selection of microarray data using simulated kalman filter with mutation. *Processes*, 11(8):2409, 2023.

[47] Thanyaluk Jirapech-Umpai and Stuart Aitken. Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. *BMC bioinformatics*, 6(1):1–11, 2005.

[48] Anthony L Komaroff and W Ian Lipkin. Me/cfs and long covid share similar symptoms and biological abnormalities: road map to the literature. *Frontiers in Medicine*, 10:1187163, 2023.

[49] Abul Abbas, Andrew Lichtman, and Shiv Pillai. *Cellular and molecular immunology E-book*. Elsevier Health Sciences, 2014.

[50] David D Chaplin. Overview of the immune response. *Journal of allergy and clinical immunology*, 125(2):S3–S23, 2010.

[51] Charles Janeway, Paul Travers, Mark Walport, Mark J Shlomchik, et al. *Immunobiology: the immune system in health and disease*, volume 2. Garland Pub. New York, 2001.

[52] Ruslan Medzhitov. Recognition of microorganisms and activation of the immune response. *Nature*, 449(7164):819–826, 2007.

[53] Niklas Engels and Jürgen Wienands. Memory control by the b cell antigen receptor. *Immunological reviews*, 283(1):150–160, 2018.

[54] Peter J Delves and Ivan M Roitt. The immune system. *New England journal of medicine*, 343(1):37–49, 2000.

[55] Xiying Chi, Yue Li, and Xiaoyan Qiu. V (d) j recombination, somatic hypermutation and class switch recombination of immunoglobulins: mechanism and regulation. *Immunology*, 160(3):233–247, 2020.

[56] David B Roth. V (d) j recombination: mechanism, errors, and fidelity. *Mobile DNA III*, pages 311–324, 2015.

[57] Gwendolyn Kaeser and Jerold Chun. Brain cell somatic gene recombination and its phylogenetic foundations. *Journal of Biological Chemistry*, 295(36):12786–12795, 2020.

[58] Istvan Berczi and Andor Szentivanyi. Immunoglobulins. In *Neuroimmune Biology*, volume 3, pages 117–127. Elsevier, 2003.

[59] Janet Stavnezer, Jeroen EJ Guikema, and Carol E Schrader. Mechanism and regulation of class switch recombination. *Annu. Rev. Immunol.*, 26:261–292, 2008.

[60] Anne Durandy. Mini-review activation-induced cytidine deaminase: a dual role in class-switch recombination and somatic hypermutation. *European journal of immunology*, 33(8):2069–2073, 2003.

[61] Paolo Casali and Hong Zan. Class switching and myc translocation: how does dna break? *Nature immunology*, 5(11):1101–1103, 2004.

[62] Michael R Lieber, Kefei Yu, and Sathees C Raghavan. Roles of nonhomologous dna end joining, v (d) j recombination, and class switch recombination in chromosomal translocations. *DNA repair*, 5(9-10):1234–1245, 2006.

[63] Catherine T Yan, Cristian Boboila, Ellen Kris Souza, Sonia Franco, Thomas R Hickernell, Michael Murphy, Sunil Gumaste, Mark Geyer, Ali A Zarrin, John P Manis, et al. Igh class switching and translocations use a robust non-classical end-joining pathway. *Nature*, 449(7161):478–482, 2007.

[64] Roald Nezlin. *The immunoglobulins: structure and function*. Academic press, 1998.

[65] G Vidarsson, G Dekkers, and T Rispens. Igg subclasses and allotypes: from structure to effector functions. front immunol. 2014; 5: 520, 2014.

[66] Blaise Corthésy. Roundtrip ticket for secretory iga: role in mucosal homeostasis? *The journal of immunology*, 178(1):27–32, 2007.

[67] Ibrahim A Darwish. Immunoassay methods and their applications in pharmaceutical analysis: basic methodology and recent advances. *International journal of biomedical science: IJBS*, 2(3):217, 2006.

[68] William Clarke and Mark Marzinke. *Contemporary practice in clinical chemistry*. Academic Press, 2020.

[69] Eric William Rogier, Emanuele Giorgi, Kevin Tetteh, and Nuno Sepúlveda. Current research on serological analyses of infectious diseases. *Frontiers in Medicine*, 10:1154584, 2023.

[70] Ziqing Chen, Tea Dodig-Crnković, Jochen M Schwenk, and Sheng-ce Tao. Current applications of antibody microarrays. *Clinical proteomics*, 15(1):1–15, 2018.

[71] Rosalyn S Yalow, Solomon A Berson, et al. Immunoassay of endogenous plasma insulin in man. *The Journal of clinical investigation*, 39(7):1157–1175, 1960.

[72] Eva Engvall. The elisa, enzyme-linked immunosorbent assay. *Clinical Chemistry*, 56(2):319–320, 2010.

[73] Rachmat Hidayat and Patricia Wulandari. Enzyme linked immunosorbent assay (elisa) technique guideline. *Bioscientia Medicina: Journal of Biomedicine and Translational Research*, 5(5):447–453, 2021.

[74] Karishma Shah and Panagiotis Maghsoudlou. Enzyme-linked immunosorbent assay (elisa): the basics. *British journal of hospital medicine*, 77(7):C98–C101, 2016.

[75] George N Konstantinou. Enzyme-linked immunosorbent assay (elisa). *Food Allergens: Methods and Protocols*, pages 79–94, 2017.

[76] Mahdis Sadat Tabatabaei and Marya Ahmed. Enzyme-linked immunosorbent assay (elisa). In *Cancer Cell Biology: Methods and Protocols*, pages 115–134. Springer, 2022.

[77] Alice V Lin. Direct elisa. *ELISA: Methods and Protocols*, pages 61–67, 2015.

[78] Mandy Alhajj, Muhammad Zubair, and Aisha Farhana. Enzyme linked immunosorbent assay. *StatPearls*, 2023.

[79] Alice V Lin. Indirect elisa. *ELISA: methods and protocols*, pages 51–59, 2015.

[80] Thomas O Kohl and Carl A Ascoli. Indirect immunometric elisa. *Cold Spring Harb Protoc*, 2017(5):396–402, 2017.

[81] Thomas O Kohl and Carl A Ascoli. Immunometric double-antibody sandwich enzyme-linked immunosorbent assay. *Cold Spring Harb Protoc*, 2017(6):458–462, 2017.

[82] Thomas O Kohl and Carl A Ascoli. Direct competitive enzyme-linked immunosorbent assay (elisa). *Cold Spring Harbor Protocols*, 2017(7):pdb–prot093740, 2017.

[83] Suleyman Aydin. A short history, principles, and types of elisa, and our laboratory experience with peptide/protein analyses using elisa. *Peptides*, 72:4–15, 2015.

[84] Hsiao-Ting Kuo, Jay Z Yeh, Po-Hua Wu, Chii-Ming Jiang, and Ming-Chang Wu. Application of immunomagnetic particles to enzyme-linked immunosorbent assay (elisa) for improvement of detection sensitivity of hcg. *Journal of Immunoassay and Immunochemistry*, 33(4):377–387, 2012.

[85] Hovhannes Hayrapetyan, Thao Tran, Eglis Tellez-Corrales, and Charitha Madiraju. Enzyme-linked immunosorbent assay: types and applications. *ELISA: Methods and Protocols*, pages 1–17, 2023.

[86] N Yoshihara. Elisa for diagnosis of infections by viruses. *Nihon Rinsho. Japanese Journal of Clinical Medicine*, 53(9):2277–2282, 1995.

[87] Leire Martin-Souto, Idoia Buldain, Maialen Areitio, Leire Aparicio-Fernandez, Aitziber Antoran, Jean-Philippe Bouchara, Maria Teresa Martin-Gomez, Aitor Rementeria, Fernando L Hernando, and Andoni Ramirez-Garcia. Elisa test for the serological detection of scedosporium/lomentospora in cystic fibrosis patients. *Frontiers in Cellular and Infection Microbiology*, 10:602089, 2020.

[88] Tiago Dias Domingues, Helena Mouriño, and Nuno Sepúlveda. Analysis of antibody data using finite mixture models based on scale mixtures of skew-normal distributions. *medRxiv*, pages 2021–03, 2021.

[89] Yi Huang and Heng Zhu. Protein array-based approaches for biomarker discovery in cancer. *Genomics, Proteomics and Bioinformatics*, 15(2):73–81, 2017.

[90] Chien-Sheng Chen and Heng Zhu. Protein microarrays. *Biotechniques*, 40(4):423–429, 2006.

[91] Heng Zhu, Metin Bilgin, Rhonda Bangham, David Hall, Antonio Casamayor, Paul Bertone, Ning Lan, Ronald Jansen, Scott Bidlingmaier, Thomas Houfek, et al. Global analysis of protein activities using proteome chips. *science*, 293(5537):2101–2105, 2001.

[92] Gavin MacBeath and Stuart L Schreiber. Printing proteins as microarrays for high-throughput function determination. *Science*, 289(5485):1760–1763, 2000.

[93] James B Delehanty. Printing functional protein microarrays using piezoelectric capillaries. *Protein Arrays: Methods and Protocols*, pages 135–143, 2004.

[94] James B Delehanty and Frances S Ligler. Method for printing functional protein microarrays. *Biotechniques*, 34(2):380–385, 2003.

[95] Brian B Haab. Antibody arrays in cancer research. *Molecular & cellular proteomics*, 4(4):377–383, 2005.

[96] Oliver Poetz, Jochen M Schwenk, Stefan Kramer, Dieter Stoll, Markus F Templin, and Thomas O Joos. Protein microarrays: catching the proteome. *Mechanisms of ageing and development*, 126(1):161–170, 2005.

[97] Ming-Wei Lin, Joshua WK Ho, Leonard C Harrison, Cristobal G dos Remedios, and Stephen Adelstein. An antibody-based leukocyte-capture microarray for the diagnosis of systemic lupus erythematosus. *PLoS One*, 8(3):e58199, 2013.

[98] Anders Carlsson, Dirk M Wuttge, Johan Ingvarsson, Anders A Bengtsson, Gunnar Sturfelt, Carl AK Borrebaeck, and Christer Wingren. Serum protein profiling of systemic lupus erythematosus and systemic sclerosis using recombinant antibody microarrays. *Molecular & cellular proteomics*, 10(5), 2011.

[99] Christer Wingren, Anna Sandström, Ralf Segersvärd, Anders Carlsson, Roland Andersson, Matthias Löhr, and Carl AK Borrebaeck. Identification of serum biomarker signatures associated with pancreatic cancer. *Cancer research*, 72(10):2481–2490, 2012.

[100] Justin E Mirus, Yuzheng Zhang, Christopher I Li, Anna E Lokshin, Ross L Prentice, Sunil R Hingorani, and Paul D Lampe. Cross-species antibody microarray interrogation identifies a 3-protein panel of plasma biomarkers for early diagnosis of pancreas cancer. *Clinical Cancer Research*, 21(7):1764–1771, 2015.

[101] Jochen M Schwenk, Ulrika Igel, Maja Neiman, Hanno Langen, Charlotte Becker, Anders Bjartell, Fredrik Ponten, Fredrik Wiklund, Henrik Grönberg, Peter Nilsson, et al. Toward next generation plasma profiling via heat-induced epitope retrieval and array-based assays. *Molecular & Cellular Proteomics*, 9(11):2497–2507, 2010.

[102] Spyros Darmanis, Tao Cui, Kimi Drobin, Su-Chen Li, Kjell Öberg, Peter Nilsson, Jochen M Schwenk, and Valeria Giandomenico. Identification of candidate serum proteins for classifying well-differentiated small intestinal neuroendocrine tumors. *PLoS One*, 8(11):e81712, 2013.

[103] Anna Häggmark, Sanna Byström, Burcu Ayoglu, Ulrika Qundos, Mathias Uhlöen, Mohsen Khademi, Tomas Olsson, Jochen M Schwenk, and Peter Nilsson. Antibody-based profiling of cerebrospinal fluid within multiple sclerosis. *Proteomics*, 13(15):2256–2267, 2013.

[104] Julia Remnestål, David Just, Nicholas Mitsios, Claudia Fredolini, Jan Mulder, Jochen M Schwenk, Mathias Uhlén, Kim Kultima, Martin Ingelsson, Lena Kilander, et al. Csf profiling of the human brain enriched proteome reveals associations of neuromodulin and neurogranin to alzheimer's disease. *PROTEOMICS–Clinical Applications*, 10(12):1242–1253, 2016.

[105] Sanna Byström, Burcu Ayoglu, Anna Häggmark, Nicholas Mitsios, Mun-Gwan Hong, Kimi Drobin, Bjoörn Forsstroöm, Claudia Fredolini, Mohsen Khademi, Sandra Amor, et al. Affinity proteomic profiling of plasma, cerebrospinal fluid, and brain tissue within multiple sclerosis. *Journal of proteome research*, 13(11):4607–4619, 2014.

[106] Anna Häggmark, Maria Mikus, Atefeh Mohsenchian, Mun-Gwan Hong, Björn Forsström, Beata Gajewska, Anna Barańczyk-Kuźma, Mathias Uhlén, Jochen M Schwenk, Magdalena Kuźma-Kozakiewicz, et al. Plasma profiling reveals three proteins associated to amyotrophic lateral sclerosis. *Annals of clinical and translational neurology*, 1(8):544–553, 2014.

[107] Arlene E Dent, Rie Nakajima, Li Liang, Elisabeth Baum, Ann M Moormann, Peter Odada Sumba, John Vulule, Denise Babineau, Arlo Randall, D Huw Davies, et al. Plasmodium falciparum protein microarray antibody profiles correlate with protection from symptomatic malaria in kenya. *The Journal of infectious diseases*, 212(9):1429–1438, 2015.

[108] Elisabeth Baum, Kingsley Badu, Douglas M Molina, Xiaowu Liang, Philip L Felgner, and Guiyun Yan. Protein microarray analysis of antibody responses to plasmodium falciparum in western kenyan highland sites with differing transmission levels. *PLOS one*, 8(12):e82246, 2013.

[109] Julie Bachmann, Florence Burté, Setia Pramana, Ianina Conte, Biobele J Brown, Adebola E Orimadegun, Wasiu A Ajetunmobi, Nathaniel K Afolabi, Francis Akinkunmi, Samuel Omokhodion, et al. Affinity proteomics reveals elevated muscle proteins in plasma of children with cerebral malaria. *PLoS pathogens*, 10(4):e1004038, 2014.

[110] Tamaki Kobayashi, Aarti Jain, Li Liang, Joshua M Obiero, Harry Hamapumbu, Jennifer C Stevenson, Philip E Thuma, James Lupiya, Mike Chaponda, Modest Mulenga, et al. Distinct antibody signatures associated with different malaria transmission intensities in zambia and zimbabwe. *Msphere*, 4(2):e00061–19, 2019.

[111] Oliver P Günther, Jennifer L Gardy, Phillip Stafford, Øystein Fluge, Olav Mella, Patrick Tang, Ruth R Miller, Shoshana M Parker, Stephen A Johnston, and David M Patrick. Immunosignature analysis of myalgic encephalomyelitis/chronic fatigue syndrome (me/cfs). *Molecular neurobiology*, 56:4249–4257, 2019.

[112] Rafael R De Assis, Aarti Jain, Rie Nakajima, Algis Jasinskas, Jiin Felgner, Joshua M Obiero, Philip J Norris, Mars Stone, Graham Simmons, Anil Bagri, et al. Analysis of sars-cov-2 antibodies in covid-19 convalescent blood using a coronavirus antigen microarray. *Nature communications*, 12(1):6, 2021.

[113] Sophie Bérubé, Tamaki Kobayashi, Douglas E Norris, Ingo Ruczinski, William J Moss, Amy Wesolowski, and Thomas A Louis. A random forest classifier uses antibody responses to plasmodium antigens to reveal candidate biomarkers of the intensity and timing of past exposure to plasmodium falciparum. *bioRxiv*, pages 2022–02, 2022.

[114] Jason A Bailey, Andrea A Berry, Mark A Travassos, Amed Ouattara, Sarah Boudova, Emmanuel Y Dotsey, Andrew Pike, Christopher G Jacob, Matthew Adams, John C Tan, et al. Microarray analyses reveal strain-specific antibody responses to plasmodium falciparum apical membrane antigen 1 variants following natural infection and vaccination. *Scientific reports*, 10(1):3952, 2020.

[115] Ramin Mazhari, Eizo Takashima, Rhea J Longley, Shazia Ruybal-Pesantez, Michael T White, Bernard N Kanoi, Hikaru Nagaoka, Benson Kiniboro, Peter Siba, Takafumi Tsuboi, et al. Identification of novel plasmodium vivax proteins associated with protection against clinical malaria. *Frontiers in Cellular and Infection Microbiology*, 13:1076150, 2023.

[116] Rhea J Longley, Michael T White, Eizo Takashima, Jessica Brewster, Masayuki Morita, Matthias Harbers, Thomas Obadia, Leanne J Robinson, Fumie Matsuura, Zoe SJ Liu, et al. Development and validation of serological markers for detecting recent plasmodium vivax infection. *Nature medicine*, 26(5):741–749, 2020.

[117] Wen-Qiang He, Stephan Karl, Michael T White, Wang Nguitragool, Wuelton Monteiro, Andrea Kuehn, Jakub Gruszczyk, Camila T Franca, Jetsumon Sattabongkot, Marcus VG Lacerda, et al. Antibodies to plasmodium vivax reticulocyte binding protein 2b are associated with protection against p. vivax malaria in populations living in low malaria transmission regions of brazil and thailand. *PLoS neglected tropical diseases*, 13(8):e0007596, 2019.

[118] Ashenafi Assefa, Ahmed Ali Ahmed, Wakgari Deressa, Heven Sime, Hussein Mohammed, Amha Kebede, Hiwot Solomon, Hiwot Teka, Kevin Gurrala, Brian Matei, et al. Multiplex serology demonstrate cumulative prevalence and spatial distribution of malaria in ethiopia. *Malaria journal*, 18:1–14, 2019.

[119] Carla Proietti, Lutz Krause, Angela Trieu, Daniel Dodoo, Ben Gyan, Kwadwo A Koram, William O Rogers, Thomas L Richie, Peter D Crompton, Philip L Felgner, et al. Immune signature against plasmodium falciparum antigens predicts clinical immunity in distinct malaria endemic communities. *Molecular & Cellular Proteomics*, 19(1):101–113, 2020.

[120] Jason Rosado, Stéphane Pelleau, Charlotte Cockram, Sarah Hélène Merkling, Narimane Nekkab, Caroline Demeret, Annalisa Meola, Solen Kerneis, Benjamin Terrier, Samira Fafi-Kremer, et al. Multiplex assays for the identification of serological signatures of sars-cov-2 infection: an antibody-based diagnostic and machine learning study. *The Lancet Microbe*, 2(2):e60–e69, 2021.

[121] Thanh Tung Khuat, Robert Bassett, Ellen Otte, Alistair Grevis-James, and Bogdan Gabrys. Applications of machine learning in antibody discovery, process development, manufacturing and formulation: Current trends, challenges, and opportunities. *Computers & Chemical Engineering*, page 108585, 2024.

[122] PGM de Mattos, SB Campos, Md A Rodrigo, HR Manoel, Ld M Jermana, LR Antonio, et al. Machine learning in medicine: Review and applicability. *Arq. Bras. Cardiol*, 118(1):95–102, 2022.

[123] Md Manjurul Ahsan, Shahana Akter Luna, and Zahed Siddique. Machine-learning-based disease diagnosis: A comprehensive review. In *Healthcare*, volume 10, page 541. MDPI, 2022.

[124] Jonathan G Richens, Ciarán M Lee, and Saurabh Johri. Improving the accuracy of medical diagnosis with causal machine learning. *Nature communications*, 11(1):3923, 2020.

[125] Samin Poudel. A study of disease diagnosis using machine learning. In *Medical Sciences Forum*, volume 10, page 8. MDPI, 2022.

[126] Yogesh Kumar, Apeksha Koul, Ruchi Singla, and Muhammad Fazal Ijaz. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *Journal of ambient intelligence and humanized computing*, 14(7):8459–8486, 2023.

[127] Snigdha Dubey, Gaurav Tiwari, Sneha Singh, Saveli Goldberg, and Eugene Pinsky. Using machine learning for healthcare treatment planning. *Frontiers in Artificial Intelligence*, 6:1124182, 2023.

[128] Monika A Myszczynska, Poojitha N Ojamies, Alix MB Lacoste, Daniel Neil, Amir Saffari, Richard Mead, Guillaume M Hautbergue, Joanna D Holbrook, and Laura Ferraiuolo. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nature Reviews Neurology*, 16(8):440–456, 2020.

[129] Majid Afshar and Hamid Usefi. Optimizing feature selection methods by removing irrelevant features using sparse least squares. *Expert Systems with Applications*, 200:116928, 2022.

[130] Nicholas Pudjihartono, Tayaza Fadason, Andreas W Kempa-Liehr, and Justin M O'Sullivan. A review of feature selection methods for machine learning-based disease risk prediction. *Frontiers in Bioinformatics*, 2:927312, 2022.

[131] Sarah Osama, Hassan Shaban, and Abdelmgeid A Ali. Gene reduction and machine learning algorithms for cancer classification based on microarray gene expression data: A comprehensive review. *Expert Systems with Applications*, 213:118946, 2023.

[132] Nashat Alrefai and Othman Ibrahim. Optimized feature selection method using particle swarm intelligence with ensemble learning for cancer classification based on microarray datasets. *Neural Computing and Applications*, 34(16):13513–13528, 2022.

[133] Sabah Sayed, Mohammad Nassef, Amr Badr, and Ibrahim Farag. A nested genetic algorithm for feature selection in high-dimensional cancer microarray datasets. *Expert Systems with Applications*, 121:233–243, 2019.

[134] David H Wolpert and William G Macready. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82, 1997.

[135] Younes Bouchlaghem, Yassine Akhiat, and Souad Amjad. Feature selection: a review and comparative study. In *E3S Web of Conferences*, volume 351, page 01046. EDP Sciences, 2022.

[136] Yvan Saeys, Inaki Inza, and Pedro Larranaga. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517, 2007.

[137] Amandeep Kaur, Kalpna Guleria, and Naresh Kumar Trivedi. Feature selection in machine learning: Methods and comparison. In *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pages 789–795. IEEE, 2021.

[138] Richard J Fox and Matthew W Dimmic. A two-sample bayesian t-test for microarray data. *BMC bioinformatics*, 7:1–11, 2006.

[139] Peyman Jafari and Francisco Azuaje. An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors. *BMC Medical Informatics and Decision Making*, 6(1):1–8, 2006.

[140] Eric Rogier, Ryan Wiegand, Delynn Moss, Jeff Priest, Evelina Angov, Sheetij Dutta, Ito Journel, Samuel E Jean, Kimberly Mace, Michelle Chang, et al. Multiple comparisons analysis of serological data from an area of low plasmodium falciparum transmission. *Malaria journal*, 14:1–12, 2015.

[141] Tae Kyun Kim. T test as a parametric statistic. *Korean journal of anesthesiology*, 68(6):540–546, 2015.

[142] Priscilla E Greenwood and Michael S Nikulin. *A guide to chi-squared testing*, volume 280. John Wiley & Sons, 1996.

[143] Patrick E McKnight and Julius Najab. Mann-whitney u test. *The Corsini encyclopedia of psychology*, pages 1–1, 2010.

[144] Nuno Sepúlveda, Gillian Stresman, Michael T White, Chris J Drakeley, et al. Current mathematical models for analyzing anti-malarial antibody data with an eye to malaria elimination and eradication. *Journal of immunology research*, 2015, 2015.

[145] NIGEL J GAY. Analysis of serological surveys using mixture models: application to a survey of parvovirus b19. *Statistics in Medicine*, 15(14):1567–1573, 1996.

[146] Irina Chis Ster. Inference for serological surveys investigating past exposures to infections resulting in long-lasting immunity–an approach using finite mixture models with concomitant information. *Journal of Applied Statistics*, 39(11):2523–2542, 2012.

[147] Tsung I Lin, Jack C Lee, and Shu Y Yen. Finite mixture modelling using the skew normal distribution. *Statistica Sinica*, pages 909–927, 2007.

[148] Rodrigo M Basso, Víctor H Lachos, Celso Rômulo Barbosa Cabral, and Pulak Ghosh. Robust mixture modeling based on scale mixtures of skew-normal distributions. *Computational Statistics & Data Analysis*, 54(12):2926–2941, 2010.

[149] Adelchi Azzalini. A class of distributions which includes the normal ones. *Scandinavian journal of statistics*, pages 171–178, 1985.

[150] Adelchi Azzalini and Antonella Capitanio. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(2):367–389, 2003.

[151] Ron Kohavi and George H John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2):273–324, 1997.

[152] Yifei Mao, Yuansheng Yang, et al. A wrapper feature subset selection method based on randomized search and multilayer structure. *BioMed research international*, 2019, 2019.

[153] Manosij Ghosh, Ritam Guha, Ram Sarkar, and Ajith Abraham. A wrapper-filter feature selection technique based on ant colony optimization. *Neural Computing and Applications*, 32:7839–7857, 2020.

[154] Giorgos Borboudakis and Ioannis Tsamardinos. Forward-backward selection with early dropping. *The Journal of Machine Learning Research*, 20(1):276–314, 2019.

[155] Verónica Bolón-Canedo, Noelia Sánchez-Marono, Amparo Alonso-Betanzos, José Manuel Benítez, and Francisco Herrera. A review of microarray datasets and applied feature selection methods. *Information sciences*, 282:111–135, 2014.

[156] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 2013.

[157] Sunil Gupta, Kamal Saluja, Ankur Goyal, Amit Vajpayee, and Vipin Tiwari. Comparing the performance of machine learning algorithms using estimated accuracy. *Measurement: Sensors*, 24:100432, 2022.

[158] Andrew A Neath and Joseph E Cavanaugh. The bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):199–203, 2012.

[159] Gary Smith. Step away from stepwise. *Journal of Big Data*, 5(1):1–12, 2018.

[160] Thomas Navin Lal, Olivier Chapelle, Jason Weston, and André Elisseeff. Embedded methods. In *Feature extraction: Foundations and applications*, pages 137–165. Springer, 2006.

[161] Nishchol Mishra and Sanjay Silakari. Predictive analytics: A survey, trends, applications, oppurtunities & challenges. *International Journal of Computer Science and Information Technologies*, 3(3):4434–4438, 2012.

[162] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[163] Bertha Hidalgo and Melody Goodman. Multivariate or multivariable regression? *American journal of public health*, 103(1):39–40, 2013.

[164] Gürol Canbek, Seref Sagiroglu, Tugba Taskaya Temizel, and Nazife Baykal. Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights. In *2017 International Conference on Computer Science and Engineering (UBMK)*, pages 821–826. IEEE, 2017.

[165] Alfred DeMaris. A tutorial in logistic regression. *Journal of Marriage and the Family*, pages 956–968, 1995.

[166] Xin-She Yang. *Nature-inspired optimization algorithms*. Academic Press, 2020.

[167] Chao-Ying Joanne Peng, Kuk Lida Lee, and Gary M Ingersoll. An introduction to logistic regression analysis and reporting. *The journal of educational research*, 96(1):3–14, 2002.

[168] Mahdi Koosha and Amirhossein Amiri. The effect of link function on the monitoring of logistic regression profiles. In *World Congress on Engineering 2011*. World Congress on Engineering 2011, 2011.

[169] Ekaba Bisong and Ekaba Bisong. Training a neural network. *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, pages 333–343, 2019.

[170] Charles Elkan. Maximum likelihood, logistic regression, and stochastic gradient training. *Tutorial notes at CIKM*, page 11, 2012.

[171] Chuanlei Zhang, Minda Yao, Wei Chen, Shanwen Zhang, Dufeng Chen, and Yuliang Wu. Gradient descent optimization in deep learning model training based on multistage and method combination strategy. *Security and Communication Networks*, 2021:1–15, 2021.

[172] Min-Yuan Cheng, Minh-Tu Cao, and Christian Kentaro Nuralim. Computer vision-based deep learning for supervising excavator operations and measuring real-time earthwork productivity. *The Journal of Supercomputing*, 79(4):4468–4492, 2023.

[173] Hana Šinkovec, Georg Heinze, Rok Blagus, and Angelika Geroldinger. To tune or not to tune, a case study of ridge logistic regression in small or sparse datasets. *BMC Medical Research Methodology*, 21:1–15, 2021.

[174] Jose Manuel Pereira, Mario Basto, and Amelia Ferreira Da Silva. The logistic lasso and ridge regression in predicting corporate failure. *Procedia Economics and Finance*, 39:634–641, 2016.

[175] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

[176] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

[177] Divy Dwivedi, Ashutosh Ganguly, and VV Haragopal. Contrast between simple and complex classification algorithms. In *Statistical Modeling in Machine Learning*, pages 93–110. Elsevier, 2023.

[178] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[179] Philipp Probst, Marvin N Wright, and Anne-Laure Boulesteix. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, 9(3):e1301, 2019.

[180] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[181] Dimitri P Solomatine and Durga L Shrestha. Adaboost. rt: a boosting algorithm for regression problems. In *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)*, volume 2, pages 1163–1168. IEEE, 2004.

[182] Junliang Fan, Xin Ma, Lifeng Wu, Fucang Zhang, Xiang Yu, and Wenzhi Zeng. Light gradient boosting machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. *Agricultural water management*, 225:105758, 2019.

[183] Ahmedbahaaaldin Ibrahem Ahmed Osman, Ali Najah Ahmed, Ming Fai Chow, Yuk Feng Huang, and Ahmed El-Shafie. Extreme gradient boosting (xgboost) model to predict the groundwater levels in selangor malaysia. *Ain Shams Engineering Journal*, 12(2):1545–1556, 2021.

[184] Eric C Polley and Mark J Van der Laan. Super learner in prediction. *Collection of Biostatistics Research Archive*, 2010.

[185] Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

[186] Claude Sammut and Geoffrey I Webb. Leave-one-out cross-validation. *Encyclopedia of machine learning*, pages 600–601, 2010.

[187] World Health Organization et al. World malaria report 2023. In *World malaria report 2023*. WHO, 2023.

[188] Jasminka Talapko, Ivana Škrlec, Tamara Alebić, Melita Jukić, and Aleksandar Včev. Malaria: the past and the present. *Microorganisms*, 7(6):179, 2019.

[189] Renu Tuteja. Malaria- an overview. *The FEBS journal*, 274(18):4670–4679, 2007.

[190] Ronald Ross. Observations on a condition necessary to the transformation of the malaria crescent. *British Medical Journal*, 1(1883):251, 1897.

[191] Shigeharu Sato. Plasmodium—a brief introduction to the parasites causing human malaria and their basic biology. *Journal of physiological anthropology*, 40(1):1–13, 2021.

[192] Malcolm J Gardner, Neil Hall, Eula Fung, Owen White, Matthew Berriman, Richard W Hyman, Jane M Carlton, Arnab Pain, Karen E Nelson, Sharen Bowman, et al. Genome sequence of the human malaria parasite plasmodium falciparum. *Nature*, 419(6906):498–511, 2002.

[193] Behailu Taye, Mohammed Seid, and Adugna Gindaba. Entomological study on species composition, behavior, longevity and probability of surviving sporogony of anopheles mosquitoes in lare district, ethiopia. *Journal of Parasitology and Vector Biology*, 9(9):137–145, 2017.

[194] Maria M Mota, Gabriele Pradel, Jerome P Vanderberg, Julius CR Hafalla, Ute Frevert, Ruth S Nussenzweig, Victor Nussenzweig, and Ana Rodrıguez. Migration of plasmodium sporozoites through cells before infection. *Science*, 291(5501):141–144, 2001.

[195] Louis H Miller, Dror I Baruch, Kevin Marsh, and Ogobara K Doumbo. The pathogenic basis of malaria. *Nature*, 415(6872):673–679, 2002.

[196] Ahmed SI Aly, Ashley M Vaughan, and Stefan HI Kappe. Malaria parasite development in the mosquito and infection of the mammalian host. *Annual review of microbiology*, 63:195–221, 2009.

[197] William O Hahn and Paul S Pottinger. Malaria in the traveler: how to manage before departure and evaluate upon return. *Medical Clinics*, 100(2):289–302, 2016.

[198] Richard-Fabian Schumacher and Elena Spinelli. Malaria in children. *Mediterranean journal of hematology and infectious diseases*, 4(1), 2012.

[199] Alessandro Bartoloni and Lorenzo Zammarchi. Clinical aspects of uncomplicated and severe malaria. *Mediterranean journal of hematology and infectious diseases*, 4(1), 2012.

[200] Sheikh Omar Bittaye, Abubacarr Jagne, Lamin ES Jaiteh, Behzad Nadjm, Alfred Amambua-Ngwa, Abdul Karim Sesay, Yankuba Singhateh, Emmanuel Effa, Ousman Nyan, and Ramou Njie. Clinical manifestations and outcomes of severe malaria in adult patients admitted to a tertiary hospital in the gambia. *Malaria journal*, 21(1):270, 2022.

[201] Andrej Trampuz, Matjaz Jereb, Igor Muzlovic, and Rajesh M Prabhu. Clinical review: Severe malaria. *Critical care*, 7:1–9, 2003.

[202] Serign J Ceesay, Lamine Koivogui, Alain Nahum, Makie Abdoulie Taal, Joseph Okebe, Muna Affara, Lama Eugène Kaman, Francis Bohissou, Carine Agbowai, Benoit Gniouma Tolno, et al. Malaria prevalence among young infants in different transmission settings, africa. *Emerging Infectious Diseases*, 21(7):1114, 2015.

[203] Denise L Doolan, Carlota Dobaño, and J Kevin Baird. Acquired immunity to malaria. *Clinical microbiology reviews*, 22(1):13–36, 2009.

[204] Katherine R Dobbs and Arlene E Dent. Plasmodium malaria and antimalarial antibodies in the first year of life. *Parasitology*, 143(2):129–138, 2016.

[205] K Artavanis-Tsakonas, JE Tongren, and EM Riley. The war between the malaria parasite and the immune system: immunity, immunoregulation and immunopathology. *Clinical & Experimental Immunology*, 133(2):145–152, 2003.

[206] Alyssa Barry and Diana Hansen. Naturally acquired immunity to malaria. *Parasitology*, 143(2):125–128, 2016.

[207] Jean Langhorne, Francis M Ndungu, Anne-Marit Sponaas, and Kevin Marsh. Immunity to malaria: more questions than answers. *Nature immunology*, 9(7):725–732, 2008.

[208] Jean-Philippe Julien and Hedda Wardemann. Antibodies against plasmodium falciparum malaria at the molecular level. *Nature Reviews Immunology*, 19(12):761–775, 2019.

[209] S Jake Gonzales, Raphael A Reyes, Ashley E Braddom, and Evelien M Bunnik. Naturally acquired humoral immunity against plasmodium falciparum malaria. *Frontiers in immunology*, 11:594653, 2020.

[210] Karen ES Hamre, Bartholomew N Ondigo, James S Hodges, Sheetij Dutta, Michael Theisen, George Ayodo, and Chandy C John. Antibody correlates of protection from clinical plasmodium falciparum malaria in an area of low and unstable malaria transmission. *The American journal of tropical medicine and hygiene*, 103(6):2174, 2020.

[211] Mark E Polhemus, Shon A Remich, Bernhards R Ogutu, John N Waitumbi, Lucas Otieno, Stella Apollo, James F Cummings, Kent E Kester, Christian F Ockenhouse, Ann Stewart, et al. Evaluation of rts, s/as02a and rts, s/as01b in adults in a high malaria transmission area. *PloS one*, 4(7):e6465, 2009.

[212] Liriye Kurtovic, Marije C Behet, Gaoqian Feng, Linda Reiling, Kiprotich Chelimo, Arlene E Dent, Ivo Mueller, James W Kazura, Robert W Sauerwein, Freya JI Fowkes, et al. Human antibodies activate complement against plasmodium falciparum sporozoites, and are associated with protection against malaria in children. *BMC medicine*, 16:1–17, 2018.

[213] Chandy C John, Aaron J Tande, Ann M Moormann, Peter O Sumba, David E Lanar, Xinan M Min, and James W Kazura. Antibodies to pre-erythrocytic plasmodium falciparum antigens and risk of clinical malaria in kenyan children. *The Journal of infectious diseases*, 197(4):519–526, 2008.

[214] Kai Matuschewski and Ann-Kristin Mueller. Vaccines against malaria–an update. *The FEBS journal*, 274(18):4680–4687, 2007.

[215] Matthew B Laurens. Rts, s/as01 vaccine (mosquirix™): an overview. *Human vaccines & immunotherapeutics*, 16(3):480–489, 2020.

[216] Ally Olotu, Gregory Fegan, Juliana Wambua, George Nyangweso, Amanda Leach, Marc Lievens, David C Kaslow, Patricia Njuguna, Kevin Marsh, and Philip Bejon. Seven-year efficacy of rts, s/as01 malaria vaccine among young african children. *New England Journal of Medicine*, 374(26):2519–2529, 2016.

[217] Daniel Dodoo, Anastasia Aikins, Kwadwo Asamoah Kusi, Helena Lamptey, Ed Remarque, Paul Milligan, Samuel Bosomprah, Roma Chilengi, Yaa Difie Osei, Bartholomew Dicky Akanmori, et al. Cohort study of the association of antibody levels to ama1, msp1 19, msp3 and glurp with protection from clinical malaria in ghanaian children. *Malaria journal*, 7:1–11, 2008.

[218] Freya JI Fowkes, Jack S Richards, Julie A Simpson, and James G Beeson. The relationship between anti-merozoite antibodies and incidence of plasmodium falciparum malaria: a systematic review and meta-analysis. *PLoS medicine*, 7(1):e1000218, 2010.

[219] Matthew B McCarra, George Ayodo, Peter O Sumba, James W Kazura, Ann M Moormann, David L Narum, and Chandy C John. Antibodies to plasmodium falciparum erythrocyte-binding antigen-175 are associated with protection from clinical malaria. *The Pediatric infectious disease journal*, 30(12):1037–1042, 2011.

[220] Bernhards R Ogutu, Odika J Apollo, Denise McKinney, Willis Okoth, Joram Siangla, Filip Dubovsky, Kathryn Tucker, John N Waitumbi, Carter Diggs, Janet Wittes, et al. Blood stage malaria vaccine eliciting high antigen-specific antibody concentrations confers no protection to young children in western kenya. *PloS one*, 4(3):e4708, 2009.

[221] Linda M Murungi, Gathoni Kamuyu, Brett Lowe, Philip Bejon, Michael Theisen, Samson M Kinyanjui, Kevin Marsh, and Faith HA Osier. A threshold concentration of anti-merozoite antibodies is required for protection from clinical episodes of malaria. *Vaccine*, 31(37):3936–3942, 2013.

[222] Faith HA Osier, Gregory Fegan, Spencer D Polley, Linda Murungi, Federica Verra, Kevin KA Tetteh, Brett Lowe, Tabitha Mwangi, Peter C Bull, Alan W Thomas, et al. Breadth and magnitude of antibody responses to multiple plasmodium falciparum merozoite antigens are associated with protection from clinical malaria. *Infection and immunity*, 76(5):2240–2248, 2008.

[223] Issa Nebie, Amidou Diarra, Alphonse Ouedraogo, Issiaka Soulama, Edith C Bougouma, Alfred B Tiono, Amadou T Konate, Roma Chilengi, Michael Theisen, Daniel Dodoo, et al. Humoral responses to plasmodium falciparum blood-stage antigens and association with incidence of clinical malaria in children living in an area of seasonal malaria transmission in burkina faso, west africa. *Infection and immunity*, 76(2):759–766, 2008.

[224] Anja Ramstedt Jensen, Yvonne Adams, and Lars Hviid. Cerebral plasmodium falciparum malaria: The role of pfemp1 in its pathogenesis and immunity, and pfemp1-based vaccines to prevent it. *Immunological reviews*, 293(1):230–252, 2020.

[225] Sofonias K Tessema, Rie Nakajima, Algis Jasinskas, Stephanie L Monk, Lea Lekieffre, Enmoore Lin, Benson Kiniboro, Carla Proietti, Peter Siba, Philip L Felgner, et al. Protective immunity against severe malaria in children is associated with a limited repertoire of antibodies to conserved pfemp1 variants. *Cell host & microbe*, 26(5):579–590, 2019.

[226] Nyamekye Obeng-Adjei, Daniel B Larremore, Louise Turner, Aissata Ongoiba, Shanping Li, Safiatou Doumbo, Takele B Yazew, Kassoum Kayentao, Louis H Miller, Boubacar Traore, et al. Longitudinal analysis of naturally acquired pfemp1 cidr domain variant antibodies identifies associations with malaria protection. *JCI insight*, 5(12), 2020.

[227] Daniel Dodoo, Trine Staalsoe, Haider Giha, Jørgen AL Kurtzhals, Bartholomew D Akanmori, Kojo Koram, Samuel Dunyo, Francis K Nkrumah, Lars Hviid, and Thor G Theander. Antibodies to variant antigens on the surfaces of infected erythrocytes are associated with protection from malaria in ghanaian children. *Infection and immunity*, 69(6):3713–3718, 2001.

[228] Vashti Irani, Paul A Ramsland, Andrew J Guy, Peter M Siba, Ivo Mueller, Jack S Richards, and James G Beeson. Acquisition of functional antibodies that block the binding of erythrocyte-binding antigen 175 and protection against plasmodium falciparum malaria in children. *Clinical Infectious Diseases*, 61(8):1244–1252, 2015.

[229] Chetan E Chitnis, Paushali Mukherjee, Shantanu Mehta, Syed Shams Yazdani, Shikha Dhawan, Ahmad Rushdi Shakri, Rukmini Bharadwaj, Puneet Kumar Gupta, Dhiraj Hans, Suman Mazumdar, et al. Phase i clinical trial of a recombinant blood stage vaccine candidate for plasmodium falciparum malaria based on msp1 and eba175. *PloS one*, 10(4):e0117820, 2015.

[230] Camila Tenorio França, Michael T White, Wen-Qiang He, Jessica B Hostetler, Jessica Brewster, Gabriel Frato, Indu Malhotra, Jakub Gruszczyk, Christele Huon, Enmoore

Lin, et al. Identification of highly-protective combinations of plasmodium vivax recombinant proteins for vaccine development. *Elife*, 6:e28673, 2017.

[231] Peter D Crompton, Matthew A Kayala, Boubacar Traore, Kassoum Kayentao, Aissata Ongoiba, Greta E Weiss, Douglas M Molina, Chad R Burk, Michael Waisberg, Algis Jasinskas, et al. A prospective analysis of the ab response to plasmodium falciparum before and after a malaria season by protein microarray. *Proceedings of the National Academy of Sciences*, 107(15):6958–6963, 2010.

[232] Noppadon Tangpukdee, Chatnapa Duangdee, Polrat Wilairatana, and Srivicha Krudsood. Malaria diagnosis: a brief review. *The Korean journal of parasitology*, 47(2):93, 2009.

[233] WARHURST DC. Laboratory diagnosis of malaria. *J Clin Pathol*, 49:533–538, 1996.

[234] David Payne. Use and limitations of light microscopy for diagnosing malaria at the primary health care level. *Bulletin of the World Health Organization*, 66(5):621, 1988.

[235] Chansuda Wongsrichanalai, Mazie J Barcus, Sinuon Muth, Awalludin Sutamihardja, and Walther H Wernsdorfer. A review of malaria diagnostic tools: microscopy and rapid diagnostic test (rdt). *Defining and Defeating the Intolerable Burden of Malaria III: Progress and Perspectives: Supplement to Volume 77 (6) of American Journal of Tropical Medicine and Hygiene*, 2007.

[236] Daniel J Kyabayinze, James K Tibenderana, George W Odong, John B Rwakimari, and Helen Counihan. Operational accuracy and comparative persistent antigenicity of hrp2 rapid diagnostic tests for plasmodium falciparum malaria in a hyperendemic region of uganda. *Malaria journal*, 7:1–11, 2008.

[237] Arsène Ratsimbasoa, Laza Fanazava, Rogelin Radrianjafy, Julien Ramilijaona, Hughes Rafanomezantsoa, and Didier Ménard. Evaluation of two new immunochromatographic assays for diagnosis of malaria. *American Journal of Tropical Medicine and Hygiene*, 79(5):670, 2008.

[238] J Russ Forney, Chansuda Wongsrichanalai, Alan J Magill, Leslie G Craig, Jeeraphat Sirichaisinthop, Christian T Bautista, R Scott Miller, Christian F Ockenhouse, Kent E Kester, Naomi E Aronson, et al. Devices for rapid diagnosis of malaria: evaluation of prototype assays that detect plasmodium falciparum histidine-rich protein 2 and a plasmodium vivax-specific antigen. *Journal of clinical microbiology*, 41(6):2358–2366, 2003.

[239] SD Fernando, ND Karunaweera, and WP Fernando. Evaluation of a rapid whole blood immunochromatographic assay for the diagnosis of plasmodium falciparum and plasmodium vivax malaria. *Ceylon Medical Jurnal*, 2004.

[240] Arthur Marx, Daniel Pewsner, Matthias Egger, Reto Nüesch, Heiner C Bucher, Blaise Genton, Christoph Hatz, and Peter Jüni. Meta-analysis: accuracy of rapid tests for malaria in travelers returning from endemic areas. *Annals of Internal Medicine*, 142(10):836–846, 2005.

[241] Joel C Mouatcho and JP Dean Goldring. Malaria rapid diagnostic tests: challenges and prospects. *Journal of medical microbiology*, 62(10):1491–1505, 2013.

[242] Nikhil Ranadive, Simon Kunene, Sarah Darteh, Nyasatu Ntshalintshali, Nomcebo Nhlabathi, Nomcebo Dlamini, Stanley Chitundu, Manik Saini, Maxwell Murphy, Adam Soble, et al. Limitations of rapid diagnostic testing in patients with suspected malaria: a diagnostic accuracy evaluation from swaziland, a low-endemicity country aiming for malaria elimination. *Clinical Infectious Diseases*, 64(9):1221–1227, 2017.

[243] Araia Berhane, Mulugeta Russom, Iyassu Bahta, Filmon Hagos, Michael Ghirmai, and Selam Uqubay. Rapid diagnostic tests failing to detect plasmodium falciparum infections in eritrea: an investigation of reported false negative rdt results. *Malaria journal*, 16:1–6, 2017.

[244] Christina T Kozycki, Noella Umulisa, Stephen Rulisa, Emil I Mwikarago, Jean Pierre Musabyimana, Jean Pierre Habimana, Corine Karema, and Donald J Krogstad. False-negative malaria rapid diagnostic tests in rwanda: impact of plasmodium falciparum isolates lacking hrp2 and declining malaria transmission. *Malaria journal*, 16:1–11, 2017.

[245] Philippe Gillet, Annelies Scheirlinck, Jocelijn Stokx, Anja De Weggheleire, Hélder S Chaúque, Oreana DJV Canhanga, Benvindo T Tadeu, Carla DD Mosse, Armindo Tiago, Samuel Mabunda, et al. Prozone in malaria rapid diagnostics tests: how many cases are missed? *Malaria journal*, 10:1–11, 2011.

[246] J Russ Forney, Alan J Magill, Chansuda Wongsrichanalai, Jeeraphat Sirichaisinthop, Christian T Bautista, D Gray Heppner, R Scott Miller, Christian F Ockenhouse, Alex Gubanov, Robyn Shafer, et al. Malaria rapid diagnostic devices: performance characteristics of the para sight f device determined in a multisite field study. *Journal of Clinical Microbiology*, 39(8):2884–2890, 2001.

[247] Baijayantimala Mishra, Jyotish Chandra Samantaray, Ashok Kumar, and Bijay Ranjan Mirdha. Study of false positivity of two rapid antigen detection tests for diagnosis of plasmodium falciparum malaria. *Journal of Clinical Microbiology*, 37(4):1233–1233, 1999.

[248] Jun Seo Oh, Jang Su Kim, Chang Hwan Lee, Deok Hwa Nam, Sun Hyung Kim, Dae Won Park, Chang Kyu Lee, Chae Seung Lim, and Gil Hong Park. Evaluation of

a malaria antibody enzyme immunoassay for use in blood screening. *Memórias do Instituto Oswaldo Cruz*, 103:75–78, 2008.

[249] Harald Noedl, Kritsanai Yingyuen, Anintita Laoboonchai, Mark Fukuda, Jeeraphat Sirichaisinthop, and R Scott Miller. Sensitivity and specificity of an antigen detection elisa for malaria diagnosis. *American Journal of Tropical Medicine and Hygiene*, 75(6):1205–1208, 2006.

[250] Clinton K Murray, Robert A Gasser Jr, Alan J Magill, and R Scott Miller. Update on rapid diagnostic testing for malaria. *Clinical microbiology reviews*, 21(1):97–110, 2008.

[251] Dennis E Bidwell and Alister Voller. Malaria diagnosis by enzyme-linked immunosorbent assays. *Br Med J (Clin Res Ed)*, 282(6278):1747–1748, 1981.

[252] Georges Snounou, Suganya Viriyakosol, William Jarra, Sodsri Thaithong, and K Neil Brown. Identification of the four human malaria parasite species in field samples by the polymerase chain reaction and detection of a high prevalence of mixed infections. *Molecular and biochemical parasitology*, 58(2):283–292, 1993.

[253] Md Tahminur Rahman, Muhammed Salah Uddin, Razia Sultana, Arumina Moue, and Muntahina Setu. Polymerase chain reaction (pcr): a short review. *Anwer Khan Modern Medical College Journal*, 4(1):30–36, 2013.

[254] B Morassin, R Fabre, A Berry, and JF Magnaval. One year's experience with the polymerase chain reaction as a routine method for the diagnosis of imported malaria. *The American journal of tropical medicine and hygiene*, 66(5):503–508, 2002.

[255] Petra F Mens, Aart Van Amerongen, Patrick Sawa, Piet A Kager, and Henk DFH Schallig. Molecular diagnosis of malaria in the field: development of a novel 1-step nucleic acid lateral flow immunoassay for the detection of all 4 human plasmodium spp. and its evaluation in mbita, kenya. *Diagnostic microbiology and infectious disease*, 61(4):421–427, 2008.

[256] Thomas Hänscheid and Martin P Grobusch. How useful is pcr in the diagnosis of malaria? *Trends in parasitology*, 18(9):395–398, 2002.

[257] Kesinee Chotivanich, Kamolrat Silamut, and Nicholas PJ Day. Laboratory diagnosis of malaria infection. *Australian Journal of Medical Science*, 27(1):11–15, 2006.

[258] Swaroop Kumar Pandey, Uttpal Anand, Waseem A Siddiqui, Renu Tripathi, et al. Drug development strategies for malaria: With the hope for new antimalarial drug discovery—an update. *Advances in Medicine*, 2023, 2023.

[259] Jane Achan, Ambrose O Talisuna, Annette Erhart, Adoke Yeka, James K Tibenderana, Frederick N Baliraine, Philip J Rosenthal, and Umberto D'Alessandro. Quinine, an old anti-malarial drug in a modern world: role in the treatment of malaria. *Malaria journal*, 10:1–12, 2011.

[260] Urban Hellgren, Orjan Ericsson, and Lars L Gustafsson. *Handbook of drugs for tropical parasitic infections*. CRC Press, 2003.

[261] Jeanne M Spudick, Lynne S Garcia, David M Graham, and David A Haake. Diagnostic and therapeutic pitfalls associated with primaquine-tolerant plasmodium vivax. *Journal of clinical microbiology*, 43(2):978–981, 2005.

[262] J Kevin Baird and Stephen L Hoffman. Primaquine therapy for malaria. *Clinical infectious diseases*, 39(9):1336–1345, 2004.

[263] Timothy ME Davis, Harin A Karunajeewa, and Kenneth F Ilett. Artemisinin-based combination therapies for uncomplicated malaria. *Medical Journal of Australia*, 182(4):181–185, 2005.

[264] Simon L Croft, Stephan Duparc, Sarah J Arbe-Barnes, J Carl Craft, Chang-Sik Shin, Lawrence Fleckenstein, Isabelle Borghini-Fuhrer, and Han-Jong Rim. Review of pyronaridine anti-malarial properties and product characteristics. *Malaria journal*, 11:1–28, 2012.

[265] SR Meshnick. Artemisinin antimalarials: mechanisms of action and resistance. *Medecine tropicale: revue du Corps de sante colonial*, 58(3 Suppl):13–17, 1998.

[266] U Eckstein-Ludwig, RJ Webb, IDA Van Goethem, JM East, AG Lee, M Kimura, PM O'neill, PG Bray, SA Ward, and S Krishna. Artemisinins target the serca of plasmodium falciparum. *Nature*, 424(6951):957–961, 2003.

[267] François Nosten and Nicholas J White. Artemisinin-based combination treatment of falciparum malaria. *Defining and Defeating the Intolerable Burden of Malaria III: Progress and Perspectives: Supplement to Volume 77 (6) of American Journal of Tropical Medicine and Hygiene*, 2007.

[268] Analigaya R Agoncillo, Kristine Ayessa Elaine B Coronacion, Julienne Theresa T Dagdag, Ma Stephanie C Matira, Niña Kashka E Pamintuan, Charles Sherwin M Soriano, Maria Sonia S Salamat, Ofelia P Saniel, and Pilarita T Rivera. Factors associated with non-compliance with anti-malarial treatment among malaria patients in puerto princesa, palawan. *Acta Medica Philippina*, 49(3), 2015.

[269] Alexandria O Amponsah, Helen Vosper, and Afia FA Marfo. Patient related factors affecting adherence to antimalarial medication in an urban estate in ghana. *Malaria Research and Treatment*, 2015, 2015.

[270] Eduardo Dias Almeida and José Luiz Fernandes Vieira. Factors associated with non-adherence to the treatment of vivax malaria in a rural community from the brazilian amazon basin. *Revista da Sociedade Brasileira de Medicina Tropical*, 49:248–251, 2016.

[271] TE Peto. Toxicity of antimalarial drugs. *Journal of the Royal Society of Medicine*, 82(Suppl 17):30, 1989.

[272] W Robert J Taylor and Nicholas J White. Antimalarial drug toxicity: a review. *Drug safety*, 27(1):25–61, 2004.

[273] Hussien O AlKadi. Antimalarial drug toxicity: a review. *Chemotherapy*, 53(6):385–391, 2007.

[274] SCTP Rts et al. Efficacy and safety of rts, s/as01 malaria vaccine with or without a booster dose in infants and children in africa: final results of a phase 3, individually randomised, controlled trial. *Lancet*, 386(9988):31–45, 2015.

[275] Ally Olotu, Gregory Fegan, Juliana Wambua, George Nyangweso, Ken O Awuondo, Amanda Leach, Marc Lievens, Didier Leboulleux, Patricia Njuguna, Norbert Peshu, et al. Four-year efficacy of rts, s/as01e and its interaction with malaria exposure. *New England Journal of Medicine*, 368(12):1111–1120, 2013.

[276] Philip Bejon, John Lusingu, Ally Olotu, Amanda Leach, Marc Lievens, Johan Vekemans, Salum Mshamu, Trudie Lang, Jayne Gould, Marie-Claude Dubois, et al. Efficacy of rts, s/as01e vaccine against malaria in children 5 to 17 months of age. *New England Journal of Medicine*, 359(24):2521–2532, 2008.

[277] Kent E Kester, James F Cummings, Opokua Ofori-Anyinam, Christian F Ockenhouse, Urszula Krzych, Philippe Moris, Robert Schwenk, Robin A Nielsen, Zufan Debebe, Evgeny Pinelis, et al. Randomized, double-blind, phase 2a trial of falciparum malaria vaccines rts, s/as01b and rts, s/as02a in malaria-naive adults: safety, efficacy, and immunologic associates of protection. *Journal of Infectious Diseases*, 200(3):337–346, 2009.

[278] Seth Owusu-Agyei, Daniel Ansong, Kwaku Asante, Sandra Kwarteng Owusu, Ruth Owusu, Naana Ayiwa Wireko Brobby, David Dosoo, Alex Osei Akoto, Kingsley Osei-Kwakye, Emmanuel Asafo Adjei, et al. Randomized controlled trial of rts, s/as02d

and rts, s/as01e malaria candidate vaccines given according to different schedules in ghanaian children. *PLoS One*, 4(10):e7302, 2009.

[279] Mehreen S Datoo, Alassane Dicko, Halidou Tinto, Jean-Bosco Ouédraogo, Mainga Hamaluba, Ally Olotu, Emma Beaumont, Fernando Ramos Lopez, Hamtandi Magloire Natama, Sophie Weston, et al. Safety and efficacy of malaria vaccine candidate r21/matrix-m in african children: a multicentre, double-blind, randomised, phase 3 trial. *The Lancet*, 403(10426):533–544, 2024.

[280] Mehreen S Datoo, Hamtandi Magloire Natama, Athanase Somé, Duncan Bellamy, Ousmane Traoré, Toussaint Rouamba, Marc Christian Tahita, N Félix André Ido, Prisca Yameogo, Daniel Valia, et al. Efficacy and immunogenicity of r21/matrix-m vaccine against clinical malaria after 2 years' follow-up in children in burkina faso: a phase 1/2b randomised controlled trial. *The Lancet Infectious Diseases*, 22(12):1728–1736, 2022.

[281] Samuel Sang, Mehreen S Datoo, Edward Otieno, Charles Muiruri, Duncan Bellamy, Emmaloise Gathuri, Omar Ngoto, Janet Musembi, Sam Provstgaard-Morys, Lisa Stockdale, et al. Safety and immunogenicity of varied doses of r21/matrix-m™ vaccine at three years follow-up: A phase 1b age de-escalation, dose-escalation trial in adults, children, and infants in kilifi-kenya. *Wellcome Open Research*, 8(450):450, 2023.

[282] Mateo Cortes Rivera, Claudio Mastronardi, Claudia T Silva-Aldana, Mauricio Arcos-Burgos, and Brett A Lidbury. Myalgic encephalomyelitis/chronic fatigue syndrome: a comprehensive review. *Diagnostics*, 9(3):91, 2019.

[283] Kjetil Gundro Brurberg, Marita Sporstøl Fønhus, Lillebeth Larun, Signe Flottorp, and Kirsti Malterud. Case definitions for chronic fatigue syndrome/myalgic encephalomyelitis (cfs/me): a systematic review. *BMJ open*, 4(2):e003973, 2014.

[284] Bruce M Carruthers, Anil Kumar Jain, Kenny L De Meirleir, Daniel L Peterson, Nancy G Klimas, A Martin Lerner, Alison C Bested, Pierre Flor-Henry, Pradip Joshi, AC Peter Powles, et al. Myalgic encephalomyelitis/chronic fatigue syndrome: clinical working case definition, diagnostic and treatment protocols. *Journal of chronic fatigue syndrome*, 11(1):7–115, 2003.

[285] Keiji Fukuda, Stephen E Straus, Ian Hickie, Michael C Sharpe, James G Dobbins, Anthony Komaroff, and International Chronic Fatigue Syndrome Study Group. The chronic fatigue syndrome: a comprehensive approach to its definition and study. *Annals of internal medicine*, 121(12):953–959, 1994.

[286] Eun-Jin Lim, Yo-Chan Ahn, Eun-Su Jang, Si-Woo Lee, Su-Hwa Lee, and Chang-Gue Son. Systematic review and meta-analysis of the prevalence of chronic fatigue syndrome/myalgic encephalomyelitis (cfs/me). *Journal of translational medicine*, 18:1–15, 2020.

[287] Cheol Hwan Kim, Ho Cheol Shin, and Chang Won Won. Prevalence of chronic fatigue and chronic fatigue syndrome in korea: community-based primary care study. *Journal of Korean medical science*, 20(4):529, 2005.

[288] Luis C Nacul, Eliana M Lacerda, Derek Pheby, Peter Campion, Mariam Molokhia, Shagufta Fayyaz, Jose CDC Leite, Fiona Poland, Amanda Howe, and Maria L Drachler. Prevalence of myalgic encephalomyelitis/chronic fatigue syndrome (me/cfs) in three regions of england: a repeated cross-sectional study in primary care. *BMC medicine*, 9:1–12, 2011.

[289] Samantha Johnston, Ekua W Brenu, Donald Staines, and Sonya Marshall-Gradisnik. The prevalence of chronic fatigue syndrome/myalgic encephalomyelitis: a meta-analysis. *Clinical epidemiology*, pages 105–110, 2013.

[290] Jesús Castro-Marrero, Mónica Faro, Luisa Aliste, Naia Sáez-Francàs, Natalia Calvo, Alba Martínez-Martínez, Tomás Fernández de Sevilla, and Jose Alegre. Comorbidity in chronic fatigue syndrome/myalgic encephalomyelitis: a nationwide population-based cohort study. *Psychosomatics*, 58(5):533–543, 2017.

[291] RA Underhill. Myalgic encephalomyelitis, chronic fatigue syndrome: an infectious disease. *Medical hypotheses*, 85(6):765–773, 2015.

[292] The Medical Staff Of The Royal and Free Hospital. An outbreak of encephalomyelitis in the royal free hospital group, london, in 1955. *British Medical Journal*, 2(5050):895, 1957.

[293] Joanna Słomko, Julia L Newton, Sławomir Kujawski, Małgorzata Tafil-Klawe, Jacek Klawe, Donald Staines, Sonya Marshall-Gradisnik, and Pawel Zalewski. Prevalence and characteristics of chronic fatigue syndrome/myalgic encephalomyelitis (cfs/me) in poland: A cross-sectional study. *BMJ open*, 9(3):e023955, 2019.

[294] Øystein Fluge, Ove Bruland, Kristin Risa, Anette Storstein, Einar K Kristoffersen, Dipak Sapkota, Halvor Næss, Olav Dahl, Harald Nyland, and Olav Mella. Benefit from b-lymphocyte depletion using the anti-cd20 antibody rituximab in chronic fatigue syndrome. a double-blind and placebo-controlled study. *PloS one*, 6(10):e26358, 2011.

[295] KA Schlauch, Svetlana F Khaiboullina, Kenny L De Meirleir, Shanti Rawat, Julia Petereit, AA Rizvanov, N Blatt, Tatjana Mijatovic, Doina Kulick, A Palotás, et al. Genome-

wide association analysis identifies genetic variations in subjects with myalgic encephalomyelitis/chronic fatigue syndrome. *Translational psychiatry*, 6(2):e730–e730, 2016.

[296] Santa Rasa, Zaiga Nora-Krukle, Nina Henning, Eva Eliassen, Evelina Shikova, Thomas Harrer, Carmen Scheibenbogen, Modra Murovska, Bhupesh K Prusty, and European Network on ME/CFS (EUROMENE). Chronic viral infections in myalgic encephalomyelitis/chronic fatigue syndrome (me/cfs). *Journal of translational medicine*, 16:1–25, 2018.

[297] Tracy Hampton. Researchers find genetic clues to chronic fatigue syndrome. *JAMA*, 295(21):2466–2467, 2006.

[298] Sarah Myhill, Norman E Booth, and John McLaren-Howard. Chronic fatigue syndrome and mitochondrial dysfunction. *International journal of clinical and experimental medicine*, 2(1):1, 2009.

[299] Michael B VanElzakker. Chronic fatigue syndrome from vagus nerve infection: a psychoneuroimmunological hypothesis. *Medical hypotheses*, 81(3):414–423, 2013.

[300] Julian AG Glassford. The neuroinflammatory etiopathology of myalgic encephalomyelitis/chronic fatigue syndrome (me/cfs). *Frontiers in physiology*, 8:233286, 2017.

[301] T Nguyen, D Staines, Bernd Nilius, Peter Smith, and S Marshall-Gradisnik. Novel identification and characterisation of transient receptor potential melastatin 3 ion channels on natural killer cells and b lymphocytes: effects on cell signalling in chronic fatigue syndrome/myalgic encephalomyelitis patients. *Biological research*, 49:1–8, 2016.

[302] Tengteng Wang, Jie Yin, Andrew H Miller, and Canhua Xiao. A systematic review of the association between fatigue and genetic polymorphisms. *Brain, behavior, and immunity*, 62:230–244, 2017.

[303] Suzanne D Vernon, Elizabeth R Unger, Irina M Dimulescu, Mangalathu Rajeevan, and William C Reeves. Utility of the blood for gene expression profiling and biomarker discovery in chronic fatigue syndrome. *Disease markers*, 18(4):193–199, 2002.

[304] Chinh Bkrong Nguyen, Lene Alsøe, Jessica M Lindvall, Dag Sulheim, Even Fagermoen, Anette Winger, Mari Kaarbø, Hilde Nilsen, and Vegard Bruun Wyller. Whole blood gene expression in adolescent chronic fatigue syndrome: an exploratory cross-sectional study suggesting altered b cell differentiation and survival. *Journal of translational medicine*, 15:1–21, 2017.

[305] Lisa D Moore, Thuc Le, and Guoping Fan. Dna methylation and its basic function. *Neuropsychopharmacology*, 38(1):23–38, 2013.

[306] Wilfred C de Vega, Suzanne D Vernon, and Patrick O McGowan. Dna methylation modifications associated with chronic fatigue syndrome. *PloS one*, 9(8):e104757, 2014.

[307] Malav S Trivedi, Elisa Oltra, Leonor Sarria, Natasha Rose, Vladimir Beljanski, Mary Ann Fletcher, Nancy G Klimas, and Lubov Nathanson. Identification of myalgic encephalomyelitis/chronic fatigue syndrome-associated dna methylation patterns. *PloS one*, 13(7):e0201066, 2018.

[308] AM Helliwell, EC Sweetman, PA Stockwell, CD Edgar, A Chatterjee, and WP Tate. Changes in dna methylation profiles of myalgic encephalomyelitis/chronic fatigue syndrome patients reflect systemic dysfunctions. *Clinical Epigenetics*, 12:1–20, 2020.

[309] Ian Hickie, Tracey Davenport, Denis Wakefield, Ute Vollmer-Conna, Barbara Cameron, Suzanne D Vernon, William C Reeves, and Andrew Lloyd. Post-infective and chronic fatigue syndromes precipitated by viral and non-viral pathogens: prospective cohort study. *Bmj*, 333(7568):575, 2006.

[310] Lily Chu, Ian J Valencia, Donn W Garvert, and Jose G Montoya. Onset patterns and course of myalgic encephalomyelitis/chronic fatigue syndrome. *Frontiers in pediatrics*, 7:12, 2019.

[311] Stephen E Straus, Giovanna Tosato, Gary Armstrong, THOMAS LAWLEY, OLIVIA T PREBLE, WERNER HENLE, RICHARD DAVEY, GARY PEARSON, JAY Epstein, IRENA BRUS, et al. Persisting illness and fatigue in adults with evidence of epstein-barr virus infection. *Annals of internal medicine*, 102(1):7–16, 1985.

[312] A Martin Lerner, Safedin H Beqaj, Robert G Deeter, and James T Fitzgerald. Igm serum antibodies to human cytomegalovirus nonstructural gene products p52 and cm2 (ul44 and ul57) are uniquely present in a subset of patients with chronic fatigue syndrome. *In Vivo (Athens, Greece)*, 16(3):153–159, 2002.

[313] Dedra Buchwald, Paul R Cheney, Daniel L Peterson, Berch Henry, Susan B Wormsley, Ann Geiger, Dharam V Ablashi, S Zaki Salahuddin, Carl Saxinger, Royce Biddle, et al. A chronic illness characterized by fatigue, neurologic and immunologic disorders, and active human herpesvirus type 6 infection. *Annals of internal medicine*, 116(2):103–113, 1992.

[314] Safak Yalcin, Hirohiko Kuratsune, Koji Yamaguchi, Teruo Kitani, and Koichi Yamanishi. Prevalence of human herpesvirus 6 variants a and b in patients with chronic fatigue syndrome. *Microbiology and immunology*, 38(7):587–590, 1994.

[315] DV Ablashi, HB Eastman, CB Owen, MM Roman, J Friedman, JB Zabriskie, DL Peterson, GR Pearson, and JE Whitman. Frequent hhv-6 reactivation in multiple sclerosis (ms) and chronic fatigue syndrome (cfs) patients. *Journal of Clinical Virology*, 16(3):179–191, 2000.

[316] Frances McGarry, John Gow, and Peter O Behan. Enterovirus in the chronic fatigue syndrome. *Annals of internal medicine*, 120(11):972–973, 1994.

[317] Michael J Holmes, Damian S Diack, Richard A Easingwood, John P Cross, and Bronwyn Carlisle. Electron microscopic immunocytological profiles in chronic fatigue syndrome. *Journal of psychiatric research*, 31(1):115–122, 1997.

[318] PA Bond and TG Dinan. Antibodies to herpes simplex types 1 and 2 in chronic fatigue syndrome. *Journal of Chronic Fatigue Syndrome*, 13(1):35–40, 2006.

[319] S-Y Tsai, T-Y Yang, H-J Chen, C-S Chen, W-M Lin, W-C Shen, C-N Kuo, and C-H Kao. Increased risk of chronic fatigue syndrome following herpes zoster: a population-based study. *European journal of clinical microbiology & infectious diseases*, 33:1653–1659, 2014.

[320] Peter Halpin, Marshall Vance Williams, Nancy G Klimas, Mary Ann Fletcher, Zachary Barnes, and Maria Eugenia Ariza. Myalgic encephalomyelitis/chronic fatigue syndrome and gulf war illness patients exhibit increased humoral responses to the herpesviruses-encoded dutpase: Implications in disease pathophysiology. *Journal of medical virology*, 89(9):1636–1645, 2017.

[321] Ke Lan and Min-Hua Luo. Herpesviruses: epidemiology, pathogenesis, and interventions, 2017.

[322] Bridgette V Rooney, Brian E Crucian, Duane L Pierson, Mark L Laudenslager, and Satish K Mehta. Herpes virus reactivation in astronauts during spaceflight and its application on earth. *Frontiers in microbiology*, 10:432964, 2019.

[323] Leen De Bolle, Johan Van Loon, Erik De Clercq, and Lieve Naesens. Quantitative analysis of human herpesvirus 6 cell tropism. *Journal of medical virology*, 75(1):76–85, 2005.

[324] Frédéric Morinet and Emmanuelle Corruble. Chronic fatigue syndrome and viral infections. *An international perspective on the future of research in chronic fatigue syndrome. Croatia: InTech*, pages 1–12, 2012.

[325] Franziska Sotzny, Julià Blanco, Enrica Capelli, Jesús Castro-Marrero, Sophie Steiner, Modra Murovska, Carmen Scheibenbogen, et al. Myalgic encephalomyelitis/chronic

fatigue syndrome–evidence for an autoimmune disease. *Autoimmunity reviews*, 17(6):601–609, 2018.

[326] Michael BA Oldstone. Molecular mimicry: its evolution from concept to mechanism as a cause of autoimmune diseases. *Monoclonal antibodies in immunodiagnosis and immunotherapy*, 33(3):158–165, 2014.

[327] Jonas Blomberg, Carl-Gerhard Gottfries, Amal Elfaitouri, and Anders Rosén. Infection elicited autoimmunity and myalgic encephalomyelitis/chronic fatigue syndrome: an explanatory model. *Frontiers in immunology*, 9:308084, 2018.

[328] Nuno Sepúlveda, Jorge Carneiro, Eliana Lacerda, and Luis Nacul. Myalgic encephalomyelitis/chronic fatigue syndrome as a hyper-regulated immune system driven by an interplay between regulatory t cells and chronic human herpesvirus infections. *Frontiers in immunology*, 10:479474, 2019.

[329] Bhargavi Sundaresan, Fatemeh Shirafkan, Kevin Ripperger, and Kristin Rattay. The role of viral infections in the onset of autoimmune diseases. *Viruses*, 15(3):782, 2023.

[330] Jan D Lünnemann, Ilijas Jelčić, Susanne Roberts, Andreas Lutterotti, Björn Tackenberg, Roland Martin, and Christian Münz. Ebna1-specific t cells from patients with multiple sclerosis cross react with myelin antigens and co-produce ifn-γ and il-2. *The Journal of experimental medicine*, 205(8):1763–1773, 2008.

[331] Katarina Tengvall, Jesse Huang, Cecilia Hellström, Patrick Kammer, Martin Biström, Burcu Ayoglu, Izaura Lima Bomfim, Pernilla Stridh, Julia Butt, Nicole Brenner, et al. Molecular mimicry between anoctamin 2 and epstein-barr virus nuclear antigen 1 associates with multiple sclerosis risk. *Proceedings of the National Academy of Sciences*, 116(34):16955–16960, 2019.

[332] Nuno Sepúlveda. Impact of genetic variation on the molecular mimicry between anoctamin-2 and epstein-barr virus nuclear antigen 1 in multiple sclerosis. *Immunology Letters*, 238:29–31, 2021.

[333] Stephanie L Grach, Jaime Seltzer, Tony Y Chon, and Ravindra Ganesh. Diagnosis and management of myalgic encephalomyelitis/chronic fatigue syndrome. In *Mayo Clinic Proceedings*, volume 98, pages 1544–1551. Elsevier, 2023.

[334] Hannah E Davis, Lisa McCorkell, Julia Moore Vogel, and Eric J Topol. Long covid: major findings, mechanisms and recommendations. *Nature Reviews Microbiology*, 21(3):133–146, 2023.

[335] Sonia Poenaru, Sara J Abdallah, Vicente Corrales-Medina, and Juthaporn Cowan. Covid-19 and post-infectious myalgic encephalomyelitis/chronic fatigue syndrome:

a narrative review. *Therapeutic advances in infectious disease*, 8:20499361211009385, 2021.

[336] Jeffrey E Gold, Ramazan A Okyay, Warren E Licht, and David J Hurley. Investigation of long covid prevalence and its relationship to epstein-barr virus reactivation. *Pathogens*, 10(6):763, 2021.

[337] Bruce M Carruthers, Marjorie I van de Sande, Kenny L De Meirleir, Nancy G Klimas, Gordon Broderick, Terry Mitchell, Don Staines, AC Peter Powles, Nigel Speight, Rosamund Vallings, et al. Myalgic encephalomyelitis: international consensus criteria. *Journal of internal medicine*, 270(4):327–338, 2011.

[338] Carmen Scheibenbogen, Helma Freitag, Julià Blanco, Enrica Capelli, Eliana Lacerda, Jerome Authier, Mira Meeus, Jesus Castro Marrero, Zaiga Nora-Krukle, Elisa Oltra, et al. The european me/cfs biomarker landscape project: an initiative of the european network euromene. *Journal of translational medicine*, 15:1–7, 2017.

[339] ME Beth Smith, Heidi D Nelson, Elizabeth Haney, Miranda Pappas, Monica Daeges, Ngoc Wasson, and Marian McDonagh. Diagnosis and treatment of myalgic encephalomyelitis/chronic fatigue syndrome. *Evidence report/technology assessment*, 98:1–433, 2014.

[340] Ellen Wright Clayton. Beyond myalgic encephalomyelitis/chronic fatigue syndrome: an iom report on redefining an illness. *Jama*, 313(11):1101–1102, 2015.

[341] Eun-Jin Lim and Chang-Gue Son. Review of case definitions for myalgic encephalomyelitis/chronic fatigue syndrome (me/cfs). *Journal of translational medicine*, 18:1–10, 2020.

[342] João Malato, Luís Graça, Luís Nacul, Eliana Lacerda, and Nuno Sepúlveda. Statistical challenges of investigating a disease with a complex diagnosis. *medRxiv*, pages 2021–03, 2021.

[343] Luis Nacul, CC Kingdon, EW Bowman, H Curran, and EM Lacerda. [accepted manuscript] differing case definitions point to the need for an accurate diagnosis of myalgic encephalomyelitis/chronic fatigue syndrome. *Fatigue: Biomedicine, Health and Behavior*, 2017.

[344] Rebekah Maksoud, Chandi Magawa, Natalie Eaton-Fitch, Kiran Thapaliya, and Sonya Marshall-Gradisnik. Biomarkers for myalgic encephalomyelitis/chronic fatigue syndrome (me/cfs): a systematic review. *BMC medicine*, 21(1):189, 2023.

[345] Lucinda Bateman, Alison C Bested, Hector F Bonilla, Bela V Chheda, Lily Chu, Jennifer M Curtin, Tania T Dempsey, Mary E Dimmock, Theresa G Dowell, Donna

Felsenstein, et al. Myalgic encephalomyelitis/chronic fatigue syndrome: essentials of diagnosis and management. In *Mayo clinic proceedings*, volume 96, pages 2861–2878. Elsevier, 2021.

[346] Helen Baxter, Nigel Speight, and William Weir. Life-threatening malnutrition in very severe me/cfs. In *Healthcare*, volume 9, page 459. MDPI, 2021.

[347] Helene Cabanas, Katsuhiko Muraki, Donald Ross Staines, and Sonya Marshall-Gradisnik. Potential therapeutic benefit of low dose naltrexone in myalgic encephalomyelitis/chronic fatigue syndrome: role of transient receptor potential melastatin 3 ion channels in pathophysiology and treatment. *Frontiers in Immunology*, 12:687806, 2021.

[348] Peter C Rowe. Myalgic encephalomyelitis/chronic fatigue syndrome: Trial fails to confirm earlier observations of rituximab's effectiveness. *Annals of internal medicine*, 170(9):656–657, 2019.

Supplementary Matherials

Taking the density function:

$$f(x)=\frac{1}{2\pi^{\frac{p}{2}}|\sum|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu)^T\sum^{-1}(x-\mu)\right)$$

and plugging into equation (7) we obtain the following equation:

$$P(Y=k|X=x)=\frac{\pi_k\left(\frac{1}{2\pi^{\frac{p}{2}}|\sum|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu)^T\sum^{-1}(x-\mu)\right)\right)}{\sum_{k=1}^{K}\pi_k\left(\frac{1}{2\pi^{\frac{p}{2}}|\sum|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu)^T\sum^{-1}(x-\mu)\right)\right)}.$$

Form such expression, we can stipulate that the proportionality relation:

$$\Pr(Y=k\,|\,X=x)\propto\pi_k\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)\right)$$

arises when we drop the denominator, as it is a constant with respect to π_k and therefore doesn't affect the comparison of the posterior probabilities. This simplification is often done to make the expression more manageable, especially when working log probabilities, which can be adopted leading to:

$$\log\mathrm{P}(Y=k\,|\,X=x)\propto-\frac{d}{2}\log(2\pi)-\frac{1}{2}\log|\Sigma|-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)+\log\pi_k$$

which can be written as:

$$\delta_k(x)=-\frac{d}{2}\log(2\pi)-\frac{1}{2}\log|\Sigma|-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)+\log\pi_k$$

Removing the terms that are constant with respect to x we finally reach to mathematical expression:

$$\delta_k(x)=-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)+\log\pi_k.$$

Part II
Benchmark analysis on low-dimensional data

Chapter 4 - Development of antibody selection strategies for multi-sera data

Andre Fonseca[1] (0000-0001-8249-0354), Clara Cordeiro[1,3] (0000-0002-1026-6078) and Nuno Sepulveda[2,3] (0000-0002-8542-1706)

[1] Faculty of Sciences and Technology, University of Algarve, Portugal
[2] Faculty of Mathematics & Information Science, Warsaw University of Technology, Poland
[3] Faculty of Sciences, University of Lisbon, Portugal

André Fonseca, Clara Cordeiro, and Nuno Sepúlveda. Identification of antibody responses predictive of protection against clinical malaria. In Regina Bispo, Lígia Henriques-Rodrigues, Russell Alpizar-Jara, and Miguel de Carvalho, editors, Recent Developments in Statistics and Data Science, pages 227–239, Cham, 2022. Springer International Publishing.

4.1 Abstract

Statistical pipelines have been proposed to discover antibody responses associated with protection against clinical malaria. However, these often produce inconsistent results due to the failure of the statistical assumptions, such as normality. In the present work, we have developed a new statistical pipeline to analyse data from IgG antibodies against 36 *Plasmodium falciparum* antigens from 121 Kenyan children. This pipeline was based on the identification of cut-off values in the antibody distributions that maximized the distinction between susceptible and protected individuals. Our pipeline enabled us to construct a classifier based on few antibodies, whose performance outperformed the previous ones based on a Random forest approach. The good performance of the pipeline suggests its applicability in antibody data analysis with the aim of identifying antimalarial vaccine candidates.

4.2 Introduction

Malaria is caused by infections of *Plasmodium* parasites with the *Plasmodium falciparum* species (*P.falciparum*) being the most lethal one. It remains a global health problem that threatens millions of people worldwide [1, 2]. Malaria is endemic to tropical and subtropical regions where children under 5 years old are the most affected by severe symptoms [1, 3]. The vulnerability of these children has been mainly attributed to the slow process in acquiring natural immunity against malaria parasites via specialized antibody responses upon repeated exposure to the infection [4, 5, 6]. Antibodies, also known as immunoglobulins, are proteins produced by B cells of the adaptive immune system upon antigen recog-

nition [7]. In turn, antigens are small protein fragments ingested and presented to B cells by other immune cells. When bound to their antigen, antibodies are typically used as molecular signals delivered to specialized immune cells (i.e., phagocytes) with the ability to remove the culprit infectious agent by a process called opsonization [7].

Given their putative protective effect, antibodies have been extensively investigated in the context of natural immunity against malaria parasites [8, 9]. However, which set of antibodies confer individual-level protection to clinical malaria is still elusive [8, 10]. A possible reason for the limited knowledge on this research topic is the lack of reproducible results across different studies, as demonstrated by different studies [8, 10, 11]. This lack of reproducibility might be attributed to the failure of the underlying statistical assumptions invoked to the data. To aggravate, there is no standard statistical pipelines to analyse immunological data consistently and reliably in order to make different studies directly comparable.

In this scenario, we propose a new methodology to analyse malaria antibody data. Our working hypothesis is that a pipeline based on strong statistical principles may increase reproducibility across studies, thus, contributing to a reliable discovery of antibodies that promote natural protection to clinical malaria.

The paper is organised as follows: the following section presents a brief description of the data, the methodologies used and the pipeline. The following section shows the results, ending with the discussion, concluding remarks and future work.

4.3 Materials and Methods

4.3.1 Data

We have analyzed a prospective cohort study of 286 children conducted in Kenya (KEN) that arbours immune profiles of ELISA-based antibody titers against 36 *P.falciparum*-specific antigens. Children were monitored for clinical episodes of malaria and classified as **Susceptible** (Sus) ($n_s = 40$) if they had at least one recorded episode of symptomatic malaria (clinical disease[25]). Children with no clinical episode were classified as **Protected** (Prt) ($n_p = 81$). Based on the article by Osier et al. [12], the analysis was performed solely on 121 children ($N = n_s + n_p$) who were infected at screening; these children had ages between 1 and 10 years old. In this way, the bias that can arise from ascertaining exposure to infectious mosquitoes was minimised.

[25] Clinical disease was defined as an auxiliary temperature >37.5ºC, plus any parasitemia for children less than 1 year, and an auxiliary temperature of >37.5ºC, plus parasitemia >2500/µl for individuals older than 1 year, during the 6-month follow-up.

4.3.2 Measuring association

The Chi-squared (χ^2) test of independence identified antibodies associated with clinical protection to malaria [13]. The latter was used to determine if individuals' seropositivity was related to clinical protection against clinical malaria.

4.3.3 Predictive methodologies

4.3.3.1 Multiple logistic and probit regression

Logistic/probit Regressions were followed by stepwise selection (forward and backward) to select the subset of immune responses most associated with the clinical malaria status response variable. The Hosmer-Lemeshow (HL) test was used to evaluate the goodness-of-fit of the estimated regressions [14]. When performing the HL test, the number of bins to calculate quantiles was set to 10. Finally, the Akaike's information criterion (AIC) was used to select the best model.

4.3.3.2 Regularization strategies

Ridge, Least Absolute Shrinkage and Selection Operator (LASSO) and Elastic-Net regressions were concomitantly used to predict immune signatures underlying clinical protection to malaria [15, 16, 17] using the `glmnet` package [18] within the R software. These regression models apply a penalty function λ to the regression model, which reduces or shrinks coefficient estimates towards 0, thus allowing the less-contributing covariates to have a coefficient close to or equal to zero [19]. To obtain the λ that provided the highest accuracy for each model, we incremented λ from 0.001 to 1 with a lag of 0.001. Then, a 10-fold cross-validation was used to compute the model accuracy for each λ [19, 20], and the process was repeated one hundred times. Usually, two distinct λ values are chosen when performing the Ridge and LASSO regressions. However, when using the `glmnet` package, a single λ value can be selected and a second tunning parameter called α that ranges from 0 to 1 can be set to adjust the tunning parameter. To perform the Ridge regression, α is set to 0 while performing the LASSO regression an α equal to 1 is established. To perform the Elastic-Net regression, we increased α from 0 and to 1 with a lag of 0.1.

4.3.3.3 Random forest

A machine learning technique known to provide good results in classification problems is Random forest. It works by constructing multiple decision trees trained on different parts of the same training set by a process called bagging or bootstrap aggregation [21]. The number of trees to grow and the number of predictors randomly sampled as candidates in each split was set to default. To obtain more robust results we performed one hundred iterations of 10-fold cross-validations.

4.3.4 Predictive accuracy

Two measures were used to assess the accuracy of the predictive approaches: Receiving Operating Characteristic (ROC) curves and confusion matrices. The area under the ROC curves (AUC) were utilized as a measure of the predictive model accuracy (or discrimination performance) [22]. In this case, ROC curves were used to assess the antibodies' inherent ability to predict individuals' protection to clinical malaria. Confusion matrices are tables used to describe the performance of a classification model on a set of data for which the true values are known. The confusion matrix is a 2×2 table in which each cell shows the frequency of a different combination of predicted and observed values [23].

4.3.5 Pipeline

Identification of antibody signatures associated with protection to clinical malaria was achieved by developing, establishing and integrating a pipeline to the KEN dataset (Figure 13). This pipeline starts by ordering the individuals according to their antibody quantity values and specifying each value as a possible cut-off to characterize patients as either seropositive or seronegative. Individuals' classified as seropositive had expression values above the cut-off point, while seronegative individuals had expression values below. Contingency tables of seropositivity against clinical malaria status were then constructed. Chi-squared tests of independence were used to determine if antibody seropositivity was associated with clinical protection to clinical malaria. Finally, the cut-off that provided the strongest association to protection (the cut-off with the smallest *p-value*) for each antibody was selected to characterize patients into seropositive and seronegative populations. This process was repeated for each of all the 36 antibodies initially present in our data set. The methodologies Logistic/Probit, Ridge, LASSO, Elastic-Net regressions, and the Random forest were then used to construct different classifiers for clinical malaria. Finally, the performance of each classifier was assessed by ROC curves.

All the analyses were performed using R version 4.0.4 [24] and their packages: `AID`[25], `caret`[26], `dplyr`[27], `ggplot2`[28], `glmnet`[18], `MASS`[29], `pROC`[30], `randomForest`[31], `stats`, `tidyr`[32]. A significance level of 0.05 was used.

4.4 Results

The analysis was performed on the 121 children who were parasite-positive at screening, in line with both the original published article [12] and by Valletta and Recker [8]. Of the 121 children, 40 were considered susceptible, and 81 were protected against clinical malaria. The immune profiles for these individuals consisted of 36 *P.falciparum*-specific antigens taken at the start of the transmission season. Selection of immune profiles against the the *P. falciparum* derives from the fact that this species is the most prevalent malaria

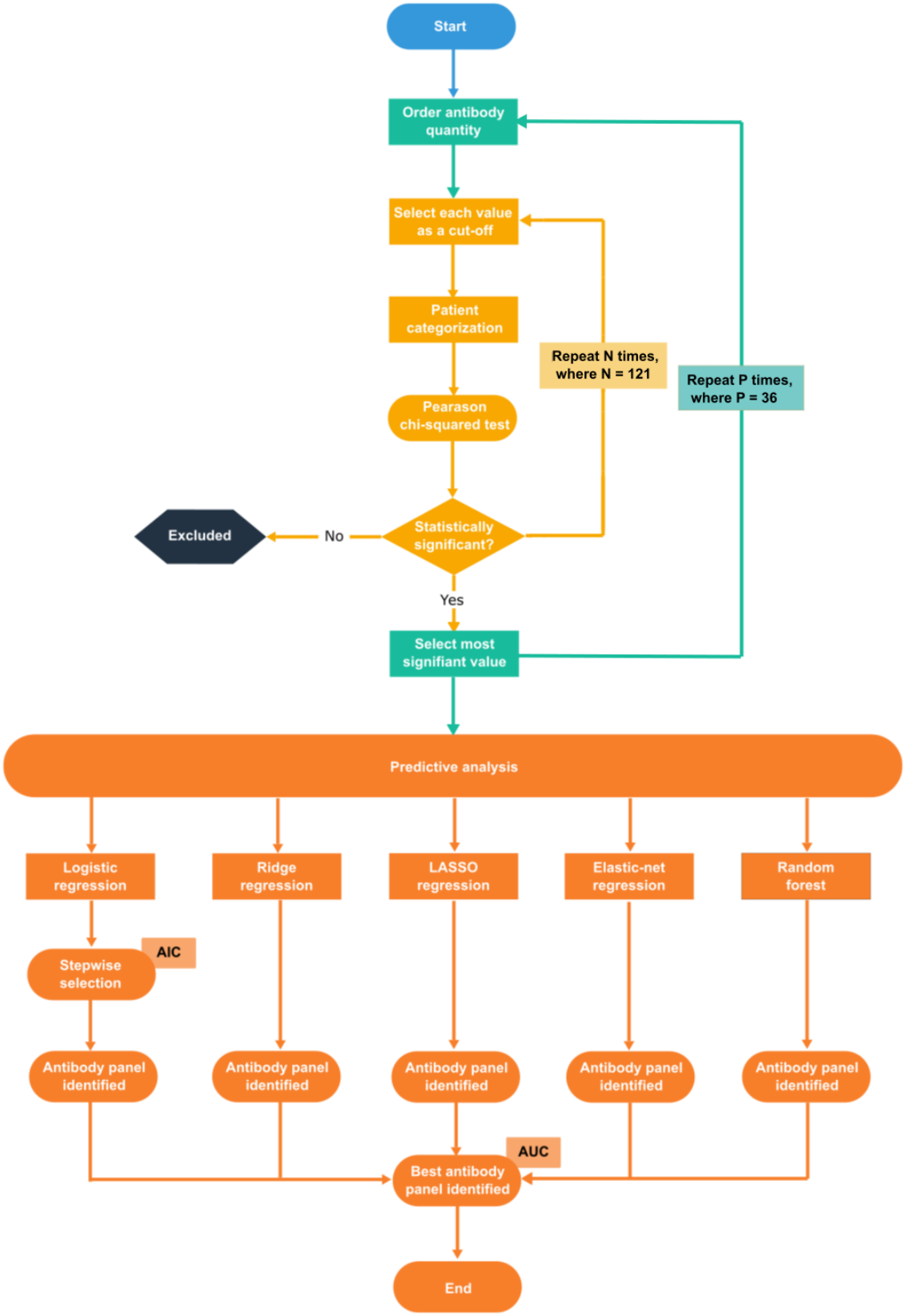


Figure 13: **Pipeline**. The different steps of the analysis are displayed on the workflow using distinct colored shapes. Blue color identifies the beginning of the pipeline. Green indicates computational steps prior to and after the loop for obtaining the χ^2 test *p-value* for each potential cut-off (light orange). Dark orange refers to the predictive performance step and dark grey indicates antibodies removed from the analysis. Additional information is provided by the faded light orange and dark orange-colored shapes.

one in the African continent, home to Kenya [33].

We started by ordering the individuals according to their antibody quantity levels and obtaining the antibody level that provided the best separation ability between the susceptible and protected group of individuals. Antibodies that were not statistically significant in the χ^2 test were removed from the analysis. The antibody data were replaced by a dichotomized seropositive/seronegative variable for the remaining antibodies, which was used later in the predictive performance analysis. According to our results, 28 out of 36 antibodies were able to differentiate susceptible from protected individuals with a 95% confidence, as seen in Table 1.

Considering the 28 antibodies, we proceeded to identify a panel of antibody signatures that could predict individuals' immune status to malaria. Therefore, five distinct methodologies: Logistic/probit, Ridge, LASSO, and Elastic-Net regressions and the Random forest were applied. The objective was to assure that the identification of the best classifier was not hindered by the predictive method selected. Regardless of the method used, individuals' status against malaria was used as the response variable. In contrast, the individuals dichotomized (seronegative/seropositive) data were used as predictive variables.

Logistic and probit regressions were performed. The subsets of antibodies with the highest association with clinical malaria were obtained by stepwise selection. Due to the similarity of the results, we present only the Logistic regression information. Our results showed that the best model was composed of antibodies against the *msp2*, *msp4*, *msp7* *msp10*, *pf11_0373* and *pf113* antigens, with an accuracy of approximately 86% (as seen in Figure 14A). Following this analysis, we included the variable "Age" into this classifier as it was statistically significant *p-value* <0.001; Mann-Whitney). In other words, Age was incorporated as a covariate owing to its statistically significant association with the outcome variable. Therefore, it was included in the model in order to understand its effect on the performance. According to the results, the accuracy of the classifier increased from 86% to 90%, reflecting the contributing effect of other antibodies that were not captured by the model (Figure 14A). Overall, our method correctly classified a total of 102 (75 + 27) individuals out of the total 121 (Figure 14B). In addition, for both the logistic models with (*p-value* = 0.450) and without Age (*p-value* = 0.9275) the *p-value* for the HL goodness of fit test was above 0.05, indicating that at a 95% confidence there was not enough evidence to say that our models were a poor fit.

Notwithstanding the predictive capability of the logistic/probit regression models, we also decided to perform the predictive analysis using regularisation strategies. Ridge regression ($\alpha = 0$) was the best performing model between the three regularisation methods utilised, reaching a predictive accuracy close to 80% when λ ranged from 0.526 to 0.839 (Figure 15A). According to the Ridge regression model, the immune responses most associated with the clinical malaria status were similar to the ones obtained with logistic regression, with the antibody against the *msp7* antigen appearing as the most important variable (im-

Table 1: **Patient Seroprevalence.** The statistically significant results of the χ^2 test for the 28 antibodies. The antibody levels that provided the best separation ability between the susceptible and protected group of individuals (Cut-off), and the proportion of seropositive individuals for all (Total), Protected (Prt) and susceptible (Sus) children, respectively.

Antibody	*P-value*	Cut-off	Total	Prt	Sus
msp1	0.01	0.15	0.85	0.91	0.73
msp2	<0.01	0.07	0.45	0.57	0.20
msp4	<0.01	0.13	0.86	0.96	0.65
msp5	0.02	0.09	0.56	0.64	0.40
msp10	<0.01	0.25	0.79	0.90	0.58
pf12	<0.01	0.10	0.65	0.75	0.45
pf92	<0.01	0.11	0.83	0.91	0.65
pf31	<0.01	0.07	0.61	0.72	0.40
pf113	0.02	0.05	0.74	0.81	0.60
gama	<0.01	0.05	0.61	0.72	0.40
ama1	<0.01	0.16	0.74	0.84	0.53
eba175	<0.01	0.14	0.71	0.84	0.45
eba140	<0.01	0.11	0.96	1.00	0.88
eba181	<0.01	0.11	0.90	0.96	0.78
mtrap	0.01	0.05	0.85	0.91	0.73
asp	<0.01	0.08	0.70	0.79	0.53
msp3	0.01	0.08	0.48	0.57	0.30
msp6	<0.01	0.12	0.78	0.86	0.60
msp7	<0.01	0.24	0.71	0.86	0.40
msrp1	<0.01	0.05	0.79	0.88	0.63
msrp3	<0.01	0.04	0.96	1.00	0.88
h101	0.03	0.05	0.74	0.80	0.60
h103	<0.01	0.07	0.50	0.60	0.28
pf41	<0.01	0.12	0.38	0.48	0.18
pff0335c	<0.01	0.05	0.88	0.95	0.75
rh5	0.04	0.16	0.39	0.46	0.25
ron6	0.02	0.04	0.81	0.88	0.68
pf11_0373	<0.01	0.08	0.21	0.28	0.05

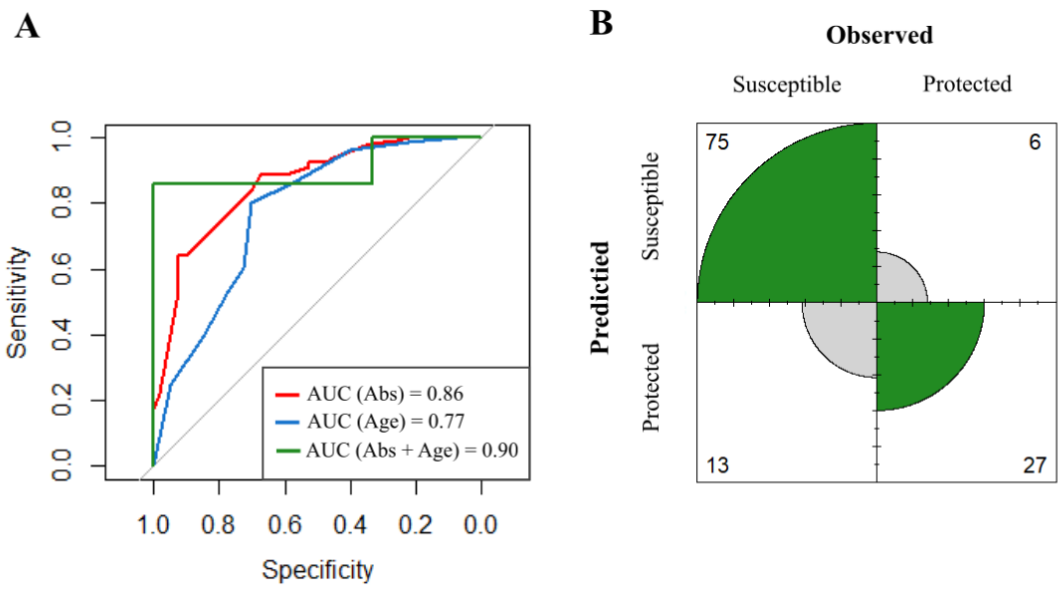


Figure 14: **Predictive performance of the best antibody signature.** (A) ROC curve for the best antibodies signature comprising only the antibodies against *msp2*, *msp4*, *msp7*, *msp10*, *pf11_0373* and *pf113* antigens (red), only Age (blue) and the best antibodies signatures together with Age (green). (B) Confusion matrices derived from the model built with the best antibodies signature together with Age (Abs and Age).

portance = 100), followed by the antibody against *msp4* (importance = 85) (Figure 15A). The antibody against the *msp10* antigen appeared on the fourth position (importance = 78), followed by the antibody against *pf11_0373* in the fifth position (importance = 70) and the antibody against *msp2* on the eight position (importance = 64). The antibody against *Pf113* appeared well below the importance scale in the twenty-seventh position, with an importance of just 6.64 (see Figure 15B). For a detailed explanation on how the importance values were determined, we refer the user to the `caret`[26] official documentation. However, Ridge Regression kept all antibodies in the final classifier with an exception for *asp*.

The results for both the LASSO and Elastic-Net Regression will not be discussed here since the main objective was to identify the method that provides the highest accuracy. To further compare the results of the traditional regression techniques with a more complex technique such as a machine learning model, the Random forest was used. This approach was able to provide an accuracy of 81% (Figure 16A). Like the Ridge Regression, the Random forest approach kept all antibodies in the final classifier except the antibody against *eba140* (Figure 16B). Even more, the immune responses most associated with the clinical malaria status resembled the ones obtained by the Ridge Regression, with the antibody against *msp7* once again appearing as the most important variable (importance = 100), the antibody against *msp2* on the third position (importance = 61), the antibody against *msp10* on the fifth position (importance = 54), followed by the antibody against *msp4* on the sixth (importance = 50). Interestingly the antibody against *pf113*, however, had more weight in

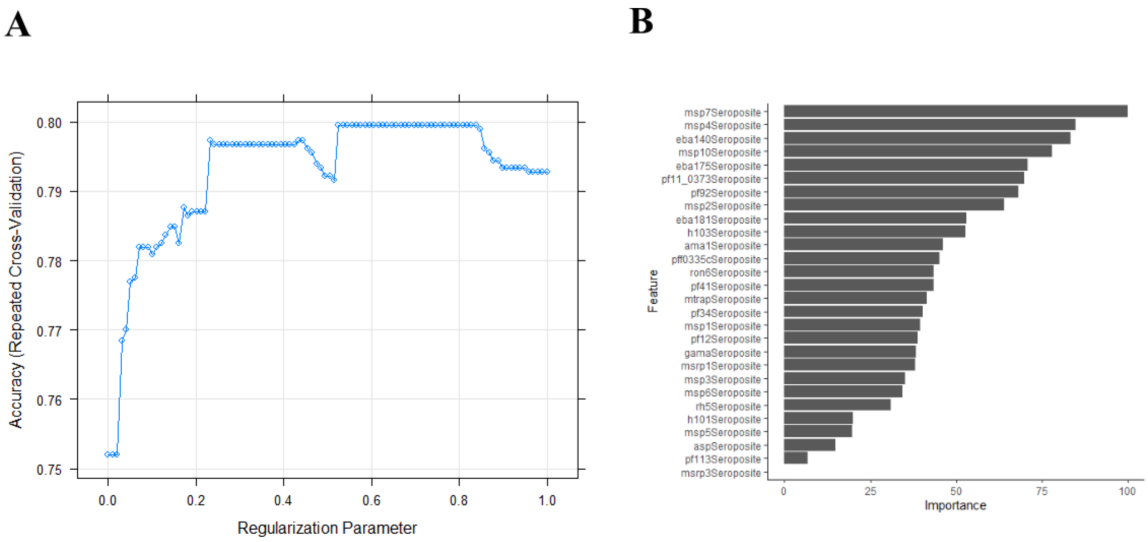


Figure 15: **Ridge Regression regularization strategy results**. (A) The mean accuracy for each regularization parameter (λ) after one hundred runs of 10-fold cross-validation are given by a blue circle. (B) Importance of each antibody in the model

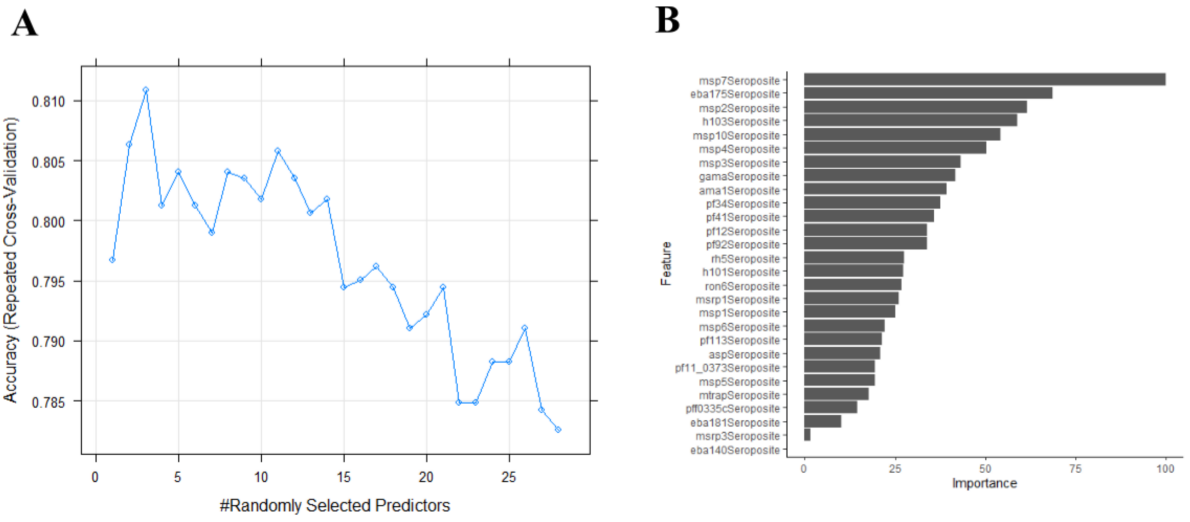


Figure 16: **Random forest results.** (A) The mean accuracy for each value of randomly selected predictors when performing each tree after one hundred runs of 10-fold cross-validation are given by a blue circle. (B) Importance of each antibody in the model

the Random forest (importance = 21) than in the Ridge Regression. Oppositely, however, the antibody against *pf11_0373* had significantly lower importance on the Random forest (importance = 19).

4.5 Discussion

Despite tremendous efforts in the malaria field, it is still unclear which antibodies are essential for developing immune responses that lead to clinical malaria protection [8, 12]. The inconsistent results among studies regarding which set of antibodies are responsible

for individual-level protection to clinical malaria highlight the need for novel approaches to analysing immunological data. Here we set to establish and implement a pipeline to analyse immunological data in a consistent and replicable manner, thus obtaining reproducible results. We hypothesise that pipelines such as the one proposed may finally help identify clear relationships between the measured immune responses and the level of protection against malaria.

Our pipeline was able to identify an immunological classifier against clinical malaria that reach a 86% accuracy using antibody information solely against 6 antigens (*msp2*, *msp4*, *msp7*, *msp10*, *pf11_0373*, *pf113*). Adding "Age" to our classifier increased its accuracy to 90%. This reveals that there were antibody responses associated with clinical protection to malaria that our model could not identify. While age itself does not confer protection against malaria, older individuals are more to have been exposed to the malaria parasite, therefore developing different antibody responses [5]. In this sense, age is a proxy of additional antibodies that the model did not capture. Nevertheless, this effect could also come from the fact that the antibodies responsible for adding this additional explainability to the model were not found in the dataset (due to the small number of features). Comparing our results with the ones obtained by Valletta and Recker [8], our pipeline systematically outperformed their approach independently of the predictive technique used. This increase in accuracy provides clear evidence that an alternative approach to just blindly applying a Random forest approach without any selection criterion may not be best suitable. The benefit of doing a more thorough data analysis before applying a predictive model becomes even more evident when we consider the performance of the Random forest technique in our analysis, which was also used by Valletta and Recker [8]. While they obtained only a predictive performance of 68%, we, on the other hand, obtained a predictive performance of 86%.

Concerning the antibody panel associated with protection to clinical malaria here identified, *msp2*, *msp4*, *msp7*, and *msp10* belong to the group of Merozoite Surface Proteins (MSPs) [34]. The MSPs are expressed on the surface of the merozoite, providing great therapeutic targets for malaria mainly because they are repeatedly and directly exposed to the host humoral immune system [34, 35]. In fact, *msp2* has been extensively associated with protection from clinical malaria in a vast number of independent studies. As an example, *msp2* has been demonstrated to be strongly associated with protection against clinical malaria in two independent cohorts of Kenyan children [10]. *Msp4* has too been already identified as a potential candidate component of the malaria vaccine. In a Senegalese community living in an area of moderate, seasonal malaria transmission, high antibody levels against msp4 constructs were associated with reduced morbidity [36]. Moreover, the protective effect of*msp4* against symptomatic malaria has been already reported in Kenyan children on two occasions [12, 37]. On the other hand, the association between *msp7* and protection against clinical malaria in the literature is less extensive, however *msp7* protec-

tion against malaria have already been identified in the Kenyan population [37]. For the *pf113* antigen, however, the literature is more prominent. Using sera from a longitudinal study in a cohort of Kenyan children, Osier et al. have identified 10 antigens among which *pf113* were associated with protection against clinical episodes of malaria [12]. Furthermore, several other studies refer *pf113* as a promising malaria vaccine candidate [38, 39]. These findings further corroborate our results, as commonly malaria vaccine candidates identified in other studies were also identified here. Interestingly, *msp10* and *pf11_0373* have not been associated with clinical malaria protection so far, as we were unable to identify a single study with such information. This evidence may suggest that there may be antibodies associated with protection against clinical malaria that has not yet been identified. Nevertheless, further studies are necessary to validate our analysis. Immune responses commonly associated with malaria protection and often referred to as potential vaccine candidates such as the merozoite surface protein-1 (*msp1*) and the apical membrane antigen-1 (*ama1*) were not among the best predictors of clinical protection malaria in children, none being incorporated in any of our panel of antibodies [34, 40].

These findings have already been reported in other studies, where antibodies against *msp1* and *ama1* have been described to show low or no associations with protection to clinical malaria [8]. These inconsistent findings further suggest the need for sturdier pipelines that may help to increase reproducibility among studies.

4.6 Concluding remarks and future work

Although we have provided a suggestive approach here, it should be noted that this pipeline is simplistic and will not provide the most sturdy results in every scenario. A situation where this pipeline may not perform well is if there are numerous antibodies related to the outcome under analysis, as a large number of antibodies will be available for the predictive analysis phase. This may reduce the strength of the analysis and consequently lead to less powerful results. One solution to overcome this problem might be to implement correction techniques (such as Bonferroni) for multiple testing. Nevertheless, the implementation of these correction techniques remains to be done. Note that, the question of multiple testing can also be raised for each 121 chi-squared tests when analyzing a given antibody. However, this question can be reframed as an estimation problem where the cut-off value that best discriminates patients from healthy control is a unknown parameter that requires to be estimated. This idea is conceptually similar to the application of the profile likelihood method with cut-off as an unknown parameter.

Work to improve our pipeline to be more suitable for a broader range of datasets is already ongoing. Implementation of other approaches has been considered, where we are trying to make our pipeline more robust. We have also developed another pipeline that relies on traditional statistical techniques after appropriate data transformation and flexi-

ble finite mixture models for determining antibody positivity. The former has also shown promising results. Additionally, it is worth mentioning that by proposing a methodology to analyse antibody data instead of just identifying the exact antibody threshold, any differences in the results may arise due to different sample handling, different sequencing instruments, or other factors that may alter the results. As experimentally conditions are complex to recreate, providing a value that would differentiate patients is a less sensible strategy than providing a methodology that can systematically reproduce the findings of the antibodies associated with antibodies in various studies.

To conclude, although promising, we propose that this pipeline should be tested in other data sets to assess its robustness in different settings. Moreover, we believe that pipelines such as the one presented here may allow the identification of the antibodies that confer protection against clinical malaria in a reproducible manner. Finally, since antibody data are an essential research component of any infectious disease, it is expected that the impact of this work transverses the field of malaria.

Acknowledgements

André Fonseca has a PhD fellowship by FCT – Fundação para a Ciência e a Tecnologia (ref. SFRH/BD/147629/2019). Clara Cordeiro and Nuno Sepúlveda are partially financed by national funds through FCT under the project UIDB/00006/2020. Nuno Sepúlveda is funded by Polish National Agency for Academic Exchange (ref. grant: PPN/ULM/2020/1/00 069/U/00001).

References

[1] Jasminka Talapko, Ivana Škrlec, Tamara Alebić, Melita Jukić, and Aleksandar Včev. Malaria: the past and the present. *Microorganisms*, 7(6):179, 2019.

[2] Elizabeth A Ashley, Aung Pyae Phyo, and Charles J Woodrow. Malaria. *The Lancet*, 391(10130):1608–1621, April 2018.

[3] Brian M Greenwood, David A Fidock, Dennis E Kyle, Stefan HI Kappe, Pedro L Alonso, Frank H Collins, Patrick E Duffy, et al. Malaria: progress, perils, and prospects for eradication. *The Journal of clinical investigation*, 118(4):1266–1276, 2008.

[4] Ann M Moormann. How might infant and paediatric immune responses influence malaria vaccine efficacy? *Parasite immunology*, 31(9):547–559, 2009.

[5] Denise L Doolan, Carlota Dobaño, and J Kevin Baird. Acquired immunity to malaria. *Clinical microbiology reviews*, 22(1):13–36, 2009.

[6] Alyssa Barry and Diana Hansen. Naturally acquired immunity to malaria. *Parasitology*, 143(2):125–128, 2016.

[7] Harry W Schroeder Jr and Lisa Cavacini. Structure and function of immunoglobulins. *Journal of allergy and clinical immunology*, 125(2):S41–S52, 2010.

[8] John Joseph Valletta and Mario Recker. Identification of immune signatures predictive of clinical protection from malaria. *PLoS computational biology*, 13(10):e1005812, 2017.

[9] Lars Hviid. Naturally acquired immunity to plasmodium falciparum malaria in africa. *Acta tropica*, 95(3):270–275, 2005.

[10] Faith HA Osier, Gregory Fegan, Spencer D Polley, Linda Murungi, Federica Verra, Kevin KA Tetteh, Brett Lowe, Tabitha Mwangi, Peter C Bull, Alan W Thomas, et al. Breadth and magnitude of antibody responses to multiple plasmodium falciparum merozoite antigens are associated with protection from clinical malaria. *Infection and immunity*, 76(5):2240–2248, 2008.

[11] Carla Proietti, Lutz Krause, Angela Trieu, Daniel Dodoo, Ben Gyan, Kwadwo A Koram, William O Rogers, Thomas L Richie, Peter D Crompton, Philip L Felgner, et al. Immune signature against plasmodium falciparum antigens predicts clinical immunity in distinct malaria endemic communities. *Molecular & Cellular Proteomics*, 19(1):101–113, 2020.

[12] Faith H Osier, Margaret J Mackinnon, Cécile Crosnier, Gregory Fegan, Gathoni Kamuyu, Madushi Wanaguru, Edna Ogada, Brian McDade, Julian C Rayner, Gavin J Wright, et al. New antigens for a multicomponent blood-stage malaria vaccine. *Science translational medicine*, 6(247):247ra102–247ra102, 2014.

[13] Mary L McHugh. The chi-square test of independence. *Biochemia medica*, 23(2):143–149, 2013.

[14] Giovanni Nattino, Michael L Pennell, and Stanley Lemeshow. Assessing the goodness of fit of logistic regression models in large samples: A modification of the hosmer-lemeshow test. *Biometrics*, 76(2):549–560, 2020.

[15] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[16] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

[17] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

[18] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[19] Daniel M McNeish. Using lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences. *Multivariate behavioral research*, 50(5):471–484, 2015.

[20] LE Melkumova and S Ya Shatskikh. Comparing ridge and lasso estimators for data analysis. *Procedia engineering*, 201:746–755, 2017.

[21] Alessia Sarica, Antonio Cerasa, and Aldo Quattrone. Random forest algorithm for the classification of neuroimaging data in alzheimer's disease: a systematic review. *Frontiers in aging neuroscience*, 9:329, 2017.

[22] Giovanni Tripepi, Kitty J Jager, Friedo W Dekker, and Carmine Zoccali. Diagnostic methods 2: receiver operating characteristic (roc) curves. *Kidney international*, 76(3):252–256, 2009.

[23] Ivo Düntsch and Günther Gediga. Confusion matrices and rough set data analysis. In *Journal of Physics: Conference Series*, volume 1229, page 012055. IOP Publishing, 2019.

[24] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[25] Özgür Asar, Ozlem Ilk, and Osman Dag. Estimating box-cox power transformation parameter via goodness-of-fit tests. *Communications in Statistics-Simulation and Computation*, 46(1):91–105, 2017.

[26] Max Kuhn. Caret: classification and regression training. *Astrophysics Source Code Library*, pages ascl–1505, 2015.

[27] Hadley Wickham, Romain François, Lionel Henry, Kirill Müller, et al. dplyr: A grammar of data manipulation. r package version 0.7. 6. *Computer software]. https://CRAN. R-project. org/package= dplyr*, 2018.

[28] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[29] W N Venables and Brian D Ripley. *Modern applied statistics with S*. Statistics and Computing. Springer, New York, 2010.

[30] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC bioinformatics*, 12(1):1–8, 2011.

[31] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[32] Hadley Wickham, Lionel Henry, et al. Tidyr: Tidy messy data. *R package version*, 1(2):397, 2020.

[33] Robert W Snow, Benn Sartorius, David Kyalo, Joseph Maina, Punam Amratia, Clara W Mundia, Philip Bejon, and Abdisalan M Noor. The prevalence of plasmodium falciparum in sub-saharan africa since 1900. *Nature*, 550(7677):515–518, 2017.

[34] Anja Jäschke, Boubacar Coulibaly, Edmond J Remarque, Hermann Bujard, and Christian Epp. Merozoite surface protein 1 from plasmodium falciparum is a major target of opsonizing antibodies in individuals with acquired immunity against malaria. *Clinical and Vaccine Immunology*, 24(11):e00155–17, 2017.

[35] Clara S Lin, Alessandro D Uboldi, Christian Epp, Hermann Bujard, Takafumi Tsuboi, Peter E Czabotar, and Alan F Cowman. Multiple plasmodium falciparum merozoite surface protein 1 complexes mediate merozoite binding to human erythrocytes. *Journal of Biological Chemistry*, 291(14):7703–7715, 2016.

[36] Ronald Perraut, Marie-Louise Varela, Charlotte Joos, Babacar Diouf, Cheikh Sokhna, Babacar Mbengue, Adama Tall, Cheikh Loucoubar, Aissatou Touré, and Odile Mercereau-Puijalon. Association of antibodies to plasmodium falciparum merozoite surface protein-4 with protection against clinical malaria. *Vaccine*, 35(48):6720–6726, 2017.

[37] Arlene E Dent, Rie Nakajima, Li Liang, Elisabeth Baum, Ann M Moormann, Peter Odada Sumba, John Vulule, Denise Babineau, Arlo Randall, D Huw Davies, et al. Plasmodium falciparum protein microarray antibody profiles correlate with protection from symptomatic malaria in kenya. *The Journal of infectious diseases*, 212(9):1429–1438, 2015.

[38] Wan Ni Chia, Yun Shan Goh, and Laurent Rénia. Novel approaches to identify protective malaria vaccine candidates. *Frontiers in microbiology*, 5:586, 2014.

[39] Roméo-Karl Imboumy-Limoukou, Sandrine Lydie Oyegue-Liabagui, Stella Ndidi, Irène Pegha-Moukandja, Charlene Lady Kouna, Francis Galaway, Isabelle Florent, and Jean Bernard Lekana-Douki. Comparative antibody responses against three antimalarial vaccine candidate antigens from urban and rural exposed individuals in gabon. *European Journal of Microbiology and Immunology*, 6(4):287–297, 2016.

[40] Kazutoyo Miura, Hong Zhou, Olga V Muratova, Andrew C Orcutt, Birgitte Giersing, Louis H Miller, and Carole A Long. In immunization with plasmodium falciparum apical membrane antigen 1, the specificity of antibodies depends on the species immunized. *Infection and immunity*, 75(12):5827–5836, 2007.

Chapter 5 - Antibody selection strategies and their impact in the analysis of malaria multi-sera data

André Fonseca[1,2], Mikolaj Spytek[3], Przemyslaw Biecek[3], Clara Cordeiro[1,2] and Nuno Sepúlveda[2,3]

[1] Faculdade de Ciências e Tecnologia, Universidade do Algarve, Faro, Portugal
[2] CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal
[3] Faculty of Mathematics & Information Science, Warsaw University of Technology, Warsaw, Poland

André Fonseca, Mikolaj Spytek, Przemyslaw Biecek, Clara Cordeiro, and Nuno Sepúlveda. Antibody selection strategies and their impact in predicting clinical malaria based on multi-sera data. BioData Mining, 17, 1 2024.

5.1 Abstract

Background: Nowadays, the chance of discovering the best antibody candidates for predicting clinical malaria has notably increased due to the availability of multisera data. The analysis of these data is typically divided into a feature selection phase followed by a predictive one where several models are constructed for predicting the outcome of interest. A key question in the analysis is to determine which antibodies should be included in the predictive stage and whether they should be included in the original or a transformed scale (i.e., binary/dichotomized).

Methods: To answer this question, we developed and implemented three approaches for antibody selection in the context of predicting clinical malaria: (i) a basic and simple approach based on selecting antibodies via the nonparametric Mann–Whitney-Wilcoxon test; (ii) an optimal dichotomization approach where each antibody was selected according to the optimal cut-off via maximization of the chi-squared (χ^2) statistic for two-way tables; (iii) a hybrid parametric/non-parametric approach that integrates Box-Cox transformation followed by a t-test, together with the use of finite mixture models and the Mann–Whitney-Wilcoxon test as a last resort. We illustrated the application of these three approaches with published serological data of 36 *Plasmodium falciparum* antigens for predicting clinical malaria in 121 Kenyan children. The predictive analysis was based on a Super Learner where predictions from multiple classifiers including the Random Forest were pooled together.

Results: Our results led to almost similar areas under the Receiver Operating Characteristic curves of 0.72 (95% CI = [0.62, 0.82]), 0.80 (95% CI = [0.71, 0.89]), 0.79 (95% CI = [0.7, 0.88])

for the simple, dichotomization and hybrid approaches, respectively. These approaches were based on 6, 20, and 16 antibodies, respectively.
Conclusions: The three feature selection strategies provided a better predictive performance of the outcome when compared to the previous results relying on Random Forest including all the 36 antibodies (AUC = 0.68, 95% CI = [0.57;0.79]). Given the similar predictive performance, we recommended that the three strategies should be used in conjunction in the same data set and selected according to their complexity.

Keywords: Multivariate Serological Data; Super Learner; Statistical modelling; Malaria outcome prediction; Random forest

5.2 Introduction

Multi-sera data, where multiple antibody targets are measured in blood samples from the same individual, are becoming widely available in malaria research due to substantial developments at the level of serological assays [1, 2, 3, 4]. This public availability has boosted basic research on the discovery of key antibodies associated with protection to malaria [5, 6, 7, 8, 9, 10]. It also motivated the development of serological-based algorithms that could predict not only past exposure to malaria parasites [11, 12] but also time since the last infection [13]. It has been suggested that these algorithms could help to design better malaria control strategies such as the serological testing and treatment (seroTAT) approach based on 8 antibodies for detecting *Plasmodium vivax* cases that should be targeted to receive an anti-hypnozoite therapy [12]. In these multi-sera studies, the total number of antibody targets varied from dozens [8, 10, 13] to thousands [6, 7, 14]. This number implies a huge computational cost for algorithms that search for the best model for the data. To overcome this problem, a brute-force approach (where every possible combination is tried out) is computationally feasible for no more than 5 antibody targets [8]. However, above that number implementation of brute-force approaches is not recommended [10, 12]. This computational drawback motivates the use of data analysis strategies that are generically divided into an antibody or feature selection stage, followed by a predictive one, in which several statistical or machine learning models are estimated from the data [7, 9, 10, 13, 15]. In this scenario, the initial antibody selection stage determines the predictive performance of the models to be constructed in the following stage.

Antibody selection can be formulated as the procedure to determine which antibodies are important to predict an outcome of interest [16, 17, 18]. However, this selection hides the question whether data transformation, including dichotomization, should be used. Data transformation is particularly relevant in multiplex serological assays because distinct data distributions can emerge due to differences in the calibration curves across antibodies, as demonstrated in assay-optimization studies [16, 17, 18]. Until now, antibody selec-

tion has been carried out using only raw or untransformed [5, 6] data or seroprevalence-like data but [10, 12] without any combination of both. Additionally, the transformation of each antibody data is typically not considered. Therefore, current antibody selection procedures for multi-sera data lack the flexibility to accommodate different data patterns. The current study tackles this issue and shows that it can potentially increase the chance of obtaining improved outcome predictions.

This paper aims at evaluating three feature selection strategies for the identification of antibody responses that could predict clinical malaria with increased accuracy. Initially, we implemented a basic approach where the statistical significance for the nonparametric Mann–Whitney-Wilcoxon test was obtained for each antibody comparing the protected individuals to susceptible ones. A second strategy is also presented in which data of each antibody is initially dichotomized using an optimal cut-off point in the antibody distribution based on the maximization of the χ^2 test statistic. Finally, we introduced a general parametric strategy for antibody selection in which a combination of transformed and dichotomized antibody data can be selected for the predictive phase. This strategy adds flexibility to feature selection by combining the Box-Cox data transformation with well-known parametric statistical tests.

To illustrate these three strategies, we have analyzed a published dataset on Immunoglobulin G (IgG) antibody responses to 36 *Plasmodium falciparum* (Pf) antigens in Kenyan children to understand protection to clinical malaria [8] and whose data analysis was previously done with Random Forest [15].

5.3 Materials and Methods

5.3.1 Data under analysis

We re-analysed published data of 121 Kenyan children (age range: 1-10 years) described in detail elsewhere [8]. All children had a documented parasitaemia (parasite-positive) at the time of sampling and were monitored for clinical episodes of malaria over a follow-up period of 6 months. As in the original publication, children were considered susceptible (Sus, $n_s = 40$) or protected (Prt, $n_p = 81$) if they had or did not have any clinical episode during the follow-up period. The serological data referred to individual IgG antibody responses to 36 *Plasmodium falciparum* antigens. These antibody responses were measured by multiple enzyme-linked immunosorbent assays (ELISA). Detailed information about recruitment, study design and experimental protocols, among other aspects of this early research, can be found in the original publication [8].

5.3.2 Preliminary antibody feature selection using Random Forest

The Random Forest (RF) works by constructing multiple decision trees trained on different parts of the same training set by a resampling process called bootstrap aggregation or bagging [19]. RFs were implemented by repeatedly fitting the model to 1000 resampled subsets of the data (100 repeats of tenfold cross-validation). For each repetition, the dataset was divided into 10 folds, of which 9-folds were used to perform an inner tenfold cross-validation [20]. The number of trees to grow and the number of predictors randomly sampled as candidates in each split was set to default [21] (number of trees = 500; number of predictors randomly selected = 2, 19 and 36), and the optimization criteria was maximization of the area under the of Receiver Operating Characteristic (ROC) curve, known as AUC [22]. Feature importance was determined by the mean decrease in accuracy [23]. Briefly, for each tree, the prediction accuracy on the out-of-bag portion of the data was recorded. Then, after permuting each predictor variable, the prediction accuracy on the out-of-bag portion of the data was once again recorded. The difference between the two accuracies was then averaged across all the generated trees, and normalized by the standard error [23].

5.3.3 Antibody selection based on non-parametric testing

The first antibody selection strategy was to select the antibodies by their statistical significance according to the non-parametric Mann-Whitney-Wilcoxon test comparing the protected and susceptible groups in each antibody data [24].

5.3.4 Antibody selection based on optimal data dichotomization

The second antibody selection strategy was based on a procedure in which the optimal cut-off to differentiate one study group from another was estimated by maximizing the chi-squared χ^2 statistic for testing independence in two-way contingency tables, as done elsewhere [25, 26] (Figure 17). In more detail, the values of each antibody were sorted by increasing order and then used to divide individuals into two latent serological groups (i.e., seronegative/seropositive individuals or high/low responders). For each value of a given antibody, the resulting data were summarized into a two-way contingency table comprising the qualitative variables serological status (below/above the cut-off) and malaria protection status (protected/non-protected). The χ^2 test statistic was then calculated for this contingency table. After repeating this procedure for all antibody values, the optimal cut-off was selected as the value that maximized that test statistic, meaning the one that provided the best discriminatory ability between both groups of patients. After selecting the optimal cut-off, we calculated the respective *p-value* associated with the χ^2 test. The dichotomized data was then used for the predictive phase. This procedure was finally repeated for each of the 36 antibodies included in the dataset. Note that this procedure is

conceptually equivalent to predict the outcome with individual decision trees using data of each antibody separately. In this procedure, we also quantified the uncertainty around each optimal cut-off by means of a 95% confidence interval. With this purpose, we used the following Bootstrap algorithm in the respective calculation: (i) generate a new sample (with the same sample size) with replacement from the observed sample of the antibody under analysis; (ii) determine the optimal cut-off value as described above; (iii) repeat points (i) and (ii) 1000 times and saving the respective optimal cut-off values; (iv) determine a 95% confidence interval by calculating the empirical 2.5% and 97.5% quantiles of the Bootstrap samples related to the estimated optimal cutoff values.

5.3.5 Antibody selection based on hybrid parametric/non-parametric approach

We adopted an alternative antibody selection approach using different parametric models or statistical tests (Figure 18). In the first step, we determined the optimal Box-Cox transformation for each antibody. This transformation was sought to obtain normal distributions with homogeneous variances in both groups. We searched the best parameter of this transformation (hereafter denoted as λ) within the interval (-4;4) by maximizing evidence for a Normal distribution using the Shapiro-Wilk (SW) test where the null hypothesis states that the data come from a normal distribution (with unknown parameters) [27]. A significance level of 5% was specified to assess wether the data of each antibody could follow a normal distribution.

In the antibodies for which there was no evidence against the normal distribution, we calculated the *p-value* for the t-test aiming at comparing the means between susceptible and protected groups. The remaining antibodies, for which the normality assumption failed, were then evaluated via finite mixture models given that it is recurrent to find latent populations in serological data [28].

Using transformed data, we estimated two-component mixture models based on Normal, Generalized t, Skew-Normal and Skew-t distributions by maximizing the likelihood function via the Expectation-Maximization algorithm [29]. We also estimated the Generalized t, Skew-Normal and Skew-t distributions to assess the evidence that the data could come from a single non-Normal serological population beyond the ones identified by the Box-Cox transformation. We also estimated the Generalized t, Skew-Normal and Skew-t distributions to assess the possibility that the data could come from a single non-Normal serological population beyond the ones identified by the Box-Cox transformation. We compared all these models using the Akaike's Information Criterion (AIC) and performed the Pearson's goodness-of-fit test by dividing the respective data into deciless (i.e., 10%-quantiles). Minimization of the AIC, together with a good fit to the data, at the significance level of 5%, was the criterion for selecting the best model. For antibodies whose data provided evidence of two latent serological populations, we divided the individuals into two latent serological

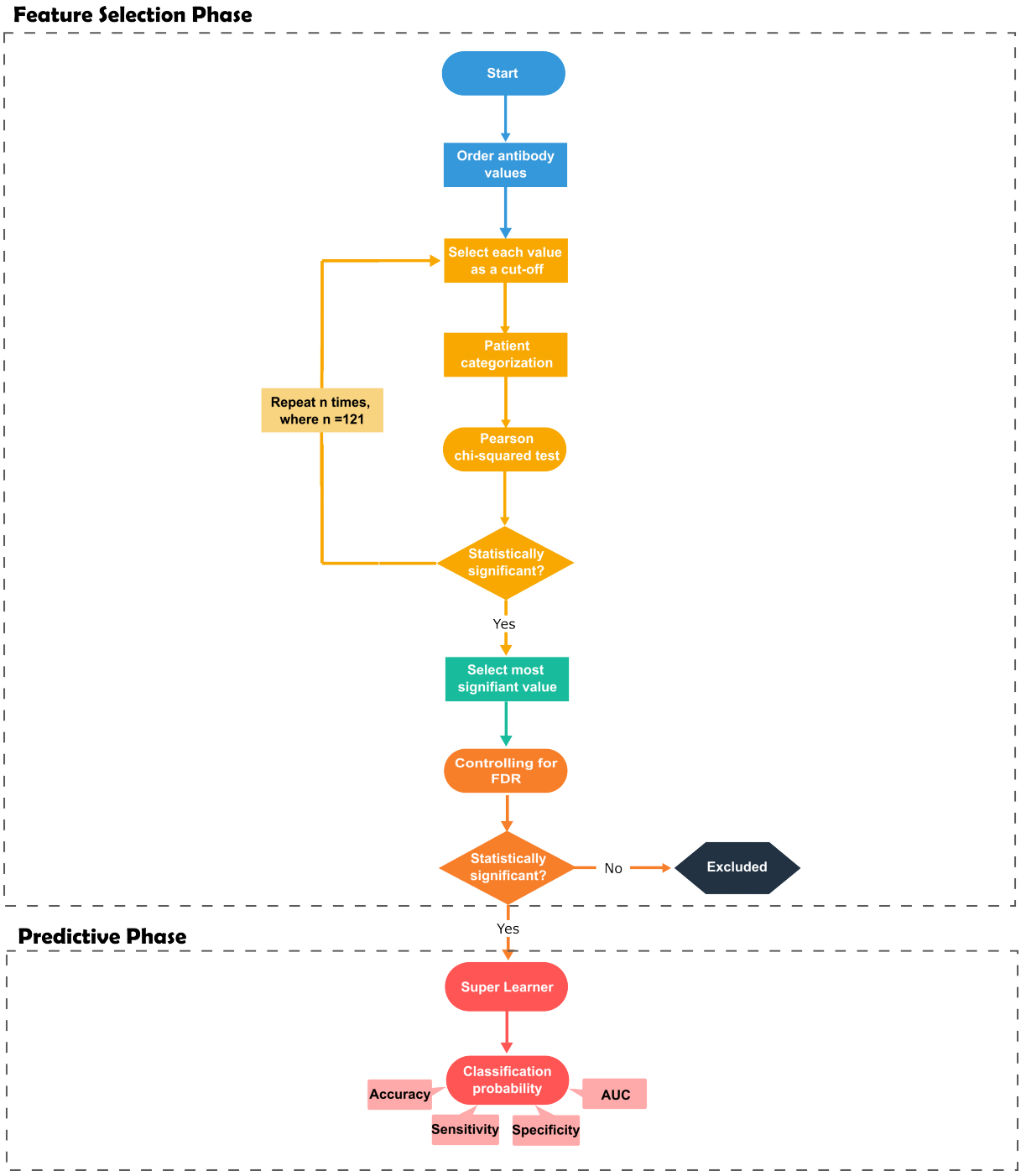


Figure 17: **Optimal data dichotomization for antibody selection.** The different steps of the analysis are displayed on the workflow using distinct colored shapes. Blue color identifies the beginning of the pipeline where the antibody values are sorted. Light orange identifies the loop for obtaining the χ^2 test *p-values* for each potential cut-off. Green indicates the selection of the most significant cut-off. Dark orange refers to the assessment of the statistical significance of the most significant cut-off after controlling for the False Discovery Rate (FDR) with the Benjamini-Yekutieli procedure. Red refers to the implementation of the Super Learner and the computation of the classification probability. Additional information is provided by the faded light orange and red colored shapes.

groups using the optimal cut-off by maximization of the χ^2 statistic (as described in the previous section). In the antibodies for which there was evidence of a single latent serological population antibody, we constructed two linear regression models using the antibody values as the response variable. The first model comprised only the intercept (i.e., not including any covariate), while the second model comprised the malaria protection status as the single covariate. We then computed the *p-value* of the Wilks' likelihood ratio test to compare the two models at the significance level of 5%. The rejection of the null hypothesis suggested statistically significant differences between the two models being compared. Finally, antibodies that could not be fitted by any of the above parametric models were analyzed using the Mann-Whitney-Wilcoxon test to compare the protected and susceptible groups.

5.3.6 Correction for multiple testing

In each antibody selection strategy, all the *p-values* obtained were adjusted to ensure a global false discovery rate (FDR) of 5%. This *p-value* adjustment was made via the Benjamini-Yekutieli procedure under a general dependence assumption between tests [30]. All antibodies whose adjusted *p-values* <0.05 were carried out to the next stage, the predictive analysis.

5.3.7 Predictive Stage

When analysing data resulting from each antibody selection strategy, we adopted a Super Learner (SL) approach to predict the malaria protection status of each individual [31, 32]. In general, this approach aims to estimate different classifiers whose individual predictions for each individual are combined into a pooled estimate via a weighted average calculated by cross-validation. To construct this pooled estimator, we used the following 5 classifiers for each set of antibodies selected: logistic regression model (LRM) with main effects only, RF, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and extreme gradient boosting (XGB). Note that the inclusion of RF in the SL model assembly algorithm allowed the comparison of the respective results with the previous one based on the same machine learning technique but using all the 36 antibodies as features. For the antibodies selected by optimal dichotomization antibody selection strategy, we did not include LDA and QDA in the SL algorithm because these classifiers are more appropriate for data containing quantitative predictors only. To assess the quality of the final predictions, we estimated the ROC curve and its area (AUC) [22, 33]. In addition, we calculated the confusion matrices where the rows and columns represented the predicted and the observed status of the individuals, respectively [34]. The predicted values in these confusion matrices were calculated using the point in the ROC curve that minimizes the distance to the point (0,1) related to the perfect classification of the individuals, here called ROC01 crite-

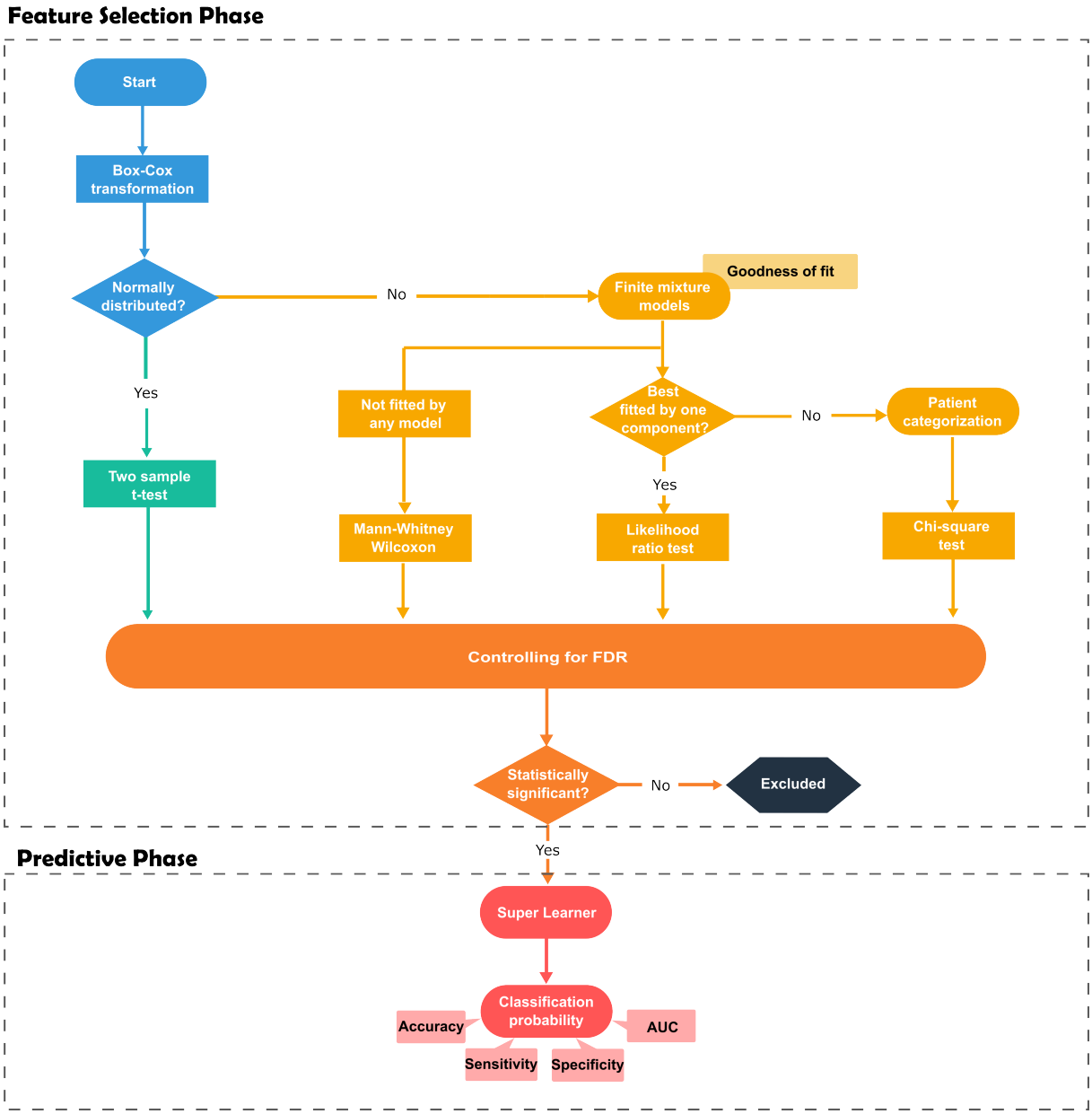


Figure 18: **Parametric antibody selection.** The different steps of the analysis are displayed on the workflow using distinct colored shapes. Blue color identifies the beginning of the pipeline where the normality assumption is verified after Box-Cox transformation. Green refers to the calculation of the t-test statistic for those antibodies for which the normality assumption was verified. Light orange refers to the implementation of the finite mixture models to those antibodies on which normality assumption failed and implementation of the different tests as according to the best fitted model, or failure to do so. Dark orange refers to the assessment of the statistical significance after correction for the FDR with the Benjamini-Yekutieli procedure and red to the implementation of the Super Learner and computation of the classification probability. Additional information is provided by the faded light orange and red colored shapes.

rion [35]. From the standpoint of constructing a fair classifier [36, 37] we also determined the predictive performance by the point in the ROC curve in which sensitivity (protected) and specificity (susceptible) were equal approximately [35]. This criterion is hereafter denoted as SpEqualSe criterion [35].

5.3.8 Statistical Software

All statistical analyses were implemented in the R [38] version 4.3.0 using the following packages: “AID” to perform Box-Cox transformation and to compute the respective Normality tests [39]; “caret” to construct the confusion matrices [23]; “doParallel” for parallel processing and faster run times [40]; “dplyr” to better manipulate the data [41]; “ggplot2” to plot the data [42]; “ggrepel" to avoid overlaid text on plots [43]; “lmtest” to perform the likelihood ratio test [44]; “MASS” for general analysis [45]; “mixsnsm” to estimate mixture models based on Skew-Normal and Skew-t distributions [46]; "OptimalCutpoints" to obtain the point in the ROC curve that minimizes of the distance to the point (0,1) [35]; “pROC” to estimate ROC curves [47]; “sn” to perform linear regression models based on Skew-Normal or Skew-t distributions for the residuals [48]; “SuperLeaner” to perform all the predictive analysis [31]; “tydir” to facilitate data manipulation [49].

5.4 Results

5.4.1 Preliminary analysis based on the Random Forest approach

Initially, an RF model was implemented using all 36 antibodies as features in order to replicate the results previously reported by Valleta and Recker [15]. We were able to reproduce the previously reported AUC of 0.68 (95% CI = (0.57;0.79))(Figure 19A). Looking at the feature importance values, we concluded that all except one of the 36 antibodies were required to achieve this predictive performance (Figure 19B). Nevertheless, a more thorough analysis of the feature’s importance values reveals that several features had very low importance values (below 20% importance) (Figure 19B). This led us to hypothesize that removing these features could improve the model’s performance. Therefore, three distinct *filter* strategies for feature selection were used.

5.4.2 Analysis based on the simple antibody selection approach

We first tested whether levels of each antibody were significantly different between susceptible and protected individuals using the Mann-Whitney-Wilcoxon test. According to this nonparametric test, 21 out of the 36 antibodies were found statistically significant before adjusting for multiple testing. This number dropped to 6 after controlling for an FDR of 5%: *msp2*, *msp4*, *msp10*, *eba175*, *msp7*, and *h103* (Figure 20A). This substantial reduction in the number of significant antibodies is likely to be explained by the positive cor-

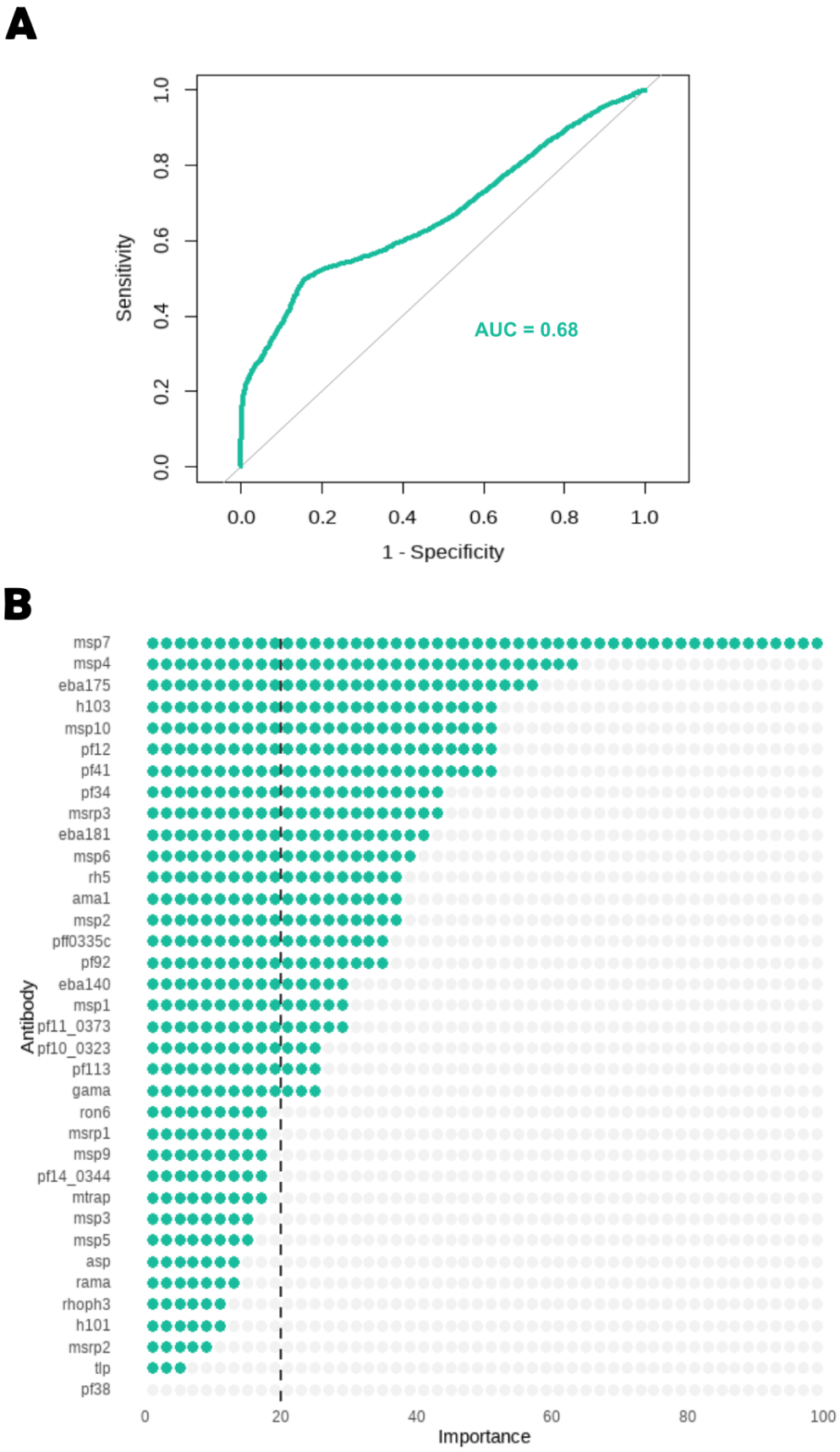


Figure 19: **Analysis of an RF using all the 36 antibodies as features.** A) ROC curve and its AUC; B) Estimated importance of each antibody in the RF

relation among different antibodies (average Spearman's correlation coefficient = 0.312; Figure 20B).

We then constructed a Super Learner classifier based on the data of these 6 antibodies. The average estimates for the AUC were 0.713, 0.703, 0.702, 0.729 and 0.728 using LRM, LDA, QDA, RF and XGB, respectively (Figure 20C). A closer examination of the RF's performance (AUC = 0.729) reveals an AUC increment over its performance prior to feature selection (Figure 19A). The average weights of these classifiers were 0.089, 0.506, 0.035, <0.001, and 0.370 in the final predictions, respectively. These weights implied an AUC of 0.719 (95% CI = [0.615, 0.824]) for the SL predictions. Moreover, the SL predictions had a sensitivity of 0.753 and a specificity of 0.625 according to the ROC01 criterion (Figure 20D). A higher number of protected individuals in the dataset could explain the fact that sensitivity was estimated at a higher value than specificity. To assess the final classifier without this potential selection bias, we determined the point at which the ROC sensitivity and specificity were similar and used it to obtain a fair classification (SpEqualSe criterion). The balanced sensitivity and specificity estimates were 0.630 and 0.625, respectively (Figure 20E).

5.4.3 Analysis based on the data dichotomization approach

In this analysis, we determined the optimal classification cut-off for each antibody according to the χ^2 statistics. The sensitivity estimates using these optimal cut-offs varied from 0.049 (*pf14_0344*) to 1 (*eba140*, *msrp3*), while the specificity varied from 0.100 (*msp9*) to 0.95 *pf11_0373*). The top 3 antibodies whose optimal cut-offs provided the sensitivity and specificity estimates closest to perfect classification (i.e., specificity = sensitivity = 1) were *msp7* (Se=0.852, Sp=0.600), *eba175* (Se = 0.827, Sp = 0.550), and msp2 (Se = 0.556, Sp = 0.800; Figure 21A).

There were 28 out of 36 antibodies whose proportions above the optimal cut-off were significantly different between protected and susceptible individuals at the 5% significance level (Table 2). The uncertainty around each optimal cut-off was highly heterogenous across these 28 antibodies. On the one extreme, the shortest 95% confidence for the optimal cut-off was obtained for the antibodies against ron6 (95% CI = [0.04;0.11]). On the other extreme, the widest 95% confidence for the optimal cut-off was obtained for the antibodies against *eba175* (95% CI = [0.10;1.81]). After controlling for an FDR of 5%, the number of statistically significant antibodies dropped to 20 (Figure 21B). The optimal dichotomization of these antibodies was used in the predictive analysis.

The AUC of the SL-based predictions was estimated at 0.801 (95% CI = (0.709, 0.892)) (Figure 21C), which showed an improvement from the previous analysis using a non-parametric antibody selection. The average AUC (and weights) estimates for each classifier were: LRM - 0.729 (<0.001), RF - 0.800 (0.973), and XGB - 0.714 (0.026). This result showed that, notwithstanding the reasonable AUC estimates for LRM and XGB, the final predictions

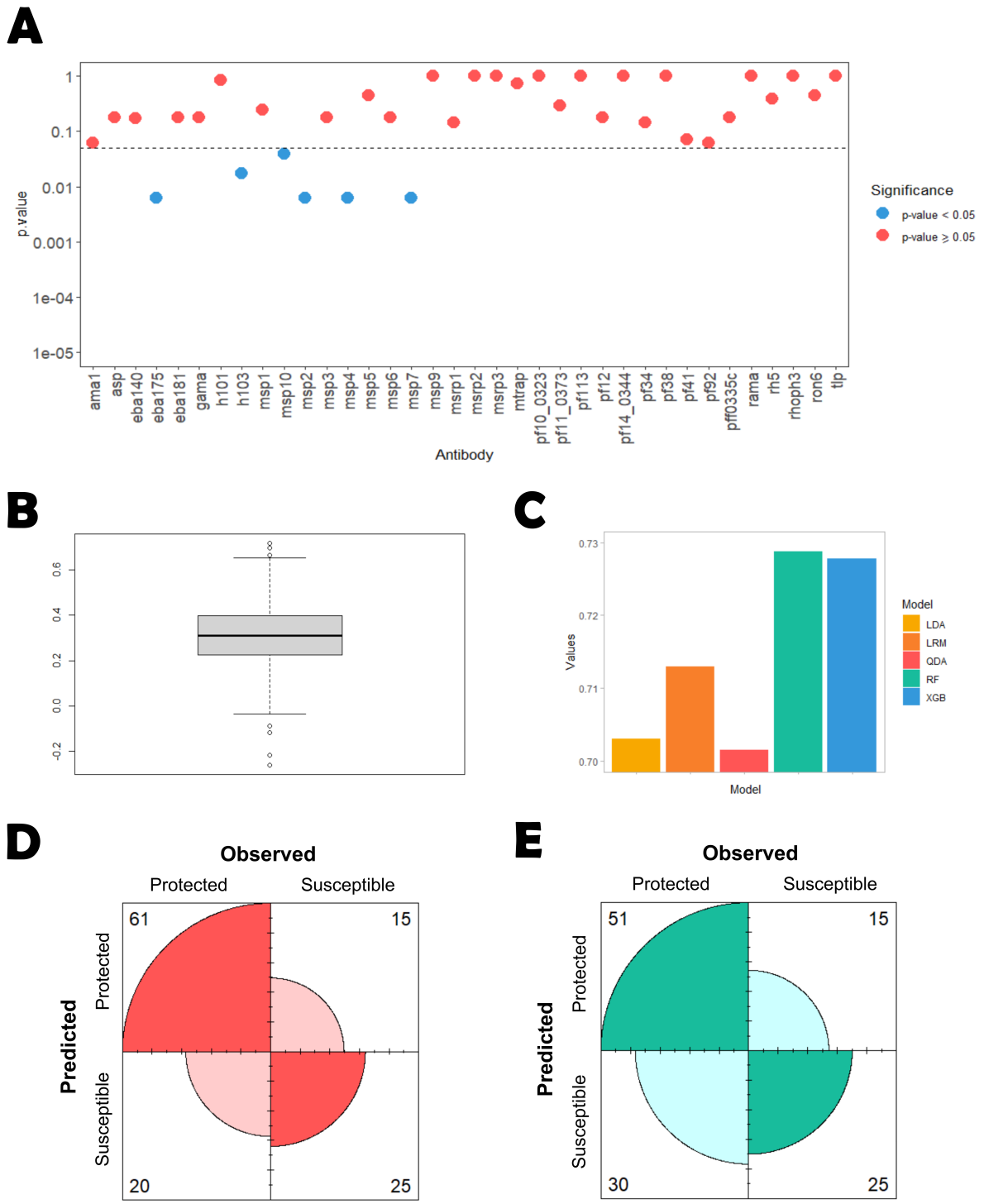


Figure 20: **Simple antibody selection results.** A) Statistical significance of each antibody according to Mann–Whitney-Wilcoxon where the *p-values* were adjusted for an FDR of 5%. B) Average Spearman's correlation concerning all the 36 antibodies. C) Average AUC estimated for each individual model embedded in the Super Learner. D) Confusion matrix of the predicted versus observed individual's classification derived from the Super Learner model using the ROC01 and E) SpEqualSe criterion.

were basically derived from the RF classifier. Not only that, but the RF's AUC also increased significantly when compared to implementation using all the variables, highlighting the value of feature selection. Moreover, note that LDA and QDA were not included in the SL algorithm, as they are more suitable for analyzing quantitative multivariate data.

According to the ROC01, the final sensitivity and specificity were estimated at 0.753 and 0.750, respectively. These estimates were identical for the SpEqualSe criterion. In conclusion, this analysis produced a combined classifier that exhibited an improved and better-balanced predictive performance than the previous one. However, this classifier had the disadvantage of including a higher number of antibodies comparing compared to the previous one (20 versus 6 antibodies).

5.4.4 Analysis based on the hybrid parametric/non-parametric approach

We first estimated the Box-Cox optimal data transformation and applied it to the antibody data. Then, we compared the protected and susceptible groups using the parametric t-tests for two independent samples. Our findings suggested that there were 6 antibodies whose data in each study group could be analysed by these tests after the Box-Cox transformation: *asp*, *pf11_0373*, *pf14_0344*, *pf34*, *rh5*, and *ron6* (Figure 22A); note that, at this stage, we did not adjust the *p-values* of the respective goodness-of-fit tests due to multiple testing, because such adjustment would increase the evidence for the null hypothesis of these tests. In these antibodies, the estimates for the parameter λ of the Box-Cox transformation varied from -3.8 (*ron6*) to -0.78 (*pf34*).

The estimates suggest that the logarithmic transformation would not be the best to reach a normal distribution. The best evidence for a Normal distribution was found for *pf34* with a *p-value* of 0.75 using the SW test (Figure 22A). The remaining 30 antibody data were then analysed by fitting finite mixture models based on Normal, Generalized T, Skew-normal, and Skew-T distributions; note that Normal and t distributions come as special cases of the latter probability distributions. For the statistical convenience of having these antibodies defined in terms of positive and negative values, we log-transformed the respective antibody data.

We found evidence that data from 7 antibodies could be described well by either Skew-Normal (*msp3* and *h103*) or Skew-t (*gama*, *h101*, *msrp2*, *msrp3*, and *pf10_0323*) distributions (Table 3). In this case, the comparison between study groups was made via regression models using these distributions for the errors. Except for the antibodies against *pf92* and *ama1*, data of the remaining antibodies were best described by a mixture of two Normal distributions (4 antibodies), two Skew-Normal distributions (16 antibodies) or two Skew-t distributions (1 antibody; see Table 3).

The best fit of these mixture models was obtained for the antibody against *pf113* using a two-component Normal mixture model (p = 0.73, Pearson's goodness-of-fit test; Table

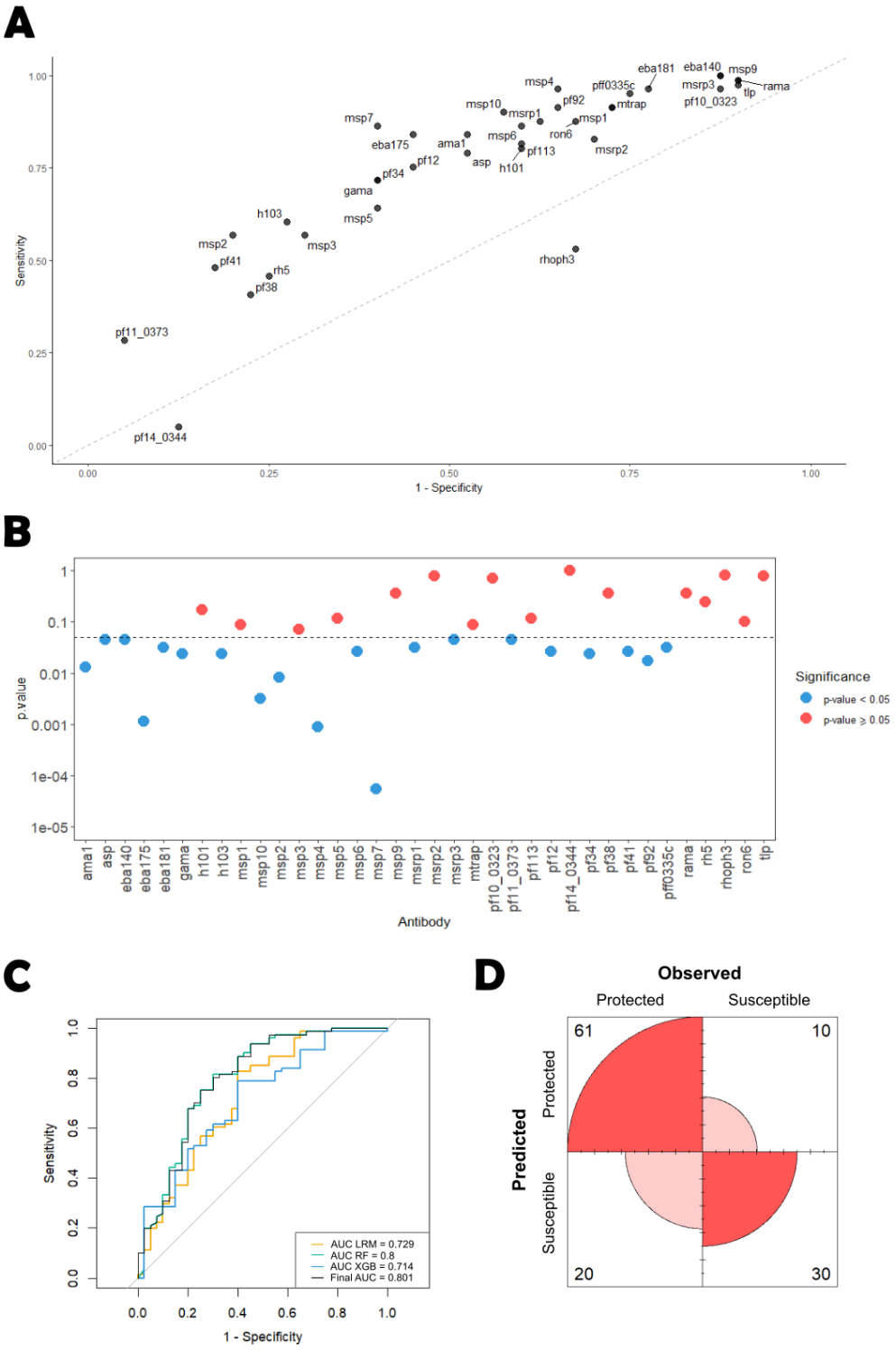


Figure 21: **Optimal data dichotomizations antibody selection results.** A) Sensitivity versus specificity plot for each antibody according to the cut-off that maximized the Pearson's χ^2 statistic. B) Statistical significance of each antibody following *p-value* correction using the Benjamini-Yekutieli procedure. C) AUCs for the individual models: Logistic regression (LRM), Random Forest (RF) and XGBoost (XGB) embedded in the Super Learner; and the overall AUC provided by the Super Learner. D) Confusion matrix of the predicted versus observed individual's classification derived from the Super Learner model using the ROC01/ SpEqualSe criterion (results were the same).

Table 2: **Results from the 28 antibodies deemed significant by the data dichotomization approach.** The antibody levels that maximized the separation between the susceptible and protected group of individuals (Cut-off) and the proportion of seropositive individuals for all (Total), Protected (Prt) and susceptible (Sus) children, respectively.

Antibody	*P-value*	Cut-off (95% CI)	Total	Prt	Sus
msp1	0.01	0.14 (0.04;0.99)	0.85	0.91	0.73
msp2	<0.01	0.07 (0.04;0.34)	0.45	0.57	0.20
msp4	<0.01	0.13 (0.10;1.36)	0.86	0.96	0.65
msp5	0.02	0.09 (0.06;0.23)	0.56	0.64	0.40
msp10	<0.01	0.25 (0.11;1.57)	0.79	0.90	0.58
pf12	<0.01	0.10 (0.07;0.45)	0.65	0.75	0.45
pf92	<0.01	0.11 (0.05;1.32)	0.83	0.91	0.65
pf34	<0.01	0.07 (0.05;0.15)	0.61	0.72	0.40
pf113	0.02	0.05 (0.04;0.13)	0.74	0.81	0.60
gama	<0.01	0.05 (0.04;0.11)	0.61	0.72	0.40
ama1	<0.01	0.16 (0.04;1.09)	0.74	0.84	0.53
eba175	<0.01	0.14 (0.10;1.81)	0.71	0.84	0.45
eba140	<0.01	0.11 (0.11;1.55)	0.96	1.00	0.88
eba181	<0.01	0.11 (0.09;1.46)	0.90	0.96	0.78
mtrap	0.01	0.05 (0.04;0.12)	0.85	0.91	0.73
asp	<0.01	0.08 (0.07;0.15)	0.70	0.79	0.53
msp3	0.01	0.08 (0.04;0.30)	0.48	0.57	0.30
msp6	<0.01	0.12 (0.10;0.32)	0.78	0.86	0.60
msp7	<0.001	0.24 (0.10;1.27)	0.71	0.86	0.40
msrp1	<0.01	0.05 (0.05;0.22)	0.79	0.88	0.63
msrp3	<0.01	0.04 (0.04;0.10)	0.96	1.00	0.88
h101	0.03	0.05 (0.04;0.11)	0.74	0.80	0.60
h103	<0.01	0.07 (0.04;0.24)	0.50	0.60	0.28
pf41	<0.01	0.12 (0.04;0.53)	0.38	0.48	0.18
pff0335c	<0.01	0.05 (0.04;0.35)	0.88	0.95	0.75
rh5	0.04	0.16 (0.09;0.25)	0.39	0.46	0.25
ron6	0.02	0.04 (0.04;0.11)	0.81	0.88	0.68
pf11_0373	<0.01	0.08 (0.05;0.14)	0.21	0.28	0.05

2). For these antibodies, we assumed the existence of a seronegative and a seropositive population. We dichotomized the respective data using the optimal cut-off by maximization of the χ^2 test statistic. Data of the antibodies against *pf92* and *ama1* could not be fitted by either Normal distribution after Box-Cox transformation or using the above mixture models. Therefore, we used the Mann-Whitney-Wilcoxon test as the last resort statistical test to compare the protected and susceptible groups. Thus, comparing the protected and susceptible groups using the different tests led to 25 significant antibodies before applying a multiple testing correction This number decreased to 16 after ensuring an FDR of 5%. These antibodies were found to be significant by the Wilks likelihood ratio test (*msp3*, *msrp3* and *h103*), the χ^2 test (*eba175*, *eba181*, *msp2*, *msp4*, *msp6*, *msp7*, *msp10*, *msrp1*, *pf12*, *pf41*, *pff0335c*) and the Mann-Whitney-Wilcoxon test (*pf92*, *ama1*) (Figure 22B). In the predictive analysis, data of each antibody were included in the SL approach according to the suggested scale by the antibody selection procedure: log-transformed data for antibodies coming from the Wilks' likelihood ratio test, dichotomized seropositive/seronegative data for antibodies coming from the χ^2 test, and the original scale for the pf92 and ama1-related antibodies coming from the Mann-Whitney-Wilcoxon test.

Before obtaining the combined predictions, we checked each individual classifier's performance. The average AUC were 0.756, 0.807, 0.768, 0.656 and 0.643 using LRM, RF, LDA, QDA, and XGB, respectively. Therefore, the best individual classifier was the RF. The average weights of these classifiers were 0.021, 0.912, 0.0132, 0.053, and 0 in the final predictions, respectively, resulting in an AUC of 0.79 CI = (0.7, 0.879)) According to the ROC01 criterion, the sensitivity and specificity were estimated at 0.703 and 0.750, respectively (Figure 22C). Moreover, based on the ROC curve, the best balance between these quantities was obtained for a sensitivity and a specificity of 0.716 and 0.725, respectively (Figure 22D).

5.5 Discussion

Multi-sera data, where thousands of antibody targets are simultaneously measured, can increase the chance of discovering the antibodies responsible for natural protection against malaria or the antibodies that can be used to detect previously exposed individuals to malaria parasites [50, 51, 52]. Nonetheless, this type of data brings novel challenges [53, 54]. One of the main drawbacks when dealing with this type of data is the difficulty of identifying the relevant features for the task at hand. Among the thousands of features screened, most will be irrelevant or redundant and will negatively impact the predictive ability of a predictive model [55]. Not only that, trying to fit a predictive model to these many features increases the computational complexity and cost, reduces the model generalization ability, and affects the explainability of the model [54]. To overcome these limitations, feature selection strategies have been proposed, where the aim is to identify and remove all the irrelevant features so that the learning algorithm focuses only on those fea-

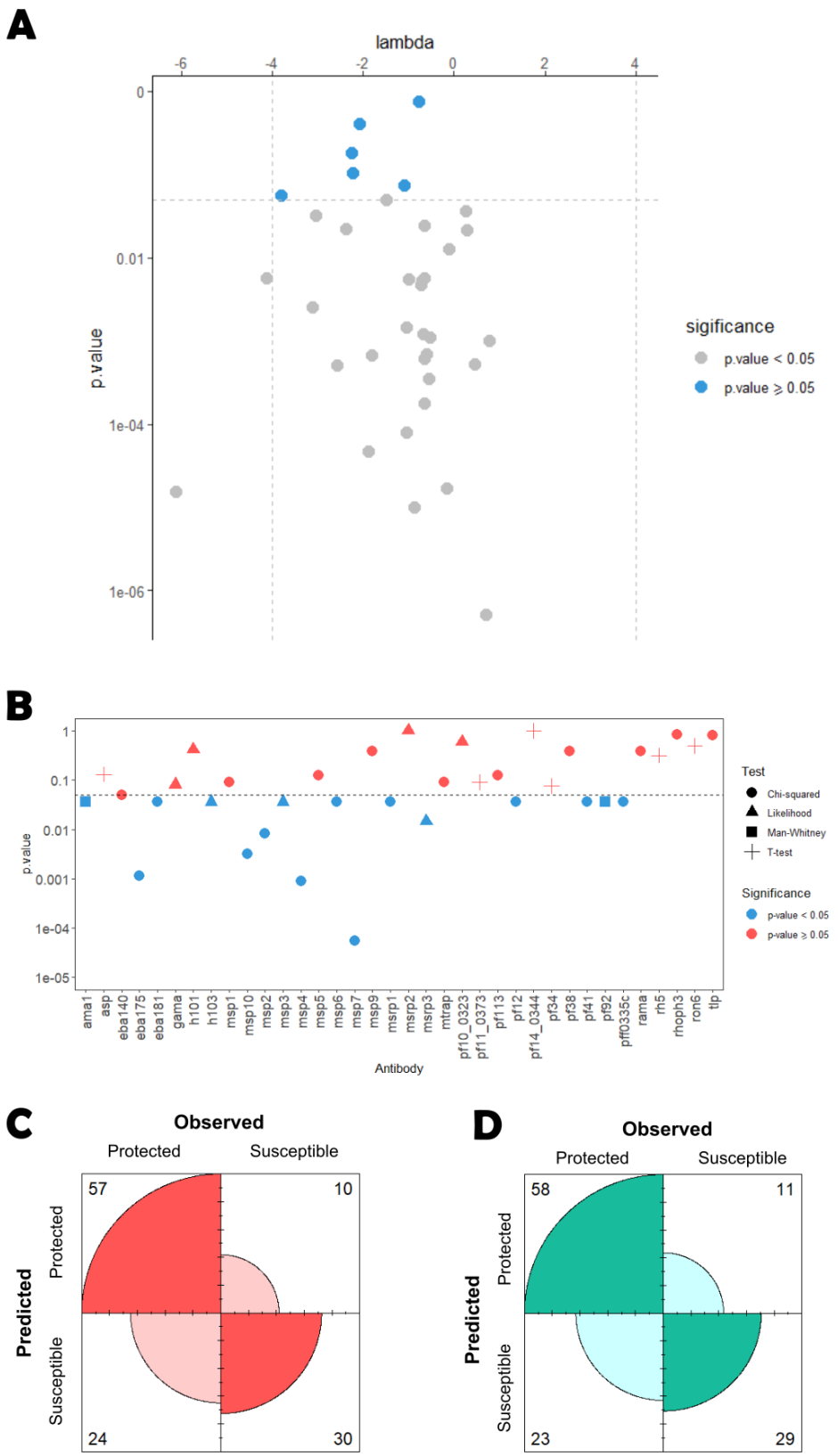


Figure 22: **Hybrid antibody selection results.** A) *P-values* for the SW normality test (y-axis) after Box-Cox transformation with the respective lambda (x-axis). B) Statistical significance of each antibody following *p-value* correction using the Benjamini-Yekutieli procedure. D) Confusion matrix of the predicted versus observed individual's classification derived from the Super Learner model using the ROC01 and D) SpEqualSe criterion.

Table 3: **Mixture Model results** Results from the analysis of 28 antibodies based on finite mixture models where AIC and GOF denote the Akaike's information criterion and Pearson's goodness-of-fit test, respectively.

Antibody	Best Mixture Model	Nº Components	AiC	*P-value*(GOF)
eba140	Skew Normal	2	23.92	0.32
eba175	Skew Normal	2	33.29	0.03
eba181	Skew Normal	2	42.9	0.03
gama	Skew-t	1	-272.19	0.24
h101	Skew-t	1	-230.91	0.33
h103	Skew Normal	1	-41.91	0.72
msp1	Skew Normal	2	23.35	0.26
msp10	Normal	2	71.52	0.07
msp2	Skew Normal	2	-24.09	0.43
msp3	Skew Normal	1	1.46	0.32
msp4	Skew Normal	2	76.23	0.04
msp5	Normal	2	-71.25	0.33
msp6	Normal	2	-168.02	0.35
msp7	Skew Normal	2	46.11	0.16
msp9	Skew Normal	2	-10.75	0.53
msrp1	Skew-t	2	-89.1	0.06
msrp2	Skew-t	1	-122.32	0.12
msrp3	Skew-t	1	-283.83	0.02
mtrap	Skew Normal	2	-213.58	0.13
pf10_0323	Skew-t	1	-344.51	0.62
pf113	Normal	2	-139,5	0,73
pf12	Skew Normal	2	-33,29	0,24
pf38	Skew Normal	2	99,41	0.05
pf41	Skew Normal	2	35.96	0.10
pff0335c	Skew Normal	2	4.83	0.04
rama	Skew Normal	2	-153.54	0.32
rhoph3	Skew Normal	2	-152.73	0.02
tlp	Skew Normal	2	-426.93	0.02

tures of the training data useful for prediction [53]. This leads to not only a simpler interpretability, as when a small number of features is selected, their biological relationship with the target disease is more easily identified, but also a lower computational cost and increased accuracy, stemmed from reducing the chance of overfitting [54, 56]. Therefore, feature selection before the implementation of a predictive model is strongly advocated [57]. Amongst the different feature selection approaches, here we opted for the use of *filter* methods [53, 56, 57]. These rely on statistical measures (e.g *p-value*, correlation coefficient), and their application precedes the predictive phase, thus being independent of any predictive model [56, 57]. For this reason, they are usually very fast to implement. Here we will discuss the advantages and drawbacks of the distinct *filter* methods employed in each proposed methodology. The simple approach relying on the Mann-Whitney Wilcoxon test for feature selection is the most scalable approach for larger datasets among the ones here proposed. It is the most straightforward and fastest approach to implement, making it a very good tool for those looking for a low-complexity model when conducting a classification task. Moreover, given its ranking intrinsic nature, this strategy represents the best option to achieve reproducible results [24]. Nevertheless, its low statistical and computational complexity comes at a cost since this oversimplistic feature selection approach led to a lower predictive performance when compared to the other strategies, as demonstrated in this study.

The best predictive performance was obtained from the feature selection strategy based on data dichotomization. This performance contradicts the general expectation of losing statistical information every time one analyses dichotomized data [58, 59, 60]. However, in serological data analysis, one typically expects the existence of a single latent seronegative population and a single latent seropositive population in a given antibody distribution [28, 61, 62]. These populations can be conceptually interpreted as noise and signal of genuine antibody responses to a given antigen, respectively. In this scenario, data dichotomization is a natural way to separate noise from a true biological signal. In other words, data dichotomization comes naturally if one intends to eliminate the effect of noise in the respective data analysis. The original study reported that the seroprevalence varied from 5 to 96% in the dataset analyzed [9]. Hence, all the antibodies contained some degree of noise in the respective data and the presence of such a noise across multiple antibodies is a likely explanation for the best performance of this feature selection method in the dataset analyzed. In the same line of thought, we speculate that a better predictive performance using this feature selection strategy could not be achieved due to a possible overlap between seronegative and seropositive populations. The detailed exploration of this point, although interesting, was beyond the scope of the present study. The data dichotomization approach also showed a great practical advantage due to its simple computational implementation. However, the performance of this approach might be dependent on the uncertainty around the optimal cut-off for each antibody. As demonstrated by our analysis, this

uncertainty varied substantially from one antibody to another. Such a variation is likely to be explained by not only a relatively small sample size of the original study but also the ratio between the proportions of susceptible and resistant individuals. Thus, the cut-offs here reported should be used with caution. Ideally, they should be confirmed with a larger data set where there is a good balance between susceptible and resistant individuals.

Notwithstanding being more complex from a statistical standpoint, our hybrid approach provides a more comprehensive analysis of the data. In this approach, feature selection is made on the basis of data transformation and dichotomization via mixture modelling, thus accommodating different data patterns. However, this feature selection strategy is expected to increase the computational time dramatically as the number of antibodies under analysis increases. The computational implementation in user-friendly packages is also not trivial in relation to the other feature strategies applied in this study. Finally, this feature selection strategy is based on complex statistical models such as finite mixture models related to Skew-Normal distributions. In this scenario, this strategy seems less appealing to the malaria research community where, despite the efforts to improve mathematical modelling capacity, the availability of qualified staff with statistical and machine learning skills remains scarce. Therefore, the use of simple *filter* methods seems a more viable solution at the moment, especially, when it comes to analyzing data featuring thousands of antibodies. Such a case is seen in Proietti et al. [7] where antibodies with a *p-value* <0.01 for the univariate logistic regression were selected after Bonferroni correction followed by sparse partial least squares discriminant analysis (sPLS-DA) and Support Vector Machine (SVM). Another example is the use of the Spearman's correlation coefficient to remove highly correlated antibodies prior to the implementation of the RF presented by Valletta and Recker [15].

A significant disadvantage of *filter* methods is the inability to detect complex relations between multiple features and the outcome of interest, which generally translates into poorer results in the predictive phase [56, 57]. Thus *wrappers* or *embedded* methods are more appealing. Wrappers are created around a particular classifier and rely on the classifier's information concerning feature relevance [56, 57]. For this reason, the computational effort they require is usually significant, becoming unusable when thousands of features are considered. Therefore, wrappers are often avoided, and their implementation for feature selection in malaria is scarce [8]. A more attractive approach are *embedded* methods that use the core of a classifier to establish a criterion to rank features [53, 56]. *Embedded* algorithms perform feature selection during the classifier training procedure while optimizing the feature set used to achieve the best accuracy. Therefore, they are less computationally costly than wrappers while still dealing with the complex interactions between multiple features and the outcome [53, 56]. Examples of *embedded* feature selection methods intending to unveil antibody immune signatures in malaria are described in the literature. Aitken et al.[63] used an elastic net-regularized logistic regression for antibody selection

followed by a partial least squares discriminant analysis to find a minimal set of antibodies that accurately classified the individuals under analysis. Helb et al. [13] used a hierarchical criterion for feature selection, where a combination of *embedded* and *filter* methods was performed before the implementation of a Super Learner for predicting past exposure to malaria. Here, the Least Absolute Shrinkage and Selection Operator (LASSO) regression was initially used to select one third of the responses. Then, using variable importance measures from Random Forest regression they iteratively selected the best responses which were than ranked by Spearman correlation's *p-values* [13]. Although not implemented here due to the relatively small number of features, we envision that *embedded* feature selection approaches will be more useful in datasets in where the number of antibody responses exceeds the number of observations, as already seen in a study from Mali [14]. A forthcoming research study will investigate this solution and its impact on variable selection.

Alternative approaches to feature selection techniques for identifying the optimal antibody combinations for the task at hand have also been proposed [10, 12]. These rely on simulated annealing algorithms that efficiently explore the vast space of feature combinations and thus identify the optimal feature combination solution given a fixed number of features defined by the user [10]. Whether this approach is preferable over feature selection techniques is an interesting discussion topic for future work.

Concerning our predictive analysis, we adopted a SL approach. The reasoning for this relied on the fact that by combining the individual predictions of each individual classifier, the SL avoids the bias created by manually choosing the best-fitting model procedure and often provides better results than each individual classifier [31, 32]. This is achieved through cross-validation, where the algorithm selects the weights used to combine the initial set of candidate models and then makes predictions based on the weighted average of estimates from each model [31, 32]. However, this was not always the case, as the RF classifier alone tended to provide better predictions than the SL. Given that RF is an *embedded* method, it performs feature selection during the classifier training procedure and thus we speculated that the removal of further features could be behind this increased performance [20, 64]. Nevertheless, our validation analysis revealed that regardless of the strategy chosen for feature selection, nearly all features were important for classification purposes. This highlights the *filter* strategy's ability to identify the most relevant features. This avoided any additional feature removal by the embedded model, including the Super Learner classifier allowing for more consistent results. However, this issue should be addressed in cases where the Super Learner comprising embedded methods is implemented after a feature selection phase, such as done in Helb et al [13], as further feature removal might occur without the user's knowledge which may affect the interpretability of the results. Hence the slight decrease in the SL performance is expected to be explained by the SL attempt to correct for a possible overfitting to the data when using RF. In this sense, these results should raise awareness concerning analysis where only RF is considered for predic-

tive purposes, as it may lead to overfitting. Thus, the implementation of techniques such as the SL may provide more consensual results across the classifiers chosen for the predictive stage.

Comparing our results with the previous ones by Valletta and Recker [15] revealed an increase in the prediction ability of up to 14% in the best-case scenario. Not only that, but feature selection also increase the RF's predictive ability compared to the one obtained by the same authors, an increase that ranged from 5% in the worst-case scenario (simple antibody selection) to 12% in the best-case scenario (data dichotomization selection). These results further emphasize the impact of feature selection prior to predictive analysis. On the one hand, this step removes antibody responses with negligible effect on clinical malaria. On the other hand, this stage decreases the number of features allowing for a more thorough feature analysis increasing the chance of finding the right transformation and dichotomization for each antibody response.

Concerning the antibodies identified, we found that the antibody responses against different Merozoite Surface Proteins (MSPs) were consistently selected across the different feature selection strategies. These proteins are expressed at the parasite surface, thus, providing promising targets for malaria immunity because they are repeatedly and directly exposed to the host humoral immune system [7, 8]. In particular, *msp2* has been associated with protection from clinical malaria in many studies and even suggested as a vaccine candidate [9, 10, 11, 12]. For example, *msp2* has been strongly associated with protection against clinical malaria in two independent cohorts of Kenyan children [13]. *Msp4* has also been reported to have a protective effect in Kenyan children [4, 18]. Additionally, high antibody levels against *msp4* constructs have been associated with reduced morbidity in a Senegalese community [16]. *Msp7* protection against malaria has also already been identified in the Kenyan population [18, 17]. Moreover, panels of antibodies comprising *msp7* have been associated with clinical protection against malaria in Kilifi, a rural district along the Kenyan coast [14]. In the same article, high antibody levels against the Erythrocyte-binding antigen-175 (*eba175*) antigen were also associated with protection from clinical malaria in children [14]. Moreover, *eba175* is associated with protection from symptomatic malaria, as demonstrated in Papua New Guinean children [15]. These findings corroborate the ability of our methodologies to identify relevant antibodies associated with protection to malaria. Interestingly, however, *msp10* and *h103* have not been associated with clinical malaria protection. This evidence suggests that there are antibodies associated with protection against clinical malaria that have not yet been identified. Nevertheless, further studies are necessary to validate our findings. Finally, none of our feature selection metrics selected *msp1*, an immune response commonly associated with malaria protection and often referred to as a potential vaccine candidate. Similar findings have been reported in other studies, where *msp1* has been described to show low or no associations with exposure or protection to clinical malaria [13, 15]. These inconsistent findings further suggest

the need for constructing robust strategies for feature selection that could help to increase reproducibility among studies.

At this moment, the pipelines are implemented in the free R software whose scripts are publicly available for consultation and improvement. However, current implementation of the pipelines is not in the form of a stand-alone and easy-to-use package. The respective adaptation to other datasets or the deployment of the tools here developed to malaria endemic countries might require the intervention of R experts to modify the available scripts. The requirement of this specific expertise might limit the applicability of these computational tools in many malaria-endemic regions with poor human resources. Therefore, setting the computational implementation of these and other tools as a top priority is likely to help in the clinic and contribute to the development of new therapeutics and a better malaria management and control.

5.6 Conclusion

In summary, we have implemented feature selection strategies to analyze multiple antibody data. These were developed with the idea of coupling classical, traditional statistical techniques for variable selection with popular machine learning techniques for predictive analysis. Considering the transformation of each antibody data individually these strategies represent a more flexible approach to accommodate different data patterns than those commonly described in the literature. Overall, these methodologies led to an improved classification over previous analysis based on the use of the RF alone, highlighting their potential to integrate future multi-sera pipelines.

Acknowledgements
Not applicable

Authors' contributions
AF and NS designed the work; AF and MS conducted the analysis; AF and NS interpreted the data; AF, PB, CC and NShave drafted or substantively revised the manuscript. All authors approved the final version of the manuscript.

Funding
AF received a PhD fellowship by FCT - Fundação para a Ciência e Tecnologia, Portugal (grant ref. SFRH/BD/147629/2019). AF, CC and NS were partially financed by national funds through FCT - Fundação para a Ciência e Tecnologia, Portugal(grant ref. UIDB/00006/2020). NS was also received funding from the Polish National Agency for Academic Exchange(grant ref.: PPN/ULM/2020/1/00069/U/00001).

Availability of data and materials The datasets used and/or analyzed during the current study are available in this published article: Valletta, J.J.& Recker,M. Identification of immune signatures predictive of clinical protection from malaria. PLoS Comput Biol 13, e1005 812(2017). https://doi.org/10.1371/journal.pcbi.1005812. The R scripts generated are freely available in the followingGitHubaddress:Immune-Stats(https://github.com/Publications/Fonseca_etal.).

Declarations

Ethics approval and consent to participate

This study is based on publicly available data.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 14 July 2023 Accepted: 16 January 2024
Published online: 25 January 2024

References

[1] Kathryn L Kellar, Rizwan R Kalwar, Kimberly A Dubois, Dennis Crouse, William D Chafin, and Beth-Ellen Kane. Multiplexed fluorescent bead-based immunoassays for quantitation of human cytokines in serum and culture supernatants. *Cytometry: The Journal of the International Society for Analytical Cytology*, 45(1):27–36, 2001.

[2] Takafumi Tsuboi, Satoru Takeo, Hideyuki Iriko, Ling Jin, Masateru Tsuchimochi, Shusaku Matsuda, Eun-Taek Han, Hitoshi Otsuki, Osamu Kaneko, Jetsumon Sattabongkot, et al. Wheat germ cell-free system-based production of malaria proteins for discovery of novel vaccine candidates. *Infection and immunity*, 76(4):1702–1708, 2008.

[3] Itziar Ubillos, Joseph J Campo, Alfons Jiménez, and Carlota Dobaño. Development of a high-throughput flexible quantitative suspension array assay for igg against multiple plasmodium falciparum antigens. *Malaria Journal*, 17(1):1–15, 2018.

[4] Gerald KK Cham, Jonathan Kurtis, John Lusingu, Thor G Theander, Anja TR Jensen, and Louise Turner. A semi-automated multiplex high-throughput assay for measur-

ing igg antibodies against plasmodium falciparum erythrocyte membrane protein 1 (pfemp1) domains in small volumes of plasma. *Malaria journal*, 7(1):1–8, 2008.

[5] Bernard N Kanoi, Eizo Takashima, Masayuki Morita, Michael T White, Nirianne MQ Palacpac, Edward H Ntege, Betty Balikagala, Adoke Yeka, Thomas G Egwang, Toshihiro Horii, et al. Antibody profiles to wheat germ cell-free system synthesized plasmodium falciparum proteins correlate with protection from symptomatic malaria in uganda. *Vaccine*, 35(6):873–881, 2017.

[6] Bernard N Kanoi, Hikaru Nagaoka, Michael T White, Masayuki Morita, Nirianne MQ Palacpac, Edward H Ntege, Betty Balikagala, Adoke Yeka, Thomas G Egwang, Toshihiro Horii, et al. Global repertoire of human antibodies against plasmodium falciparum rifins, surfins, and stevors in a malaria exposed population. *Frontiers in Immunology*, 11:893, 2020.

[7] Carla Proietti, Lutz Krause, Angela Trieu, Daniel Dodoo, Ben Gyan, Kwadwo A Koram, William O Rogers, Thomas L Richie, Peter D Crompton, Philip L Felgner, et al. Immune signature against plasmodium falciparum antigens predicts clinical immunity in distinct malaria endemic communities. *Molecular & Cellular Proteomics*, 19(1):101–113, 2020.

[8] Faith H Osier, Margaret J Mackinnon, Cécile Crosnier, Gregory Fegan, Gathoni Kamuyu, Madushi Wanaguru, Edna Ogada, Brian McDade, Julian C Rayner, Gavin J Wright, et al. New antigens for a multicomponent blood-stage malaria vaccine. *Science translational medicine*, 6(247):247ra102–247ra102, 2014.

[9] Faith HA Osier, Gregory Fegan, Spencer D Polley, Linda Murungi, Federica Verra, Kevin KA Tetteh, Brett Lowe, Tabitha Mwangi, Peter C Bull, Alan W Thomas, et al. Breadth and magnitude of antibody responses to multiple plasmodium falciparum merozoite antigens are associated with protection from clinical malaria. *Infection and immunity*, 76(5):2240–2248, 2008.

[10] Camila Tenorio França, Michael T White, Wen-Qiang He, Jessica B Hostetler, Jessica Brewster, Gabriel Frato, Indu Malhotra, Jakub Gruszczyk, Christele Huon, Enmoore Lin, et al. Identification of highly-protective combinations of plasmodium vivax recombinant proteins for vaccine development. *Elife*, 6:e28673, 2017.

[11] Lotus L Van den Hoogen, Gillian Stresman, Jacquelin Présumé, Ithamare Romilus, Gina Mondélus, Tamara Elismé, Alexandre Existe, Karen ES Hamre, Ruth A Ashton, Thomas Druetz, et al. Selection of antibody responses associated with plasmodium falciparum infections in the context of malaria elimination. *Frontiers in immunology*, 11:928, 2020.

[12] Rhea J Longley, Michael T White, Eizo Takashima, Jessica Brewster, Masayuki Morita, Matthias Harbers, Thomas Obadia, Leanne J Robinson, Fumie Matsuura, Zoe SJ Liu, et al. Development and validation of serological markers for detecting recent plasmodium vivax infection. *Nature medicine*, 26(5):741–749, 2020.

[13] Danica A Helb, Kevin KA Tetteh, Philip L Felgner, Jeff Skinner, Alan Hubbard, Emmanuel Arinaitwe, Harriet Mayanja-Kizza, Isaac Ssewanyana, Moses R Kamya, James G Beeson, et al. Novel serologic biomarkers provide accurate estimates of recent plasmodium falciparum exposure for individuals and communities. *Proceedings of the National Academy of Sciences*, 112(32):E4438–E4447, 2015.

[14] Peter D Crompton, Matthew A Kayala, Boubacar Traore, Kassoum Kayentao, Aissata Ongoiba, Greta E Weiss, Douglas M Molina, Chad R Burk, Michael Waisberg, Algis Jasinskas, et al. A prospective analysis of the ab response to plasmodium falciparum before and after a malaria season by protein microarray. *Proceedings of the National Academy of Sciences*, 107(15):6958–6963, 2010.

[15] John Joseph Valletta and Mario Recker. Identification of immune signatures predictive of clinical protection from malaria. *PLoS computational biology*, 13(10):e1005812, 2017.

[16] Lindsey Wu, Tom Hall, Isaac Ssewanyana, Tate Oulton, Catriona Patterson, Hristina Vasileva, Susheel Singh, Muna Affara, Julia Mwesigwa, Simon Correa, et al. Optimisation and standardisation of a multiplex immunoassay of diverse plasmodium falciparum antigens to assess changes in malaria transmission using sero-epidemiology. *Wellcome open research*, 4, 2019.

[17] Elena Ambrosino, Chloé Dumoulin, Eve Orlandi-Pradines, Franck Remoue, Aissatou Toure-Baldé, Adama Tall, Jean Biram Sarr, Anne Poinsignon, Cheikh Sokhna, Karine Puget, et al. A multiplex assay for the simultaneous detection of antibodies against 15 plasmodium falciparum and anopheles gambiae saliva antigens. *Malaria journal*, 9(1):1–12, 2010.

[18] Lotus L van den Hoogen, Jacquelin Présumé, Ithamare Romilus, Gina Mondélus, Tamara Elismé, Nuno Sepúlveda, Gillian Stresman, Thomas Druetz, Ruth A Ashton, Vena Joseph, et al. Quality control of multiplex antibody detection in samples from large-scale surveys: the example of malaria in haiti. *Scientific reports*, 10(1):1135, 2020.

[19] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[20] FY Ahmed, Yasir Hassan Ali, and Siti Mariyam Shamsuddin. Using k-fold cross validation proposed models for spikeprop learning enhancements. *International Journal of Engineering & Technology*, 7(4.11):145–151, 2018.

[21] Marvin N Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in c++ and r. *arXiv preprint arXiv:1508.04409*, 2015.

[22] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[23] Max Kuhn. Caret: classification and regression training. *Astrophysics Source Code Library*, pages ascl–1505, 2015.

[24] Nadim Nachar et al. The mann-whitney u: A test for assessing whether two independent samples come from the same distribution. *Tutorials in quantitative Methods for Psychology*, 4(1):13–20, 2008.

[25] Tiago Dias Domingues, Anna D Grabowska, Ji-Sook Lee, Jose Ameijeiras-Alonso, Francisco Westermeier, Carmen Scheibenbogen, Jacqueline M Cliff, Luis Nacul, Eliana M Lacerda, Helena Mouriño, et al. Herpesviruses serology distinguishes different subgroups of patients from the united kingdom myalgic encephalomyelitis/chronic fatigue syndrome biobank. *Frontiers in medicine*, 8:686736, 2021.

[26] Katarina Tengvall, Jesse Huang, Cecilia Hellström, Patrick Kammer, Martin Biström, Burcu Ayoglu, Izaura Lima Bomfim, Pernilla Stridh, Julia Butt, Nicole Brenner, et al. Molecular mimicry between anoctamin 2 and epstein-barr virus nuclear antigen 1 associates with multiple sclerosis risk. *Proceedings of the National Academy of Sciences*, 116(34):16955–16960, 2019.

[27] Özgür Asar, Ozlem Ilk, and Osman Dag. Estimating box-cox power transformation parameter via goodness-of-fit tests. *Communications in Statistics-Simulation and Computation*, 46(1):91–105, 2017.

[28] Nuno Sepúlveda, Gillian Stresman, Michael T White, Chris J Drakeley, et al. Current mathematical models for analyzing anti-malarial antibody data with an eye to malaria elimination and eradication. *Journal of immunology research*, 2015, 2015.

[29] Tiago Dias Domingues, Helena Mouriño, and Nuno Sepúlveda. Analysis of antibody data using finite mixture models based on scale mixtures of skew-normal distributions. *medRxiv*, pages 2021–03, 2021.

[30] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

[31] Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

[32] Eric C Polley and M SuperLearner van der Laan. super learner prediction: R package, version 2.0-21, 2017.

[33] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

[34] Ivo Düntsch and Günther Gediga. Confusion matrices and rough set data analysis. In *Journal of Physics: Conference Series*, volume 1229, page 012055. IOP Publishing, 2019.

[35] Mónica López-Ratón, María Xosé Rodríguez-Álvarez, Carmen Cadarso-Suárez, and Francisco Gude-Sampedro. Optimalcutpoints: an r package for selecting optimal cutpoints in diagnostic tests. *Journal of statistical software*, 61:1–36, 2014.

[36] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

[37] Xiaomeng Wang, Yishi Zhang, and Ruilin Zhu. A brief review on algorithmic fairness. *Management System Engineering*, 1(1):7, 2022.

[38] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[39] Osman Dag and Ozlem Ilk. An algorithm for estimating box–cox transformation parameter in anova. *Communications in Statistics-Simulation and Computation*, 46(8):6424–6435, 2017.

[40] Microsoft Corporation and S Weston. doparallel: Foreach parallel adaptor for the "parallel" package. *r package version 1.0. 16*, 2020.

[41] Hadley Wickham, Romain François, Lionel Henry, Kirill Müller, et al. dplyr: A grammar of data manipulation. r package version 0.7. 6. *Computer software]. https://CRAN. R-project. org/package= dplyr*, 2018.

[42] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[43] Kamil Slowikowski, Alicia Schep, Sean Hughes, Saulius Lukauskas, Jean-Olivier Irisson, Zhian N Kamvar, Thompson Ryan, Dervieux Christophe, Yutani Hiroaki, and Pierre Gramme. Package ggrepel. *Automatically position non-overlapping text labels with 'ggplot2*, 2018.

[44] Torsten Hothorn, Achim Zeileis, Richard W Farebrother, Clint Cummins, Giovanni Millo, David Mitchell, and Maintainer Achim Zeileis. Package 'lmtest'. *Testing linear*

regression models. https://cran. r-project. org/web/packages/lmtest/lmtest. pdf. Accessed, 6, 2015.

[45] W N Venables and Brian D Ripley. *Modern applied statistics with S*. Statistics and Computing. Springer, New York, 2010.

[46] Marcos Oliveira Prates, Victor Hugo Lachos, and Celso Rômulo Barbosa Cabral. mixsmsn: Fitting finite mixture of scale mixture of skew-normal distributions. *Journal of Statistical Software*, 54:1–20, 2013.

[47] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC bioinformatics*, 12(1):1–8, 2011.

[48] Adelchi Azzalini. sn: The skew-normal and related distributions such as the skew-t and the sun. *R package version*, 2(0), 2022.

[49] Hadley Wickham, Lionel Henry, et al. Tidyr: Tidy messy data. *R package version*, 1(2):397, 2020.

[50] Michelle J Boyle, Linda Reiling, Faith H Osier, and Freya JI Fowkes. Recent insights into humoral immunity targeting plasmodium falciparum and plasmodium vivax malaria. *International journal for parasitology*, 47(2-3):99–104, 2017.

[51] Will JR Stone, Joseph J Campo, André Lin Ouédraogo, Lisette Meerstein-Kessel, Isabelle Morlais, Dari Da, Anna Cohuet, Sandrine Nsango, Colin J Sutherland, Marga van de Vegte-Bolmer, et al. Unravelling the immune signature of plasmodium falciparum transmission-reducing immunity. *Nature communications*, 9(1):558, 2018.

[52] Tate Oulton, Joshua Obiero, Isabel Rodriguez, Isaac Ssewanyana, Rebecca A Dabbs, Christine M Bachman, Bryan Greenhouse, Chris Drakeley, Phil L Felgner, Will Stone, et al. Plasmodium falciparum serology: A comparison of two protein production methods for analysis of antibody responses by protein microarray. *PloS one*, 17(8):e0273106, 2022.

[53] Verónica Bolón-Canedo, Noelia Sánchez-Marono, Amparo Alonso-Betanzos, José Manuel Benítez, and Francisco Herrera. A review of microarray datasets and applied feature selection methods. *Information sciences*, 282:111–135, 2014.

[54] Roberto Ruiz, Jose C Riquelme, and Jesus S Aguilar-Ruiz. Incremental wrapper-based gene selection from microarray data for cancer classification. *Pattern Recognition*, 39(12):2383–2392, 2006.

[55] Gregory Piatetsky-Shapiro and Pablo Tamayo. Microarray data mining: facing the challenges. *ACM SIGKDD Explorations Newsletter*, 5(2):1–5, 2003.

[56] Yvan Saeys, Inaki Inza, and Pedro Larranaga. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517, 2007.

[57] Inaki Inza, Pedro Larranaga, Rosa Blanco, and Antonio J Cerrolaza. Filter versus wrapper gene selection approaches in dna microarray domains. *Artificial intelligence in medicine*, 31(2):91–103, 2004.

[58] Valerii Fedorov, Frank Mannino, and Rongmei Zhang. Consequences of dichotomization. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry*, 8(1):50–61, 2009.

[59] Bongin Yoo. The impact of dichotomization in longitudinal data analysis: a simulation study. *Pharmaceutical Statistics*, 9(4):298–312, 2010.

[60] Robert C MacCallum, Shaobo Zhang, Kristopher J Preacher, and Derek D Rucker. On the practice of dichotomization of quantitative variables. *Psychological methods*, 7(1):19, 2002.

[61] Irene Kyomuhangi and Emanuele Giorgi. A threshold-free approach with age-dependency for estimating malaria seroprevalence. *Malaria Journal*, 21:1–12, 2022.

[62] Emilie Pothin, Neil M Ferguson, Chris J Drakeley, and Azra C Ghani. Estimating malaria transmission intensity from plasmodium falciparum serological data using antibody density models. *Malaria journal*, 15:1–11, 2016.

[63] Elizabeth H Aitken, Timon Damelang, Amaya Ortega-Pajares, Agersew Alemu, Wina Hasang, Saber Dini, Holger W Unger, Maria Ome-Kaius, Morten A Nielsen, Ali Salanti, et al. Developing a multivariate prediction model of antibody features associated with protection of malaria-infected pregnant women from placental malaria. *Elife*, 10:e65776, 2021.

[64] Markus Loecher. Unbiased variable importance for random forests. *Communications in Statistics-Theory and Methods*, 51(5):1413–1425, 2022.

Part III
Analysis on high-dimensional data

Chapter 6 - The SARS-CoV-2 receptor ACE2 in ME/CFS: A meta-analysis of public DNA methylation and gene expression data

J. Malato[1,2], F. Sotzny[3], S. Bauer[3], H. Freitag[3], A. Fonseca[4], A.D. Grabowska[5], L. Graça[1], C. Cor- deiro[2,4], L. Nacul[6,7], E.M. Lacerda[6], J. Castro-Marrero[8], C. Scheibenbogen[3], F. Westermeier[9,10], and N. Sepúlveda[2,3,11].

[1]Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal
[2]CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal
[3]Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt Universität zu Berlin and Berlin Institute of Health, Institute of Medical Immunology, Berlin, Germany
[4]Faculdade de Ciências e Tecnologia, Universidade do Algarve, Faro, Portugal
[5]Department of Biophysics, Physiology, and Pathophysiology, Medical University of Warsaw, Warsaw, Poland
[6]Department of Clinical Research, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom
[7]Complex Chronic Diseases Program, British Columbia Women's Hospital and Health Centre, Vancouver, British Columbia, Canada
[8]Vall d'Hebron Hospital Research Institute, Division of Rheumatology, ME/CFS Unit, Universitat Autònoma de Barcelona, Barcelona, Spain
[9]Institute of Biomedical Science, Department of Health Studies, FH Joanneum University of Applied Sciences, Graz, Austria
[10]Centro Integrativo de Biología y Química Aplicada (CIBQA), Universidad Bernardo O´Higgins, Santiago, Chile
[11]Department of Infection Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom

João Malato, Franziska Sotzny, Sandra Bauer, Helma Freitag, André Fonseca, Anna D Grabowska, Luís Graça, Clara Cordeiro, Luís Nacul, Eliana M Lacerda, et al. The sars-cov-2 receptor angiotensin-converting enzyme 2 (ace2) in myalgic encephalomyelitis/chronic fatigue syndrome: A meta analysis of public dna methylation and gene expression data. Heliyon, 7(8),2021.

6.1 Abstract

People with myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) often report a high frequency of viral infections and flu-like symptoms during their disease course. Given that this reporting agrees with different immunological abnormalities and altered gene expression profiles observed in the disease, we aimed at answering whether the expression of the human angiotensin-converting enzyme 2 (ACE2), the major cell entry receptor for SARS-CoV-2, is also altered in these patients. In particular, a low expression of *ACE2* could be indicative of a high risk of developing Covid-19. We then performed a meta-analysis of public data on CpG DNA methylation and gene expression of this enzyme and its homologous ACE protein in peripheral blood mononuclear cells and related subsets. We found that patients with ME/CFS have decreased methylation levels of four CpG probes in the *ACE* locus (cg09920557, cg19802564, cg21094739, and cg10468385) and of another probe in the promoter region of the *ACE2* gene (cg08559914). We also found a decreased expression of *ACE2* but not of *ACE* in patients when compared to healthy controls. Accordingly, in newly collected data, there was evidence for a significant higher proportion of samples with an *ACE2* expression below the limit of detection in patients than healthy controls. Altogether, patients with ME/CFS can be at a higher Covid-19 risk and, if so, they should be considered a priority group for vaccination by public health authorities. To further support this conclusion, similar research is recommended for other human cell entry receptors and cell types, namely, those cells targeted by the virus.

Keywords: Myalgic encephalomyelitis/Chronic fatigue syndrome; SARS-CoV-2; ACE2; Gene expression; DNA methylation

6.2 Introduction

Myalgic encephalomyelitis/Chronic fatigue syndrome (ME/CFS) is a multifactorial and complex disease characterised by two key symptoms: (1) persistent but unexplained fatigue that is not alleviated by rest; and (2) post-exertional malaise upon minimal physical or even mental effort [1, 2]. Although its cause remains unknown, a growing body of evidence strongly associates ME/CFS with several microbial and viral infections, as potential triggering factors [3, 4]. In addition, it is currently hypothesised that reactivations of dormant viral infections also play a role [5, 6] due to several immunological abnormalities [7, 8, 9]. On the molecular basis of the disease, peripheral blood mononuclear cells (PBMCs) have altered gene expression profiles [10], including a decreased abundance of the human angiotensin-converting enzyme 2 (ACE2) [11], the main receptor of the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) for cell invasion [12, 13, 14]. Altogether, this evidence raises the question about the Covid-19 risk in patients with ME/CFS.

As basic information, ACE2 is encoded by the X-linked *ACE2* gene whose expression is predominant in the lungs, heart, skin, and kidneys [15, 16, 17, 18]. Its expression can also be detected in monocytes [19] and activated macrophages [20]. However, the percentage of *ACE2*-expressing cells is below 5% in the main immune-cell populations [20]. Accordingly, current RNA-Seq studies suggest a residual *ACE2* expression in PBMCs from healthy controls [18]. ACE2 has an amino-acid sequence identity of 41% with its homologous angiotensin-converting enzyme (ACE) [21]. This sequence similarity increases to 61% at the nucleotide level [21]. The enzymes ACE and ACE2 are members of the renin-angiotensin-aldosterone system (RAAS), which regulates blood pressure and vascular resistance [22]. In particular, ACE and ACE2 have vasoconstriction and vasodilation effects, respectively. Given this counteracting effect, high ACE:ACE2 ratios are possible indicators of severe Covid-19 outcomes, linked to increased reactive oxygen species (ROS) production, vasoconstriction, and inflammation [23].

To answer our research question, we performed a meta-analysis of public DNA methylation and gene expression data of *ACE2* and *ACE* in PBMCs. Similar study was conducted on the DNA methylation pattern of *ACE2* in the same cell type from patients with systemic lupus erythematosus [24], an autoimmune disease whose symptoms overlap with the ones from ME/CFS [25]. To complement our findings, we also compared the mRNA levels of these two genes in PBMCs from a new cohort of female patients with ME/CFS and healthy women.

6.3 Materials and methods

6.3.1 Eligible diagnostic criteria of ME/CFS

In our meta-analysis, we selected public data from studies using either the 1994 US Center for Disease Control and Prevention criteria (CDC-1994) [1] or the 2003 Canadian Consensus Criteria (CCC-2003) [2] for the disease diagnosis. These criteria are defined by the presence of several key symptoms while excluding known medical conditions (i.e., multiple sclerosis or lupus) that can also explain fatigue. The choice of using these two criteria for study selection complies with the research standards set by the European Network on ME/CFS [26].

6.3.2 Analysis of published DNA methylation association studies

Our meta-analysis was based on six genome-wide DNA methylation association studies (Table 4), four of which [27, 28, 29, 30] were previously reviewed [31], and other two published after this review [32, 33]. Briefly, these studies aimed at identifying differentially methylated CpG dinucleotide sites between patients and healthy controls. Illumina methylation arrays were used to measure the respective DNA methylation levels with the excep-

tion of a single study (Table 4). In this study, the measurements were made by the reduced representation bisulfite sequencing [33].

Table 4: **DNA methylation studies.** Summary of the six DNA methylation studies under analysis.

Reference	Sample type	ME/CFS patients			Healthy controls, n	Technology (manufacturer)	NCBI GEO Accession number
		n	Sample characteristics	Case definition			
[27]	$CD4^+$ T cells	25	Female/male adults Mean age: 50 years old Mean BMI: not reported	CDC-1994	18	Infinium HumanMethylation450K Array (Illumina)	NA
[28]	PBMC	12	Female adults Mean age: 41 years old Mean BMI: 23 kg/m^2	CDC-1994 & CCC-2003	12	Infinium HumanMethylation450K Array (Illumina)	GSE59489
[29]	PBMC	49	Female adults Mean age: 50 years old Mean BMI: 23 kg/m^2	CDC-1994 & CCC-2003	25	Infinium HumanMethylation450K Array (Illumina)	GSE93266
[30]	PBMC	13	Female adults Mean age: 50 years old Mean BMI: 26 kg/m^2	CDC-1994 & CCC-2003	12	Methylation EPIC Array (Illumina)	GSE111183
[32]	T lymphocytes	61	Female/male adults Mean age: 32 years old Mean BMI: 27 kg/m^2	CDC-1994 & CCC-2003	48	Infinium HumanMethylation450K Array (Illumina)	GSE156792
[33]	PBMC	10	Female/male adults Mean age: not reported Mean BMI: not reported	CCC-2003	10	Reduced representation Bisulfite sequencing	GSE153667

With respect to the exclusion criteria, one study excluded individuals who were taking beta-blockers or ACE inhibitors [30]. Three studies excluded participants who were treated with immunomodulatory effects or affecting the underlying DNA methylation levels at the time of data collection [28, 29, 32].

In four of the published DNA methylation studies, patients and healthy controls were matched for age, gender, and body mass index (Table 4) [28, 29, 30]. In two other studies, the matching was only based on age and gender [27, 33]. Ethnicity was also used for further matching [30, 32] or the same matching could be assumed in studies that only recruited white females [28, 29]. The DNA methylation levels were quantified in $CD4^+$ T cells [27], PBMCs [28, 29, 30, 33], and T lymphocytes [32].

We conducted a joint analysis of the four array-based studies which made the data available [28, 29, 30, 32]. We first retrieved the data from all the CpG probes located in the coding regions and the transcription starting sites (TSS) of *ACE* and *ACE2*, respectively. We then restricted our data analysis to the 27 probes shared between the Infinium HumanMethylation450K and the Infinium HumanMethylationEPIC arrays (Supplementary Table 1).

Before conducting the statistical analysis itself, we checked whether (1) the selected probes showed a high probability of detection, (2) they were not cross-reactive with other genomic regions, and (3) they were not affected by single nucleotide polymorphisms (SNPs) with high minor allele frequencies [34]. In the latter criterion, the SNPs included in the selected probes had a minor allele frequency less than 5% in Europeans and North Americans (Figure 23A; Supplementary Table 2) referring to the sampled populations of the studies. All probes passed the remaining basic quality control checks.

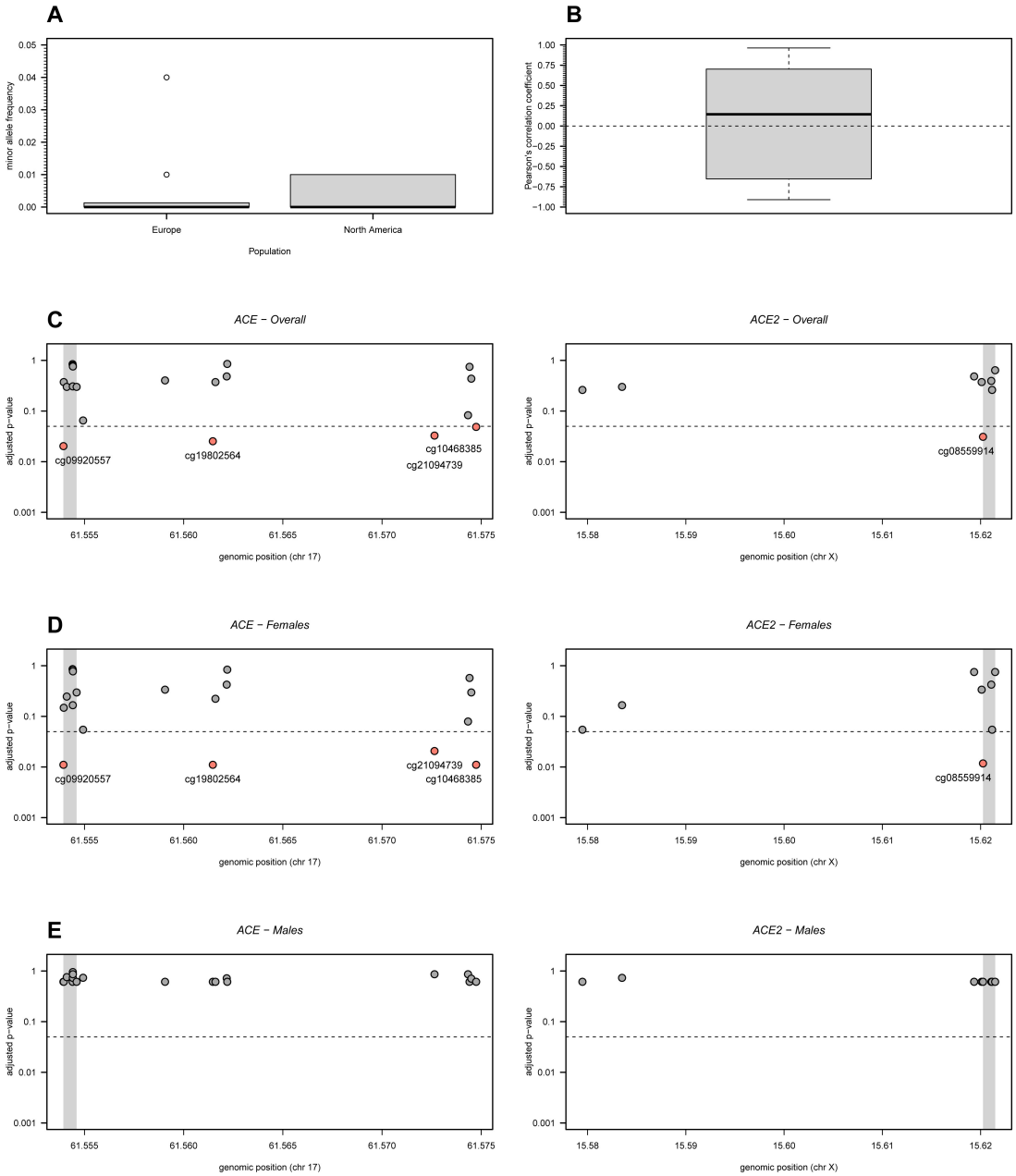


Figure 23: **DNA methylation analysis of 19 and 8 CpG probes located in the ACE and ACE2 genes, respectively**. DNA methylation analysis of 19 and 8 CpG probes located in the ACE and ACE2 genes, respectively. (A) Minor allele frequency in European and North American populations of SNPs located in the probes under analysis (see the respective data in Supplementary Table 2). (B) Boxplot of all possible Pearson's correlation coefficients (y axis) between the M-values of the probes under analysis. Horizontal dashed line represents the situation of lack of correlation. (C) Adjusted *p-values* for the overall association between each probe and ME/CFS. Adjusted *p-values* were calculated according to the Benjamini-Hochberg procedure with a false discovery rate of 5% (dashed line). Grey areas in the plots represent the TSS of the genes. (D) and (E) The same analyses as shown in C but for women and men separately.

We analyzed the M-values of a given probe instead of the respective β-values to ensure a good approximation of the Normal distribution to the data [35]. Briefly, the β-values were calculated as the proportion of the methylation signal relative to the total signal for a given probe. The M-values were finally obtained by applying a logit transformation to the β-values.

To analyze the M-values of each probe, we initially estimated a linear regression model where the respective covariates were the study indicator and the disease status of the participants. In this model, we included the main effects of the covariates and the interaction. The model parameters were then estimated by the maximum likelihood method. Note that the main effect of the disease status is usually seen as the pooled effect of this covariate across all studies, as done in meta-analysis.

We then simplified the model using a backward stepwise procedure based on Akaike's information criteria. Since the effect of the study indicator was significant for the data of each probe, we tested the association between ME/CFS and a given probe using a likelihood ratio test. In this test, we compared the model including the study indicator only with the best model including that covariate and the one associated with disease status (i.e., either the model only including the main effects or the model including both main effects and the interaction term).

To control for multiple testing, we adjusted the raw *p-values* using the Benjamini-Hochberg procedure [36]. This adjustment ensured a false discovery rate of 5% under the assumption of independent tests. Pearson's correlation coefficient was used to check the validity of this assumption (Figure 23B).

We also repeated the same association analysis for women and men separately. Note that three studies only recruited women [28, 29, 30] while the remaining study recruited both men and women [32]. In the latter study, there was no information available about the gender of each participant. In this case, we estimated this missing information using the function *getSex* of the R package *minfi* applied to the genome-wide DNA methylation data [37]. The resulting frequencies of men and women matched with those reported in the original study.

In the women-specific analysis, we performed the same association analysis as described above. In the men-specific analysis, we compared a linear regression model with the disease status as the single covariate against another model without that covariate, when analysing data from each probe. The comparison was done by the likelihood ratio test whose *p-values* were then adjusted for multiple testing in the same way as described above.

Finally, for the study which did not share the respective data [9], we checked whether the reported differentially methylated CpG probes were located in either *ACE* or *ACE2* (see Table 4 from this study). We did the same for the study based on the reduced representation bisulfite sequencing technology [33] (see the Supplementary files from this study).

6.3.3 Analysis of gene expression studies

Our meta-analysis of gene expression studies was focused on eight reports using microarray technology (Table 5) [11, 38, 39, 40, 41, 42, 43, 44]. These studies complied with the Minimum Information about a Microarray Experiment (MIAME) standard [45] and, therefore, they were considered to have sufficient quality for their inclusion in the meta-analysis. In particular, these studies normalized the data which ensured comparability between different samples and between different measurements of the same genes.

Table 5: **Microarray-based gene expression.**Summary of the 8 microarray-based gene expression studies under analysis, ordered by the year of publication.

Reference	Sample type	ME/CFS patients			Healthy controls, n	Technology (manufacturer)	ACE/ACE2 available	Data availability (NCBI GEO Assession number)
		n	Sample characteristics	Case definition				
[38]	PBMC	5	Female adults Mean age: 42 years old Mean BMI: not reported	CDC-1994	5	Atlas Glass Human 3.8 I Microarray (BD Biosciences Clontech)	No/No	No (NA)
[39]	PBMC	25	Female/male adults Mean age: 41 years old Mean BMI: not reported	CDC-1994	25	Custom microarray (Nimblegen)	Unclear	No (NA)
[40]	Whole blood	25	Female/male adults Mean age: 43 years old Mean BMI: not reported	CDC-1994	50	GeneChip Human Genome U133 Plus 2.0 (Affymetrix)	Yes/Yes	No (NA)
[41]	Whole blood	11	Female/male adults Mean age: 34 years old Mean BMI: 20.3 kg/m^2	CDC-1994	11	Custom microarray (NA)	Yes/No	Yes (NA)[1]
[42]	Muscle biopsies	4	Female/male adults Mean age: 45/37 years old Mean BMI: not reported	CDC-1994	5	Operon V2.0 (CRIBI University of Padova)	Yes/Yes	No (NA)
[43]	PBMC	8	Male adults Median age: 36 years old Mean BMI: not reported	CDC-1994	7	GeneChip Human Genome U133 (Affymetrix)	Yes/Yes	Yes (GSE14577)
[11]	PBMC	37	Female/male adults Mean age: 51 years old Mean BMI:29.4 kg/m^2	CDC-1994	25	MWG 20K human Array (Biotech MWG)	Yes/Yes	No (NA)
[44]	PBMC	33	Female/male adults Mean age: not reported Mean BMI: not reported	CDC-1994	21	GeneChip Human Gene ST (Affymetrix)	Yes/No	No (NA)

[1]Data shared as a supplementary file in the online version of the study.

Gene expression of these studies was performed in PBMCs (5 studies), whole blood (2 studies) and muscle biopsies (one study). One study excluded participants who were taking any regular medication [45]. Another study reviewed the medications taken by the participants [11]. However, it was unclear which medications were considered as a part of the exclusion criteria. A third study reported that healthy controls were free from any medication at the time of sampling [41].

Three additional studies using microarray technology [46, 47, 48] were excluded from our meta-analysis due to unclear or ineligible case definitions of ME/CFS. We also excluded four RNA-seq studies [49, 50, 51, 52], because of insufficient reporting on the basic quality control checks. In particular, these studies did not report the percentage of reads that could be mapped onto the reference transcriptome, the percentage of the transcriptome covered, the average number of mapped reads per transcript, the relationship between the GC content and the mapped read distribution, as recommended elsewhere [53]. More importantly,

given the high sequence homology between ACE and ACE2, these studies did not explain how their mapping algorithms dealt with reads that could be ambiguously mapped onto different locations in the transcriptome.

The selected studies were conducted in small cohorts of patients with ME/CFS (mean sample size = 18.5; range = 4–37) and healthy controls (mean sample size = 18.6; range = 5–50 individuals) (Table 5). In these studies, the patients and healthy controls were matched for age and gender. Different commercial and custom microarray technologies were used for the respective gene expression quantification. There was only one study in which the microarray did not include any probe in the genes of interest [38]. Another study used a custom array based on 9,522 genes from the RefSeq database, as available in August 2002 [39]. However, this study did not provide the list of genes included in the respective microarray. In terms of data sharing, one study made the data available in the GEO database [43] and another one within the respective publication [43]. The latter study used a custom microarray that measured the expression of stress-related genes including *ACE* but excluding *ACE2*.

Before conducting a meta-analysis of the available data, we first re-analysed two studies where the normalized data were available [41, 43]. In the first study [41], we calculated the mean of the $\log_2$(fold-change) for *ACE* and the respective standard error. Note that the microarray used in this study did not include any probe in *ACE2*. In the second study [43], we initially calculated the mean and the respective standard error of the $\log_2$(fold-change) for each probe located in *ACE* and *ACE2*. We then pooled each pair of means for the same gene using the inverse-variance weighting method [54]. A third study reported the mean of the $\log_2$(fold-change) for *ACE2* and the respective *p-value* using a two-tailed Student's test [11]. In this case, we determine the quantile of the t-distribution associated with half of the reported *p-value*, equated it to the test statistic, and solved the resulting equation as a function of the standard error. No information was available from this study concerning the expression levels of *ACE*.

Finally, we pooled the different estimates for the same gene from different studies using the inverse-variance weighting method [54].

6.3.4 Analysis of new RNA data on the *ACE/ACE2* gene expression in ME/CFS

6.3.4.1 Study participants

Thirty-seven women with ME/CFS were recruited in 2020 from the outpatient clinic for immunodeficiencies at the Institute for Medical Immunology at the Charité-Universitätsmedizin Berlin, Germany. These patients were diagnosed according to the CCC-2003 while excluding other medical or neurological diseases which could explain fatigue [2]. Thirty-four women with self-reported healthy status were recruited from staff.

6.3.4.2 Experimental procedure for RNA isolation and expression

Consistently with previous studies of ME/CFS, the gene expression quantification was performed in PBMCs. These cells were isolated from heparinized whole blood by density gradient centrifugation using Biocoll Separating Solution (Merck Millipore). Total RNA was isolated and extracted from 2×10^6 PBMCs according to the manufacturer's instructions (NucleoSpin RNA Kit, Macherey-Nagel, cat. nr. 740955.50). Afterwards cDNA was prepared by reverse transcription (High-Capacity cDNA Reverse Transcription Kit, Applied Biosystems, cat. nr. 4368814) and real-time PCR was performed using TaqMan™ Universal PCR Master Mix (cat. nr. 4305719) and TaqMan™ Gene Expression Assays (cat. nr. 4331182) for *ACE* (Hs00174179_m1), *ACE2* (Hs01085333_m1) and the housekeeping gene *HPRT1* (Hs02800695_m1) (Applied Biosystems). The amplification of *ACE* and *HPRT1* was based on 20 ng template cDNA. For the amplification of *ACE2*, this quantity was increased to 100 ng. All measurements were performed with the ABI7200 and software Step One Plus as absolute quantification according to manufacturer's instruction. Relative gene expression was analysed using the ΔCT method.

6.3.4.3 Statistical analysis

We first tested whether patients and healthy controls were matched for age using the Kolgomorov-Smirnov test for two independent samples. For statistical convenience, gene expression values were independently transformed for ACE and ACE2 using a Box-Cox transformation [55]. The parameter estimates of this transformation were 0.303 and 0.225 for ACE and ACE2, respectively. The transformed values for each gene were then analysed as the outcome variable of a linear regression model specifying age and disease status of the participants as the respective covariates. The linear regression model was estimated using the maximum likelihood method. After estimating the models, we tested the Normal distribution in the resulting residuals using the Shapiro-Wilk test. We also visually inspected the assumption of constant variance of the same residuals as a function of the covariates.

Note that we were unable to quantify the *ACE2* expression in 11 patients due to cDNA material below the limit of detection. These problematic samples could be due to a lower expression of *ACE2* in ME/CFS patients than in healthy controls. To test this hypothesis, we compared the respective proportion of samples below the limit of detection using the Pearson's χ^2 test for two-way frequency tables.

The significance level of the statistical analysis was set at 5%.

6.3.4.4 Ethical approval The protocol of this study was approved by the Ethics Committee of Charité-Universitä- tsmedizin Berlin in accordance with the 1964 Declaration of Helsinki and its later amendments (reference number EA2/067/20). All patients and healthy controls gave written informed consent to participate in the study.

6.3.5 Statistical software

We performed our statistical analysis in the R software version 4.0.3. In this analysis, we used the following Bioconductor packages: *hgu133a.db*, *hgu133plus2.db*, *IlluminaHumanMethylation450kanno.ilmn12.hg19*, and *IlluminaHumanMethylationEPICanno.ilm10b2.hg19* to retrieve the annotation of the GeneChip HG-U133A, GeneChip U133+2, Infinium HumanMethylation450K Array and HumanMethylationEPIC arrays, respectively; *minfi* to estimate the sex of each individual from DNA methylation data [37]. The R scripts are freely available from the first and last authors upon request.

6.4 Results

6.4.1 Meta-analysis of ACE/ACE2 DNA methylation in ME/CFS patients

The oldest DNA methylation study [27] did not make the data available and hence, we screened the list of 120 differentially methylated probes (see Table 1 from this study). Although located in 70 genes, these probes were neither located in *ACE* nor *ACE2*. We also screened the list of differentially methylated probes reported by the study based on the reduced representation bisulfite sequencing technology (see Additional File 1 from [33]). Again, none of these probes was in the ACE or ACE2 loci.

For the four array-based studies [28, 29, 30, 32], we conducted a joint analysis of the respective data in accordance with a meta-analysis. We first observed that the M-values of the 27 probes under investigation tended to be uncorrelated with each other (Figure 23B). This observation supported the use of the Benjamini-Hochberg procedure to adjust the raw *p-values* under a multiple testing scenario.

The subsequent analysis suggested four CpG probes in *ACE* to be associated with ME/CFS (Figure 23C). The probe cg09920557 belongs to the TSS region of the gene while the remaining probes (cg19802564, cg21094739, and cg10468385) are located in the gene body. The best linear regression models for each probe included both the main effects of the study indicator and of the disease status and the respective interaction term (Supplementary Table 3). The statistical interaction between these two covariates could be seen when plotting the whole data set (Figure 24A). Although not significant, the estimated main effect of the disease status was negative for each of the significantly associated probes.

Concerning the probes in *ACE2*, the only significant association with ME/CFS was obtained for cg08559914 located in the TSS region of the gene (Figure 23C). According to the best linear regression model for this probe, there was a negative association between the respective M-values and ME/CFS (coefficient estimate $=-0.141$ with a standard error of 0.048; Figure 24B and Supplementary Table 3). Given that a hypomethylated promoter region is typically indicative of an increased expression of the respective gene, this finding suggested an increased *ACE2* expression in patients with ME/CFS.

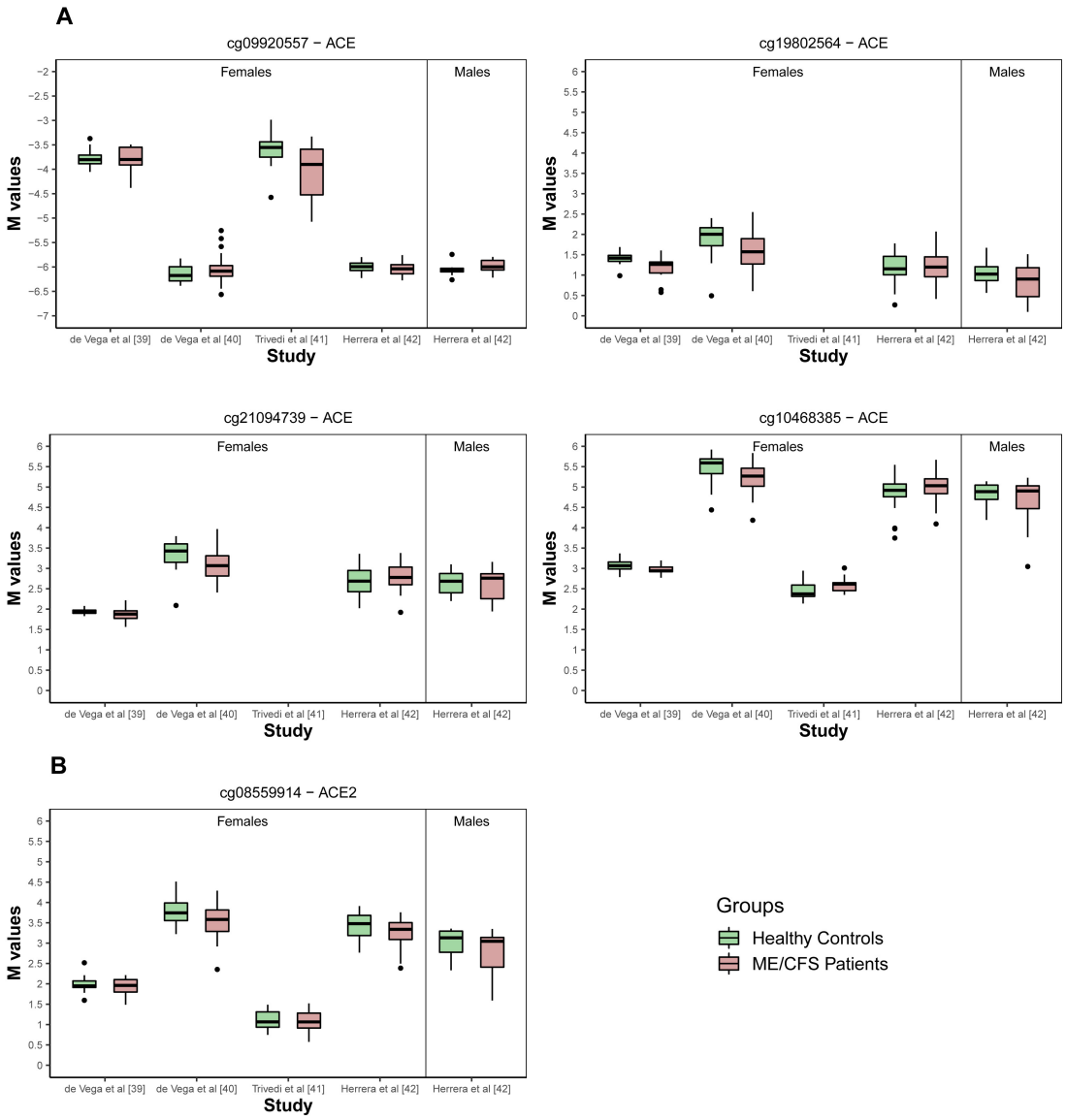


Figure 24: **Boxplots per study, group and gender of the M-values referring to probes identified in Figure 23C and Figure 23D**. (A) Significant probes located in *ACE*. (B) Significant probe located in *ACE2*.

We then repeated the same analysis for women and men separately. For women, we obtained the same disease associations, as described above (Figure 23D and Supplementary Table 3). For men, we did not find any significant associations, probably due to data from a single study [32] (Figure 23E).

6.4.2 Meta-analysis of ACE/ACE2 gene expression in ME/CFS patients

We first conducted a re-analysis of the two studies in which the expression levels of *ACE* or *ACE2* were available for each participant (Figure 25A) [41, 43]. In the first study [41], there was evidence for an increased expression of *ACE* in patients with ME/CFS (mean of the $\log_2$(fold-change) $= 0.265$; 95% CI $= [0.089, 0.441]$). In the second study [43], the means of the $\log_2$(fold-change) were estimated at 0.012 (95% CI $= [-0.012, 0.036]$) and 0.004 (95% CI $= [-0.014, 0.022]$) for the two probes in *ACE*. The corresponding estimates for the two probes in *ACE2* were -0.038 (95% CI $= [-0.085, 0.009]$) and -0.037 (95% CI $= [-0.083, 0.008]$) (Figure 25A). The pooled estimates for this study were 0.007 (95% CI $= [-0.006, 0.020]$) and -0.038 (95% CI $= [-0.067, -0.008]$) for *ACE* and *ACE2*, respectively.

Although not sharing the data, there was a study [11] that reported a significant negative association between ME/CFS and *ACE2* expression (see Supplementary Table 2 of this study). In this case, we obtained the following mean of the $\log_2$(fold-change) $= -2.396$ and 95% CI $= (-4.518, -0.273)$.

We then pooled the estimates from different studies for the same gene: 0.008 (95% CI $= [-0.005, 0.021]$) and -0.038 (95% CI $= [-0.068, -0.009]$) for *ACE* and *ACE2*, respectively (Figure 25B). Therefore, our meta-analysis suggested a reduced expression of *ACE2* but not of *ACE* in patients with ME/CFS when comparing to healthy controls.

Finally, the remaining gene expression studies neither shared the respective data nor reported any differential *ACE*/*ACE2* expression between patients and healthy controls.

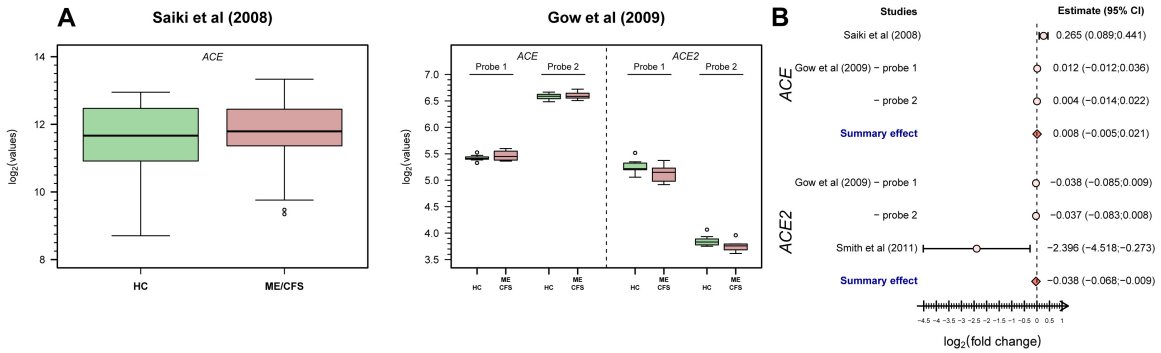


Figure 25: **Analysis of *ACE*/*ACE2*-related data from eligible microarray-based gene expression studies.** (A) Boxplots of the data from these studies [41, 43]. (B) Forest plot for the study-specific and pooled estimate of the mean of the $\log_2$(fold-change) between patients with ME/CFS and healthy controls using data shown in A.

6.4.3 Analysis of ACE/ACE2 gene expression from a new female cohort

To complement our findings from the above meta-analysis, we measured the *ACE* and *ACE2* mRNA levels in PBMCs from 37 women with ME/CFS (mean age = 41.1 years old) and 34 healthy women (mean age = 37.4 years old) (Table 6). Patients and healthy participants were matched for age (Kolmogorov-Smirnov test, p = 0.38). There was no information about the disease duration for 4 patients. The average disease duration for the remaining patients was 5.4 months in relation to the time of diagnosis (range = 0–24 months).

Table 6: Summary statistics for the gene expression of *ACE* and *ACE2* from the German female study participants where data of *ACE2* were only available for 26 affected patients.

Summary statistic	Healthy controls	ME/CFS patients
N	34	37
Mean age (range), years	37.4 (23, 65)	41.1 (19, 60)
Mean disease duration since diagnostic (range), months	—	5.4 (0, 24)
ACE		
Geometric mean	0.153	0.144
Interquartile range	0.087	0.073
ACE2		
Geometric mean	0.002	0.001
Interquartile range	0.005	0.004

We observed higher mRNA levels of *ACE* than of *ACE2* (Table 7, Figure 26A). There was no evidence for a significant correlation between *ACE* and *ACE2* expression levels (Spearman's correlation coefficient = −0.120) (Figure 26B). In contrast to the above meta-analysis, we could not find a reduced expression of *ACE2* in patients with ME/CFS using the complete case scenario (Table 7). However, there were 11 (29.7%) of the 37 samples from patients in which the expression level of *ACE2* was below the limit of detection. This proportion of samples was significantly higher than that for healthy controls given that the expression of *ACE2* could be quantified in all the samples (29.7% versus 0%; Pearson's χ^2 test, p = 0.002). Consequently, we could not rule out that the patients with ME/CFS from this cohort have a decreased expression of *ACE2* when compared to healthy controls. Finally, in accordance with our meta-analysis, there was no evidence of differential expression of *ACE* between patients and healthy controls from this cohort.

6.5 Discussion

In this work, we investigated potential differences in *ACE*/*ACE2* DNA methylation and expression levels between patients with ME/CFS and healthy controls. With the identification of these differences, we expected to determine the health risk of patients with ME/CFS

Table 7: Analysis of the linear regression models for the Box-Cox-transformed *ACE* and *ACE2* mRNA levels where data were only available for 26 ME/CFS patients.

Analyses	Estimate (SE)	*P-value*
Box-Cox transformed ACE		
`Intercept`	0.541 (0.032)	≤ 0.001
`Age`	0.001 (0.001)	0.328
`Disease status (ME/CFS)`	−0.013 (0.018)	0.481
Box-Cox transformed ACE2		
`Intercept`	0.307 (0.038)	≤ 0.001
`Age`	−0.001 (0.001)	0.137
`Disease status (ME/CFS)`	−0.006 (0.021)	0.789

if infected by SARS-CoV-2. However, we stumbled upon hurdles related to (i) data unavailability for a possible re-analysis, (ii) availability of data derived from PBCMs and related subsets in which *ACE2* is not particularly expressed, (iii) studies with unclear data quality, and (iv) studies using disease case definitions that are not recommended for research. As a consequence, we could not provide a more definite answer to our main research question.

Notwithstanding these difficulties, we could identify four CpG probes on *ACE* and another one on *ACE2* with decreased DNA methylation levels in patients with ME/CFS. This finding suggested an increased expression of the respective genes. However, our meta-analysis of public data suggested the opposite. Such decrease in *ACE2* expression was partially confirmed by new data in which there was a significant higher proportion of samples below the limit of detection in patients with ME/CFS than in healthy controls. Nonetheless, it was clear that *ACE2* is not particularly expressed in PBMCs from both patients with ME/CFS and healthy controls, as mentioned in the introduction.

In general, *ACE2* downregulation is known to occur after host-cell entry by SARS-CoV-2 [56]. This downregulation is particularly problematic in individuals affected by cardiovascular diseases, diabetes, and other medical conditions, due to their low *ACE2* levels before the infection [57]. SARS-CoV-2 infection is then expected to further increase the ACE:ACE2 ratio, thus, promoting vasoconstriction, increased production of ROS and inflammation in patients with these co-morbidities [23]. In this scenario, a putative reduction of the *ACE2* expression makes patients with ME/CFS similar to these patients with a high risk for Covid-19. As a consequence, patients with ME/CFS could be considered a priority group for vaccination by public health authorities. The fundamental question is then to know whether our findings based on PBMCs could recreate what occurs in pulmonary epithelial and endothelial cells, the main targets of SARS-CoV-2. Future research should be conducted to answer this question, as similarly done in past studies aiming at understanding how the gene expression profiles from PBMCs could mimic those present in other tissues affecting

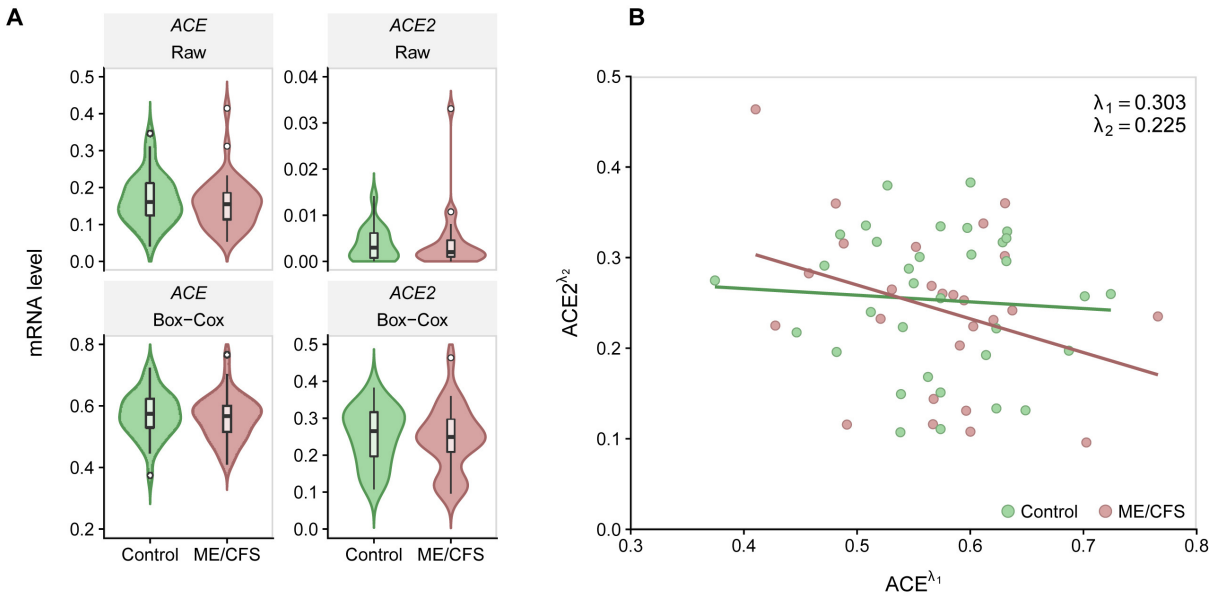


Figure 26: **Analysis of *ACE* and *ACE2* expression levels from the German study. Analysis of *ACE* and *ACE2* expression levels from the German study**. (A) Violin plots of *ACE* (left side) and *ACE2* (right side) mRNA raw data (upper row) and transformed data using a Box-Cox transformation (lower row). (B) Scatterplot between the transformed *ACE* and *ACE2* expression levels (Spearman's correlation coefficient = −0.120).

by a given disease [58, 59, 60].

Given the residual *ACE2* expression in PBMCs under normal conditions, one is tempted to say that SARS-CoV-2 does not infect these cells. However, earlier studies on SARS-CoV-1 found this virus within T lymphocytes, macrophages, and dendritic cells [61]. More recently, an *in vitro* study was able to infect PBMCs with SARS-CoV-2 [62]. Monocytes are particularly susceptible to such infections. In this context, one cannot rule out that SARS-CoV-2 might use alternative receptors when infecting PBMCs.

Among the alternative receptors for SARS-CoV-2, the human transmembrane protease serine 2 (TMPRSS2) was suggested as a strong candidate [63] due to its role on SARS-CoV-1 infection [64, 65]. This protease seems to induce SARS-CoV-2 cell entry through endocytosis via a mechanism of ACE2 cleavage [14]. Another candidate receptor is the A disintegrin and metallopeptidase domain 17 protein (ADAM17) recognised by the immune system as a stress-response signal [66]. Like TMPRSS2, ADAM17 can also cleave ACE2 but with a reduced viral invasion efficiency [67].

With respect to the role of these proteases in ME/CFS, a targeted gene expression study analysed ADAM17 and other stress-response proteins [41]. This study did not report any differential expression of this protease between patients with ME/CFS and healthy controls. However, this study is likely to be affected by a low statistical power due to small sample sizes for both groups. In addition, one of the selected DNA methylation studies suggested a decrease in the DNA methylation levels of one ADAM17-related CpG probe in patients with ME/CFS [30].

Dipeptidyl peptidase-4 (DPP4), also known as the lymphocyte cell surface protein CD26, was found to be the main receptor for the Middle East respiratory syndrome–related coronavirus [68, 69]. In contrast to ACE2, this surface protein is highly abundant in PBMCs including $CD4^+$ and $CD8^+$ T cells [18]. Bioinformatic analysis also suggested a strong interaction potential between this protein and SARS-CoV-2 [70, 71]. Finally, DPP4 inhibitors were found to be protective against severe Covid-19 in patients with diabetes mellitus when compared to RAAS blockers [72]. After initial concerns, this finding combined with others suggested an interesting therapeutic avenue against Covid-19 using DPP4 blockers [73].

Interestingly, there is evidence for an increased proportion of natural killer cells and T cells expressing DPP4/$CD26^+$ in patients with ME/CFS [7, 74]. However, the number of DPP4/CD26 molecules was significantly reduced in T lymphocytes and natural killer cells of these patients [74]. If DPP4 is indeed a relevant receptor for immune-cell invasion by SARS-CoV-2, research about this receptor should be prioritised when analysing PBMCs from patients with ME/CFS.

Sialic acids were also hypothesised as binding receptors used by SARS-CoV-2, as reported for other human coronaviruses [75]. These acids are highly expressed in the epithelium cells of the lungs and oral cavity [76]. *In vitro* and *in silico* studies demonstrated the same binding potential for SARS-CoV-2 [77, 78]. However, the ACE2 glycosylation inhibition studies suggested that sialic acids on ACE2 receptor prevent ACE2-virus interaction [79, 80]. Again, detailed research on these putative receptors could help to determine the health risk of patients with ME/CFS when infected by SARS-CoV-2.

It was suggested that the arousal state experienced by patients with ME/CFS protects them against microbial infections [81]. This suggestion came from a clinical trial where patients were treated with clonidine to decrease such a state. Treated patients got their symptoms worsened and had their inflammation markers increased during the trial. In contrast, basic epidemiological studies reported many patients with frequent viral infections and flu-like symptoms [3, 4, 82]. The question is how an infection by SARS-CoV-2 lies in this contrasting evidence. A possible answer can be given with the assistance of the so-called sustained arousal model of ME/CFS [83]. According to this model, a sustained arousal state promotes in the long-run deleterious alterations of different body systems, including the immune system. Similar prediction was made by a recent study discussing the natural history of ME/CFS [84]. If so, patients with longer disease durations are more likely to show these immunological alterations than patients at the early stages of the disease. However, we could not analyse the effect of disease duration on our results, because this variable was not available in the public data sets included in our meta-analyses.

Finally, our original idea was also to include a meta-analysis of *ACE*/*ACE2* data from published genome-wide association studies on ME/CFS [11, 32, 85, 86, 87]. However, we could not materialise this idea, because such studies did not make their data publicly available. Nevertheless, evidence is scarce for a putative role of *ACE*/*ACE2* polymorphisms on

ME/CFS. Two studies reported many candidate SNPs for such association, but none was located in *ACE* or *ACE2* [11, 85]. Two other studies did not find any significant SNPs associated with ME/CFS [32, 87]. The most optimistic study reported thousands of SNPs related to the disease [86]. However, this study did not perform all the basic quality control checks [88].

6.6 Conclusions

Notwithstanding the low expression of *ACE2* in PBMCs in general, there is evidence for a decreased expression of the gene in these cells from patients with ME/CFS. If PBMCs can qualitatively recreate what is occurring in the main cellular targets of SARS-CoV-2, then patients with this disease could be at a higher Covid-19 risk. In this regard, a recent preliminary report suggested that patients with ME/CFS got their symptoms worsened upon SARS-CoV-2 infection [89]. Altogether, these patients could be considered a priority group for vaccination against Covid-19, even though vaccines could trigger ME/CFS [90, 91] or even exacerbate ME/CFS symptoms as the case of the natural immunisation by SARS-CoV-2. To further consolidate the existing evidence, future research should prioritise the collection of data from the main cellular targets in patients with ME/CFS. Further investigation should be also conducted on alternative SARS-CoV-2 receptors (i.e., DPP4 and sialic acids). At last, future research should also consider investigating putative sex differences in patients with ME/CFS given that, in general, men are more affected by Covid-19 than women [92].

Declarations

Author contribution statement

João Malato, André Fonseca, Anna D Grabowska, Luís Graça, Clara Cordeiro, Luís Nacul, Eliana M Lacerda, Jesus Castro-Marrero, Francisco Westermeier: Analyzed and interpreted the data; Wrote the paper. Franziska Sotzny: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper. Sandra Bauer, Helma Freitag: Performed the experiments; Contributed reagents, materials, analysis tools or data; Wrote the paper. Carmen Scheibenbogen: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper. Nuno Sepúlveda: Conceived and designed the experiments; Analyzed and interpreted the data; Wrote the paper.

Funding statement

João Malato and André Fonseca were fully funded by FCT – Fundação para a Ciência e Tecnologia, Portugal (ref.grant: SFRH/BD/149758/2019 and SFRH/BD/147629/2019, respectively). Nuno Sepúlveda and Clara Cordeiro were partially funded by FCT – Fundação para

a Ciência e a Tecnologia, Portugal (ref. grant: UIDB/00006/2020). Luís Nacul and Eliana M Lacerda acknowledge the funding from the National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health (NIH -Award Number: R01AI103629), and from the ME Association (Award number: PF8947) for their studies on ME/CFS.

Data availability statement
Data are available from the GEO database under the accession number GSE59489, GSE93266, GSE111183, GSE156792, GSE153667, GSE14577.

Declaration of interests statement
The authors declare no conflict of interest.

Additional information
No additional information is available for this paper.

References

[1] Keiji Fukuda, Stephen E Straus, Ian Hickie, Michael C Sharpe, James G Dobbins, Anthony Komaroff, and International Chronic Fatigue Syndrome Study Group. The chronic fatigue syndrome: a comprehensive approach to its definition and study. *Annals of internal medicine*, 121(12):953–959, 1994.

[2] Bruce M Carruthers, Anil Kumar Jain, Kenny L De Meirleir, Daniel L Peterson, Nancy G Klimas, A Martin Lerner, Alison C Bested, Pierre Flor-Henry, Pradip Joshi, AC Peter Powles, et al. Myalgic encephalomyelitis/chronic fatigue syndrome: clinical working case definition, diagnostic and treatment protocols. *Journal of chronic fatigue syndrome*, 11(1):7–115, 2003.

[3] Samantha C Johnston, Donald R Staines, and Sonya M Marshall-Gradisnik. Epidemiological characteristics of chronic fatigue syndrome/myalgic encephalomyelitis in australian patients. *Clinical epidemiology*, pages 97–107, 2016.

[4] Lily Chu, Ian J Valencia, Donn W Garvert, and Jose G Montoya. Onset patterns and course of myalgic encephalomyelitis/chronic fatigue syndrome. *Frontiers in pediatrics*, 7:12, 2019.

[5] Santa Rasa, Zaiga Nora-Krukle, Nina Henning, Eva Eliassen, Evelina Shikova, Thomas Harrer, Carmen Scheibenbogen, Modra Murovska, Bhupesh K Prusty, and European Network on ME/CFS (EUROMENE). Chronic viral infections in myalgic encephalomyelitis/chronic fatigue syndrome (me/cfs). *Journal of translational medicine*, 16:1–25, 2018.

[6] ME Ariza. Myalgic encephalomyelitis/chronic fatigue syndrome: The human herpesviruses are back! biomolecules 2021, 11, 185, 2021.

[7] Nancy G Klimas, Fernando R Salvato, Robert Morgan, and Mary Ann Fletcher. Immunologic abnormalities in chronic fatigue syndrome. *Journal of clinical microbiology*, 28(6):1403–1410, 1990.

[8] Lorenzo Lorusso, Svetlana V Mikhaylova, Enrica Capelli, Daniela Ferrari, Gaelle K Ngonga, and Giovanni Ricevuti. Immunological aspects of chronic fatigue syndrome. *Autoimmunity reviews*, 8(4):287–291, 2009.

[9] Ekua W Brenu, Mieke L van Driel, Don R Staines, Kevin J Ashton, Sandra B Ramos, James Keane, Nancy G Klimas, and Sonya M Marshall-Gradisnik. Immunological abnormalities as potential biomarkers in chronic fatigue syndrome/myalgic encephalomyelitis. *Journal of translational medicine*, 9(1):1–9, 2011.

[10] Jonathan R Kerr. Gene profiling of patients with chronic fatigue syndrome/myalgic encephalomyelitis. *Current rheumatology reports*, 10(6):482–491, 2008.

[11] Alicia K Smith, Hong Fang, Toni Whistler, Elizabeth R Unger, and Mangalathu S Rajeevan. Convergent genomic studies identify association of grik2 and npas2 with chronic fatigue syndrome. *Neuropsychobiology*, 64(4):183–194, 2011.

[12] Wenhui Li, Michael J Moore, Natalya Vasilieva, Jianhua Sui, Swee Kee Wong, Michael A Berne, Mohan Somasundaran, John L Sullivan, Katherine Luzuriaga, Thomas C Greenough, et al. Angiotensin-converting enzyme 2 is a functional receptor for the sars coronavirus. *Nature*, 426(6965):450–454, 2003.

[13] Xing-Yi Ge, Jia-Lu Li, Xing-Lou Yang, Aleksei A Chmura, Guangjian Zhu, Jonathan H Epstein, Jonna K Mazet, Ben Hu, Wei Zhang, Cheng Peng, et al. Isolation and characterization of a bat sars-like coronavirus that uses the ace2 receptor. *Nature*, 503(7477):535–538, 2013.

[14] Markus Hoffmann, Hannah Kleine-Weber, Simon Schroeder, Nadine Krüger, Tanja Herrler, Sandra Erichsen, Tobias S Schiergens, Georg Herrler, Nai-Huei Wu, Andreas Nitsche, et al. Sars-cov-2 cell entry depends on ace2 and tmprss2 and is blocked by a clinically proven protease inhibitor. *cell*, 181(2):271–280, 2020.

[15] Inge Hamming, Wim Timens, MLC Bulthuis, AT Lely, GJ van Navis, and Harry van Goor. Tissue distribution of ace2 protein, the functional receptor for sars coronavirus. a first step in understanding sars pathogenesis. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, 203(2):631–637, 2004.

[16] KF To and Anthony WI Lo. Exploring the pathogenesis of severe acute respiratory syndrome (sars): the tissue distribution of the coronavirus (sars-cov) and its putative receptor, angiotensin-converting enzyme 2 (ace2). *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, 203(3):740–743, 2004.

[17] Meng-Yuan Li, Lin Li, Yue Zhang, and Xiao-Sheng Wang. Expression of the sars-cov-2 cell receptor gene ace2 in a wide variety of human tissues. *Infectious diseases of poverty*, 9(02):23–29, 2020.

[18] Urszula Radzikowska, Mei Ding, Ge Tan, Damir Zhakparov, Yaqi Peng, Paulina Wawrzyniak, Ming Wang, Shuo Li, Hideaki Morita, Can Altunbulakli, et al. Distribution of ace2, cd147, cd26, and other sars-cov-2 associated molecules in tissues and immune cells in health and in asthma, copd, obesity, hypertension, and covid-19 risk factors. *Allergy*, 75(11):2829–2845, 2020.

[19] Magdalena Rutkowska-Zapała, Maciej Suski, Rafał Szatanek, Marzena Lenart, Kazimierz Węglarczyk, Rafał Olszanecki, Tomasz Grodzicki, Magdalena Strach, Jerzy Gąsowski, and Maciej Siedlar. Human monocyte subsets exhibit divergent angiotensin i-converting activity. *Clinical & Experimental Immunology*, 181(1):126–132, 2015.

[20] Xiang Song, Wei Hu, Haibo Yu, Laura Zhao, Yeqian Zhao, Xin Zhao, Hai-Hui Xue, and Yong Zhao. Little to no expression of angiotensin-converting enzyme-2 on most human peripheral blood immune cells but highly expressed on tissue macrophages. *Cytometry Part A*, 103(2):136–145, 2023.

[21] Sarah R Tipnis, Nigel M Hooper, Ralph Hyde, Eric Karran, Gary Christie, and Anthony J Turner. A human homolog of angiotensin-converting enzyme: cloning and functional expression as a captopril-insensitive carboxypeptidase. *Journal of Biological Chemistry*, 275(43):33238–33243, 2000.

[22] Francisco Westermeier, Mario Bustamante, Mario Pavez, Lorena García, Mario Chiong, María Paz Ocaranza, and Sergio Lavandero. Novel players in cardioprotection: Insulin like growth factor-1, angiotensin-(1–7) and angiotensin-(1–9). *Pharmacological research*, 101:41–55, 2015.

[23] Pasquale Pagliaro and Claudia Penna. Ace/ace2 ratio: a key also in 2019 coronavirus disease (covid-19)? *Frontiers in medicine*, 7:335, 2020.

[24] Amr H Sawalha, Ming Zhao, Patrick Coit, and Qianjin Lu. Epigenetic dysregulation of ace2 and interferon-regulated genes might suggest increased covid-19 susceptibility and severity in lupus patients. *Clinical Immunology*, 215:108410, 2020.

[25] Jonas Blomberg, Carl-Gerhard Gottfries, Amal Elfaitouri, Muhammad Rizwan, and Anders Rosén. Infection elicited autoimmunity and myalgic encephalomyelitis/chronic fatigue syndrome: an explanatory model. *Frontiers in immunology*, 9:229, 2018.

[26] Derek FH Pheby, Diana Araja, Uldis Berkis, Elenka Brenna, John Cullinan, Jean-Dominique de Korwin, Lara Gitto, Dyfrig A Hughes, Rachael M Hunter, Dominic Trepel, et al. The development of a consistent europe-wide approach to investigating the economic impact of myalgic encephalomyelitis (me/cfs): A report from the european network on me/cfs (euromene). In *Healthcare*, volume 8, page 88. MDPI, 2020.

[27] EW Brenu, DR Staines, and SM Marshall-Gradisnik. Methylation profile of cd4+ t cells in chronic fatigue syndrome/myalgic encephalomyelitis. *J Clin Cell Immunol*, 5(228):10–4172, 2014.

[28] Wilfred C de Vega, Suzanne D Vernon, and Patrick O McGowan. Dna methylation modifications associated with chronic fatigue syndrome. *PloS one*, 9(8):e104757, 2014.

[29] Wilfred C de Vega, Santiago Herrera, Suzanne D Vernon, and Patrick O McGowan. Epigenetic modifications and glucocorticoid sensitivity in myalgic encephalomyelitis/chronic fatigue syndrome (me/cfs). *BMC medical genomics*, 10:1–14, 2017.

[30] Malav S Trivedi, Elisa Oltra, Leonor Sarria, Natasha Rose, Vladimir Beljanski, Mary Ann Fletcher, Nancy G Klimas, and Lubov Nathanson. Identification of myalgic encephalomyelitis/chronic fatigue syndrome-associated dna methylation patterns. *PloS one*, 13(7):e0201066, 2018.

[31] Eloy Almenar-Perez, Tamara Ovejero, Teresa Sanchez-Fito, Jose A Espejo, Lubov Nathanson, and Elisa Oltra. Epigenetic components of myalgic encephalomyelitis/chronic fatigue syndrome uncover potential transposable element activation. *Clinical Therapeutics*, 41(4):675–698, 2019.

[32] Santiago Herrera, Wilfred C de Vega, David Ashbrook, Suzanne D Vernon, and Patrick O McGowan. Genome-epigenome interactions associated with myalgic encephalomyelitis/chronic fatigue syndrome. *Epigenetics*, 13(12):1174–1190, 2018.

[33] AM Helliwell, EC Sweetman, PA Stockwell, CD Edgar, A Chatterjee, and WP Tate. Changes in dna methylation profiles of myalgic encephalomyelitis/chronic fatigue syndrome patients reflect systemic dysfunctions. *Clinical Epigenetics*, 12:1–20, 2020.

[34] Sarah Dedeurwaerder, Matthieu Defrance, Martin Bizet, Emilie Calonne, Gianluca Bontempi, and François Fuks. A comprehensive overview of infinium humanmethylation450 data processing. *Briefings in bioinformatics*, 15(6):929–941, 2014.

[35] Pan Du, Xiao Zhang, Chiang-Ching Huang, Nadereh Jafari, Warren A Kibbe, Lifang Hou, and Simon M Lin. Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*, 11:1–9, 2010.

[36] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[37] Martin J Aryee, Andrew E Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P Feinberg, Kasper D Hansen, and Rafael A Irizarry. Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium dna methylation microarrays. *Bioinformatics*, 30(10):1363–1369, 2014.

[38] Toni Whistler, James F Jones, Elizabeth R Unger, and Suzanne D Vernon. Exercise responsive genes measured in peripheral blood of women with chronic fatigue syndrome and matched control subjects. *BMC physiology*, 5(1):1–9, 2005.

[39] N Kaushik, D Fear, SCM Richards, CR McDermott, EF Nuwaysir, P Kellam, TJ Harrison, RJ Wilkinson, DAJ Tyrrell, ST Holgate, et al. Gene expression in peripheral blood mononuclear cells from patients with chronic fatigue syndrome. *Journal of clinical pathology*, 58(8):826–832, 2005.

[40] Jonathan R Kerr, Robert Petty, Beverley Burke, John Gough, David Fear, Lindsey I Sinclair, Derek L Mattey, Selwyn C M Richards, Jane Montgomery, Don A Baldwin, et al. Gene expression subtypes in patients with chronic fatigue syndrome/myalgic encephalomyelitis. *The Journal of infectious diseases*, 197(8):1171–1184, 2008.

[41] Takuya Saiki, Tomoko Kawai, Kyoko Morita, Masayuki Ohta, Toshiro Saito, Kazuhito Rokutan, and Nobutaro Ban. Identification of marker genes for differential diagnosis of chronic fatigue syndrome. *Molecular Medicine*, 14(9):599–607, 2008.

[42] Tiziana Pietrangelo, Rosa Mancinelli, Luana Toniolo, G Montanari, Jacopo Vecchiet, G Fanò, and Stefania Fulle. Transcription profile analysis of vastus lateralis muscle from patients with chronic fatigue syndrome. *International Journal of Immunopathology and Pharmacology*, 22(3):795–807, 2009.

[43] John W Gow, Suzanne Hagan, Pawel Herzyk, Celia Cannon, Peter O Behan, and Abhijit Chaudhuri. A gene signature for post-infectious chronic fatigue syndrome. *BMC medical genomics*, 2(1):1–11, 2009.

[44] Mary G Jeffrey, Lubov Nathanson, Kristina Aenlle, Zachary M Barnes, Mirza Baig, Gordon Broderick, Nancy G Klimas, Mary Ann Fletcher, and Travis JA Craddock.

Treatment avenues in myalgic encephalomyelitis/chronic fatigue syndrome: a split-gender pharmacogenomic study of gene-expression modules. *Clinical Therapeutics*, 41(5):815–835, 2019.

[45] Alvis Brazma, Pascal Hingamp, John Quackenbush, Gavin Sherlock, Paul Spellman, Chris Stoeckert, John Aach, Wilhelm Ansorge, Catherine A Ball, Helen C Causton, et al. Minimum information about a microarray experiment (miame)—toward standards for microarray data. *Nature genetics*, 29(4):365–371, 2001.

[46] Sally Galbraith, Barbara Cameron, Hui Li, Diana Lau, Ute Vollmer-Conna, and Andrew R Lloyd. Peripheral blood gene expression in postinfective fatigue syndrome following from three different triggering infections. *Journal of Infectious Diseases*, 204(10):1632–1640, 2011.

[47] Suzanne D Vernon, Elizabeth R Unger, Irina M Dimulescu, Mangalathu Rajeevan, and William C Reeves. Utility of the blood for gene expression profiling and biomarker discovery in chronic fatigue syndrome. *Disease markers*, 18(4):193–199, 2002.

[48] Chinh Bkrong Nguyen, Lene Alsøe, Jessica M Lindvall, Dag Sulheim, Even Fagermoen, Anette Winger, Mari Kaarbø, Hilde Nilsen, and Vegard Bruun Wyller. Whole blood gene expression in adolescent chronic fatigue syndrome: an exploratory cross-sectional study suggesting altered b cell differentiation and survival. *Journal of translational medicine*, 15(1):1–21, 2017.

[49] Jerome Bouquet, Jennifer L Gardy, Scott Brown, Jacob Pfeil, Ruth R Miller, Muhammad Morshed, Antonio Avina-Zubieta, Kam Shojania, Mark McCabe, Shoshana Parker, et al. Rna-seq analysis of gene expression, viral pathogen, and b-cell/t-cell receptor signatures in complex chronic disease. *Clinical Infectious Diseases*, 64(4):476–481, 2017.

[50] Jerome Bouquet, Tony Li, Jennifer L Gardy, Xiaoying Kang, Staci Stevens, Jared Stevens, Mark VanNess, Christopher Snell, James Potts, Ruth R Miller, et al. Whole blood human transcriptome and virome analysis of me/cfs patients experiencing post-exertional malaise following cardiopulmonary exercise testing. *PloS one*, 14(3):e0212193, 2019.

[51] Eiren Sweetman, Margaret Ryan, Christina Edgar, Angus MacKay, Rosamund Vallings, and Warren Tate. Changes in the transcriptome of circulating immune cells of a new zealand cohort with myalgic encephalomyelitis/chronic fatigue syndrome. *International journal of immunopathology and pharmacology*, 33:2058738418820402, 2019.

[52] Ruud PH Raijmakers, Anne FM Jansen, Stephan P Keijmel, Rob Ter Horst, Megan E Roerink, Boris Novakovic, Leo AB Joosten, Jos WM van der Meer, Mihai G Netea, and

Chantal P Bleeker-Rovers. A possible role for mitochondrial-derived peptides humanin and mots-c in patients with q fever fatigue syndrome and chronic fatigue syndrome. *Journal of Translational Medicine*, 17:1–10, 2019.

[53] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J Gaffney, Laura L Elo, Xuegong Zhang, et al. A survey of best practices for rna-seq data analysis. *Genome biology*, 17(1):1–19, 2016.

[54] Joachim Hartung, Guido Knapp, and Bimal K Sinha. *Statistical meta-analysis with applications*. John Wiley & Sons, 2011.

[55] Özgür Asar, Ozlem Ilk, and Osman Dag. Estimating box-cox power transformation parameter via goodness-of-fit tests. *Communications in Statistics-Simulation and Computation*, 46(1):91–105, 2017.

[56] Prasun K Datta, Fengming Liu, Tracy Fischer, Jay Rappaport, and Xuebin Qin. Sars-cov-2 pandemic and research gaps: Understanding sars-cov-2 interaction with the ace2 receptor and implications for therapy. *Theranostics*, 10(16):7448, 2020.

[57] Paolo Verdecchia, Claudio Cavallini, Antonio Spanevello, and Fabio Angeli. The pivotal link between ace2 deficiency and sars-cov-2 infection. *European journal of internal medicine*, 76:14–20, 2020.

[58] Ivan C Gerling, Robert A Ahokas, German Kamalov, Wenyuan Zhao, Syamal K Bhattacharya, Yao Sun, and Karl T Weber. Gene expression profiles of peripheral blood mononuclear cells reveal transcriptional signatures as novel biomarkers of cardiac remodeling in rats with aldosteronism and hypertensive heart disease. *JACC: Heart Failure*, 1(6):469–476, 2013.

[59] Toshinari Takamura, Masao Honda, Yoshio Sakai, Hitoshi Ando, Akiko Shimizu, Tsuguhito Ota, Masaru Sakurai, Hirofumi Misu, Seiichiro Kurita, Naoto Matsuzawa-Nagata, et al. Gene expression profiles in peripheral blood mononuclear cells reflect the pathophysiology of type 2 diabetes. *Biochemical and biophysical research communications*, 361(2):379–384, 2007.

[60] Fernanda S Manoel-Caetano, Danilo J Xavier, Adriane F Evangelista, Paula Takahashi, Cristhianna V Collares, Denis Puthier, Maria C Foss-Freitas, Milton C Foss, Eduardo A Donadi, Geraldo A Passos, et al. Gene expression profiles displayed by peripheral blood mononuclear cells from patients with type 2 diabetes mellitus focusing on biological processes implicated on the pathogenesis of the disease. *Gene*, 511(2):151–160, 2012.

[61] Matthew Zirui Tay, Chek Meng Poh, Laurent Rénia, Paul A MacAry, and Lisa FP Ng. The trinity of covid-19: immunity, inflammation and intervention. *Nature Reviews Immunology*, 20(6):363–374, 2020.

[62] Ana Campos Codo, Gustavo Gastão Davanzo, Lauar de Brito Monteiro, Gabriela Fabiano de Souza, Stéfanie Primon Muraro, João Victor Virgilio-da Silva, Juliana Silveira Prodonoff, Victor Corasolla Carregari, Carlos Alberto Oliveira de Biagi Junior, Fernanda Crunfli, et al. Elevated glucose levels favor sars-cov-2 infection and monocyte response through a hif-1α/glycolysis-dependent axis. *Cell metabolism*, 32(3):437–446, 2020.

[63] Waradon Sungnak, Ni Huang, Christophe Bécavin, Marijn Berg, Rachel Queen, Monika Litvinukova, Carlos Talavera-López, Henrike Maatz, Daniel Reichart, Fotios Sampaziotis, et al. Sars-cov-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. *Nature medicine*, 26(5):681–687, 2020.

[64] Ilona Glowacka, Stephanie Bertram, Marcel A Müller, Paul Allen, Elizabeth Soilleux, Susanne Pfefferle, Imke Steffen, Theodros Solomon Tsegaye, Yuxian He, Kerstin Gnirss, et al. Evidence that tmprss2 activates the severe acute respiratory syndrome coronavirus spike protein for membrane fusion and reduces viral control by the humoral immune response. *Journal of virology*, 85(9):4122–4134, 2011.

[65] Shutoku Matsuyama, Noriyo Nagata, Kazuya Shirato, Miyuki Kawase, Makoto Takeda, and Fumihiro Taguchi. Efficient activation of the severe acute respiratory syndrome coronavirus spike protein by the transmembrane protease tmprss2. *Journal of virology*, 84(24):12658–12664, 2010.

[66] Stefan Düsterhöft, Juliane Lokau, and Christoph Garbers. The metalloprotease adam17 in inflammation and cancer. *Pathology-Research and Practice*, 215(6):152410, 2019.

[67] Adeline Heurich, Heike Hofmann-Winkler, Stefanie Gierer, Thomas Liepold, Olaf Jahn, and Stefan Pöhlmann. Tmprss2 and adam17 cleave ace2 differentially and only proteolysis by tmprss2 augments entry driven by the severe acute respiratory syndrome coronavirus spike protein. *Journal of virology*, 88(2):1293–1307, 2014.

[68] Neeltje van Doremalen, Kerri L Miazgowicz, Shauna Milne-Price, Trenton Bushmaker, Shelly Robertson, Dana Scott, Joerg Kinne, Jason S McLellan, Jiang Zhu, and Vincent J Munster. Host species restriction of middle east respiratory syndrome coronavirus through its receptor, dipeptidyl peptidase 4. *Journal of virology*, 88(16):9220–9232, 2014.

[69] Widagdo Widagdo, Nisreen MA Okba, Wentao Li, Alwin De Jong, Rik L de Swart, Lineke Begeman, Judith MA van den Brand, Berend-Jan Bosch, and Bart L Haagmans. Species-specific colocalization of middle east respiratory syndrome coronavirus attachment and entry receptors. *Journal of virology*, 93(16):10–1128, 2019.

[70] Yu Li, Ziding Zhang, Li Yang, Xianyi Lian, Yan Xie, Shen Li, Shuyu Xin, Pengfei Cao, and Jianhong Lu. The mers-cov receptor dpp4 as a candidate binding target of the sars-cov-2 spike. *Iscience*, 23(6), 2020.

[71] Naveen Vankadari and Jacqueline A Wilce. Emerging covid-19 coronavirus: glycan shield and structure prediction of spike glycoprotein and its interaction with human cd26. *Emerging microbes & infections*, 9(1):601–604, 2020.

[72] Sang Youl Rhee, Jeongwoo Lee, Hyewon Nam, Dae-Sung Kyoung, Dong Wook Shin, and Dae Jung Kim. Effects of a dpp-4 inhibitor and ras blockade on clinical outcomes of patients with diabetes and covid-19. *Diabetes & metabolism journal*, 45(2):251–259, 2021.

[73] André J Scheen. Dpp-4 inhibition and covid-19: from initial concerns to recent expectations. *Diabetes & metabolism*, 47(2):101213, 2021.

[74] Mary A Fletcher, Xiao R Zeng, Kevin Maher, Silvina Levis, Barry Hurwitz, Michael Antoni, Gordon Broderick, and Nancy G Klimas. Biomarkers in chronic fatigue syndrome: evaluation of natural killer cell function and dipeptidyl peptidase iv/cd26. *PloS one*, 5(5):e10817, 2010.

[75] Xue-Long Sun. The role of cell surface sialic acids for sars-cov-2 infection. *Glycobiology*, 31(10):1245–1253, 2021.

[76] Benjamin W Cross and Stefan Ruhl. Glycan recognition at the saliva–oral microbiome interface. *Cellular immunology*, 333:19–33, 2018.

[77] Mayanka Awasthi, Sahil Gulati, Debi P Sarkar, Swasti Tiwari, Suneel Kateriya, Peeyush Ranjan, and Santosh Kumar Verma. The sialoside-binding pocket of sars-cov-2 spike glycoprotein structurally resembles mers-cov. *Viruses*, 12(9):909, 2020.

[78] Alexander N Baker, Sarah-Jane Richards, Collette S Guy, Thomas R Congdon, Muhammad Hasan, Alexander J Zwetsloot, Angelo Gallo, Jozef R Lewandowski, Phillip J Stansfeld, Anne Straube, et al. The sars-cov-2 spike protein binds sialic acids and enables rapid detection in a lateral flow point of care diagnostic device. *ACS central science*, 6(11):2046–2052, 2020.

[79] Hin Chu, Bingjie Hu, Xiner Huang, Yue Chai, Dongyan Zhou, Yixin Wang, Huiping Shuai, Dong Yang, Yuxin Hou, Xi Zhang, et al. Host and viral determinants for efficient sars-cov-2 infection of the human lung. *Nature communications*, 12(1):134, 2021.

[80] Qi Yang, Thomas A Hughes, Anju Kelkar, Xinheng Yu, Kai Cheng, Sheldon J Park, Wei-Chiao Huang, Jonathan F Lovell, and Sriram Neelamegham. Inhibition of sars-cov-2 viral entry in vitro upon blocking n-and o-glycan elaboration. *bioRxiv*, pages 2020–10, 2020.

[81] Dag Sulheim, Even Fagermoen, Anette Winger, Anders Mikal Andersen, Kristin Godang, Fredrik Müller, Peter C Rowe, J Philip Saul, Eva Skovlund, Merete Glenne Øie, et al. Disease mechanisms and clonidine treatment in adolescent chronic fatigue syndrome: a combined cross-sectional and randomized clinical trial. *JAMA pediatrics*, 168(4):351–360, 2014.

[82] Joanna Słomko, Julia L Newton, Sławomir Kujawski, Małgorzata Tafil-Klawe, Jacek Klawe, Donald Staines, Sonya Marshall-Gradisnik, and Pawel Zalewski. Prevalence and characteristics of chronic fatigue syndrome/myalgic encephalomyelitis (cfs/me) in poland: A cross-sectional study. *BMJ open*, 9(3):e023955, 2019.

[83] Vegard B Wyller, Hege R Eriksen, and Kirsti Malterud. Can sustained arousal explain the chronic fatigue syndrome? *Behavioral and Brain Functions*, 5(1):1–10, 2009.

[84] Luis Nacul, Shennae O'Boyle, Luigi Palla, Flavio E Nacul, Kathleen Mudie, Caroline C Kingdon, Jacqueline M Cliff, Taane G Clark, Hazel M Dockrell, and Eliana M Lacerda. How myalgic encephalomyelitis/chronic fatigue syndrome (me/cfs) progresses: the natural history of me/cfs. *Frontiers in neurology*, 11:826, 2020.

[85] KA Schlauch, Svetlana F Khaiboullina, Kenny L De Meirleir, Shanti Rawat, Julia Petereit, AA Rizvanov, N Blatt, Tatjana Mijatovic, Doina Kulick, A Palotas, et al. Genome-wide association analysis identifies genetic variations in subjects with myalgic encephalomyelitis/chronic fatigue syndrome. *Translational psychiatry*, 6(2):e730–e730, 2016.

[86] Melanie Perez, Rajeev Jaundoo, Kelly Hilton, Ana Del Alamo, Kristina Gemayel, Nancy G Klimas, Travis JA Craddock, and Lubov Nathanson. Genetic predisposition for immune system, hormone, and metabolic dysfunction in myalgic encephalomyelitis/chronic fatigue syndrome: a pilot study. *Frontiers in Pediatrics*, page 206, 2019.

[87] Joshua J Dibble, Simon J McGrath, and Chris P Ponting. Genetic risk factors of me/cfs: a critical review. *Human Molecular Genetics*, 29(R1):R117–R124, 2020.

[88] Anna D Grabowska, Eliana M Lacerda, Luís Nacul, and Nuno Sepúlveda. Review of the quality control checks performed by current genome-wide and targeted-genome association studies on myalgic encephalomyelitis/chronic fatigue syndrome. *Frontiers in Pediatrics*, page 293, 2020.

[89] Report on the impact of covid-19 on me. Accessed: 5-3-2021.

[90] Romain K Gherardi, Guillemette Crépeaux, and François-Jérome Authier. Myalgia and chronic fatigue syndrome following immunization: macrophagic myofasciitis and animal studies support linkage to aluminum adjuvant persistency and diffusion in the immune system. *Autoimmunity Reviews*, 18(7):691–705, 2019.

[91] Jody Phelan, Anna D Grabowska, and Nuno Sepúlveda. A potential antigenic mimicry between viral and human proteins linking myalgic encephalomyelitis/chronic fatigue syndrome (me/cfs) with autoimmunity: The case of hpv immunization. *Autoimmunity Reviews*, 19(4):102487, 2020.

[92] Nirupa Gadi, Samantha C Wu, Allison P Spihlman, and Vaishali R Moulton. What's sex got to do with covid-19? gender-based differences in the host immune response to coronaviruses. *Frontiers in immunology*, 11:2147, 2020.

Supplementary Materials

Supplementary Table 1: ***ACE* and *ACE2* CpG probes.** Nineteen and eight CpG probes located in ACE and ACE2 and shared between Infinium HumanMethylation450K and Infinium HumanMethylationEPIC arrays by Illumina. The annotation was obtained from the packages "IlluminaHumanMethylation450kanno.ilmn12.hg19" and "IlluminaHumanMethylationEPICanno.ilm10b2.hg19" available from Bioconductor.

Gene (chr)	Probes	Position	Annotation
ACE (17)	cg09920557	61553938	TSS1500
	cg25054907	61553954	TSS1500
	cg02131967	61554106	TSS1500
	cg02440279	61554400	TSS200
	cg02040921	61554411	TSS200
	cg19354750	61554413	TSS200
	cg05952120	61554416	TSS200
	cg24877195	61554604	1stExon
	cg06751221	61554929	Body
	cg02261408	61559061	Body
	cg19802564	61561470	Body;TSS1500
	cg19826045	61561602	Body;TSS1500
	cg21796427	61562170	TSS200;Body
	cg04199256	61562197	Body;5'UTR;1stExon
	cg21094739	61572645	Body;Body
	cg01489398	61574335	Body;Body
	cg21881537	61574411	Body;Body
	cg21657705	61574500	Body;Body
	cg10468385	61574744	3'UTR;3'UTR
ACE2 (X)	cg23232263	15579482	3'UTR
	cg05039749	15583512	Body
	cg05748796	15619337	5'UTR
	cg16734967	15620103	5'UTR
	cg08559914	15620240	TSS200
	cg18877734	15621084	TSS1500
	cg21598868	15621167	TSS1500
	cg18458833	15621477	TSS1500

Supplementary Table 2: Summary data of the CpG probes including SNP or coincided with a polymorphic SNP.

Annotation	Probe ID	Chromosome	Position	SNP ID	Ref allele/ Alt allele	Minor allele frequency	
						North America	Europe
SNP within probes	cg04199256	17	61562197	rs191697444	C/T	0.00	0.00
	cg04199256	17	61562197	rs12720723	G/A	0.12	<0.01
	cg21881537	17	61574411	rs117135474	C/T	0.00	0.00
	cg21881537	17	61574411	rs200695691	TTGCCC/T	0.03	0.00
	cg10468385	17	61574744	rs4365	G/A	0.00	0.04
	cg02481451	17	61594924	rs149678437	C/T	<0.01	<0.01
	cg16734967	X	15620103	rs182809041	A/C	<0.01	0.00
Polymorphic SNP	cg08559914	X	15620240	rs186143966	C/T	0.00	0.00

Supplementary Table 3: **Linear models estimates.** Estimates of the best linear regression models for 5 significant CpG probes shown in Figure 27C. Healthy controls and the study of de Vega et al (2014) were considered the reference effects of the disease status and of the study indicator variable, respectively. The respective data are shown in Figure 24.

Analysis	Coefficient	*ACE*								*ACE2*	
		cg09920557		cg19802564		cg21094739		cg10468385		cg08559914[1]	
		Estimate (SE)	*P-value*	Estimate (SE)	*P-value*	Estimate (SE)	*P-value*	Estimate (SE)	*P-value*	Estimate (SE)	*P-value*
Overall	`Intercept`	−3.769 (0.065)	<0.001	1.551 (0.081)	<0.001	1.938 (0.096)	<0.001	3.082 (0.100)	<0.001	2.035 (0.078)	<0.001
	`Disease status:ME/CFS`	−0.060 (0.092)	0.519	−0.016 (0.115)	0.889	−0.064 (0.136)	0.638	−0.118 (0.142)	0.409	−0.141 (0.048)	0.004
	`Study:de Vega et al.(2017)`	−2.367 (0.079)	<0.001	0.7071 (0.098)	<0.001	1.419 (0.116)	<0.001	2.437 (0.121)	<0.001	1.703 (0.085)	<0.001
	`Study:Trivedi et al.(2018)`[2]	0.145 (0.092)	0.117	—	—	—	—	−0.625 (0.142)	<0.001	−0.88672 (0.10334)	<0.001
	`Study:Herrera et al.(2018)`	−2.245 (0.073)	<0.001	0.425 (0.091)	<0.001	0.740 (0.107)	<0.001	1.772 (0.112)	<0.001	1.270 (0.082)	<0.001
	`Disease status:ME/CFS× Study:de Vega et al.(2017)`	0.131 (0.108)	0.224	−0.107 (0.133)	0.425	−0.198 (0.158)	0.210	−0.159 (0.165)	0.336	—	—
	`Disease status:ME/CFS× Study:Trivedi et al.(2018)`	−0.315 (0.130)	0.016	—	—	—	—	0.264 (0.199)	0.186	—	—
	`Disease status:ME/CFS× Study:Herrera et al.(2018)`	0.048 (0.102)	0.638	0.072 (0.127)	0.571	0.157 (0.150)	0.295	0.181 (0.157)	0.250	—	—
Females	`Intercept`	−3.769 (0.068)	<0.001	1.406 (0.107)	<0.001	1.938 (0.093)	<0.001	3.082 (0.092)	<0.001	2.035 (0.068)	<0.001
	`Disease status:ME/CFS`	−0.060 (0.096)	0.535	−0.224 (0.152)	0.143	−0.064 (0.132)	0.629	−0.117 (0.130)	0.366	−0.141 (0.045)	0.002
	`Study:de Vega et al.(2017)`	−2.367 (0.082)	<0.001	0.499 (0.130)	<0.001	1.419 (0.113)	<0.001	2.437 (0.111)	<0.001	1.703 (0.075)	<0.001
	`Study:Trivedi et al.(2018)`[2]	0.145 (0.096)	0.131	—	—	—	—	−0.625 (0.130)	<0.001	−0.887 (0.090)	<0.001
	`Study:Herrera et al.(2018)`	−2.238 (0.079)	<0.001	−0.233 (0.124)	0.062	0.770 (0.108)	<0.001	1.786 (0.106)	<0.001	1.379 (0.073)	<0.001
	`Disease status:ME/CFS× Study:de Vega et al.(2017)`	0.131 (0.112)	0.241	−0.112 (0.177)	0.526	−0.198 (0.153)	0.198	−0.159 (0.151)	0.293	—	—
	`Disease status:ME/CFS× Study:Trivedi et al.(2018)`	−0.315 (0.134)	0.020	—	—	—	—	0.264 (0.182)	0.148	—	—
	`Disease status:ME/CFS× Study:Herrera et al.(2018)`	0.029 (0.109)	0.791	0.258 (0.173)	0.138	0.166 (0.150)	0.270	0.241 (0.148)	0.104	—	—

[1]The best regression model for cg08559914 included the main effects of disease status and study indicator covariate.

[2]The study of Trivedi et al (2018) discarded cg19802564 and cg21094739 from their analysis and, therefore, the respective data were not available in the NCBI GEO data repository.

Chapter 7 - Revisiting IgG antibody reactivity to EBV in ME/CFS and its potential application to disease diagnosis

N. Sepúlveda[1,2], J. Malato[2,3], F. Sotzny[4], A.D. Grabowska[5], A. Fonseca[2,6], C. Cordeiro[2,6], L. Graça[3], P. Biecek[1], U. Behrends[7,8], J. Mautner[7,8], F. Westermeier[9,10], E.M. Lacerda[11], and C. Scheibenbogen[4]

[1]Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland
[2]CEAUL – Centro de Estatística e Aplicações da Universidade de Lisboa, Lisbon, Portugal
[3]Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal
[4]Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal
[5]Department of Biophysics, Physiology, and Pathophysiology, Medical University of Warsaw, Warsaw, Poland
[6]Faculdade de Ciências e Tecnologia, Universidade do Algarve, Faro, Portugal
[7]Technical University of Munich, School of Medicine, Childrens' Hospital, Munich, Germany
[8]German Center for Infection Research (DZIF), Braunschweig, Germany
[9]Department of Health Studies, Institute of Biomedical Science, FH Joanneum University of Applied Sciences, Graz, Austria
[1]0Centro Integrativo de Biología y Química Aplicada (CIBQA), Universidad Bernardo O'Higgins, Santiago, Chile
[1]1Department of Clinical Research, Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, United Kingdom

Nuno Sepúlveda, João Malato, Franziska Sotzny, Anna D Grabowska, André Fonseca, Clara Cordeiro, Luís Graça, Przemyslaw Biecek, Uta Behrends, Josef Mautner, et al. Revisiting igg antibody reactivity to epstein-barr virus in myalgic encephalomyelitis/chronic fatigue syndrome and its potential application to disease diagnosis. Frontiers in Medicine, 9:921101, 2022.

7.1 Abstract

Infections by the Epstein-Barr virus (EBV) are often at the disease onset of patients suffering from myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS). However, serological analyses of these infections remain inconclusive when comparing patients with

healthy controls. In particular, it is unclear if certain EBV-derived antigens eliciting antibody responses have a biomarker potential for disease diagnosis. With this purpose, we re-analysed a previously published microarray data on the IgG antibody responses against 3,054 EBV-related antigens in 92 patients with ME/CFS and 50 healthy controls. This re-analysis consisted of constructing different regression models for binary outcomes with the ability to classify patients and healthy controls. In these models, we tested for a possible interaction of different antibodies with age and gender. When analysing the whole data set, there were no antibody responses that could distinguish patients from healthy controls. A similar finding was obtained when comparing patients with non-infectious or unknown disease trigger with healthy controls. However, when data analysis was restricted to the comparison between healthy controls and patients with a putative infection at their disease onset, we could identify stronger antibody responses against two candidate antigens (EBNA4_0529 and EBNA6_0070). Using antibody responses to these two antigens together with age and gender, the final classification model had an estimated sensitivity and specificity of 0.833 and 0.720, respectively. This reliable case-control discrimination suggested the use of the antibody levels related to these candidate viral epitopes as biomarkers for disease diagnosis in this subgroup of patients. To confirm this finding, a follow-up study will be conducted in a separate cohort of patients.

Keywords: Epstein-Barr virus; Myalgic encephalomyelitis/Chronic fatigue syndrome; antigen mimicry; biomarker discovery; patient stratification

7.2 Introduction

Infections by the ubiquitous Epstein-Barr virus (EBV) are linked to multiple sclerosis, rheumatoid arthritis, systemic erythematosus lupus, lymphomas, among other known diseases [1, 2, 3]. A less-known disease where EBV infections are also important is myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) [4, 5, 6]. The hallmark symptom of this condition is an unexplained but persistent fatigue that cannot be alleviated by rest and that can increase upon minimal physical and emotional effort [7, 8]. In ME/CFS, acute EBV infections are reported by a subset of patients at the onset of their symptoms [9, 10]. Reactivation of latent EBV infections has also been described during the disease course [11]. However, current evidence remains inconclusive on whether the prevalence of these reactivations is either higher or lower in patients than in healthy controls [12]. This conflicting evidence notwithstanding, ME/CFS patients show deficient B- and T-cell responses against EBV and altered antibody profiles when compared with healthy controls [10, 13, 14, 15]. Finally, CD4+ T cells recognising self-peptides on HLA-DR15, the strongest genetic risk factor for multiple sclerosis, have been shown to cross-react with peptides derived from EBV [16]. Multiple sclerosis patients share many symptoms with the ones suffering from ME/CFS [17, 18, 19]. EBV antigens were also reported to share sequence homology with

human peptides derived from the myelin basic protein [20, 21, 21], lactoperoxidase [22], and anoctamin-2 [23, 24]. These observations suggest that molecular mimicry between human and EBV-derived antigens could play a role in the pathogenesis of ME/CFS. This suggestion is in line with our recent hypothesis that links the pathogenesis of ME/CFS to chronically activated immune responses [25]. Our assumption raises the possibility that the immune system of some ME/CFS patients is oscillating between an activation state that attempts controlling latent herpesviruses infections and the suppression of deleterious autoimmune responses via the activation of regulatory T cells [25]. Thus, considering the growing body of evidence that links EBV infection to the pathogenesis of ME/CFS, studies that aim at elucidating underlying mechanisms are needed.

A major problem in investigating ME/CFS is the non-existence of a robust biomarker that could ascertain the disease diagnosis. In the past, different discovery studies suggested certain cytokines, antibodies against self and non-self-antigens, microRNAs, and methylation markers as potential disease biomarkers [26]. Antibodies against EBV antigens are of particular interest as disease biomarkers given the above evidence connecting this virus with the disease and routine application of serological assays in the clinical practice. However, EBV antigens included in commercial kits are mostly markers of exposure to the infection and are unable to distinguish between patients with ME/CFS and healthy controls [27]. This distinction can only be made when comparing a subset of clinically diagnosed ME/CFS patients with an EBV infection trigger to healthy controls [10]. A serological evaluation of antibodies against less-studied EBV antigens did not identify any that could be used as a specific disease biomarker [28]. However, this antibody evaluation was done using a limited number of EBV-derived antigens and no subgroup analysis was performed. The lack of patient stratification in ME/CFS studies reduces the chance of reproducing the same findings in follow-up studies [26, 29]. Therefore, it is still possible to identify alternative antigens whose antibody responses could be used as disease biomarkers for a subgroup of patients.

Recently, we analysed antibody responses against more than 3,000 overlapping antigens derived from 14 EBV proteins [22]. The aim of this study was to extract an antibody signature against EBV in ME/CFS patients when compared to healthy controls. In the present study, we extended the analysis of the obtained data with the specific objective of optimising biomarker discovery. In particular, we compared patients with or without an infectious trigger at disease onset to healthy controls in order to discover EBV-derived antigens whose antibody responses could be used for ME/CFS diagnosis.

7.3 Materials and methods

7.3.1 Study participants

Ninety-two ME/CFS patients were recruited between 2011 and 2015 at the Charité outpatient clinic for immunodeficiencies at the Institute of Medical Immunology in the Charité Universitatsmedizin Berlin, Germany. Additional fifty individuals were recruited from the employees of the same clinic, who self-reported to be healthy and to not suffer from fatigue. However, neither clinical nor laboratory assessment was performed to confirm the healthy status of those individuals. ME/CFS patients and healthy controls were matched for gender and age (Table 8) with 50% of women and an overall average of ~43 years of age. Fifty-four out of 92 patients (58.7%) reported an acute infection at their disease onset, whilst the remaining 38 patients (41.3%) reported either a disease trigger other than an infection, did not know their disease onset or the information about the disease trigger was missing. These two subgroups were also matched for age and gender (Table 8).

Table 8: **Basic characteristics of ME/CFS patients and healthy controls.** Basic characteristics of ME/CFS patients and healthy controls, where p-values refer to the comparison between ME/CFS groups and healthy controls.

Group	Female			Age, years	
	N	%	P-value	Mean (age range)	P-value
Healthy controls	50	50.0	—	42.4 (25–61)	—
ME/CFS (all)	92	51.1	0.901	43.7 (25–66)	0.453
With infectious trigger	54	50.0	~1.000	43.2 (17–66)	0.585
Unknown trigger or without infectious trigger	38	52.6	0.807	44.4 (24–66)	0.679

7.3.2 Peptide array

Data under analyses refer to the signal intensities derived from IgG antibody responses to 3,054 EBV-associated peptides measured by a seroarray described in detail in the original study [22]. These peptides consisted of partially overlapping 15 amino acids (15-mer) and covered the full length of the following proteins (Supplementary Table 4): BALF-2, BALF-5, BFRF-3, BLLF-1, BLLF-3, BLRF-2, BMRF-1, BZLF-1, EBNA-1, EBNA-3, EBNA-4, EBNA-6, LMP-1, and LMP-2. The 15-mer peptides overlapped in 11 amino acids. The amino-acid sequences of these peptides were representative of the following EBV strains: AG876 (West Africa, EBV type 2), B95.8 (USA, EBV type 1), GD1 (China, EBV type 1), Cao (China, EBV type 1), Raji (Nigeria, EBV type 1), and P3HR-1 (Nigeria, EBV type 2). These data are freely available in Supplementary File S1 of the original study [22].

7.3.3 Statistical analysis

We used the Pearson's χ^2 test to compare ME/CFS patients to healthy controls in terms of gender distribution. The non-parametric Mann-Whitney test was used to compare the medians of the respective age distributions. There was evidence for age- and gender-matched distributions if the p-values of these tests were greater than the significance level of 5%.

We first performed a multivariate analysis using (i) the classical principal component analysis (PCA) and (ii) computing different correlation matrices using Spearman's correlation coefficient (which is invariant to monotonic changes in the scale of the data, is robust against the presence of outliers, and does not depend on the normality assumption). We then performed linear discriminant analyses (LDA) to determine the best linear combination of all the antibody responses that could distinguish ME/CFS patients and their subgroups from healthy individuals. A similar analysis was done to compare the two subgroups of ME/CFS patients.

The outcome of each LDA was the estimated classification probability for each individual. These estimated probabilities were then analyzed by the respective receiver operating characteristic (ROC) curve where 1−specificity and sensitivity are plotted against each other as a function of the cutoff of the underlying classification probability. After computing each ROC curve, we calculated the respective area under the curve (AUC) and its 95% confidence interval to determine the accuracy of the classification irrespective of the cutoff used. In general, an AUC = 0.50 is indicative of a complete random classification of the individuals, while AUC = 1.00 implies that the constructed classifier perfectly predicts the true class membership of each individual.

We performed further antibody-wide association analyses related to the following comparisons (or classification exercises): (i) healthy controls versus all the ME/CFS patients; (ii) healthy controls versus ME/CFS patients with an infectious trigger; (iii) healthy controls versus ME/CFS patients with a non-infectious or unknown trigger; and (iv) ME/CFS patients with an infectious trigger versus the remaining ME/CFS patients. In each association analysis, we first estimated three regression models: logistic model, probit model, and complementary log-log model. In these models, the disease status was the outcome variable, age and gender were the respective covariates. To determine the best link function for the outcome variable, we selected the model with the lowest Akaike's information criterion (AIC). For the best link function ("the null model"), we estimated the respective ROC and its AUC as described above.

We fitted five different logistic models, including the main effects and all the interaction terms related to age, gender, and the antibody response under analysis: (i) a model with main effects only and no interaction terms; (ii) a model with an interaction term between age and the antibody response; (iii) a model with an interaction term between gender and the antibody response; (iv) a model with two interaction terms between age and

the antibody response and between gender and the antibody response; (v) a model with all two-way and three-way interaction terms related to age, gender, and the antibody response. We compared each of these models with the null one using Wilks's likelihood ratio test, where low p-values provide evidence for these models, including effects of an antibody response. We reported the minimum p-value obtained from these model comparisons. Finally, we adjusted the minimum p-values of each analysis. This adjustment was made using the Benjamini-Yekutieli procedure ensuring a global false discovery rate (FDR) of 5% under the assumption of dependent tests [30]. In this analysis, adjusted p-values < 0.05 indicated statistically significant results.

To filter out redundant antibody responses, we pooled all the significant antibody responses in a single model. The effect and interaction terms of these antibody responses were defined according to the most significant model obtained in the previous stage of analysis. We performed a backward stepwise model selection. The resulting model was finally evaluated in terms of predictive performance using ROC analysis as described above.

The above analysis was primarily done for the whole data set irrespective of the ME/CFS subgroups. We repeated the same analysis to compare each subgroup of ME/CFS patients (with infectious and non-infectious or unknown disease trigger) with the healthy controls. Finally, we repeated the analysis to compare the two subgroups of ME/CFS patients.

7.3.4 Statistical software

The statistical analysis was performed in the R software version 4.0.3 with core functions and the following packages: *MASS* v7.3-56 to perform stepwise model selection [31], *pROC* v1.18.0 to estimate the ROC curve and the respective AUC [32], *OptimalCutpoints* v1.1-5 to estimate the optimal cut-off and the associated sensitivity/specificity [33]. The full reproducible code is freely available from NS or JMal upon request.

7.4 Results

7.4.1 Principal component and linear discriminant analyses

We first performed a PCA to discriminate patients with ME/CFS and their subgroups from healthy controls (Figures 27A–C). A similar analysis was done for discriminating patients with an infectious trigger from the remaining patients (Figure 27D).

The proportion of variance explained by the first principal component varied from 35.4% (Figure 27D) to 44.6% (Figure 27C) referring to the comparisons between the two subgroups of ME/CFS patients, and between healthy controls and patients with non-infectious or unknown disease trigger, respectively. These high estimates suggested that different antibody levels were correlated with each other. This interpretation was confirmed by determining the distributions of Spearman's correlation coefficient between all possible pairs of anti-

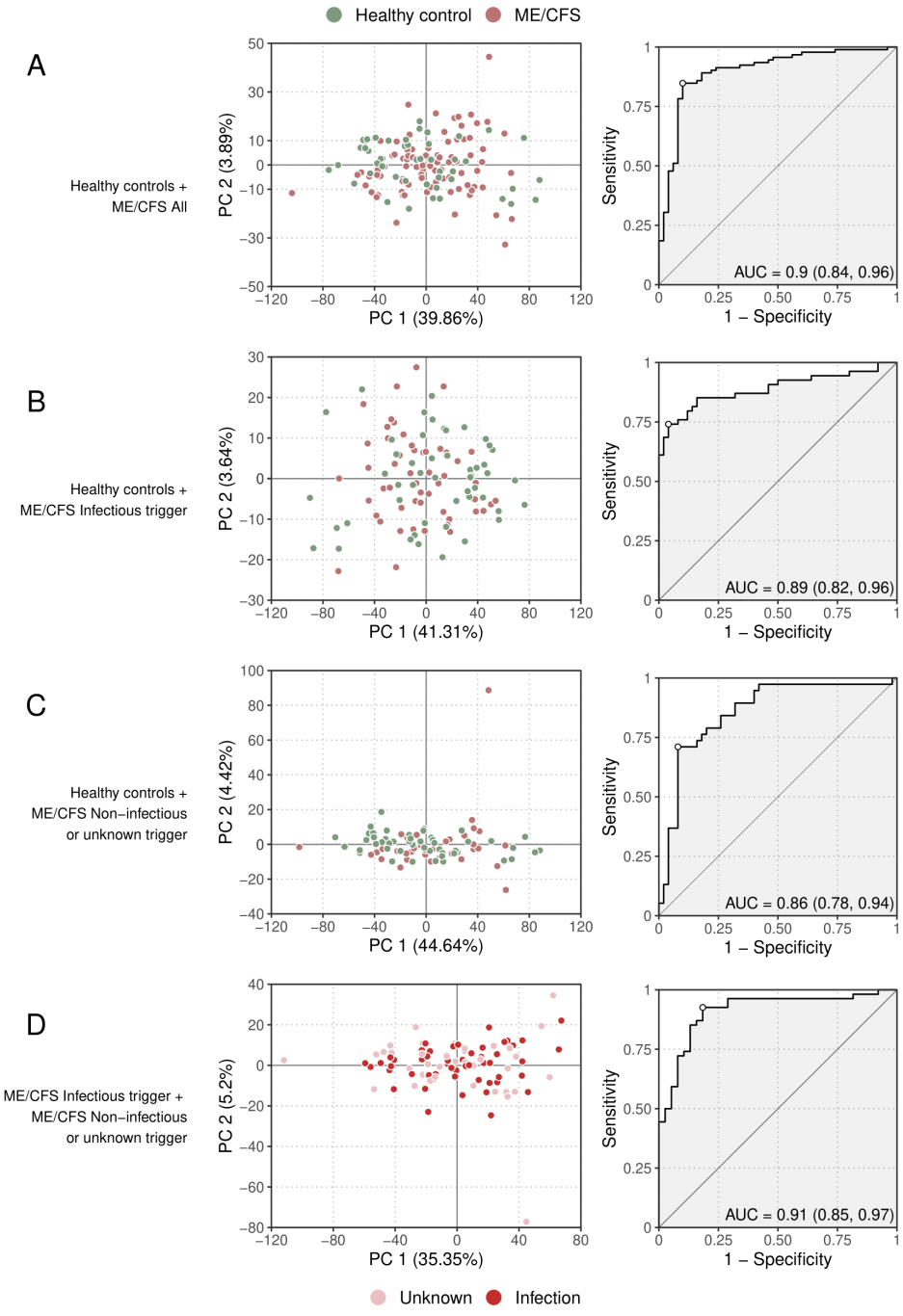


Figure 27: **Preliminary multivariate analysis of the data.** Preliminary multivariate analysis of the data. Scatterplots of the first two principal components (left plots) and the ROC curve and its AUC of the respective LDA (right plots) when comparing all the ME/CFS patients to healthy controls (A), ME/CFS patients with an infectious trigger to healthy controls (B), ME/CFS patients with a non-infectious or unknown trigger to healthy controls (C), and ME/CFS patients with an infectious trigger to the remaining patients (D). The percentage of the variance explained by each principal component is shown in each axis within brackets.

bodies using data from each study group (Supplementary Figure 1). In particular, the antibody levels were positively correlated with each other with median correlation estimates of 0.56, 0.56, 0.40, and 0.48 for healthy controls, all the ME/CFS patients, ME/CFS patients with an infectious disease trigger, and the remaining ME/CFS patients, respectively. Interestingly, the median correlation estimate was decreased in ME/CFS patients with an infectious trigger when compared to other study groups. This finding suggested that the production of the antibodies against the EBV-derived antigens could be reduced in these patients when compared to healthy controls or patients with non-infectious or unknown disease trigger.

The visualisation of the first two components did not reveal a clear discrimination between healthy controls and ME/CFS patients (or their subgroups). To improve this analysis, we then performed different LDAs in search of a linear combination of the antibody measurements that could be used for disease diagnosis. The performance of the constructed classifiers ranged from 0.86 (Figure 27C) to 0.91 (Figure 27D) referring to the classification of healthy controls and ME/CFS patients with non-infectious or unknown disease trigger and the classification of the two subgroups of ME/CFS patients, respectively. Therefore, the results of this analysis indicate that the antibody data could discriminate different study groups.

7.4.2 Antibody-wide association analysis

The next step of the analysis was to identify specific antibody responses that could be used to discriminate the different study groups. With this purpose, we first determined the best "null" model among the logistic, probit, and complementary log-log models. All of them included age and gender and their interaction as covariates for each comparison between any two study groups (Supplementary Table 5). The best "null" models were the following: (i) complementary log-log – comparison between healthy controls and all the ME/CFS patients (AUC = 0.574; 95% CI = [0.475, 0.672]); (ii) probit – comparison between healthy controls and ME/CFS patients with an infectious trigger (AUC = 0.606; 95% CI = [0.496, 0.715]); (iii) complementary log-log – comparison between healthy controls and ME/CFS patients with a non-infectious or unknown trigger (AUC = 0.556; 95% CI = [0.429, 0.683]); and (iv) logit – comparison between the two subgroups of ME/CFS groups (AUC = 0.596; 95% CI = [0.471, 0.720]). The 95% confidence interval for the AUC of these null models included 0.50 and therefore, the respective predicted classification was consistent with a random guess. Such a result was in agreement with the age and gender matching between different study groups and healthy controls (Table 8).

We performed further antibody-wide association analyses controlling for a global FDR of 5%. The comparison between healthy controls and all the ME/CFS patients did not identify any significant antibody associations with the disease (Figure 28A). The top 5 antibod-

ies, although not statistically significant, were EBNA6_0066, BLRF2_0005, EBNA4_0392, EBNA4_0497, and EBNA4_0529 (adjusted p-values = 0.181, 0.326, 0.326, 0.326, and 0.326, respectively).

When the comparison was limited to healthy controls and ME/CFS patients with an infectious trigger, we identified three significant antibodies related to the following antigens (Figure 28B): EBNA6_0066, EBNA6_0070, and EBNA4_0529 (adjusted *p-values* = 0.005, 0.005, and 0.038, respectively). The first two antigens were shared between AG876, B95.8, and GD1 strains, while the third one was derived from the B95.8 strain. We compared ME/CFS patients with non-infectious or unknown disease trigger to healthy controls, and found no significant differences in the antibody responses (Figure 28C). The same finding was obtained when we compared the two subgroups of ME/CFS patients (Figure 28D). The top 5 antibodies related to these analyses can be found in Supplementary Table 6.

7.4.3 Analysis of candidate antigens for classifying ME/CFS patients with infectious trigger

We then analysed in detail the impact of the antibody levels against the three candidate antigens on the classification of ME/CFS patients with an infectious trigger. Antibody levels were increased in this subgroup of ME/CFS patients when compared to healthy controls (Figure 29A). The same evidence could not be found when comparing all the ME/CFS patients to healthy controls (Figure 29A). Data related to EBNA4_0529, EBNA6_0066 and EBNA6_0070 were significantly correlated with each other (Spearman's correlation coefficients higher than 0.58; Figure 29B). The correlation between the levels of antibodies against EBNA6_0066 and EBNA6_0070 could be explained by the fact that these two peptides are 15-mers overlapping 11 amino acids with each other [22]. In contrast, it was unclear why the levels of antibodies against EBNA4_0529 and EBNA6_0066 were highly correlated (Spearman's correlation coefficient = 0.79), considering that these antigens did not share a high sequence homology (Figure 29C).

Given the high correlation between antibody levels related to these antigens, a statistical redundancy was expected when using their data for patients' classification purpose. This redundancy was confirmed when the three candidate antibodies were included as covariates in the same model. A stepwise variable selection procedure led to the exclusion of the antibody levels related to EBNA6_0066 from the final classification model.

The final model included the main effects of antibodies to EBNA4_0529 and EBNA6_0070 and the two-way interaction of the latter with age and gender (Table 9). On the one hand, the $\log_{10}$-levels of antibodies related to EBNA4 increased the probability of being a patient (coefficient estimate = 2.25, standard error = 1.09). In particular, the odds of being a patient were estimated to increase ∼9.5 ($e^{2.25}$) times per fold-change in the levels of these antibodies. On the other hand, the effects of antibody levels related to EBNA6_0070 on the

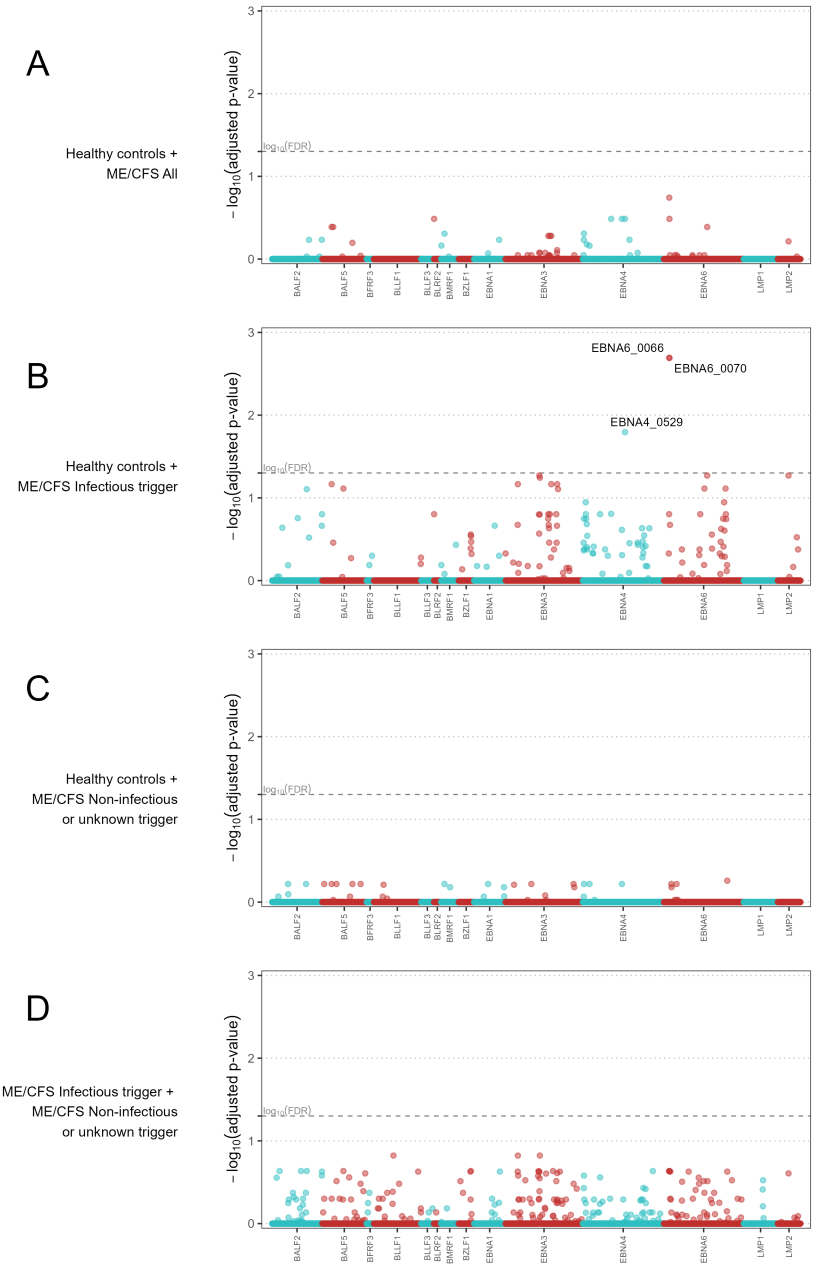


Figure 28: **Antibody-wide association analyses.** Antibody-wide association analyses when comparing all the ME/CFS patients to healthy controls (A), ME/CFS patients with an infectious trigger to healthy controls (B), ME/CFS patients with a noninfectious or unknown trigger to healthy controls (C), and ME/CFS patients with an infectious trigger to the remaining patients (D). The x-axes comprise each antibody while the y-axes represent the $-\log_{10}$(adjusted p-value) of the respective association. In the x-axes, the antibodies were ordered alphabetically first by the protein name and then by the starting point of the antigen within the protein. Adjusted p-values were calculated according to the Benjamini-Yekutieli procedure for a global FDR of 5% under the assumption of dependent data. Dashed line represents the threshold for statistical significance (i.e., $-\log_{10}(\text{FDR} = 0.05)$) and $-\log_{10}(\text{adjusted p-value}) > 1.30$ were considered statistically significant.

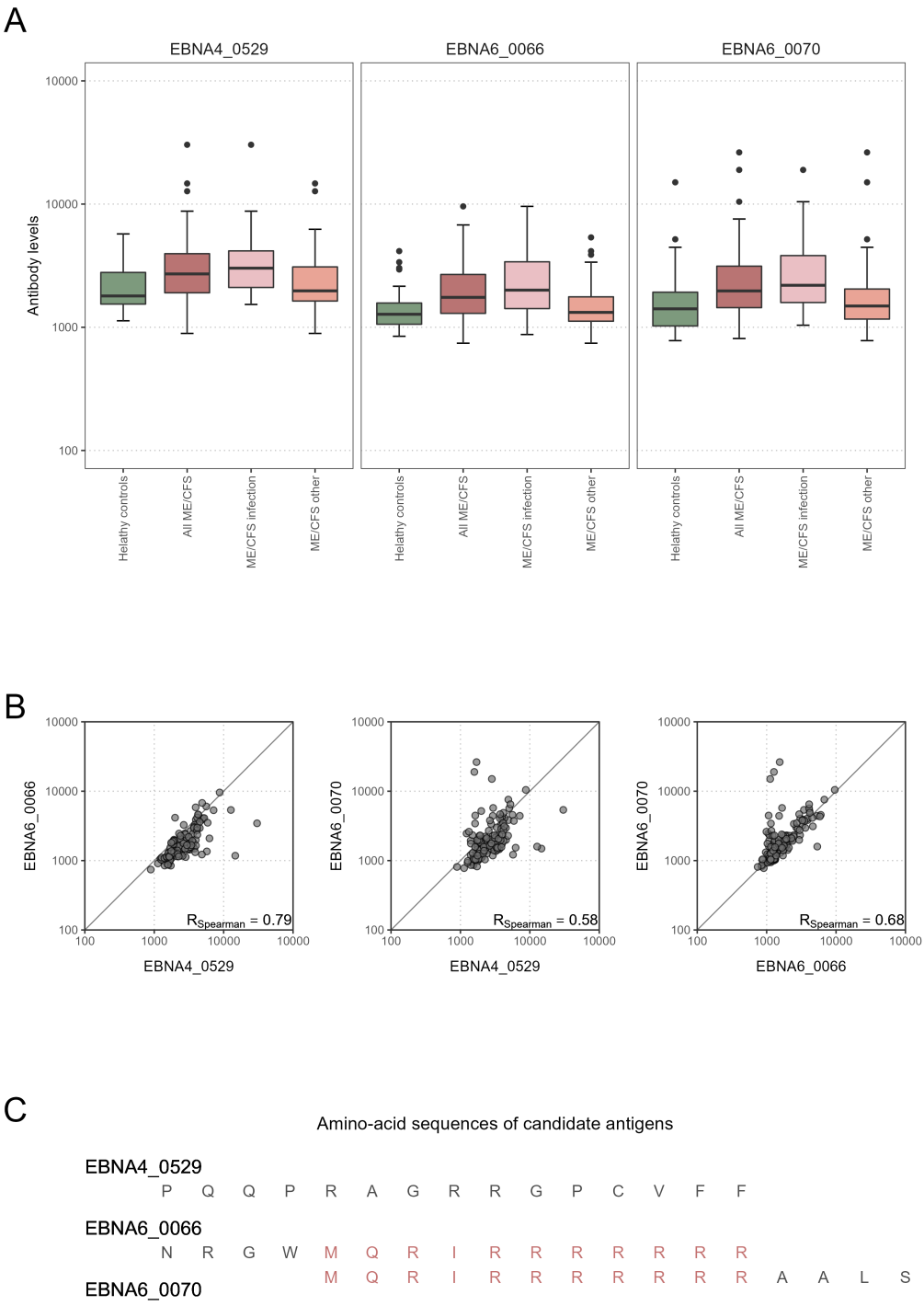


Figure 29: **Statistical analysis of the antibody levels related to EBNA4_0529, EBNA6_0066, and EBNA6_0070.** (A) Boxplots of the data per study group. (B) Scatterplots and the respectively Spearman's correlation coefficients (R) in the whole dataset. (C) Amino acid sequences of EBNA4_0529, EBNA6_0066, and EBNA6_0070.

probability of an individual being an ME/CFS patients were not so trivial to ascertain (Figure 30A). In particular, women with high EBNA6_0070 antibody levels showed an increasing estimated probability of being a patient with increasing age. In contrast, the probability profile of being patient was different in men. In that case, younger men with low EBNA6_0070 antibody levels or older men with high EBNA6 antibody levels had a higher probability of being a patient.

The AUC of the classification predicted by the final model was estimated at 0.835 with a 95% CI = (0.759, 0.911) (Figure 30B). This estimate suggested that the combination of these two antibodies together with age and gender could be used for the diagnosis of patients with an infectious trigger. The optimal sensitivity and specificity were estimated at 0.833 and 0.720, respectively. Therefore, ME/CFS patients were better discriminated than healthy controls by this model.

When the same classification model was applied to the whole cohort of ME/CFS patients, the AUC decreased to 0.731 with a 95% CI = (0.648, 0.814). This could be explained by the cohort of patients with a non-infectious or unknown trigger in which the performance of the classification model was close to a random guess (AUC = 0.583; 95% CI = [0.461, 0.705]).

Table 9: **Complmentary log-log estimates.**Estimates of the final complementary log-log model to discriminate ME/CFS patients with an infectious disease trigger from healthy controls.

Model term	Coefficient estimate (SE)	P-value
`Intercept`	10.67 (10.33)	0.302
`Age (in years)`	−0.49 (0.26)	0.060
`Gender (Woman)`	−17.33 (6.85)	0.011
`EBNA4_0529`	2.25 (1.09)	0.039
`EBNA6_0070`	−5.62 (3.09)	0.069
`Age×Gender`	0.07 (0.04)	0.070
`Gender×EBNA6_0070`	4.05 (1.75)	0.021
`Age×EBNA6_0070`	0.15 (0.08)	0.062

7.5 Discussion

This study, based on previously published data, aimed to discover EBV-derived antigens that could elicit distinct antibody responses in ME/CFS patients when compared to healthy controls. The key finding was the identification of two candidate antigens inducing increased antibody responses in ME/CFS patients with an infectious trigger. The high sensitivity and specificity of our classification model including these antibodies suggest

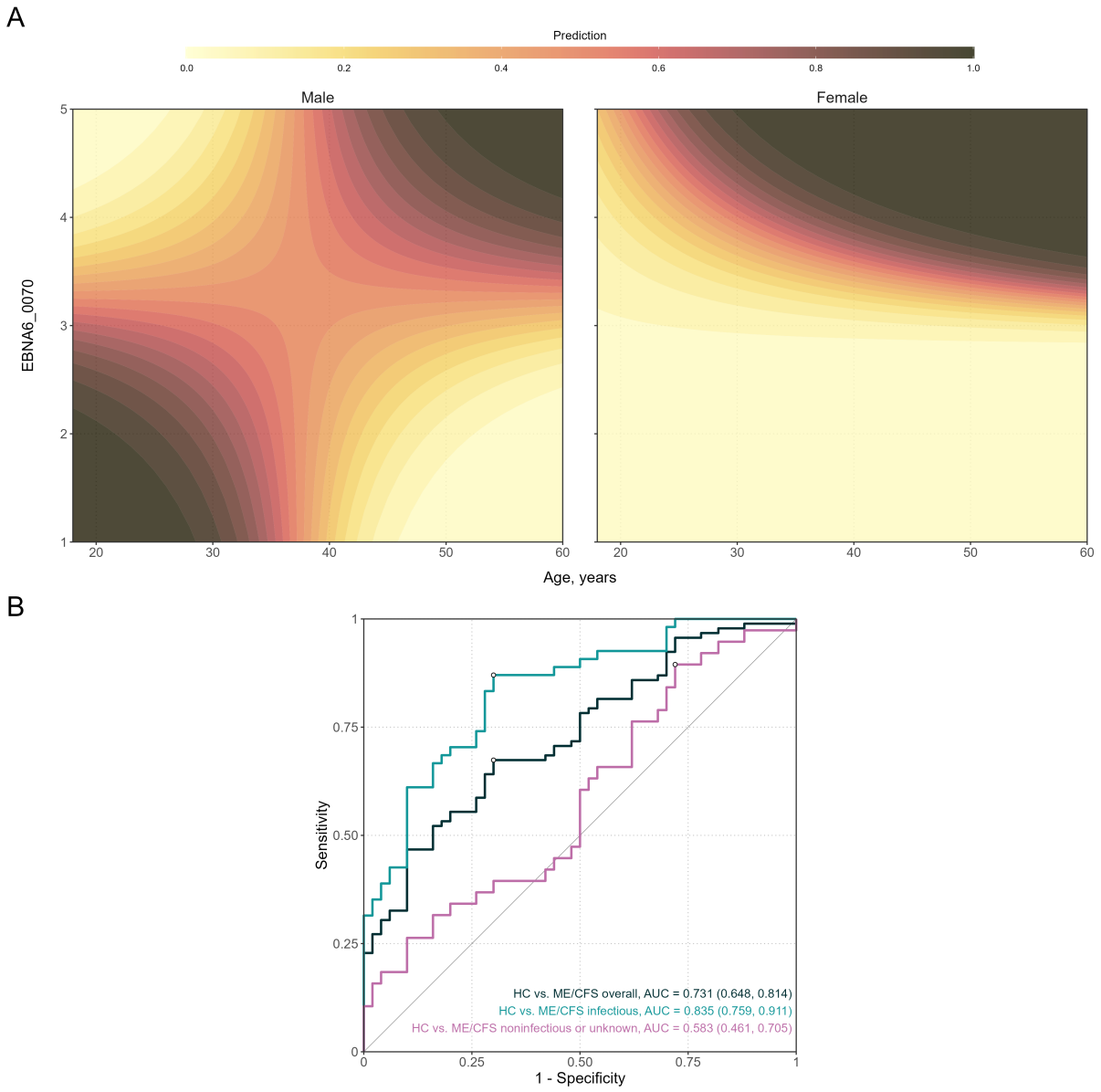


Figure 30: **Analysis of the final classification model for predicting ME/CFS patients with an infectious trigger when compared to healthy controls.** (A) Contour plots of the probability of being a patient as a function of age and EBNA6_0070 antibody levels, for men and women, respectively. The prediction values were calculated by fixing $\log_{10}(EBNA4\_0529)$ at the respective mean value. (B) ROC curves and the respective AUC (95% confidence interval shown within brackets) when using the model to compare different groups of ME/CFS patients to healthy controls.

their potential for diagnosis of this subgroup of affected individuals. For ME/CFS patients without an infectious trigger, we could not find any antigens causing antibody responses that could be used for diagnostic purposes. This finding is in agreement with an extensive serological investigation of different herpesviruses in ME/CFS patients [28]. This negative finding supports the hypothesis that EBV plays a role in the group of ME/CFS patients with an infectious trigger. In a subset of patients, infectious mononucleosis caused by primary

EBV infection can be documented as a trigger [10]. In many others, no infection with a specific pathogen could be associated with the disease onset [5]. A tempting hypothesis from our finding is that EBV reactivation which can occur during other infections may play until now an underestimated role in triggering ME/CFS. In line with this concept, a recent study showed that EBV reactivation during Covid-19 is a risk factor for Long Covid which also includes ME/CFS [34]. Alternatively, the responses to the EBNA6 peptides are due to a cross-reactivity to other pathogens, as outlined below.

Other findings of this study pointed to three key challenges associated with the discovery of a biomarker. Firstly, it is difficult to identify a disease-specific biomarker for all the ME/CFS patients. Thus, given the heterogeneous nature of ME/CFS, it is pivotal to stratify patients adequately [29], based on age, gender, and disease trigger for biomarker discovery [26]. In this regard, the identification of antibody patterns specific to ME/CFS patients with an infectious trigger was in agreement with other studies where significant results could be found for the same subgroup of patients [10, 35, 36]. However, given the vast number of infectious agents associated with ME/CFS [5, 37], it is worth noting that this subgroup of patients could be further subdivided according to the nature of the causative infection. In this regard, the data about the infectious agents that could have initiated ME/CFS are either inconclusive or simply based on self-reported history in most patients, as demonstrated by the data from the United Kingdom ME/CFS Biobank, where only a minority of patients had their infection confirmed with the lab test [10]. Secondly, the final classification model included non-trivial statistical interactions of antibodies against EBNA6_0070 with both age and gender. This finding implies that significant interactions between candidate biomarkers and confounding factors might be overlooked by analysts or, even when tested, they are likely to be discarded due to the small sample sizes to detect them. The presence of these interactions might be yet another factor that contributes to the lack of reproducibility between biomarker studies on ME/CFS. A proposed strategy to overcome this limitation is to conduct more advanced statistical analyses including the application of machine learning techniques which intrinsically consider the complexity of a large set of clinical and biological data, as demonstrated in drug discovery [38]. Thirdly, the interaction between the candidate antibodies against EBNA6_0070 and gender implied a remarkable distinct antibody signature between male and female patients. Again, this finding is in line with gender differences in immunity to viral infection [39]. In particular, men have typically lower antibody responses when vaccinated and are more susceptible to infections than women [40]. In this regard, our study suggested that the higher probability of younger man being an ME/CFS patient is associated with lower levels of antibodies against the antigen EBNA6_0070. In contrast, female and male patients seemed to be at higher risk with higher antibodies at increasing age suggesting that at least a subset develop these antibody responses later in life. An implication of having a different antibody profiling between men and women is that analysis of each gender should be performed separately. At

the same time, it is important to note that epidemiological data on ME/CFS suggested approximately a disease ratio of three women to one man [41, 42, 43]. Therefore, if gender is an important stratification factor for biomarker discovery, studies should be designed toward a more balanced gender ratio. Similar sample sizes between male and female cohorts ensure comparable statistical power when analysing data from each sex separately.

Both EBNA4_0529 and EBNA6_0070 antigens are derived from proteins whose genetic expression typically occurs during the EBV type III latency. Therefore, the acquisition of the respective antibodies might have occurred during initial B-cell transformation and immortalization. It could also be acquired slowly over time, given that the type III latency pattern can be detected sporadically in lymphoid follicles where EBV-infected B cells can proliferate and mimic a germinal center reaction program [44]. We can hypothesize from our data that both male and female patients developing higher antibody responses against this antigen later in life are at an increased risk of developing ME/CFS suggesting that reactivation of EBV plays a role. In male patients a subgroup with lower EBNA6 antibodies early in live is at risk of developing ME/CFS, too. Using the recent analytical framework of ME/CFS natural progression [45], antibodies against these antigens are more likely to be biomarkers of patients suffering from ME/CFS more than 2 years of disease rather than the ones either in prodromal period or at early stages in line with our findings. Based on that assumption, these antibodies seemed more appropriate for diagnosing putative patients with delayed disease diagnosis rather than early suspected cases. However, it is known that the delay of ME/CFS diagnosis is a recurrent problem in the clinic [8, 46]. As such, we anticipate a higher utility of these antibodies when redeployed to real-world screening. Another practical implication of using these antibodies as biomarkers is the possibility of developing routine ELISA kits that can be standardised across different laboratories and easily scalable for large population screenings. Notwithstanding these promising practical expectations, it is important to emphasise that past studies also suggested potential disease biomarkers [26] and, therefore, it is imperative to replicate the findings of this study with different cohorts of patients.

An interesting observation is that both EBNA6_0066 and EBNA6_0070 contain an arginine-repeat sequence. Such a sequence has homologies with putative epitopes from several human proteins [47]. Such homologies suggest a potential molecular mimicry between the viral and human antigens. Molecular mimicry can trigger deleterious autoimmune responses as hypothesised for ME/CFS pathogenesis [37, 48]. Molecular mimicry between human and microbial antigens has been also hypothesised for several autoimmune diseases [49], such as multiple sclerosis and rheumatoid arthritis, and Long Covid, whose patients share similar symptoms with ME/CFS ones [19, 50, 51, 52]. Interestingly, T cell clones recognising such arginine-repeat sequences were isolated from a patient with multiple sclerosis supporting our concept of epitope mimicry [47]. Finally, arginine-repeat sequences are found in various other pathogens including enteroviruses and human papil-

lomavirus which are also triggers of ME/CFS [5].

Further we can hypothesise that peptides highly enriched in arginine residues might be particularly susceptible to citrullination, in which arginine residues are post-translationally converted to citrulline. These post-translational modifications occur during cell death under normal physiological conditions. However, under chronic inflammation, the accumulation of citrullinated (auto)antigens in inflamed sites might lead to deleterious autoimmune responses, thus, promoting the onset of different autoimmune diseases [53]. A potential cross-reactivity between microbial and citrullinated human antigens could also be a mechanism by which an autoimmune disease can be triggered. In rheumatoid arthritis, antibodies against EBNA-1 peptides were shown to cross-react with denatured collagen and keratin [54]. However, in the present study, we could not find any antibodies against EBNA-1-derived peptides to be associated with ME/CFS. Interestingly, the serum levels of citrulline were reported to be elevated in ME/CFS patients when compared to healthy controls [55]. However, another study could not confirm this finding, but instead provided evidence for increased plasma levels of arginine residues [56]. Another source of antigen modification is the process of generating new and more immunogenic epitopes from ubiquitous molecules upon oxidative and nitrosative stress. In ME/CFS, IgM antibodies against several of these neoepitopes, including NO-Arginine, were increased in patients [57]. In all of these possible scenarios, it is imperative to investigate the stability of this candidate biomarker antigen to post-translational modifications that could be occurred and eventually increased during the disease course.

In conclusion, this study identified two candidate antigens whose antibodies could be used to identify ME/CFS patients with an infectious trigger. To strengthen our findings, two other cohorts of patients are currently studied, including the well-characterised ME/CFS patients with different disease triggers and healthy controls from the United Kingdom ME/CFS biobank [10].

References

[1] Gunnar Houen and Nicole Hartwig Trier. Epstein-barr virus and systemic autoimmune diseases. *Frontiers in immunology*, 11:587380, 2021.

[2] Claire Shannon-Lowe, Alan B Rickinson, and Andrew I Bell. Epstein–barr virus-associated lymphomas. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1732):20160271, 2017.

[3] Kjetil Bjornevik, Marianna Cortese, Brian C Healy, Jens Kuhle, Michael J Mina, Yumei Leng, Stephen J Elledge, David W Niebuhr, Ann I Scher, Kassandra L Munger, et al. Longitudinal analysis reveals high prevalence of epstein-barr virus associated with multiple sclerosis. *Science*, 375(6578):296–301, 2022.

[4] DENISE Koo. Chronic fatigue syndrome. a critical appraisal of the role of epstein-barr virus. *Western Journal of Medicine*, 150(5):590, 1989.

[5] Santa Rasa, Zaiga Nora-Krukle, Nina Henning, Eva Eliassen, Evelina Shikova, Thomas Harrer, Carmen Scheibenbogen, Modra Murovska, Bhupesh K Prusty, and European Network on ME/CFS (EUROMENE). Chronic viral infections in myalgic encephalomyelitis/chronic fatigue syndrome (me/cfs). *Journal of translational medicine*, 16:1–25, 2018.

[6] Manuel Ruiz-Pablos, Bruno Paiva, Rosario Montero-Mateo, Nicolas Garcia, and Aintzane Zabaleta. Epstein-barr virus and the origin of myalgic encephalomyelitis or chronic fatigue syndrome. *Frontiers in Immunology*, page 4637, 2021.

[7] M Cortes Rivera, C Mastronardi, CT Silva-Aldana, M Arcos-Burgos, and BA Lidbury. Myalgic encephalomyelitis/chronic fatigue syndrome: a comprehensive review, diagnostics 9 (2019) 91.

[8] Lucinda Bateman, Alison C Bested, Hector F Bonilla, Bela V Chheda, Lily Chu, Jennifer M Curtin, Tania T Dempsey, Mary E Dimmock, Theresa G Dowell, Donna Felsenstein, et al. Myalgic encephalomyelitis/chronic fatigue syndrome: essentials of diagnosis and management. In *Mayo clinic proceedings*, volume 96, pages 2861–2878. Elsevier, 2021.

[9] Ian Hickie, Tracey Davenport, Denis Wakefield, Ute Vollmer-Conna, Barbara Cameron, Suzanne D Vernon, William C Reeves, and Andrew Lloyd. Post-infective and chronic fatigue syndromes precipitated by viral and non-viral pathogens: prospective cohort study. *Bmj*, 333(7568):575, 2006.

[10] Tiago Dias Domingues, Anna D Grabowska, Ji-Sook Lee, Jose Ameijeiras-Alonso, Francisco Westermeier, Carmen Scheibenbogen, Jacqueline M Cliff, Luis Nacul, Eliana M Lacerda, Helena Mouriño, et al. Herpesviruses serology distinguishes different subgroups of patients from the united kingdom myalgic encephalomyelitis/chronic fatigue syndrome biobank. *Frontiers in medicine*, 8:686736, 2021.

[11] Evelina Shikova, Valentina Reshkova, Antoniya Kumanova, Sevdalina Raleva, Dora Alexandrova, Natasa Capo, and Modra Murovska. Cytomegalovirus, epstein-barr virus, and human herpesvirus-6 infections in patients with myalgic encephalomyelitis/chronic fatigue syndrome. *Journal of medical virology*, 92(12):3682–3688, 2020.

[12] Ji-Sook Lee, Eliana M Lacerda, Luis Nacul, Caroline C Kingdon, Jasmin Norris, Shennae O'Boyle, Luigi Palla, Eleanor M Riley, Jacqueline M Cliff, et al. Salivary dna loads for human herpesviruses 6 and 7 are correlated with disease phenotype in myalgic encephalomyelitis/chronic fatigue syndrome. *Frontiers in medicine*, page 1129, 2021.

[13] Jonathan R Kerr. Epstein-barr virus induced gene-2 upregulation identifies a particular subtype of chronic fatigue syndrome/myalgic encephalomyelitis. *Frontiers in Pediatrics*, page 59, 2019.

[14] A Martin Lerner, Maria E Ariza, Marshall Williams, Leonard Jason, Safedin Beqaj, James T Fitzgerald, Stanley Lemeshow, and Ronald Glaser. Antibody to epstein-barr virus deoxyuridine triphosphate nucleotidohydrolase and deoxyribonucleotide polymerase in a chronic fatigue syndrome subset. *PloS one*, 7(11):e47891, 2012.

[15] Madlen Loebel, Kristin Strohschein, Carolin Giannini, Uwe Koelsch, Sandra Bauer, Cornelia Doebis, Sybill Thomas, Nadine Unterwalder, Volker von Baehr, Petra Reinke, et al. Deficient ebv-specific b-and t-cell response in patients with chronic fatigue syndrome. *PloS one*, 9(1):e85387, 2014.

[16] Jian Wang, Ivan Jelcic, Lena Mühlenbruch, Veronika Haunerdinger, Nora C Toussaint, Yingdong Zhao, Carolina Cruciani, Wolfgang Faigle, Reza Naghavian, Magdalena Foege, et al. Hla-dr15 molecules jointly shape an autoreactive t cell repertoire in multiple sclerosis. *Cell*, 183(5):1264–1281, 2020.

[17] João Malato, Luís Graça, Luís Nacul, Eliana Lacerda, and Nuno Sepúlveda. Statistical challenges of investigating a disease with a complex diagnosis. *medRxiv*, pages 2021–03, 2021.

[18] Gerwyn Morris and Michael Maes. Myalgic encephalomyelitis/chronic fatigue syndrome and encephalomyelitis disseminata/multiple sclerosis show remarkable levels of similarity in phenomenology and neuroimmune characteristics. *BMC medicine*, 11(1):1–23, 2013.

[19] Tarek A-ZK Gaber, Wah Wah Oo, and Hollie Ringrose. Multiple sclerosis/chronic fatigue syndrome overlap: when two common disorders collide. *NeuroRehabilitation*, 35(3):529–534, 2014.

[20] Kai W Wucherpfennig and Jack L Strominger. Molecular mimicry in t cell-mediated autoimmunity: viral peptides activate human t cell clones specific for myelin basic protein. *Cell*, 80(5):695–705, 1995.

[21] Trygve Holmøy, Espen Østhagen Kvale, and Frode Vartdal. Cerebrospinal fluid cd4+ t cells from a multiple sclerosis patient cross-recognize epstein-barr virus and myelin basic protein. *Journal of neurovirology*, 10(5):278–283, 2004.

[22] Madlen Loebel, Maren Eckey, Franziska Sotzny, Elisabeth Hahn, Sandra Bauer, Patricia Grabowski, Johannes Zerweck, Pavlo Holenya, Leif G Hanitsch, Kirsten Wittke, et al.

Serological profiling of the ebv immune response in chronic fatigue syndrome using a peptide microarray. *PloS one*, 12(6):e0179124, 2017.

[23] Katarina Tengvall, Jesse Huang, Cecilia Hellström, Patrick Kammer, Martin Biström, Burcu Ayoglu, Izaura Lima Bomfim, Pernilla Stridh, Julia Butt, Nicole Brenner, et al. Molecular mimicry between anoctamin 2 and epstein-barr virus nuclear antigen 1 associates with multiple sclerosis risk. *Proceedings of the National Academy of Sciences*, 116(34):16955–16960, 2019.

[24] Nuno Sepulveda. Impact of genetic variation on the molecular mimicry between anoctamin-2 and epstein-barr virus nuclear antigen 1 in multiple sclerosis. *Immunology Letters*, 238:29–31, 2021.

[25] Nuno Sepúlveda, Jorge Carneiro, Eliana Lacerda, and Luis Nacul. Myalgic encephalomyelitis/chronic fatigue syndrome as a hyper-regulated immune system driven by an interplay between regulatory t cells and chronic human herpesvirus infections. *Frontiers in immunology*, 10:2684, 2019.

[26] Carmen Scheibenbogen, Helma Freitag, Julià Blanco, Enrica Capelli, Eliana Lacerda, Jerome Authier, Mira Meeus, Jesus Castro Marrero, Zaiga Nora-Krukle, Elisa Oltra, et al. The european me/cfs biomarker landscape project: an initiative of the european network euromene. *Journal of translational medicine*, 15(1):1–7, 2017.

[27] Jacqueline M Cliff, Elizabeth C King, Ji-Sook Lee, Nuno Sepúlveda, Asia-Sophia Wolf, Caroline Kingdon, Erinna Bowman, Hazel M Dockrell, Luis Nacul, Eliana Lacerda, et al. Cellular immune function in myalgic encephalomyelitis/chronic fatigue syndrome (me/cfs). *Frontiers in immunology*, page 796, 2019.

[28] Jonas Blomberg, Muhammad Rizwan, Agnes Böhlin-Wiener, Amal Elfaitouri, Per Julin, Olof Zachrisson, Anders Rosén, and Carl-Gerhard Gottfries. Antibodies to human herpesviruses in myalgic encephalomyelitis/chronic fatigue syndrome patients. *Frontiers in Immunology*, 10:1946, 2019.

[29] Leonard A Jason, Karina Corradi, Susan Torres-Harding, Renee R Taylor, and Caroline King. Chronic fatigue syndrome: the need for subtypes. *Neuropsychology review*, 15:29–58, 2005.

[30] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

[31] William N Venables and Brian D Ripley. *Modern applied statistics with S-PLUS*. Springer Science & Business Media, 2013.

[32] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC bioinformatics*, 12(1):1–8, 2011.

[33] Mónica López-Ratón, María Xosé Rodríguez-Álvarez, Carmen Cadarso-Suárez, and Francisco Gude-Sampedro. Optimalcutpoints: an r package for selecting optimal cutpoints in diagnostic tests. *Journal of statistical software*, 61:1–36, 2014.

[34] Yapeng Su, Dan Yuan, Daniel G Chen, Rachel H Ng, Kai Wang, Jongchan Choi, Sarah Li, Sunga Hong, Rongyu Zhang, Jingyi Xie, et al. Multiple early factors anticipate post-acute covid-19 sequelae. *Cell*, 185(5):881–895, 2022.

[35] Sophie Steiner, Sonya C Becker, Jelka Hartwig, Franziska Sotzny, Sebastian Lorenz, Sandra Bauer, Madlen Löbel, Anna B Stittrich, Patricia Grabowski, and Carmen Scheibenbogen. Autoimmunity-related risk variants in ptpn22 and ctla4 are associated with me/cfs with infectious onset. *Frontiers in Immunology*, 11:578, 2020.

[36] Marvin Szklarski, Helma Freitag, Sebastian Lorenz, Sonya C Becker, Franziska Sotzny, Sandra Bauer, Jelka Hartwig, Harald Heidecke, Kirsten Wittke, Claudia Kedor, et al. Delineating the association between soluble cd26 and autoantibodies against g-protein coupled receptors, immunological and cardiovascular parameters identifies distinct patterns in post-infectious vs. non-infection-triggered myalgic encephalomyelitis/chronic fatigue syndrome. *Frontiers in Immunology*, page 1077, 2021.

[37] Jonas Blomberg, Carl-Gerhard Gottfries, Amal Elfaitouri, Muhammad Rizwan, and Anders Rosén. Infection elicited autoimmunity and myalgic encephalomyelitis/chronic fatigue syndrome: an explanatory model. *Frontiers in immunology*, 9:229, 2018.

[38] Rohan Gupta, Devesh Srivastava, Mehar Sahu, Swati Tiwari, Rashmi K Ambasta, and Pravir Kumar. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Molecular diversity*, 25:1315–1360, 2021.

[39] Henning Jacobsen and Sabra L Klein. Sex differences in immunity to viral infections. *Frontiers in immunology*, 12:720952, 2021.

[40] Peter Aaby, Christine Stabell Benn, Katie L Flanagan, Sabra L Klein, Tobias R Kollmann, David J Lynn, and Frank Shann. The non-specific and sex-differential effects of vaccines. *Nature Reviews Immunology*, 20(8):464–470, 2020.

[41] Lily Chu, Ian J Valencia, Donn W Garvert, and Jose G Montoya. Onset patterns and course of myalgic encephalomyelitis/chronic fatigue syndrome. *Frontiers in pediatrics*, 7:12, 2019.

[42] Samantha C Johnston, Donald R Staines, and Sonya M Marshall-Gradisnik. Epidemiological characteristics of chronic fatigue syndrome/myalgic encephalomyelitis in australian patients. *Clinical epidemiology*, pages 97–107, 2016.

[43] Luis C Nacul, Eliana M Lacerda, Derek Pheby, Peter Campion, Mariam Molokhia, Shagufta Fayyaz, Jose CDC Leite, Fiona Poland, Amanda Howe, and Maria L Drachler. Prevalence of myalgic encephalomyelitis/chronic fatigue syndrome (me/cfs) in three regions of england: a repeated cross-sectional study in primary care. *BMC medicine*, 9(1):1–12, 2011.

[44] David A Thorley-Lawson. Ebv persistence—introducing the virus. *Epstein Barr Virus Volume 1: One Herpes Virus: Many Diseases*, pages 151–209, 2015.

[45] Luis Nacul, Shennae O'Boyle, Luigi Palla, Flavio E Nacul, Kathleen Mudie, Caroline C Kingdon, Jacqueline M Cliff, Taane G Clark, Hazel M Dockrell, and Eliana M Lacerda. How myalgic encephalomyelitis/chronic fatigue syndrome (me/cfs) progresses: the natural history of me/cfs. *Frontiers in neurology*, 11:826, 2020.

[46] Luis Nacul, François Jérôme Authier, Carmen Scheibenbogen, Lorenzo Lorusso, Ingrid Bergliot Helland, Jose Alegre Martin, Carmen Adella Sirbu, Anne Marit Mengshoel, Olli Polo, Uta Behrends, et al. European network on myalgic encephalomyelitis/chronic fatigue syndrome (euromene): expert consensus on the diagnosis, service provision, and care of people with me/cfs in europe. *Medicina*, 57(5):510, 2021.

[47] Mireia Sospedra, Yingdong Zhao, Harald zur Hausen, Paolo A Muraro, Christa Hamashin, Ethel-Michele de Villiers, Clemencia Pinilla, and Roland Martin. Recognition of conserved amino acid motifs of common viruses and its role in autoimmunity. *PLoS pathogens*, 1(4):e41, 2005.

[48] Jody Phelan, Anna D Grabowska, and Nuno Sepúlveda. A potential antigenic mimicry between viral and human proteins linking myalgic encephalomyelitis/chronic fatigue syndrome (me/cfs) with autoimmunity: The case of hpv immunization. *Autoimmunity Reviews*, 19(4):102487, 2020.

[49] Manuel Rojas, Paula Restrepo-Jiménez, Diana M Monsalve, Yovana Pacheco, Yeny Acosta-Ampudia, Carolina Ramírez-Santana, Patrick SC Leung, Aftab A Ansari, M Eric Gershwin, and Juan-Manuel Anaya. Molecular mimicry and autoimmunity. *Journal of autoimmunity*, 95:100–123, 2018.

[50] Sheila Ali, Faith Matcham, Katherine Irving, and Trudie Chalder. Fatigue and psychosocial variables in autoimmune rheumatic disease and chronic fatigue syndrome: a cross-sectional comparison. *Journal of Psychosomatic Research*, 92:1–8, 2017.

[51] Rona Moss-Morris and Trudie Chalder. Illness perceptions and levels of disability in patients with chronic fatigue syndrome and rheumatoid arthritis. *Journal of psychosomatic research*, 55(4):305–308, 2003.

[52] Anthony L Komaroff and W Ian Lipkin. Insights from myalgic encephalomyelitis/chronic fatigue syndrome may help unravel the pathogenesis of postacute covid-19 syndrome. *Trends in Molecular Medicine*, 27(9):895–906, 2021.

[53] Mohammed Alghamdi, Doaa Alasmari, Amjad Assiri, Ehab Mattar, Abdullah A Aljaddawi, Sana G Alattas, Elrashdy M Redwan, et al. An overview of the intrinsic role of citrullination in autoimmune disorders. *Journal of immunology research*, 2019, 2019.

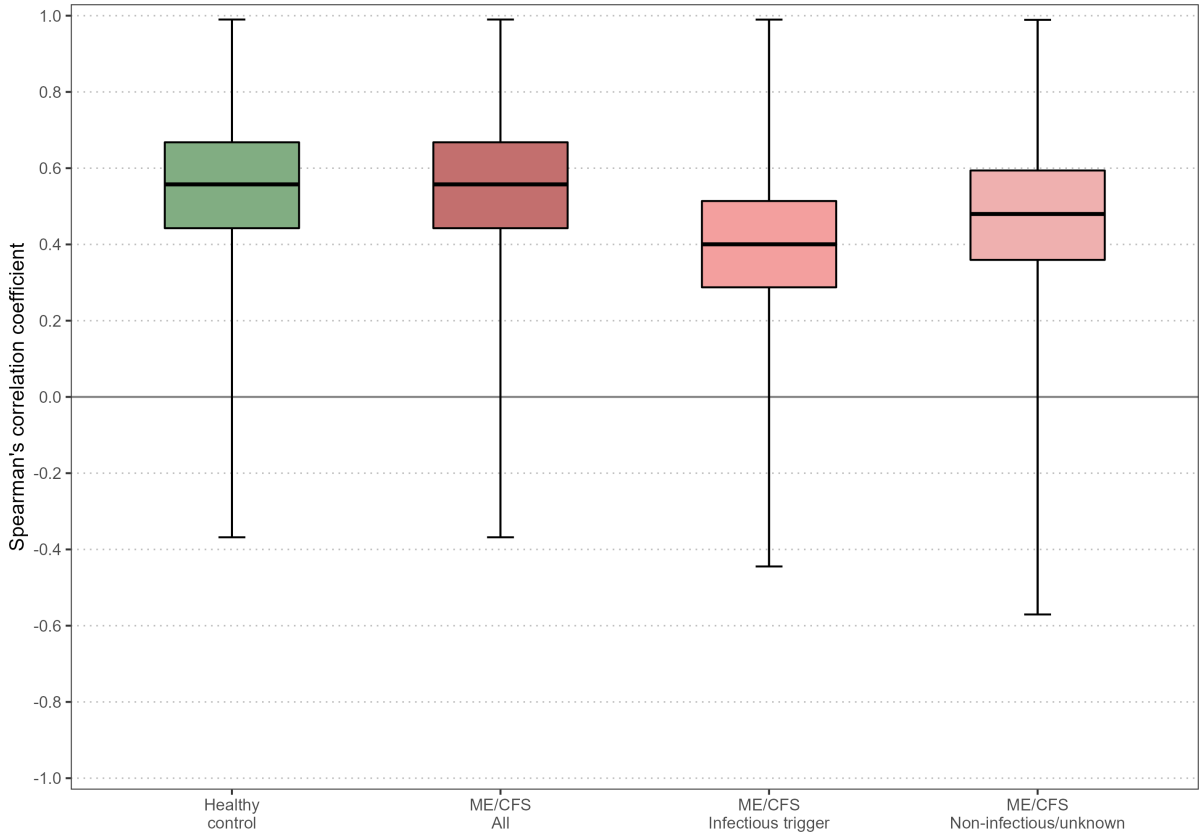
[54] Pninit Birkenfeld, Noam Haratz, George Klein, and Dov Sulitzeanu. Cross-reactivity between the ebna-1 p107 peptide, collagen, and keratin: implications for the pathogenesis of rheumatoid arthritis. *Clinical immunology and immunopathology*, 54(1):14–25, 1990.

[55] Martin L Pall. Levels of nitric oxide synthase product citrulline are elevated in sera of chronic fatigue syndrome patients. *Journal of Chronic Fatigue Syndrome*, 10(3-4):37–41, 2002.

[56] Robert K Naviaux, Jane C Naviaux, Kefeng Li, A Taylor Bright, William A Alaynick, Lin Wang, Asha Baxter, Neil Nathan, Wayne Anderson, and Eric Gordon. Metabolic features of chronic fatigue syndrome. *Proceedings of the National Academy of Sciences*, 113(37):E5472–E5480, 2016.

[57] Michael Maes, Ivanka Mihaylova, and Jean-Claude Leunis. Chronic fatigue syndrome is accompanied by an igm-related immune response directed against neopitopes formed by oxidative or nitrosative damage to lipids and proteins. *Neuroendocrinology Letters*, 27(5):615–622, 2006.

Supplementary Materials



Supplementary Figure 1: **Spearman's correlation coefficient.** Distributions of the Spearman's correlation coefficient between all the possible pairs of EBV-derived antibodies in healthy controls, all the ME/CFS patients, ME/CFS patients with an infectious trigger, and ME/CFS patients with a non-infectious or unknown trigger.

Supplementary Table 4: **EBV peptides**.The overall and per-EBV-strain number of 15-mer peptides (antigens) whose antibody responses were analyzed.

EBV Protein	Associated stage	Number of 15-mer peptides per EBV strain						
		Overall	AG876	B95.8	GD1	Cao	Raji	P3HR.1
BALF-2	Early lytic	290	278	278	278	0	0	0
BALF-5	Early lytic	256	250	250	250	0	0	0
BFRF-3	Late lytic	42	0	42	0	0	0	0
BLLF-1	Late lytic	273	204	202	199	0	0	204
BLLF-3	Early lytic	74	66	67	66	0	0	0
BLRF-2	Late lytic	41	38	38	38	0	0	0
BMRF-1	Early lytic	102	99	99	99	0	0	0
BZLF-1	Immediate early lytic	89	57	57	58	0	0	0
EBNA-1	Latency I, II, and III	182	98	107	111	0	0	0
EBNA-3	Latency III	446	223	226	224	0	0	0
EBNA-4	Latency III	469	229	221	224	0	0	0
EBNA-6	Latency III	461	254	234	230	0	0	0
LMP-1	Latency II and III	197	79	85	80	77	84	0
LMP-2	Latency II and	III	132	120	120	120	0	0

Supplementary Table 5: **Null models' results**. Comparison among different null models (including the covariates age and gender and their interaction) using the Akaike's information criterion (AIC). The best model for each analysis/comparison is shown in bold. ME/CFS_all, ME/CFS_inf and ME/CFS_noninf represent all the ME/CFS patients, ME/CFS patients with an infectious trigger, and ME/CFS patients with a non-infectious trigger, respectively.

Analysis/Comparison	Model (link function)	AIC	ROC (95% CI)
ME/CFS_{all} vs. Healthy controls	Logit	189.973	0.577 (0.478, 0.676)
	Probit	189.964	0.576 (0.478, 0.675)
	Clog-log	**189.936**	**0.574 (0.475, 0.672)**
ME/CFS_{inf} vs. Healthy controls	Logit	147.055	0.610 (0.500, 0.719)
	Probit	**147.029**	**0.606 (0.496, 0.715)**
	Clog-log	147.220	0.609 (0.499, 0.718)
ME/CFS_{noinf} vs. Healthy controls	Logit	127.619	0.556 (0.429, 0.683)
	Probit	127.629	0.559 (0.432, 0.687)
	Clog-log	**127.547**	**0.556 (0.429, 0.683)**
ME/CFS_{inf} vs. ME/CFS_{noinf}	**Logit**	**129.205**	**0.596 (0.471, 0.720)**
	Probit	129.236	0.597 (0.472, 0.721)
	Clog-log	129.529	0.596 (0.472, 0.721)

Supplementary Table 6: **Most significant antibodies**. The top 5 most significant antibodies for each association analysis where ME/CFS_all, ME/CFS_inf and ME/CFS_noninf represent all ME/CFS patients, ME/CFS patients with an infectious trigger, and ME/CFS patients with a non-infectious trigger, respectively. For simplicity, the antibodies were identified by their peptide. Statistically significant findings were obtained for -log10(adjusted p-value) > 1.30 (=−log10(0.05)) controlling for false discovery rate of 5% using the Benjamini-Yekutieli procedure.

Analysis/Comparison	Peptide	$-\log_{10}$(adjusted p-value)
ME/CFS_{all} vs. Healthy controls	EBNA6_0066	0.743
	BLRF2_0005	0.486
	EBNA4_0392	0.486
	EBNA4_0497	0.486
	EBNA4_0529	0.486
ME/CFS_{inf} vs. Healthy controls	EBNA6_0066	2.693
	EBNA6_0070	2.693
	EBNA4_0529	1.794
	EBNA3_0380	1.270
	EBNA6_0569	1.270
ME/CFS_{noinf} vs. Healthy controls	EBNA6_0782	1.193
	BALF2_0358	1.153
	BALF2_0765	1.153
	BALF5_0041	1.153
	BALF5_0206	1.153

Chapter 8 - IgG Antibody Responses to Epstein-Barr Virus in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome: Their Effective Potential for Disease Diagnosis and Pathological Antigenic Mimicry

André Fonseca[1,2,*], Mateusz Szysz[3],Hoang Thien Ly[3], Clara Cordeiro[1,2] and Nuno Sepúlveda[2,3]

[1] Faculty of Sciences and Technology, University of Algarve, 8005-139 Faro, Portugal
[2] CEAUL - Center of Statistics and its Applications, Faculty of Sciences, University of Lisbon, 1749-016 Lisbon, Portugal
[3] Faculty of Mathematics & Information Science, Warsaw University of Technology, 00-662 Warsaw, Poland

André Fonseca, Mateusz Szysz, Hoang Thien Ly, Clara Cordeiro, and Nuno Sepúlveda. Igg antibody responses to epstein-barr virus in myalgic encephalomyelitis/chronic fatigue syndrome: Their effective potential for disease diagnosis and pathological antigenic mimicry. Medicina, 60(1):161, 2024.

8.1 Abstract

Background and Objectives: The diagnosis and pathology of myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) remain under debate. However, there is a growing body of evidence for an autoimmune component in ME/CFS caused by the Epstein-Barr virus (EBV) and other viral infections. Therefore we set to identify antibodies against EBV-derived antigens to understand whether these immune responses could help diagnose patients and trigger pathological autoimmunity.
Materials and Methods: Therefore here, we have analyzed a large public dataset on the IgG antibodies to 3054 EBV peptides in ME/CFS patients and healthy controls (HCs) as a comparator cohort. An ensemble of Random Forests was initially implemented with the objective of ranking the feature's importance. Subsequently classifiers of increasing size were then constructed on the topmost important antibodies and their performance accessed using a Super Learner. Here we aimed at predicting the disease status of the study participants targeting an accuracy of 85% when splitting data into train and test datasets.
Results: When we compared the data of all ME/CFS patients or the data of a subgroup of those patients with non-infectious or unknown disease triggers to the data of the HC, we could not find an antibody-based classifier that would meet the desired accuracy in the test dataset. However, we could identify a 26-antibody classifier that could distinguish ME/CFS

patients with an infectious disease trigger from the HCs with 100% and 90% accuracies in the train and test sets, respectively. We finally performed a bioinformatic analysis of the EBV peptides associated with these 26 antibodies. We found no correlation between the importance metric of the selected antibodies in the classifier and the maximal sequence homology between human proteins and each EBV peptide recognized by these antibodies. *Conclusions:* In conclusion, these 26 antibodies against EBV have an effective potential for disease diagnosis in the subset of patients where an infection triggered the disease. However, the peptides associated with these antibodies are less likely to induce autoimmune B-cell responses that could explain the pathogenesis of ME/CFS.

Keywords: biomarker discovery, disease pathogenesis, autoimmunity, antigenic mimicry, machine learning

8.2 Introduction

The clinical manifestation of myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) is typically a post-exertional malaise upon minimal physical and mental effort, a persistent fatigue that is not alleviated by rest, together with other symptoms related to neurologic, autonomic, and immunologic systems [1, 2]. Several pathological mechanisms have been proposed to explain the origin of the disease and its progression over time [1, 3, 4, 5, 6]. Among these mechanisms, deleterious autoimmunity mostly driven by viruses is gaining traction in the literature [7, 8, 9]. SARS-CoV-2 is the newest causative agent of ME/CFS, as some long-COVID patients also comply with the diagnostic criteria for this disease [10, 11, 12, 13]. However, the Epstein-Barr virus (EBV) along with other herpesviruses remain the usual suspects for causing ME/CFS and now long-COVID [14, 15, 16, 17]. EBV is a particularly interesting virus given the growing evidence that, in some patients with ME/CFS, it enhances T follicular helper differentiation and promotes the formation of abnormal germinal centers that are essential for the generation of long-lived plasma cells and high-affinity antibodies [18]. The cause of these altered immune activities was hypothesized as an increase in activin A and IL-21 serum levels stimulated by EBV deoxyuridine triphosphate nucleotidohydrolase (dUTPase) in these patients [18]. EBV has also a strong potential for antigenic mimicry with human proteins, especially the EBNA1 protein, which contains highly repetitive glycine-alanine motifs [19, 20]. This potential for eliciting autoimmunity has motivated serological investigations in patients with ME/CFS to identify key pathological EBV antigens and peptides [21, 22, 23, 24]. However, these efforts did not lead to the identification of specific anti-EBV antibody signatures with a high accuracy in distinguishing patients from healthy controls (HCs).

These disappointing findings are typically explained by: (i) a very heterogenous clinical population; (ii) the presence of selection bias when recruiting patients; and (iii) the possi-

bility of misdiagnosis where cases suspected of suffering from ME/CFS are actually genuine patients with another disease with a known cause [25, 26]. An alternative explanation is an inadequate choice of the anti-EBV antibodies under analysis [27].

To avoid this problem, a recent study performed a large screening of IgG antibody responses to more than 3000 EBV peptides in patients with ME/CFS and HCs [28]. In a subsequent study on the same data, antibody responses to two peptides (EBNA4_0529 and EBNA6_00 70) were identified as candidate biomarkers for the subgroup of ME/CFS patients whose disease started with an infection [29]. However, these antibody responses were included in simple statistical models based on linear relationships between the antibodies (i.e., covariates) and the disease status. Therefore, the previous study could have failed to detect alternative antibody responses with more complex statistical relationships with the disease status. Also, the same study did not evaluate possible problems concerning data overfitting[29].

The present paper aims at re-analyzing the same dataset with the objective of using a machine learning approach, where the above analytical limitations could be tackled. We have also re-evaluated the role of eventual molecular mimicry between EBV and human antigens in the pathogenesis of ME/CFS.

8.3 Materials and methods

8.3.1 Study participants

Given that this study is a re-analysis of previously published data, the reader is recommended to consult the description of the original study in the respective reference [28]. In brief, 92 patients with ME/CFS were recruited from the Charité outpatient clinic for immunodeficiencies at the Institute of Medical Immunology in the Charité Universitatsmedizin Berlin, Germany. Fifty-four of these patients reported an acute infection at the beginning of their disease symptoms. The remaining patients (n = 38) reported not knowing their disease trigger or a disease trigger other than an infection. Fifty self-reported HCs were recruited from the staff of the same clinic. Age and gender distributions of the ME/CFS cohort as a whole or divided into its two subgroups were matched with the ones of the HC cohort. See the corresponding analysis in Ref. [29].

8.3.2 Basic Description of Serological Data

The serological dataset under analysis is publicly available (see Supplementary File of Ref. [28]). In a nutshell, the serological dataset was generated by a seroarray that measured the signal intensities induced by individual IgG antibody responses to each one of the 3054 EBV peptides. These peptides were derived from 14 EBV proteins: BALF2, BALF5, BFRF3, BLLF1, BLLF3, BLRF2, BMRF1, BZLF1, EBNA1, EBNA3, EBNA4, EBNA6, LMP1, and

LMP2. The peptides had a length of 15 amino acids (15-mer) and they could overlap within the same protein. To denote each peptide, we used the protein name and its starting position within the corresponding protein, using the reference strain AG876. When the peptide name included *, it referred to the starting position of the reference strain B95-8.

8.3.3 Statistical Analysis for Predicting the Disease Status

8.3.3.1 Dividing the Dataset into Train and Test Sets

Before conducting any analyses, the original dataset was divided into train and test subsets using a 9:1 ratio while maintaining the proportions of ME/CFS patients with subgroups, and the HCs. This ratio was approximately the optimal splitting ratio when applying a linear regression model that explains the variability of the data at the cost of 81 covariates [30]

8.3.3.2 Ranking Antibodies by Their Importance for Predicting the Disease Status

We performed an initial step where we ranked the antibodies according to their importance in discriminating ME/CFS patients from HCs (Supplementary Figure 2). In this step, we estimated 2500 Random Forests (RFs) using different hyperparameters: the number of trees used to construct each Random Forest (100, 500, 750, 1000, 2000), the number of features that could be used to split each node (fifty randomly generated values between 1 and 100), and the minimal node size, meaning the minimal number of observations after a split required to keep growing the trees (1, 2, 3, ..., 10). These hyperparameters were modified in each run through a grid approach. In each run, the mean decrease in the Gini index was used to determine the importance of each antibody in predicting the disease status. After the 2500 runs, the importance value was determined by calculating the mean for each antibody and sorting them in descending order, identifying the most to the least essential antibodies for disease prediction.

In general, the Gini index measures the inequality of a given probability distribution. In the context of RF, the mean decrease in the Gini index is a measure of how each covariate contributes to the homogeneity of the nodes and leaves in the resulting RF [31]. In this scenario, a higher mean decrease in the Gini index indicates that a given antibody is essential for the respective classification .

8.3.3.3 Individual Statistical and Machine Learning Methods for Predicting the Disease Status from the Anti-EBV Antibodies We applied 5 statistical techniques to predict the disease status based on the anti-EBV antibodies: elastic-net logistic regression (GLMNet), Random Forest (RF), support vector machine (SVM), linear discriminant analysis (LDA), and extreme gradient boosting (XGB). These methods were chosen due to their capacity for capturing different data patterns. On the one hand, GLMNet and LDA are based on

linear combinations of the antibody values for predicting the disease status. On the other hand, RF, SVM, and XGB are particularly appropriate for detecting non-linear relationships between a set of covariates and the outcome. It is worth noting that probit regression could have been chosen. However, the probit and logistic regression models usually provide similar results due to their symmetric link functions. Regression models based on alternative link functions (e.g., log link) were also excluded from this analysis because there was no computational implementation for pooling their different predictions.

We increased the number of the most important antibodies to be included in each of the above 5 statistical techniques.

8.3.3.4 Construction of Final Models for Predicting the Disease Status by Assembling Predictions from Individual Models For a given number of antibodies, the results from the 5 individual classifiers were combined by the Super Learner (SL) algorithm, which assigns different weights to each individual classifier estimated for the same data [32]. Under generic assumptions, this algorithm typically improves the prediction of an outcome when compared to the accuracy of the predictions generated by each model individually.

The accuracy of the resulting classifier was evaluated by the proportion of individuals correctly classified using the ROC01 criterion [33]. This criterion dictates that the optimal accuracy is the one generated from a cut-off in the estimated classification probabilities that minimizes the distance between sensitivity/specificity to the perfect classification scenario (i.e., both sensitivity and specificity equal to 1).

We started our SL-based analysis with the two most important antibodies as the respective features/covariates. Every time the data from a new antibody response were added to the SL-based classifier (and its subclassifiers), we calculated the Spearman correlation coefficient r_s and removed highly correlated antibody responses ($r_S > |0.8|$), as done elsewhere [34]. This step was conducted in order to avoid redundancy and multicollinearity. We kept adding new antibody responses until we reached the maximum number of 100 antibody responses. The best classifier was the SL-based classifier with the lowest antibodies reaching the target accuracy of 85% in both the train and test sets; this accuracy is regarded as the optimal value for classification problems [35].

The above analysis was performed to compare the cohorts of all ME/CFS patients, ME/CFS patients with reported infectious disease triggers, and ME/CFS patients with non-infectious or unknown disease triggers against HCs.

8.3.4 Bioinformatic Analysis to Test the Importance of Antigen Mimicry in Predicting the Disease Status

When we found an SL-based classifier with the target accuracy in both train and test datasets, we then performed protein-protein alignments between the EBV peptide asso-

ciated with each selected antibody and the human proteins included in the RefSeq reference protein database [36], as available in the National Centre for Biotechnology Information (https://blast.ncbi.nlm.nih.gov/, accessed on 1 August 2023). The quality of the alignments was based on the E-score statistic [37]. In our analysis, we focused on the maximal E-score associated with the alignments obtained for each EBV peptide under analysis. Subsequently, we calculated the Spearman non-parametric correlation coefficient between the importance of each selected antibody for disease prediction and the respective maximal E-score of the peptide associated with that antibody. We also calculated the respective 95% confidence interval. The same analysis was repeated using the human proteins included in the RefSeq non-redundant (nr) protein database.

8.3.5 Statistical Software

The statistical analyses were performed in the R [38] software version 4.3.0 using the following packages: caret for multicollinearity analysis [39], OptimalCutpoints to obtain the accuracy based on the ROC01 criterion of each predictive model [33], pROC for the AUC estimation [40], ranger to perform the Random Forest [41], and SuperLearner for the SL-based analysis [42].

8.4 Results

8.4.1 Construction of a Predictive Model to Distinguish All ME/CFS Patients from HCs

A comparison between all 92 patients with ME/CFS and the 50 HCs was carried out to develop a classifier that could predict the disease status of these study participants. The train dataset was composed of forty-five HCs and eighty-three ME/CFS patients, while the test dataset comprised five HCs and nine ME/CFS patients.

The overall average antibody importance distribution is presented in Figure 31A as a density plot. Our results showed that the overall average antibody importance was around 0.018, with the antibody against EBNA6_0066 being the most important (0.36). Furthermore, seven out of the ten topmost antibodies are associated with peptides belonging to the family of the Epstein-Barr nuclear antigen (EBNA) proteins.

It is worth noting that the levels of antibodies against EBNA1_430 — a peptide with a potential molecular mimicry with the human Anoctamin-2 protein [43, 44] — were similar in both ME/CFS patients and HCs (Supplementary Figure 3). Consequently, they only had an average importance of 0.010 (ranked in the 1866^{th} place of the most important antibodies). Therefore, this finding suggested a negligible role of these antibodies in predicting disease status.

In the train subset, the target accuracy of 85% was already achieved by an SL classifier including only two antibodies (Figure 31B). The corresponding sensitivity and specificity

were close or equal to 1). This finding resulted from the high accuracy of the RF irrespective of the number of antibodies used as features (Figure 31C).

In the test subset, the accuracy estimates fluctuated around 50% and were at best 64%, using an SL classifier including 36 antibodies as features. This poor performance was explained mainly by the low sensitivity of the classifiers (Figure 31D). Hence, the target accuracy of 85% was not achieved for the overall dataset, largely due to poor performance in predicting ME/CFS patients in this data subset.

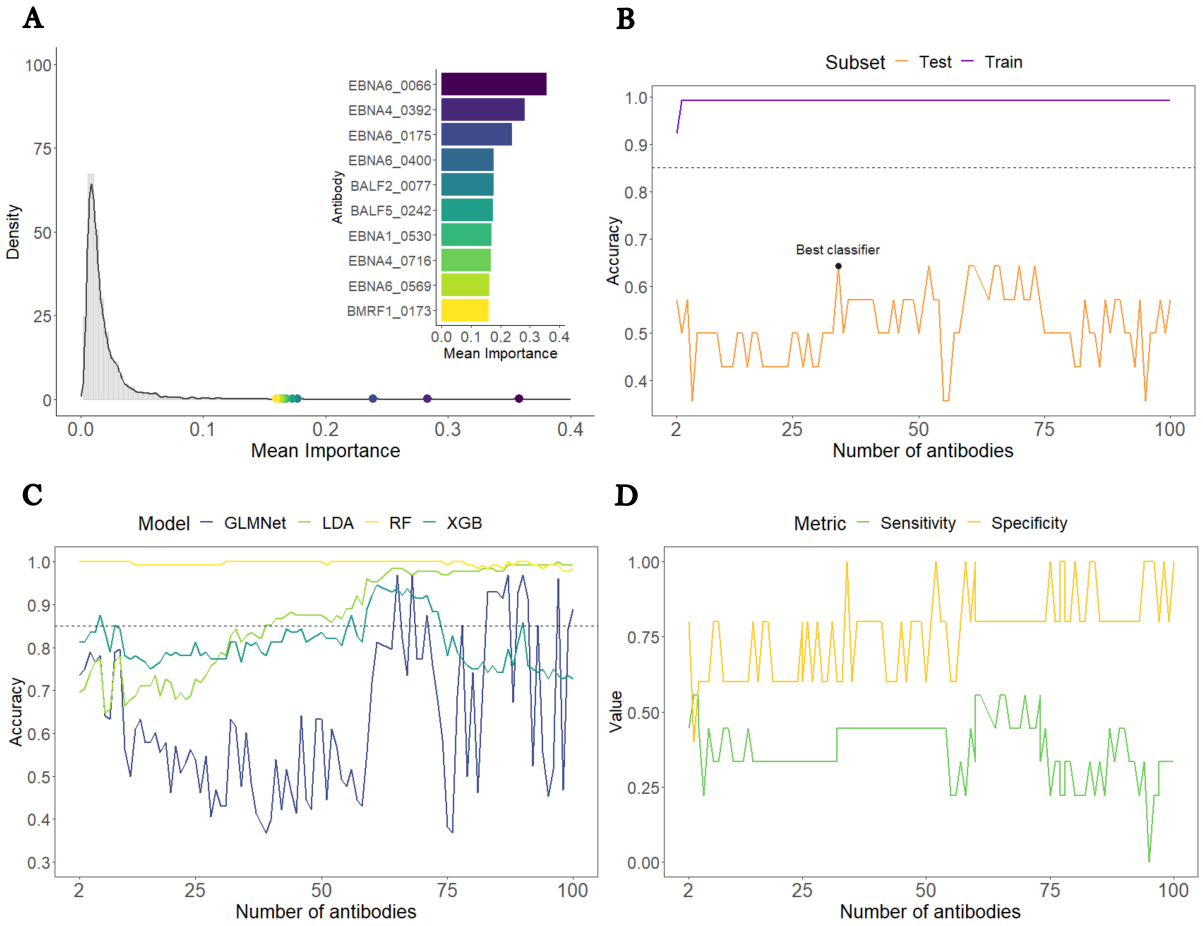


Figure 31: **Analysis of all ME/CFS patients versus HCs.** A) Density plot of each antibody's average importance distribution obtained by the RF with the top 10 most important antibodies highlighted. B) Accuracy of the SL classifier in the train (purple) and test (orange) subsets as a function of the number of antibodies included. The black and blue horizontal dashed lines indicate 85% (target accuracy). The best classifier is highlighted with a black dot. C) Accuracy of the different classifiers assembled by the SL in the train subset. D) Sensitivity and specificity of the SL classifiers in the test subset as a function of the number of antibodies included.

8.4.2 Construction of a Predictive Model to Distinguish ME/CFS Patients with Non-Infectious or Unknown Disease Triggers from HCs

We then compared the 38 ME/CFS patients with non-infectious or unknown disease triggers to the 50 HCs. This time, the overall average antibody importance was around 0.012. The most important antibodies were EBNA1_0595, EBNA1_0530 and EBNA6_0400, with the mean importance of 0.13, 0.12, and 0.11, respectively (Figure 32A). Once more, seven out of the ten topmost important antibodies recognized antigens from the EBNA protein group. In line with the analysis based on the whole cohort of ME/CFS, the antibodies against EBNA1_430 had an average importance of 0.005, which translated into poor importance ranking (2384^{th} place) among all the antibodies.

In the train subset, based on the top three antibodies (EBNA1_0595, EBNA1_0530 and EBNA6_0400) an SL classifier reached an accuracy of 94%, a value higher than the target accuracy of 85% (Figure 32B). In contrast, the same classifier reached only 78% in the test subset (Figure 32B). Such a value was the best accuracy that could be achieved for this analysis.

In the test subset, the target accuracy of the top three antibodies SL classifiers was not achieved by a relatively modest sensitivity (Figure 32C). Hence, this analysis suggested that these EBV antibodies were unable to discriminate this subset of ME/CFS patients from the HCs with high sensitivity and high specificity.

8.4.3 Construction of a Predictive Model to Distinguish ME/CFS Patients with a Putative Infectious Disease Trigger from HCs

We finally conducted a comparison between the 54 ME/CFS patients who reported an infection at their disease onset and the 50 HCs. This time the antibodies recognizing the EBNA4_0529, EBNA3_0139, and EBNA3_0577 antigens had the highest importance values in the RF (0.26, 0.25, and 0.24, respectively; Figure 33A). Eight of the ten topmost important antibodies belonged to the group of EBNA proteins. Once again, the antibodies against EBNA1_430 had a low average importance (0.009) and a poor ranking (1447^{th} place) in terms of predictive importance.

In this analysis, we found an SL classifier including 26 antibodies that could reach an accuracy above the target value of 85% in both train and test datasets (99% and 90%, respectively; Figure 33B). These antibodies were associated with antigens derived from nine different EBV proteins: BALF2 (n = 2), BALF5 (n = 3), BLLF1 (n = 1), BMRF1 (n = 1), EBNA1 (n = 2), EBNA3 (n = 4), EBNA4 (n = 4), EBNA6 (n = 7), and LMP2 (n = 2). Twenty-two out of the twenty-six selected antibodies had increased levels in this subset of ME/CFS patients compared to the HCs (Figure 33C). The estimated classifier had a sensitivity of 100% and a specificity of 80% (Figure 31D). The corresponding AUCs for both the train and test datasets were 1.00 and 0.88, respectively (Figure 33E).

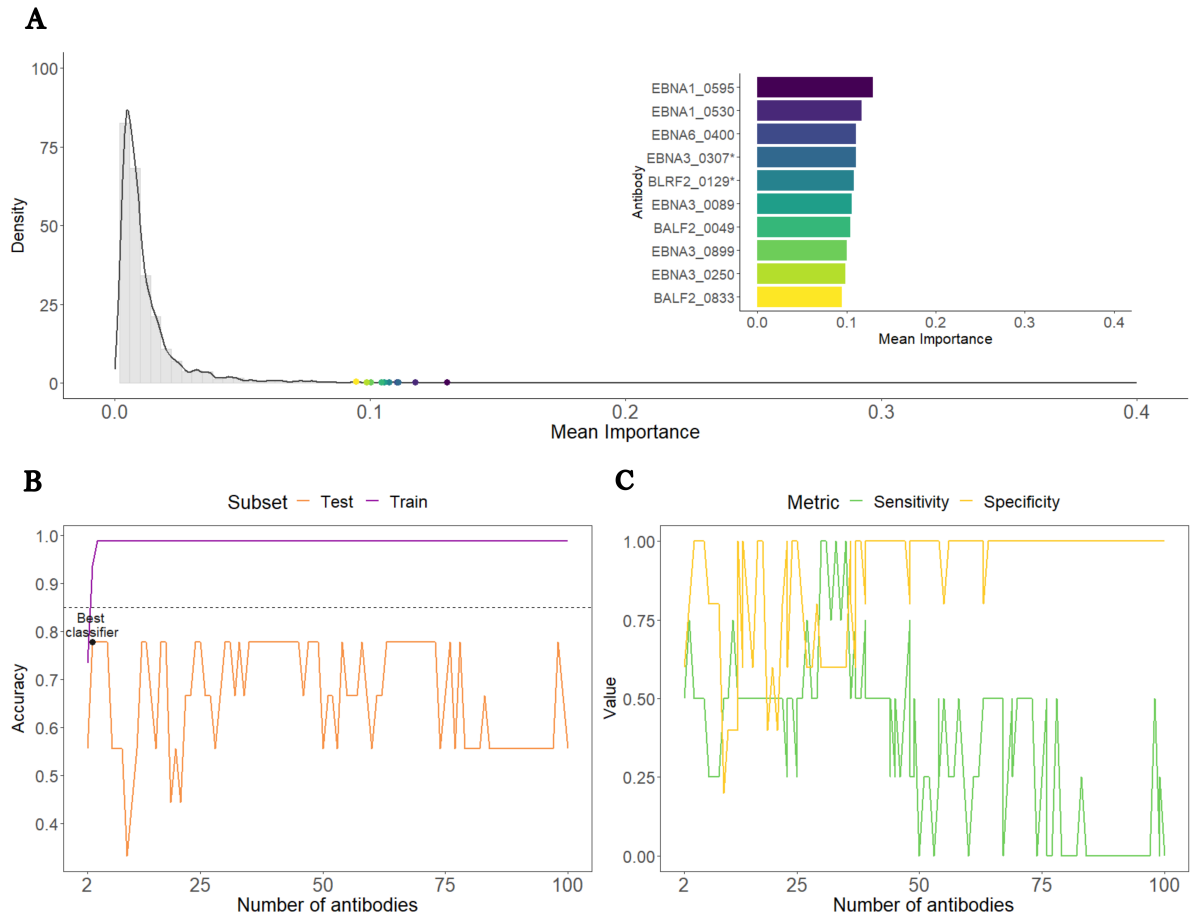


Figure 32: **Analysis of ME/CFS patients with non-infectious or unknown disease triggers against HCs.** A) Density plot of each antibody's average importance distribution obtained by the RF with the top 10 most important antibodies highlighted. B) Accuracy of the SL classifier in the train (purple) and test (orange) subsets as a function of the number of antibodies included. The black and blue horizontal dashed lines referred to the 85% target accuracy. The best classifier is highlighted with a black dot. C) Sensitivity and specificity of the SL classifiers in the test subset as a function of number of antibodies included.

Note that an SL classifier including 42 antibodies predicted disease status almost perfectly in both train and test datasets (Figure 33B). However, this perfect classification was achieved at the cost of approximately 2.5 antibodies per study participant.

8.4.4 Testing the Importance of Antigen Mimicry on Disease Prediction Using a Bioinformatic Approach

The final analysis aimed at testing the hypothesis whether the EBV peptides associated with the above 26 antibodies selected in the SL classifier might help explain the pathology of ME/CFS via a mechanism of molecular mimicry. Under this hypothesis, we expected a positive correlation between the importance of each antibody selected and the best alignment score between the associated peptides and human proteins.

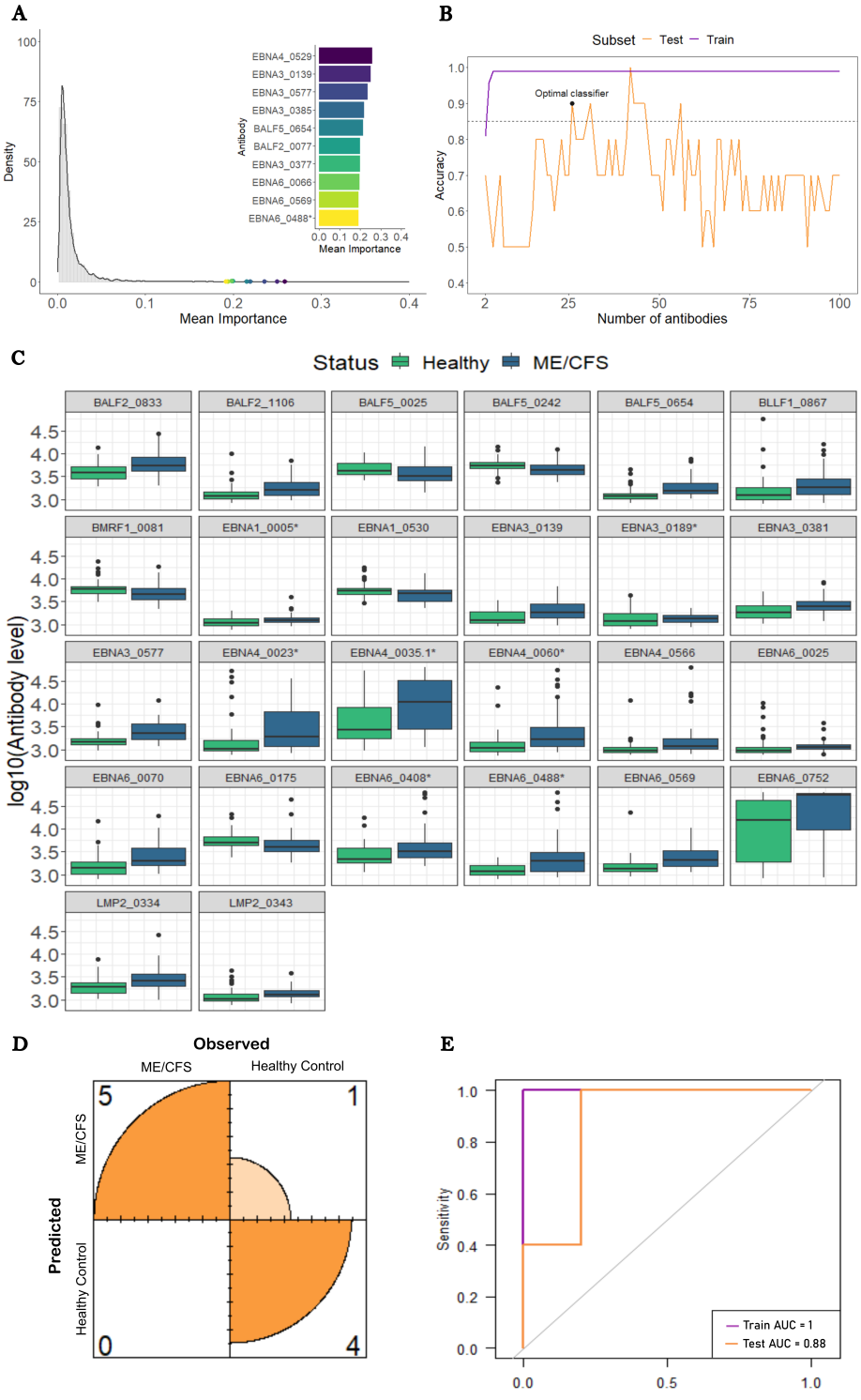


Figure 33: **Analysis of ME/CFS patients with an infectious disease trigger against HCs.** A) Density plot of each antibody's average importance distribution obtained by the RF with the top 10 most important antibodies highlighted. B) Accuracy of the SL classifier in the train and test subsets as a function of the number of antibodies included. The black and blue horizontal dashed lines indicate the target accuracy of 85% (i.e., 0.85). The best classifier is highlighted with a black dot. C) Boxplots of the log10-levels of the selected 26 antibodies in HCs and ME/CFS patients. D) Confusion matrix concerning the optimal classifier performance on the test subset. E) ROC curve of the optimal classifier for both train and test subsets.

The best alignment score per peptide varied from 22.7 (BALF5_0025) to 32.9 (EBNA6_0070) when conducting the protein alignments in the human RefSeq reference protein dataset. In the case of EBNA6_0070, the highest alignment scores were associated with alignments against the Homeobox-9A (HOXA-91) and adrenergic receptor alpha (ADRA1B) proteins (Figure 34A). The peptide with the second-best alignment was EBNA6_0488. In this case, this peptide had an extensive sequence homology with the human proteins CCCTC-binding factor (CTCF) and adipocyte enhancer-binding proteins (AEBP1) (Figure 34B).

When we analyzed the average importance of these 26 antibodies against the best alignment scores of the respective peptides, we found a slight negative association between these two quantities, but this association was not statistically significant (r_s = -0.161, 95% CI [-0.473; 0.185]; Figure 34C). We obtained the same lack of correlation using the non-redundant protein database (r_s = -0.018, 95% CI [-0.355; 0.330]; Supplementary Figure 4). Hence, these data did not support the hypothesis that antibodies recognizing important EBV peptides for disease prediction had a high potential for cross-reactivity with human proteins.

8.5 Discussion

8.5.1 General Comments

This paper has demonstrated the recurrent difficulty of finding anti-EBV antibodies that could be used as general markers of ME/CFS. Similar difficulty was encountered in studies on IgG antibodies against common pathogens [45] or multiple peptides derived from different herpesviruses [21]. In our case, such a difficulty was mainly due to patients who reported a non-infectious disease trigger or did not know their disease trigger. Therefore, stratifying patients by the respective disease trigger was sufficient to make specific EBV antibody signatures emerge for patients with an infectious disease trigger. A more detailed discussion of the utility of stratifying patients by their disease trigger can be found in our previous works [24, 29]. A more general discussion of the issue of patient stratification can be found in Jason et al. [46].

In the present work, the best-case scenario was obtained for the patients with an infectious disease trigger where a set of 26 EBV-related antibodies led to good accuracy in both train and test datasets. This finding is not surprising, given that our previous study also led to a similar conclusion but using a different statistical methodology [29].

Similar large-scale antibody screening was performed on patients suffering from multiple sclerosis (MS), a disease strongly correlated with EBV infections. In one of these screenings, the levels of the 26 identified antibodies were not significantly different in patients with MS when compared to healthy donors [47]. This finding suggests that our 26-antibody signature is specific to ME/CFS patients. However, in one of the largest studies of MS,

the levels of antibodies related to EBNA1_0005, EBNA3_0577, EBNA4_0566, EBNA6_0025, EBNA6_0488, and EBNA6_0752 peptides were significantly elevated in MS patients [48]. This previous finding is in line with the recurrent observation that some patients with ME/CFS share the same symptomatology [49, 50] and antibody alterations with patients with MS [51]

8.5.2 Clinical and Diagnostic Implications

The identification of multiple elevated anti-EBV antibodies linked to the cohort of ME/CFS patients with an infectious disease trigger suggests that treatments such as immuno-adsorption, rituximab, and cyclophosphamide could be deployed to treat this clinical group specifically. Interestingly, the clinical value of these three treatment options was at the heart of a recent meeting entirely dedicated to ME/CFS research [52].

In general, immuno-adsorption is an aphaeretic procedure that removes specific proteins (e.g., antibodies) from a patient's plasma. This treatment has already been tested successfully in a small cohort of ME/CFS patients with an infectious disease trigger whose autoantibodies against adrenergic receptors were elevated [53, 54]. The same treatment was recently tested in long COVID ME/CFS patients, with similar successful outcomes [55]. Given that immuno-adsorption induces a significant depletion of total IgG, we speculate that the clinical benefit of this treatment comes from the removal of not only autoantibodies against adrenergic receptors from patients but also the EBV antibodies here identified.

The administration of rituximab, an anti-CD20 monoclonal antibody, aims at inducing a slow depletion of hypothetical autoreactive B cells in ME/CFS patients [3]. This depletion also has the advantage of removing EBV-infected B cells. At the same time, the depletion of the peripheral B-cell pool induces the renewal of the B-cell pool by a healthy one coming from the bone marrow. The initial clinical trials on the deployment of this drug to ME/CFS treatment were promising [56, 57]. However, similar results were not observed when the clinical trial was scaled to a larger cohort of ME/CFS patients [58]. A possible explanation for this finding is the use of rituximab to treat all patients irrespective of their disease. A better treatment strategy is to use this drug only on patients with an infectious disease trigger, mainly those who reported an infectious mononucleosis at the disease onset.

Cyclophosphamide is an immunosuppressive drug that was already tested, with promising results, in ME/CFS patients [59]. This drug generally acts on both T and B cells [60]. Interestingly, early studies on cyclophosphamide demonstrated a strong suppressive effect on antibody formation in animal models [61, 62]. In addition, there is evidence of a decrease in the B-cell numbers after cyclophosphamide administration in patients suffering from rheumatoid arthritis [63]. Given that EBV can trigger both T- and B-cell responses and the antibody production requires the activation of the CD4+ T helper cells recognizing a given antigen, this drug is expected to have a higher clinical benefit than rituximab due to

A

EBNA6_0070	M Q R I R R R R R R R A A L S
HOXA-9A (11, 20)	Q R R R R R R R R A
ADRA1B (370, 378)	R R R R R R R R R R

B

EBNA6_0488*	P V K P T P P P S R R R R G A
CTCF (640, 653)	P V T P A P P P A K K R R G
AEBP1 (256, 265)	P R P P P S R R R R

C

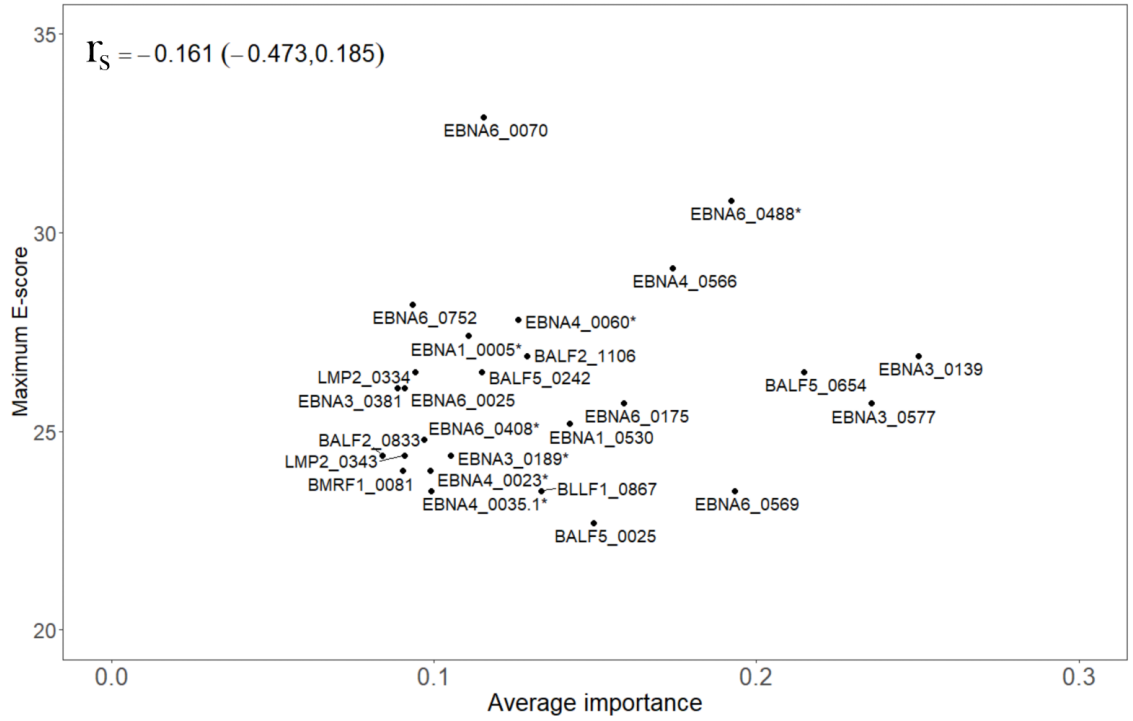


Figure 34: **Bioinformatic analysis of the EBV peptides associated with the 26 antibodies for predicting ME/CFS patients with an infectious disease trigger.** A) Alignments between EBNA6_0070 and the human proteins HOXA-9A and ADRA1B_371 with the corresponding amino-acid coordinates within brackets. (B) Alignments between EBNA6_0488* and the human proteins CTCF and AEBP1 with the corresponding protein coordinates within brackets. (C) Scatterplot between the average importance of each EBV peptide and the maximum E-score alignment score with human proteins using the RefSeq reference protein database, where R is the Spearman correlation coefficient with the respective 95% confidence interval in brackets.

a broader suppression of the adaptive immune responses towards potentially pathogenic EBV peptides. In light of this perspective, it would be interesting to confirm whether CD4+ T helper cells from the same cohort of ME/CFS patients can be activated by the same EBV peptides associated with the identified 26-antibody signature. If such an activation occurs, it provides a strong rationale for using cyclophosphamide instead of rituximab to treat these patients.

Our results suggest that 26 antibodies could be used jointly as a diagnostic tool for suspected cases who reported an infection at the onset of their symptoms. This is of great clinical value, given that more than 64% of ME/CFS patients report an infectious disease trigger [24, 64, 65]. However, further studies should be conducted to understand the effect of disease duration, the infectious agent, and other factors on sensitivity. On the other hand, the diagnostic potential of these 26 antibodies requires going beyond the routine use of ELISA for a single antibody testing. In this regard, the practical use of a 26-antibody diagnostic tool demands the use of multiplex and other high-throughput serological platforms in which these antibodies could be quantified simultaneously, as done for testing malaria exposure [66, 67]. If so, the 26-antibody diagnostic tool could be conducted in reference clinical centers dedicated to ME/CFS, where high-throughput serological platforms are available.

On the other hand, we could not find any significant EBV antibody signatures for patients who did not know or report a non-infectious disease trigger. The same negative result was obtained in our previous analysis of the same data [29]. It is worth noting that non-infectious disease triggers might be, among others, pregnancy, surgery, personal and work-related stress, or exposure to chemicals [24, 65]. These triggers seem not to have as direct and strong an impact on the immune system as infections do. Hence, it is perfectly conceivable that patients with non-infectious disease triggers have pathological mechanisms that are more related to dysfunctions of the body's systems other than the immune one. As such, alternative approaches (e.g., metabolomics, DNA methylation, or genetics) should be sought to tackle the pathogenesis of patients who did not know their disease trigger or reported a non-infectious one, and then to discover the respective biomarker.

8.5.3 EBV Antigenic Mimicry and Its Putative Role in ME/CFS Pathogenesis

8.5.3.1 Replication of Previous Finding on EBNA6_0070 Peptide

Among the peptides recognized by the selected antibodies, EBNA6_0070 had the highest sequence homology with a human protein. This antigen had already been discovered and amply discussed in our previous study [29]. The replication of this finding using a different statistical approach provided further support for the hypothetical role of this peptide in ME/CFS. However, in contrast with our previous work, antibodies against this candidate antigen were not among the top most important antibodies for disease prediction.

This result suggests that the potential pathological effect of this EBV antigen via molecular mimicry is not as so straightforward as our initial study suggested.

To resolve this question, one could purify IgG antibodies against this EBV peptide from ME/CFS patients and then transfer them to recipient humanized mice. Alternatively, one could synthesize the EBNA6_0070 peptide artificially and then inject it into humanized mice. One can then measure the motor activity of these mice after these challenges. Classical measures of motor activity are distance travelled, voluntary wheel running, or time standing still [68]. Suppose this peptide or the respective recognizing antibodies are indeed in the causal pathway of ME/CFS. In that case, one should observe a significant reduction of motor activity in these challenged mice over the time course of the experiment. It is worth noting that animal models are already in use to understand the role of antibodies on fibromyalgia [69], a co-morbidity of many ME/CFS patients. Animal models also provide proof-of-principle evidence of a hypothetical pathological mechanism for ME/CFS [18]. Efforts were also conducted to develop a mouse model for ME/CFS, using lipopolysaccharide challenges [70, 71]. However, it is unclear how these models really mimic the main disease symptoms, especially post-exertional malaise.

8.5.3.2 EBNA6_0488 Peptide and the Antigenic Mimicry with CTCF and AEBP1

The EBNA6_0488 peptide had the second-highest sequence homology with human proteins. This homology was related to two possible human 10-mer peptides belonging to CTCF and AEBP1. The former protein is a master transcription factor due to its more than 50,000 possible binding sites and its role as a chromatin barrier element [72, 73]. In addition, the level of this transcription factor is inversely correlated with the levels of DNA methylation [73]. CTCF in partnership with cohesion molecules is also important in many immunological pathways, such as the interferon gamma production in Th1 cells and the establishment and maintenance of regulatory T cells in visceral adipose tissue and skeletal muscle [74, 75]. In this scenario, we speculate that the increased quantity of antibodies against EBNA6_0488 results in a cross-reactive antibody response to the CTCF peptide, thus reducing the abundance of this transcription factor. This putative reduction could lead to altered gene expression and DNA methylation patterns, and abnormal immunological processes, including the maintenance of deleterious autoimmunity in check. This speculation is in line with findings from altered gene expression and DNA methylation profiles in ME/CFS patients [76, 77, 78, 79]. As an extreme case, one study identified more than 12,000 CpG sites with altered DNA methylation levels in patients with ME/CFS compared to HCs [80]. Immunological abnormalities are also reported by many studies in ME/CFS patients (reviewed in Refs. [8, 81]). An alternative interpretation is that the increased levels of antibodies against EBNA6_0488 resulted from a putative CTCF overexpression during the disease progression in the cohort of ME/CFS patients with an infectious disease trigger. An overexpression of this transcription factor could be the result of a stress-induced response

to restore homeostatic equilibrium within cells. However, altered gene expression was not reported for CTCF by any gene expression studies published so far. This negative reporting could be explained by not performing any patient stratification when analyzing data from these studies.

With respect to AEBP1, this protein is a ubiquitous transcriptional repressor involved in the regulation of adipogenesis, mammary gland development, inflammation, macrophage cholesterol homeostasis, and atherogenesis [82]. Interestingly, mutations on the AEBP1-encoding gene were implicated with the onset of Ehlers-Danlos syndrome (EDS) [83, 84]. Patients with EDS hypermobility type can also receive a diagnosis of ME/CFS [85]. On the other hand, patients with a diagnosis of ME/CFS also show EDS as a co-morbidity [86]. In fact, the presence of EDS in a suspected case of ME/CFS has not been considered as an exclusionary condition for the respective disease diagnosis [87]. However, genome-wide association studies of ME/CFS did not report any genetic markers located in the AEBP1 gene [88, 89, 90, 91, 92]. In this scenario, antibody responses to EBNA6_0488 with the potential of being cross-reactive with AEBP1 should alter the regulation of biological processes where this protein is involved. In particular, the deficient regulation of inflammatory processes is particularly relevant for ME/CFS, given the general idea that established ME/CFS translates into a persistent low-grade inflammatory process in patients [6]. Given that endothelial dysfunction is also observed in patients with ME/CFS [93, 94, 95], such a dysfunction could result from damaged endothelial cells via persistent low-grade inflammation in response to EBNA6_0488 mimicking an AEBP1 peptide. Hence, the identification of this molecular mimicry brings an unexpected link between EBV and AEBP1. As alluded above for CTCF, current gene expression studies did not highlight AEBP1 at the top of the proteins with the most significant differential abundance between patients with ME/CFS and HCs. The lack of patient stratification is once again a possible reason for not detecting an altered abundance of AEBP1 in ME/CFS patients when compared to HCs.

Interestingly, the maximum sequence homology of the peptides recognized by the 26 selected antibodies and human proteins was not associated with the importance of the same antibodies in disease prediction. Moreover, antibodies against EBNA1_430, which contains a peptide mimicking a peptide from the human Anoctamin-2 protein[43, 44], had low importance in predicting ME/CFS patients. These results suggest that potential molecular mimicries due to antibody reactivity between EBV and human antigens have a minor role in the underlying pathological mechanism in such a subset of patients. However, we cannot rule out the possibility of a potential mimicry based on the three-dimensional molecular structure of the respective peptides, but not at the level of their amino-acid sequence. We cannot also rule out that molecular mimicries based on sequence homology might be elicited by peptides from EBV proteins other than the ones evaluated in this study. This might be the case of two EBV peptides from BPLF1 and BHRF1 proteins that were able to elicit an immune response by self-reactive T-cell clones derived from patients with MS

[96].

8.5.4 Interpretation of the Findings under the Lens of the Danger Theory

An interesting perspective on the above results can be given by the so-called danger theory [97]. The theory is based on the premise that the immune system is activated by danger or damage signals sent by infected (or stressed) cells to the immune system. These danger signals are independent of the intrinsic nature of antigens (self or non-self) seen by the immune system. As a corollary, autoimmune responses and autoimmune diseases are then understood as unintended consequences of persistent danger signals that ultimately include chronic presentation of multiple self and non-self antigens. This explains why chronic and low-grade infections by herpesviruses are among the most documented triggers of autoimmune diseases. In this scenario, the theory exactly predicts the lack of correlation between the importance of the selected antibodies in predicting ME/CFS and the degree of molecular mimicry between the EBV peptides and human proteins. The basic question of applying the danger theory to ME/CFS pathogenesis lies in understanding which danger signals are at the core of the disease. According to the original proponents of the danger theory, general danger signals are the heat shock proteins (HSP), the vasoactive intestinal polypeptide (VIP), and the cytokines TNFα and IL1β, among others [98]. A brief discussion about some of these danger signals in the context of ME/CFS is given below; a more comprehensive discussion of this topic will be conducted in the near future.

8.5.5 Potential Danger Signals in ME/CFS Pathogenesis

HSPs are highly conservative proteins in nature and are produced in response to many different cellular stresses. In theory, antigens derived from these proteins were thought to belong to the so-called immunological homunculus, a limited set of dominant self antigens that allow the immune system to have a picture of the self [99]. However, there is no consensus on whether HSPs are indeed signaling danger or are simply key regulatory and resolution elements of a stress or immune response [97, 100]. This alternative interpretation of the functional role of HSPs might explain the lack of consistency in HSP-related responses across studies where patients with ME/CFS and HCs were challenged with physical exercise [101, 102, 103]. In addition, antibodies against endogenous and microbial HSP65 peptides were at the same level in patients with ME/CFS and HCs, with the exception of a higher seroprevalence to an HSP65 peptide derived from Chlamydia pneumoniae in the former [104].

VIP is a neuromodulator present in the gut and the anterior chamber of the eye [98]. On the one hand, it can activate dendritic cells [105] (thus its suggestion as a potential danger signal). Conversely, the binding of VIP to its receptor in immune cells also leads to anti-inflammatory actions [106]. In this line of thought, a loss of tolerance to VIP, other va-

soactive neuropeptides, or their receptors was hypothesized to be at the genesis of ME/CFS [107]. However, a follow-up study showed an elevated expression of VPAC2—the VIP receptor—in immune cells and an increased frequency of the regulatory Foxp3+CD4+ T cells in ME/CFS patients in comparison with HCs [108]. Given that the generation of these regulatory cells can be induced by VIP [109], this finding is more in line with this neuromodulator being a mediator of regulation in the context of ME/CFS.

TNFα and IL1β are two classical pro-inflammatory cytokines. According to a systematic review [110], it was found that 20–25% of the studies reported elevated levels of these cytokines in patients with ME/CFS when compared to HCs. However, the same systematic review did not perform a meta-analysis of the published data. Therefore, it is unclear whether the lack of significant findings related to these two cytokines results from insufficient statistical power due to reduced sample sizes used in the respective studies. It is worth noting that ME/CFS patients from Italy had a higher frequency of an allele variant associated with elevated levels of TNFα (rs1800629:G >A) [111]. However, this finding was not replicated by another study with German patients [112].

TNFα and IL1β are also known to bridge the adaptive and innate arms of the immune system via the so-called CD40/CD40 ligand (CD40L) pathway. In particular, CD40L and its mutations trigger different immunological signaling cascades on B cells [113] that might be important for the establishment of EBV latency and its reactivation. The fundamental question is to understand how TNFα, IL1β, and CD40L are balanced under normal and disease-related conditions. On the one hand, there is ample evidence that TNFα and CD40L influence the immunological activity of each other [114, 115, 116]. In the context of Crohn's disease, TNFα can show anti-inflammatory activity by down-regulating the CD40/CD40L pathway [117]. This capacity might be an explanation for the observation that CD40L levels were significantly decreased in ME/CFS patients with shorter disease duration [118, 119]. From this perspective, the increased levels of TNFα in ME/CFS might be seen as an anti-inflammatory response to an exacerbated immune response to an ongoing infection (caused by EBV or another virus) that initiated ME/CFS. On the other hand, IL1β is known to cooperate with CD40L to increase the production of pro-inflammatory cytokines and activate dendritic cells [120, 121]. In light of the above evidence, it is an interesting research question to know the interplay between TNFα, IL1β, and CD40L at the early stages of ME/CFS.

8.6 Conclusions

In summary, this study provided a list of possible EBV peptides whose associated IgG antibody responses could be used in the diagnosis of suspected ME/CFS cases who reported an infection at their symptoms' onset. Two of these peptides had a high sequence homology with human proteins, but the corresponding antibody responses were not the

most important ones for disease prediction. This finding suggested that the role of EBV on eventual ME/CFS-related autoimmunity should be reconsidered under the lens of danger theory.

Declarations

Author Contributions: N.S. designed this research; A.F. conceptualized and performed all the computational implementation of the methodology; M.S. and H.T.L. performed preliminary analyses concerning the molecular mimicry between EBNA1 and Anoctamin-2; A.F. conducted the analysis; A.F., C.C. and N.S. interpreted the data; A.F., C.C. and N.S. participated in the original draft preparation, writing, reviewing, editing, and revising the manuscript during peer-review. All authors have read and agreed to the published version of the manuscript.

Funding: André Fonseca received funding from FCT—Fundação para a Ciência e Tecnologia, Portugal (refs. SFRH/BD/147629/2019 and UIDB/00006/2020). Clara Cordeiro and Nuno Sepúlveda received partial funding from FCT—Fundação para a Ciência e Tecnologia, Portugal (ref. UIDB/00006/2020).

Institutional Review Board Statement Ethical review and approval were waived for this study given that it is a re-analysis of a public data set. The original study was approved by the Ethics Committee of Charité Universitätsmedizin Berlin in accordance with the 1964 Declaration of Helsinki [28]. This paper deals with open-access data previously published in the following paper: https://doi.org/10.1371/journal.pone.0179124.

Informed Consent Statement Informed consent was obtained from all subjects involved in the original study [28].

Data Availability Statement The dataset is freely available as Supplementary Materials of Loebel et al. [28] (accessed on 1 March 2023). The R scripts are available without restriction at the following address: https://github.com/Publications/Fonseca_etal. The authors declare no conflict of interest.

Conflicts of Interest The authors declare no conflicts of interest.

References

[1] Mateo Cortes Rivera, Claudio Mastronardi, Claudia T Silva-Aldana, Mauricio Arcos-Burgos, and Brett A Lidbury. Myalgic encephalomyelitis/chronic fatigue syndrome: a comprehensive review. *Diagnostics*, 9(3):91, 2019.

[2] Undine-Sophie Deumer, Angelica Varesi, Valentina Floris, Gabriele Savioli, Elisa Mantovani, Paulina López-Carrasco, Gian Marco Rosati, Sakshi Prasad, and Giovanni Ricevuti. Myalgic encephalomyelitis/chronic fatigue syndrome (me/cfs): an overview. *Journal of Clinical Medicine*, 10(20):4786, 2021.

[3] Øystein Fluge, Karl J Tronstad, Olav Mella, et al. Pathomechanisms and possible interventions in myalgic encephalomyelitis/chronic fatigue syndrome (me/cfs). *The Journal of Clinical Investigation*, 131(14), 2021.

[4] Dominic Stanculescu, Lars Larsson, and Jonas Bergquist. Hypothesis: Mechanisms that prevent recovery in prolonged icu patients also underlie myalgic encephalomyelitis/chronic fatigue syndrome (me/cfs). *Frontiers in Medicine*, page 41, 2021.

[5] Dominic Stanculescu, Nuno Sepúlveda, Chin Leong Lim, and Jonas Bergquist. Lessons from heat stroke for understanding myalgic encephalomyelitis/chronic fatigue syndrome. *Frontiers in Neurology*, 12:2258, 2021.

[6] Luis Nacul, Shennae O'Boyle, Luigi Palla, Flavio E Nacul, Kathleen Mudie, Caroline C Kingdon, Jacqueline M Cliff, Taane G Clark, Hazel M Dockrell, and Eliana M Lacerda. How myalgic encephalomyelitis/chronic fatigue syndrome (me/cfs) progresses: the natural history of me/cfs. *Frontiers in neurology*, 11:826, 2020.

[7] Jonas Blomberg, Carl-Gerhard Gottfries, Amal Elfaitouri, Muhammad Rizwan, and Anders Rosén. Infection elicited autoimmunity and myalgic encephalomyelitis/chronic fatigue syndrome: an explanatory model. *Frontiers in immunology*, 9:229, 2018.

[8] Franziska Sotzny, Julià Blanco, Enrica Capelli, Jesús Castro-Marrero, Sophie Steiner, Modra Murovska, Carmen Scheibenbogen, et al. Myalgic encephalomyelitis/chronic fatigue syndrome–evidence for an autoimmune disease. *Autoimmunity reviews*, 17(6):601–609, 2018.

[9] Gerwyn Morris, Michael Berk, Piotr Galecki, and Michael Maes. The emerging role of autoimmunity in myalgic encephalomyelitis/chronic fatigue syndrome (me/cfs). *Molecular neurobiology*, 49:741–756, 2014.

[10] Lindsay S Petracek, Stacy J Suskauer, Rebecca F Vickers, Neel R Patel, Richard L Violand, Renee L Swope, and Peter C Rowe. Adolescent and young adult me/cfs after confirmed or probable covid-19. *Frontiers in Medicine*, 8:668944, 2021.

[11] Hannah E Davis, Gina S Assaf, Lisa McCorkell, Hannah Wei, Ryan J Low, Yochai Re'em, Signe Redfield, Jared P Austin, and Athena Akrami. Characterizing long covid in an

international cohort: 7 months of symptoms and their impact. *EClinicalMedicine*, 38, 2021.

[12] Leonard A Jason and Joseph A Dorri. Me/cfs and post-exertional malaise among patients with long covid. *Neurology International*, 15(1):1–11, 2022.

[13] Rodrigo Vélez-Santamaría, Jessica Fernández-Solana, Fátima Méndez-López, Marta Domínguez-García, Jerónimo J González-Bernal, Rosa Magallón-Botaya, Bárbara Oliván-Blázquez, Josefa González-Santos, and Mirian Santamaría-Peláez. Functionality, physical activity, fatigue and quality of life in patients with acute covid-19 and long covid infection. *Scientific Reports*, 13(1):19907, 2023.

[14] Manuel Ruiz-Pablos, Bruno Paiva, Rosario Montero-Mateo, Nicolas Garcia, and Aintzane Zabaleta. Epstein-barr virus and the origin of myalgic encephalomyelitis or chronic fatigue syndrome. *Frontiers in Immunology*, page 4637, 2021.

[15] Maria Eugenia Ariza. Myalgic encephalomyelitis/chronic fatigue syndrome: the human herpesviruses are back! *Biomolecules*, 11(2):185, 2021.

[16] Nuno Sepúlveda, Jorge Carneiro, Eliana Lacerda, and Luis Nacul. Myalgic encephalomyelitis/chronic fatigue syndrome as a hyper-regulated immune system driven by an interplay between regulatory t cells and chronic human herpesvirus infections. *Frontiers in immunology*, 10:2684, 2019.

[17] Manuel Ruiz-Pablos, Bruno Paiva, and Aintzane Zabaleta. Epstein–barr virus-acquired immunodeficiency in myalgic encephalomyelitis—is it present in long covid? *Journal of Translational Medicine*, 21(1):1–30, 2023.

[18] Brandon S Cox, Khaled Alharshawi, Irene Mena-Palomo, William P Lafuse, and Maria Eugenia Ariza. Ebv/hhv-6a dutpases contribute to myalgic encephalomyelitis/chronic fatigue syndrome pathophysiology by enhancing tfh cell differentiation and extrafollicular activities. *JCI insight*, 7(11), 2022.

[19] Giovanni Capone, Michele Calabrò, Guglielmo Lucchese, Candida Fasano, Bruna Girardi, Lorenzo Polimeno, and Darja Kanduc. Peptide matching between epstein–barr virus and human proteins. *Pathogens and disease*, 69(3):205–212, 2013.

[20] C Baboonian, PJ Venables, DG Williams, RO Williams, and RN Maini. Cross reaction of antibodies to a glycine/alanine repeat sequence of epstein-barr virus nuclear antigen-1 with collagen, cytokeratin, and actin. *Annals of the rheumatic diseases*, 50(11):772, 1991.

[21] Jonas Blomberg, Muhammad Rizwan, Agnes Böhlin-Wiener, Amal Elfaitouri, Per Julin, Olof Zachrisson, Anders Rosén, and Carl-Gerhard Gottfries. Antibodies to

human herpesviruses in myalgic encephalomyelitis/chronic fatigue syndrome patients. *Frontiers in Immunology*, 10:1946, 2019.

[22] A Martin Lerner, Maria E Ariza, Marshall Williams, Leonard Jason, Safedin Beqaj, James T Fitzgerald, Stanley Lemeshow, and Ronald Glaser. Antibody to epstein-barr virus deoxyuridine triphosphate nucleotidohydrolase and deoxyribonucleotide polymerase in a chronic fatigue syndrome subset. *PloS one*, 7(11):e47891, 2012.

[23] Jonathan R Kerr. Epstein-barr virus induced gene-2 upregulation identifies a particular subtype of chronic fatigue syndrome/myalgic encephalomyelitis. *Frontiers in Pediatrics*, page 59, 2019.

[24] Tiago Dias Domingues, Anna D Grabowska, Ji-Sook Lee, Jose Ameijeiras-Alonso, Francisco Westermeier, Carmen Scheibenbogen, Jacqueline M Cliff, Luis Nacul, Eliana M Lacerda, Helena Mouriño, et al. Herpesviruses serology distinguishes different subgroups of patients from the united kingdom myalgic encephalomyelitis/chronic fatigue syndrome biobank. *Frontiers in medicine*, 8:686736, 2021.

[25] João Malato, Luís Graça, and Nuno Sepúlveda. Impact of misdiagnosis in case-control studies of myalgic encephalomyelitis/chronic fatigue syndrome. *Diagnostics*, 13(3):531, 2023.

[26] Luis Nacul, Eliana M Lacerda, Caroline C Kingdon, Hayley Curran, and Erinna W Bowman. How have selection bias and disease misclassification undermined the validity of myalgic encephalomyelitis/chronic fatigue syndrome studies? *Journal of health psychology*, 24(12):1765–1769, 2019.

[27] Maria Eugenia Ariza. Commentary: antibodies to human herpesviruses in myalgic encephalomyelitis/chronic fatigue syndrome patients. *Frontiers in Immunology*, 11:1400, 2020.

[28] Madlen Loebel, Maren Eckey, Franziska Sotzny, Elisabeth Hahn, Sandra Bauer, Patricia Grabowski, Johannes Zerweck, Pavlo Holenya, Leif G Hanitsch, Kirsten Wittke, et al. Serological profiling of the ebv immune response in chronic fatigue syndrome using a peptide microarray. *PloS one*, 12(6):e0179124, 2017.

[29] Nuno Sepúlveda, João Malato, Franziska Sotzny, Anna D Grabowska, André Fonseca, Clara Cordeiro, Luís Graça, Przemyslaw Biecek, Uta Behrends, Josef Mautner, et al. Revisiting igg antibody reactivity to epstein-barr virus in myalgic encephalomyelitis/chronic fatigue syndrome and its potential application to disease diagnosis. *Frontiers in Medicine*, page 1775, 2022.

[30] V Roshan Joseph. Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 15(4):531–538, 2022.

[31] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[32] Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

[33] Mónica López-Ratón, María Xosé Rodríguez-Álvarez, Carmen Cadarso-Suárez, and Francisco Gude-Sampedro. Optimalcutpoints: an r package for selecting optimal cutpoints in diagnostic tests. *Journal of statistical software*, 61:1–36, 2014.

[34] John Joseph Valletta and Mario Recker. Identification of immune signatures predictive of clinical protection from malaria. *PLoS computational biology*, 13(10):e1005812, 2017.

[35] Robert C Wilson, Amitai Shenhav, Mark Straccia, and Jonathan D Cohen. The eighty five percent rule for optimal learning. *Nature communications*, 10(1):4646, 2019.

[36] NA O'Leary, MW Wright, JR Brister, S Ciufo, D Haddad, R McVeigh, et al. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Europe PMC free article][Abstract][Google Scholar]*, pages 733–745, 2016.

[37] Samuel Karlin and Stephen F Altschul. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proceedings of the National Academy of Sciences*, 90(12):5873–5877, 1993.

[38] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[39] Max Kuhn. Building predictive models in r using the caret package. *Journal of statistical software*, 28:1–26, 2008.

[40] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC bioinformatics*, 12(1):1–8, 2011.

[41] Marvin N Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in c++ and r. *arXiv preprint arXiv:1508.04409*, 2015.

[42] Eric Polley, Erin LeDell, Chris Kennedy, Sam Lendle, and Mark van der Laan. Package 'superlearner'. *CRAN*, 2019.

[43] Katarina Tengvall, Jesse Huang, Cecilia Hellström, Patrick Kammer, Martin Biström, Burcu Ayoglu, Izaura Lima Bomfim, Pernilla Stridh, Julia Butt, Nicole Brenner, et al. Molecular mimicry between anoctamin 2 and epstein-barr virus nuclear antigen 1 associates with multiple sclerosis risk. *Proceedings of the National Academy of Sciences*, 116(34):16955–16960, 2019.

[44] Nuno Sepulveda. Impact of genetic variation on the molecular mimicry between anoctamin-2 and epstein-barr virus nuclear antigen 1 in multiple sclerosis. *Immunology Letters*, 238:29–31, 2021.

[45] Adam J O'Neal, Katherine A Glass, Christopher J Emig, Adela A Vitug, Steven J Henry, Dikoma C Shungu, Xiangling Mao, Susan M Levine, and Maureen R Hanson. Survey of anti-pathogen antibody levels in myalgic encephalomyelitis/chronic fatigue syndrome. *Proteomes*, 10(2):21, 2022.

[46] Leonard A Jason, Karina Corradi, Susan Torres-Harding, Renee R Taylor, and Caroline King. Chronic fatigue syndrome: the need for subtypes. *Neuropsychology review*, 15:29–58, 2005.

[47] Klemens Ruprecht, Benjamin Wunderlich, René Gieß, Petra Meyer, Madlen Loebel, Klaus Lenz, Jörg Hofmann, Berit Rosche, Oliver Wengert, Friedemann Paul, et al. Multiple sclerosis: The elevated antibody response to epstein–barr virus primarily targets, but is not confined to, the glycine–alanine repeat of epstein–barr nuclear antigen-1. *Journal of neuroimmunology*, 272(1-2):56–61, 2014.

[48] Kjetil Bjornevik, Marianna Cortese, Brian C Healy, Jens Kuhle, Michael J Mina, Yumei Leng, Stephen J Elledge, David W Niebuhr, Ann I Scher, Kassandra L Munger, et al. Longitudinal analysis reveals high prevalence of epstein-barr virus associated with multiple sclerosis. *Science*, 375(6578):296–301, 2022.

[49] Leonard A Jason, D Ohanian, A Brown, M Sunnquist, S McManimen, L Klebek, P Fox, and M Sorenson. Differentiating multiple sclerosis from myalgic encephalomyelitis and chronic fatigue syndrome. *Insights in biomedicine*, 2(2), 2017.

[50] Diana Ohanian, Abigail Brown, Madison Sunnquist, Jacob Furst, Laura Nicholson, Lauren Klebek, and Leonard A Jason. Identifying key symptoms differentiating myalgic encephalomyelitis and chronic fatigue syndrome from multiple sclerosis. *Neurology (E-Cronicon)*, 4(2):41, 2016.

[51] Tiago Dias Domingues, João Malato, Anna D Grabowska, Ji-Sook Lee, Jose Ameijeiras-Alonso, Przemysław Biecek, Luís Graça, Helena Mouriño, Carmen

Scheibenbogen, Francisco Westermeier, et al. Association analysis between symptomology and herpesvirus igg antibody concentrations in myalgic encephalomyelitis/chronic fatigue syndrome (me/cfs) and multiple sclerosis. *Heliyon*, 9(7), 2023.

[52] Sophie Steiner, Annick Fehrer, Friederike Hoheisel, Simon Schoening, Anna Aschenbrenner, Nina Babel, Judith Bellmann-Strobl, Carsten Finke, Øystein Fluge, Laura Froehlich, et al. Understanding, diagnosing, and treating myalgic encephalomyelitis/chronic fatigue syndrome–state of the art: Report of the 2nd international meeting at the charité fatigue center. *Autoimmunity reviews*, page 103452, 2023.

[53] Carmen Scheibenbogen, Madlen Loebel, Helma Freitag, Anne Krueger, Sandra Bauer, Michaela Antelmann, Wolfram Doehner, Nadja Scherbakov, Harald Heidecke, Petra Reinke, et al. Immunoadsorption to remove ß2 adrenergic receptor antibodies in chronic fatigue syndrome cfs/me. *PLoS One*, 13(3):e0193672, 2018.

[54] Markus Tölle, Helma Freitag, Michaela Antelmann, Jelka Hartwig, Mirjam Schuchardt, Markus van der Giet, Kai-Uwe Eckardt, Patricia Grabowski, and Carmen Scheibenbogen. Myalgic encephalomyelitis/chronic fatigue syndrome: efficacy of repeat immunoadsorption. *Journal of Clinical Medicine*, 9(8):2443, 2020.

[55] Elisa Stein, Cornelia Heindrich, Kirsten Wittke, Claudia Kedor, Laura Kim, Helma Freitag, Anne Krueger, Markus Toelle, and Carmen Scheibenbogen. Observational study of repeat immunoadsorption (ria) in post-covid me/cfs patients with elevated ß2-adrenergic receptor autoantibodies—an interim report. *Journal of Clinical Medicine*, 12(19):6428, 2023.

[56] Øystein Fluge, Kristin Risa, Sigrid Lunde, Kine Alme, Ingrid Gurvin Rekeland, Dipak Sapkota, Einar Kleboe Kristoffersen, Kari Sørland, Ove Bruland, Olav Dahl, et al. B-lymphocyte depletion in myalgic encephalopathy/chronic fatigue syndrome. an open-label phase ii study with rituximab maintenance treatment. *PloS one*, 10(7):e0129898, 2015.

[57] Øystein Fluge, Ove Bruland, Kristin Risa, Anette Storstein, Einar K Kristoffersen, Dipak Sapkota, Halvor Næss, Olav Dahl, Harald Nyland, and Olav Mella. Benefit from b-lymphocyte depletion using the anti-cd20 antibody rituximab in chronic fatigue syndrome. a double-blind and placebo-controlled study. *PloS one*, 6(10):e26358, 2011.

[58] Øystein Fluge, Ingrid G Rekeland, Katarina Lien, Hanne Thürmer, Petter C Borchgrevink, Christoph Schäfer, Kari Sørland, Jörg Aßmus, Irini Ktoridou-Valen, Ingrid Herder, et al. B-lymphocyte depletion in patients with myalgic encephalomyelitis/chronic fatigue syndrome: a randomized, double-blind, placebo-controlled trial. *Annals of internal medicine*, 170(9):585–593, 2019.

[59] Ingrid G Rekeland, Alexander Fosså, Asgeir Lande, Irini Ktoridou-Valen, Kari Sørland, Mari Holsen, Karl J Tronstad, Kristin Risa, Kine Alme, Marte K Viken, et al. Intravenous cyclophosphamide in myalgic encephalomyelitis/chronic fatigue syndrome. an open-label phase ii study. *Frontiers in medicine*, page 162, 2020.

[60] Francesco Patti, Salvatore Lo Fermo, et al. Lights and shadows of cyclophosphamide in the treatment of multiple sclerosis. *Autoimmune Diseases*, 2011, 2011.

[61] JM Willers and E Sluis. The influence of cyclophosphamide on antibody formation in the mouse. In *Annales D'immunologie*, volume 126, pages 267–279, 1975.

[62] Martina Ahlmann and Georg Hempel. The effect of cyclophosphamide on the immune system: implications for clinical cancer therapy. *Cancer chemotherapy and pharmacology*, 78:661–671, 2016.

[63] Eric R Hurd and VJ Giuliano. The effect of cyclophosphamide on b and t lymphocytes in patients with connective tissue diseases. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, 18(1):67–75, 1975.

[64] Luis C Nacul, Eliana M Lacerda, Derek Pheby, Peter Campion, Mariam Molokhia, Shagufta Fayyaz, Jose CDC Leite, Fiona Poland, Amanda Howe, and Maria L Drachler. Prevalence of myalgic encephalomyelitis/chronic fatigue syndrome (me/cfs) in three regions of england: a repeated cross-sectional study in primary care. *BMC medicine*, 9(1):1–12, 2011.

[65] Lily Chu, Ian J Valencia, Donn W Garvert, and Jose G Montoya. Onset patterns and course of myalgic encephalomyelitis/chronic fatigue syndrome. *Frontiers in pediatrics*, 7:12, 2019.

[66] Rhea J Longley, Michael T White, Eizo Takashima, Jessica Brewster, Masayuki Morita, Matthias Harbers, Thomas Obadia, Leanne J Robinson, Fumie Matsuura, Zoe SJ Liu, et al. Development and validation of serological markers for detecting recent plasmodium vivax infection. *Nature medicine*, 26(5):741–749, 2020.

[67] Danica A Helb, Kevin KA Tetteh, Philip L Felgner, Jeff Skinner, Alan Hubbard, Emmanuel Arinaitwe, Harriet Mayanja-Kizza, Isaac Ssewanyana, Moses R Kamya, James G Beeson, et al. Novel serologic biomarkers provide accurate estimates of recent plasmodium falciparum exposure for individuals and communities. *Proceedings of the National Academy of Sciences*, 112(32):E4438–E4447, 2015.

[68] Simon P Brooks and Stephen B Dunnett. Tests to assess motor phenotype in mice: a user's guide. *Nature Reviews Neuroscience*, 10(7):519–529, 2009.

[69] Andreas Goebel, Emerson Krock, Clive Gentry, Mathilde R Israel, Alexandra Jurczak, Carlos Morado Urbina, Katalin Sandor, Nisha Vastani, Margot Maurer, Ulku Cuhadar, et al. Passive transfer of fibromyalgia symptoms from patients to mice. *The Journal of clinical investigation*, 131(13), 2021.

[70] Adam Janowski, Joseph Lesnak, Ashley Plumb, Lynn Rasmussen, and Kathleen Sluka. Development of a mouse model for chronic fatigue syndrome. *The Journal of Pain*, 23(5):12, 2022.

[71] Yasuhisa Tamura, Masanori Yamato, and Yosky Kataoka. Animal models for neuroinflammation and potential treatment methods. *Frontiers in Neurology*, 13:890217, 2022.

[72] Sjoerd Johannes Bastiaan Holwerda and Wouter de Laat. Ctcf: the protein, the binding partners, the binding sites and their chromatin loops. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1620):20120369, 2013.

[73] Bondita Dehingia, Małgorzata Milewska, Marcin Janowski, and Aleksandra Pękowska. Ctcf shapes chromatin structure and gene expression in health and disease. *EMBO reports*, 23(9):e55146, 2022.

[74] Joanna R DiSpirito, David Zemmour, Deepshika Ramanan, Jun Cho, Rapolas Zilionis, Allon M Klein, Christophe Benoist, and Diane Mathis. Molecular diversification of regulatory t cells in nonlymphoid tissues. *Science immunology*, 3(27):eaat5861, 2018.

[75] Venkataragavan Chandrasekaran, Nina Oparina, Maria-Jose Garcia-Bonete, Caroline Wasén, Malin C Erlandsson, Eric Malmhäll-Bah, Karin ME Andersson, Maja Jensen, Sofia T Silfverswärd, Gergely Katona, et al. Cohesin-mediated chromatin interactions and autoimmunity. *Frontiers in Immunology*, 13:840002, 2022.

[76] Jonathan R Kerr. Gene profiling of patients with chronic fatigue syndrome/myalgic encephalomyelitis. *Current rheumatology reports*, 10(6):482–491, 2008.

[77] N Kaushik, D Fear, SCM Richards, CR McDermott, EF Nuwaysir, P Kellam, TJ Harrison, RJ Wilkinson, DAJ Tyrrell, ST Holgate, et al. Gene expression in peripheral blood mononuclear cells from patients with chronic fatigue syndrome. *Journal of clinical pathology*, 58(8):826, 2005.

[78] Eiren Sweetman, Margaret Ryan, Christina Edgar, Angus MacKay, Rosamund Vallings, and Warren Tate. Changes in the transcriptome of circulating immune cells of a new zealand cohort with myalgic encephalomyelitis/chronic fatigue syndrome. *International journal of immunopathology and pharmacology*, 33:2058738418820402, 2019.

[79] Eloy Almenar-Perez, Tamara Ovejero, Teresa Sanchez-Fito, Jose A Espejo, Lubov Nathanson, and Elisa Oltra. Epigenetic components of myalgic encephalomyelitis/chronic fatigue syndrome uncover potential transposable element activation. *Clinical Therapeutics*, 41(4):675–698, 2019.

[80] Wilfred C de Vega, Santiago Herrera, Suzanne D Vernon, and Patrick O McGowan. Epigenetic modifications and glucocorticoid sensitivity in myalgic encephalomyelitis/chronic fatigue syndrome (me/cfs). *BMC medical genomics*, 10:1–14, 2017.

[81] Lorenzo Lorusso, Svetlana V Mikhaylova, Enrica Capelli, Daniela Ferrari, Gaelle K Ngonga, and Giovanni Ricevuti. Immunological aspects of chronic fatigue syndrome. *Autoimmunity reviews*, 8(4):287–291, 2009.

[82] Amin F Majdalawieh, Mariam Massri, and Hyo-Sung Ro. Aebp1 is a novel oncogene: mechanisms of action and signaling pathways. *Journal of oncology*, 2020, 2020.

[83] Chloe Angwin, Neeti Ghali, and Fleur Stephanie van Dijk. Case report: Two individuals with aebp1-related classical-like eds: Further clinical characterisation and description of novel aebp1 variants. *Frontiers in Genetics*, 14:1148224, 2023.

[84] Marco Ritelli, Valeria Cinquina, Marina Venturini, Letizia Pezzaioli, Anna Maria Formenti, Nicola Chiarelli, and Marina Colombi. Expanding the clinical and mutational spectrum of recessive aebp1-related classical-like ehlers-danlos syndrome. *Genes*, 10(2):135, 2019.

[85] M Castori, C Celletti, F Camerota, and P Grammatico. Chronic fatigue syndrome is commonly diagnosed in patients with ehlers-danlos syndrome hypermobility type/joint hypermobility syndrome. *Clinical and Experimental Rheumatology-Incl Supplements*, 29(3):597, 2011.

[86] Alan Hakim, Inge De Wandele, Chris O'Callaghan, Alan Pocinki, and Peter Rowe. Chronic fatigue in ehlers–danlos syndrome—hypermobile type. In *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, volume 175, pages 175–180. Wiley Online Library, 2017.

[87] Luis Nacul, François Jérôme Authier, Carmen Scheibenbogen, Lorenzo Lorusso, Ingrid Bergliot Helland, Jose Alegre Martin, Carmen Adella Sirbu, Anne Marit Mengshoel, Olli Polo, Uta Behrends, et al. European network on myalgic encephalomyelitis/chronic fatigue syndrome (euromene): expert consensus on the diagnosis, service provision, and care of people with me/cfs in europe. *Medicina*, 57(5):510, 2021.

[88] KA Schlauch, Svetlana F Khaiboullina, Kenny L De Meirleir, Shanti Rawat, Julia Petereit, AA Rizvanov, N Blatt, Tatjana Mijatovic, Doina Kulick, A Palotas, et al. Genome-

wide association analysis identifies genetic variations in subjects with myalgic encephalomyelitis/chronic fatigue syndrome. *Translational psychiatry*, 6(2):e730–e730, 2016.

[89] Joshua J Dibble, Simon J McGrath, and Chris P Ponting. Genetic risk factors of me/cfs: a critical review. *Human Molecular Genetics*, 29(R1):R117–R124, 2020.

[90] Santiago Herrera, Wilfred C de Vega, David Ashbrook, Suzanne D Vernon, and Patrick O McGowan. Genome-epigenome interactions associated with myalgic encephalomyelitis/chronic fatigue syndrome. *Epigenetics*, 13(12):1174–1190, 2018.

[91] Riad Hajdarevic, Asgeir Lande, Jesper Mehlsen, Anne Rydland, Daisy D Sosa, Elin B Strand, Olav Mella, Flemming Pociot, Øystein Fluge, Benedicte A Lie, et al. Genetic association study in myalgic encephalomyelitis/chronic fatigue syndrome (me/cfs) identifies several potential risk loci. *Brain, Behavior, and Immunity*, 102:362–369, 2022.

[92] Sayoni Das, Krystyna Taylor, James Kozubek, Jason Sardell, and Steve Gardner. Genetic risk factors for me/cfs identified using combinatorial analysis. *Journal of Translational Medicine*, 20(1):1–20, 2022.

[93] Milan Haffke, Helma Freitag, Gordon Rudolf, Martina Seifert, Wolfram Doehner, Nadja Scherbakov, Leif Hanitsch, Kirsten Wittke, Sandra Bauer, Frank Konietschke, et al. Endothelial dysfunction and altered endothelial biomarkers in patients with post-covid-19 syndrome and chronic fatigue syndrome (me/cfs). *Journal of Translational Medicine*, 20(1):1–11, 2022.

[94] Nadja Scherbakov, Marvin Szklarski, Jelka Hartwig, Franziska Sotzny, Sebastian Lorenz, Antje Meyer, Patricia Grabowski, Wolfram Doehner, and Carmen Scheibenbogen. Peripheral endothelial dysfunction in myalgic encephalomyelitis/chronic fatigue syndrome. *ESC heart failure*, 7(3):1064–1071, 2020.

[95] J Blauensteiner, Romina Bertinat, Luis E León, Monika Riederer, Nuno Sepúlveda, and Francisco Westermeier. Altered endothelial dysfunction-related mirs in plasma from me/cfs patients. *Scientific Reports*, 11(1):10604, 2021.

[96] Jian Wang, Ivan Jelcic, Lena Mühlenbruch, Veronika Haunerdinger, Nora C Toussaint, Yingdong Zhao, Carolina Cruciani, Wolfgang Faigle, Reza Naghavian, Magdalena Foege, et al. Hla-dr15 molecules jointly shape an autoreactive t cell repertoire in multiple sclerosis. *Cell*, 183(5):1264–1281, 2020.

[97] Thomas Pradeu and Edwin L Cooper. The danger theory: 20 years later. *Frontiers in immunology*, 3:287, 2012.

[98] Stefania Gallucci and Polly Matzinger. Danger signals: Sos to the immune system. *Current opinion in immunology*, 13(1):114–119, 2001.

[99] Irun R Cohen and Douglas B Young. Autoimmunity, microbial immunity and the immunological homunculus. *Immunology today*, 12(4):105–110, 1991.

[100] Francisco J Quintana, Avishai Mimran, Pnina Carmi, Felix Mor, and Irun R Cohen. Hsp60 as a target of anti-ergotypic regulatory t cells. *PLoS One*, 3(12):e4026, 2008.

[101] Y Jammes, JG Steinberg, and S Delliaux. Chronic fatigue syndrome: acute infection and history of physical activity affect resting levels and response to exercise of plasma oxidant/antioxidant status and heat shock proteins. *Journal of internal medicine*, 272(1):74–84, 2012.

[102] Anita A Thambirajah, Kenna Sleigh, H Grant Stiver, and Anthony W Chow. Differential heat shock protein responses to strenuous standardized exercise in chronic fatigue syndrome patients and matched healthy controls. *Clinical and Investigative Medicine*, 31(6):E319–E327, 2008.

[103] Ellen Wright Clayton. Beyond myalgic encephalomyelitis/chronic fatigue syndrome: an iom report on redefining an illness. *Jama*, 313(11):1101–1102, 2015.

[104] Amal Elfaitouri, Björn Herrmann, Agnes Bölin-Wiener, Yilin Wang, Carl-Gerhard Gottfries, Olof Zachrisson, R$\ddot{v}$diger Pipkorn, Lars Rönnblom, and Jonas Blomberg. Epitopes of microbial and human heat shock protein 60 and their recognition in myalgic encephalomyelitis. *PLoS One*, 8(11):e81155, 2013.

[105] Yves Delneste, Nathalie Herbault, Brice Galea, Giovanni Magistrelli, Ingrid Bazin, Jean-Yves Bonnefoy, and Pascale Jeannin. Vasoactive intestinal peptide synergizes with tnf-α in inducing human dendritic cell maturation. *The Journal of Immunology*, 163(6):3071–3075, 1999.

[106] RP Gomariz, Y Juarranz, Ce ABAD, A Arranz, J Leceta, and C Martinez. Vip–pacap system in immunity: new insights for multitarget therapy. *Annals of the New York Academy of Sciences*, 1070(1):51–74, 2006.

[107] Donald R Staines. Is chronic fatigue syndrome an autoimmune disorder of endogenous neuropeptides, exogenous infection and molecular mimicry? *Medical hypotheses*, 62(5):646–652, 2004.

[108] Ekua W Brenu, Mieke L van Driel, Don R Staines, Kevin J Ashton, Sandra B Ramos, James Keane, Nancy G Klimas, and Sonya M Marshall-Gradisnik. Immunological abnormalities as potential biomarkers in chronic fatigue syndrome/myalgic encephalomyelitis. *Journal of translational medicine*, 9(1):1–9, 2011.

[109] Elena Gonzalez-Rey and Mario Delgado. Vasoactive intestinal peptide and regulatory t-cell induction: a new mechanism and therapeutic potential for immune homeostasis. *Trends in molecular medicine*, 13(6):241–251, 2007.

[110] S Blundell, KK Ray, M Buckland, and PD White. Chronic fatigue syndrome and circulating cytokines: a systematic review. *Brain, behavior, and immunity*, 50:186–195, 2015.

[111] Nicoletta Carlo-Stella, C Badulli, A De Silvestri, LAURA Bazzichi, M Martinetti, L Lorusso, S Bombardieri, Laura Salvaneschi, and Mariaclara Cuccia. A first study of cytokine genomic polymorphisms in cfs: Positive association of tnf-857 and ifngamma 874 rare alleles. *Clinical and experimental rheumatology*, 24(2):179, 2006.

[112] Sophie Steiner, Sonya C Becker, Jelka Hartwig, Franziska Sotzny, Sebastian Lorenz, Sandra Bauer, Madlen Löbel, Anna B Stittrich, Patricia Grabowski, and Carmen Scheibenbogen. Autoimmunity-related risk variants in ptpn22 and ctla4 are associated with me/cfs with infectious onset. *Frontiers in Immunology*, 11:578, 2020.

[113] Aditya Yashwant Sarode, Mukesh Kumar Jha, Shubhranshu Zutshi, Soumya Kanti Ghosh, Hima Mahor, Uddipan Sarma, and Bhaskar Saha. Residue-specific message encoding in cd40-ligand. *Iscience*, 23(9), 2020.

[114] Danielle Burger, Nicolas Molnarfi, Lyssia Gruaz, and Jean-Michel Dayer. Differential induction of il-1β and tnf by cd40 ligand or cellular contact with stimulated t cells depends on the maturation stage of human monocytes. *The Journal of Immunology*, 173(2):1292–1297, 2004.

[115] Rosa M Andrade, Matthew Wessendarp, and Carlos S Subauste. Cd154 activates macrophage antimicrobial activity in the absence of ifn-γ through a tnf-α-dependent mechanism. *The Journal of Immunology*, 171(12):6750–6756, 2003.

[116] Li-Fan Lu, W James Cook, Ling-Li Lin, and Randolph J Noelle. Cd40 signaling through a newly identified tumor necrosis factor receptor-associated factor 2 (traf2) binding site. *Journal of Biological Chemistry*, 278(46):45414–45418, 2003.

[117] Silvio Danese, Miquel Sans, Franco Scaldaferri, Alessandro Sgambato, Sergio Rutella, Achille Cittadini, Josep M Piqué, Julian Panes, Jeffry A Katz, Antonio Gasbarrini, et al. Tnf-α blockade down-regulates the cd40/cd40l pathway in the mucosal microcirculation: a novel anti-inflammatory mechanism of infliximab in crohn's disease. *The Journal of Immunology*, 176(4):2617–2624, 2006.

[118] Andrea T White, Alan R Light, Ronald W Hughen, Lucinda Bateman, Thomas B Martins, Harry R Hill, and Kathleen C Light. Severity of symptom flare after moderate

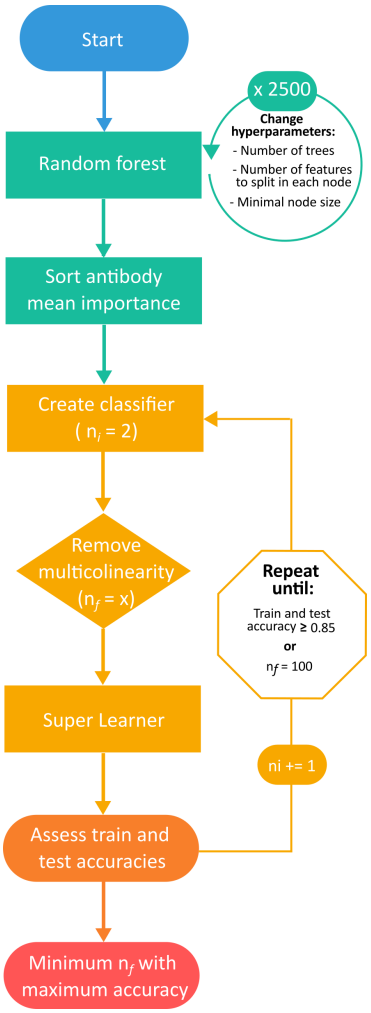
exercise is linked to cytokine activity in chronic fatigue syndrome. *Psychophysiology*, 47(4):615–624, 2010.

[119] Mady Hornig, José G Montoya, Nancy G Klimas, Susan Levine, Donna Felsenstein, Lucinda Bateman, Daniel L Peterson, C Gunnar Gottschalk, Andrew F Schultz, Xiaoyu Che, et al. Distinct plasma immune signatures in me/cfs are present early in the course of illness. *Science advances*, 1(1):e1400121, 2015.

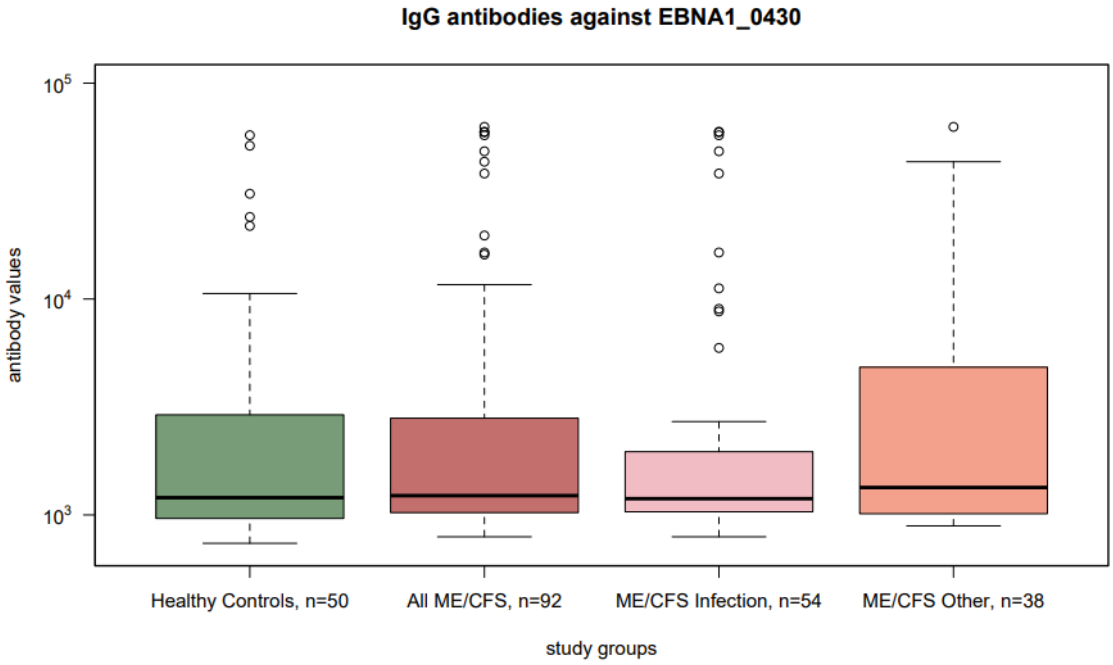
[120] Thomas Luft, Michael Jefford, Petra Luetjens, Hubertus Hochrein, Kelly-Anne Masterman, Charlie Maliszewski, Ken Shortman, Jonathan Cebon, and Eugene Maraskovsky. Il-1β enhances cd40 ligand-mediated cytokine secretion by human dendritic cells (dc): a mechanism for t cell-independent dc activation. *The Journal of Immunology*, 168(2):713–722, 2002.

[121] Amy Wesa and Anne Galy. Increased production of pro-inflammatory cytokines and enhanced t cell responses after activation of human dendritic cells with il-1 and cd40 ligand. *BMC immunology*, 3(1):1–11, 2002.

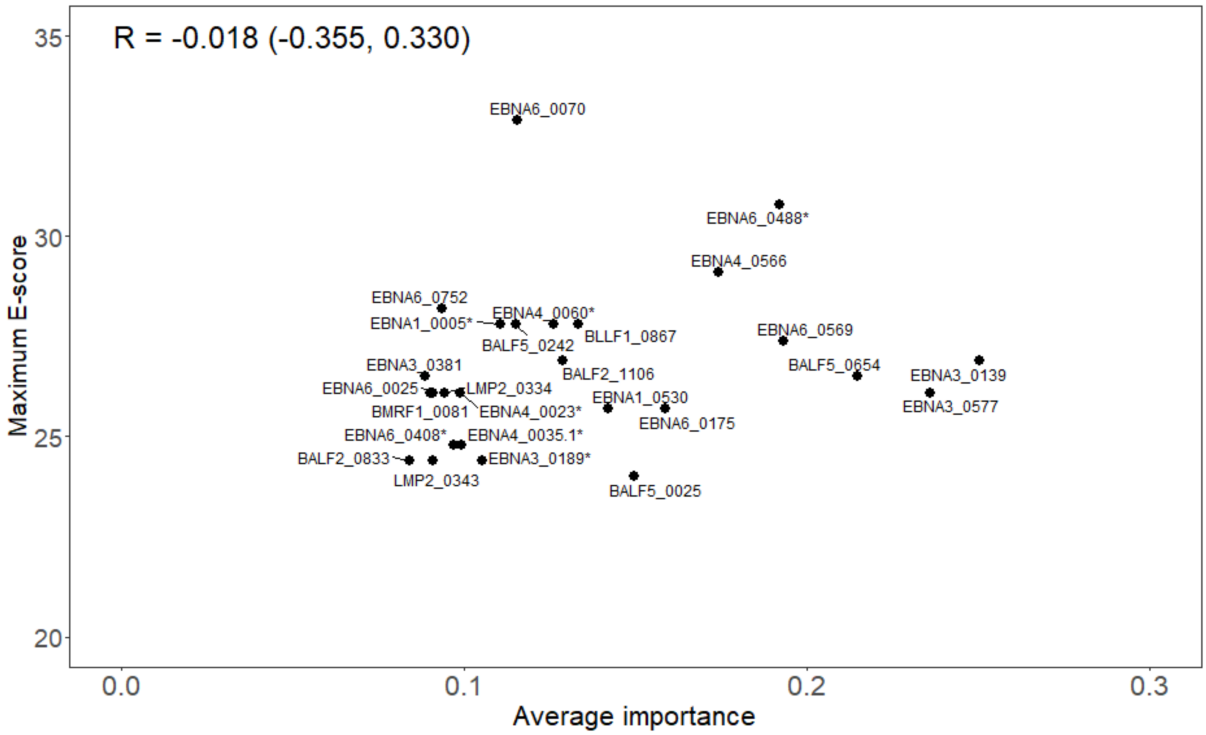
Supplementary Materials



Supplementary Figure 2: **Pipeline for data analysis where different steps are shown in the flowchart using distinct coloured shapes.** Green refers to the RF step, where two-thousand and five-hundred RF runs were conducted while changing the model hyperparameters. The average antibody importance was then obtained, and antibodies were sorted according to their average importance. Light orange refers to the steps concerning the classifier construction and assessment of its train and test predictive accuracies using the SL. The process of creating a classifier and assessing its predictive performance through the SL was repeated, each time adding a new feature, until the stopping criterium (octagonal shape) was overified. Once a new feature was added to the construct, the Spearman coefficient was obtained to remove highly correlated features. The flowchart ended by selecting the optimal classifier, the one with the minimum number of antibody responses (kmin), while simultaneously achieving a train and test accuracy above or equal to 0.85. If the accuracy of 0.85 was not met, the classifier with the minimum number of antibody responses and highest accuracy would be considered the best classifier.



Supplementary Figure 3: **Data concerning to the IgG antibody against EBNA1_0430.** Boxplot of data related to IgG antibody against EBNA1_0430 in the whole ME/CFS group (All ME/CFS), the ME/CFS subgroup whose patients reported an infectious disease trigger (ME/CFS Infection), the ME/CFS subgroup whose patients reported a non-infectious disease trigger or did not know their disease trigger (ME/CFS Other), and their healthy controls. Differences in antibody distributions between ME/CFS group/subgroups and healthy controls were not statistically significant according to the Mann-Whitney-Wilcoxon test with continuity correction (All ME/CFS versus HC, p-value = 0.410; ME/CFS Infection versus HC, p-value = 0.626; ME/CFS Infection versus HC, p-value = 0.322).



Supplementary Figure 4: **Bioinformatic analysis of the EBV peptides associated with the 26 antibodies for predicting ME/CFS patients with an infectious disease trigger.** Scatterplot between the average importance of each EBV peptide associated with the 26 selected antibodies and the maximum E-score alignment score with human proteins using the *nr* protein database where R is the Spearman's correlation coefficient with the respective 95% confidence interval in brackets.

Chapter 9 - Assessing model reliability: The impact of train-test split proportions on the accuracy and reliability of biomarkers against clinical Malaria

André Fonseca[1,2,*], Clara Cordeiro[1,2] and Nuno Sepúlveda[2,3]

[1] FCT - Faculty of Sciences and Technology, University of Algarve, 8005-139 Faro, Portugal
[2] CEAUL - Centre of Statistics and its Applications, Faculty of Sciences, University of Lisbon, Portugal
[3] Faculty of Mathematics & Information Science, Warsaw University of Technology, 00-662 Warsaw, Poland
Being drafted for submission in BioMedInformatics

9.1 Abstract

Background: Machine Learning (ML) models have become an essential tool in analysis aiming at identifying antibodies against clinical malaria. These techniques have allowed for a better understanding of the complex interactions between the host's immune responses and the malaria parasite and improved the chance of identifying novel biomarkers against clinical disease. However, a critical aspect when deploying these models is the careful consideration of the train and test split proportions, which can significantly affect the models' robustness and generalizability, leading to unreliable results. Moreover, in the absence of designated methods to evaluate the reliability of the results provided by such models, assessing their trustworthiness becomes inherently challenging. Together, these factors may explain the variability of findings across studies and the difficulty in unveilling the exact identity of the antibodies that confer malaria protection.
Objective: The objective of this paper is two-folded. First this paper delves into the impact of varying the train-test split ratios on the generalization ability of the model, shedding light on the pivotal role this parameter plays in the soundness of model outcome. Next, we have provided a benchmark study for understanding the reliability of such models in datasets with low sample size, similar to the ones traditionally used for the identification of antibody biomarkers against clinical malaria.
Methods: Here, we propose an approach anchored on recent research advocating a specific target accuracy of 85% as a benchmark for model evaluation to estimate the inherent power of the models' performances obtained providing a means to evaluate their reliability. Such models relied on the use of the Random Forest for variable selection and the Super Learner (SL) to assess the predictive performance, where different split ratios (9:1, 8:2, 7:3, 5:5) were considered for data partitioning. To illustrate this strategy, we used published data on IgG

antibody responses to 2054 antigens obtained for 186 individuals living in Mali obtained, using microarrays.
Results: Our results clearly denote the impact of the split ratio on the initial selection of variables, which translated into the identification of distinct classifiers by the predictive SL-model. Furthermore, decreasing the available data to train the model significantly impacted its ability to generalized, which was highlighted by a reduction in the test set accuracy, that dropped below the 0.85 target when using the lower split ratios to partition the data. Notwithstanding, using the 9:1 and 8:2 split ratios we were able to identify a 35-antibody and a 75-antibody classifier, whose respective test set accuracies reached close to 0.90 (95% CI [0.668, 0.987]) and 0.86 (95% CI [0.71, 0.95]), respectively. Finally, our results demonstrate that a much large number of samples (individuals) than the ones typically used in this type of data are required to attain reliable results. Within our framework, we stablished a minimum accuracy boundary of 80% to claim enough statistical power to the target 85% accuracy. Our analysis suggests that a minimum of 1420 observations would be necessary to admit this difference with 95% certainty, a number higher than the one traditionally found in most of these studies.
Conclusions: In conclusion, this research underscores the importance of a deliberate train-test split consideration and introduces a practical framework for assessing the reliability of ML-based models through a target accuracy benchmark. Such approach may help researchers to enhance the credibility of their models, contributing to the establishment of more reliable and reproducible machine learning practices, ultimately aiding in the identification of the antibodies that confer protection against malaria.

Keywords: Random Forest, Accuracy, Sample Size, Power, Type II error

9.2 Introduction

The wide availability of high-throughput antibody data has led to the introduction of machine learning approaches for the identification of antibodies against clinical malaria [1, 2, 3, 4]. Within such approaches, decision tree-based statistical methods, such as the Random Forest [5] have gained particular prominence [1, 4, 6, 7]. However, as these models continue to be leveraged across diverse domains, it becomes necessary to evaluate the reliability of the results that they provide. Central to this is the pivotal role played by the train-test split ratio, a parameter that inherently influences the model's performance, affecting its robustness and ability to generalize to unseen data [8, 9, 10]. Data splitting or train-test split refers to the process of portioning the data into subsets for model training and evaluation of its predictive capability. This technique is indispensable to reduce the model's bias towards the training data, preventing machine learning algorithms from overfitting on the data used to train the model and perform poorly on the new unseen data (test

data) [11]. The delicate balance in the allocation of data for training and testing purposes is well-acknowledged in the machine learning community, however it is often overlooked in malaria studies aiming at identifying biomarkers against malaria, posing a major problem to the reliability of the machine learning models and the consequent findings.

Adding to this, the absence of a pre-designated approach to assess the reliability of the results obtained by such models further aggravates this situation. This becomes even more concerning when we consider the low sample size of traditional high-throughput data malaria studies (usually a few hundred) [1, 2, 3, 4, 12, 13, 14] which directly impacts the model's ability to generalize. Combined, these factors may explain why, despite their outstanding performance, to this date, these techniques have failed to identify the exact identity of the antibodies that confer protection to clinical malaria. Thus here, we contend that the performance of machine learning models should not solely rest on its accuracy but should also consider the statistical power inherent in achieving a specific target accuracy. Recently, Wilson R. et al.[15] have proposed the adoption of a standardized accuracy benchmark of 85%, to enhance the evaluative criteria for model performance. Here, we argue that this target accuracy may serve not only as a benchmark for model performance but also become a focal point for assessing the statistical power of such model. By introducing a specific target accuracy, researchers can thus establish a measurable standard for assessing the reliability of their models.

Therefore, this paper introduces the concept of statistical power analysis, elucidating how it can be leveraged to ascertain whether the model's accuracy significantly deviates from this benchmark value. This, while scrutinising the intricacies of the train-test split and its consequential impact on the performance and generalizability of the models which directly impacts their reliability. Here, the models' accuracies were obtained through a two-step approach. First, we proceeded to identify the variables (antibodies) with a stronger association with the target variable through the Random Forest algorithm[5]. Then classifiers with an increasing number of the most important variables were constructed, and their performances were obtained via the implementation of a Super Learner [16]. This strategy was implemented using the most described train/test split ratio in the literature (9:1, 8:2, 7:3, 5:5) [8, 10, 11, 17] for data partitioning and illustrated using published data on IgG antibody responses to 2054-antigens obtained for 186 individuals living in Mali obtained using microarrays [18].

9.3 Materials and methods

9.3.1 Data

Here we have analysed published data on 186 Malian individuals (age range: 2-25 years of age) described in detail elsewhere [18]. Briefly, this data comprises information con-

cerning protein microarray-based antibody reactivity against a panel of 2320 P.*falciparum*-specific epitopes of the 3D7 line, representing 1204 unique proteins (≈ 23% of the P. falciparum proteome) collected before the start of the transmission season. The response variable is the incidence of clinical malaria, defined as axillary temperature >37.5°C plus a parasitaemia >5000/μl, over an 8-month follow-up period [18].

9.3.2 Data Partitioning / Split Ratio:

Before conducting any analyses, the original dataset was divided according to the 9:1, 8:2, 7:3 and 5:5 split ratios while maintaining the proportions of Protected and Susceptible patients in each split. These consisted of using 90% of the data as training data and 10% as test data (9:1), 80% training and 20% test (8:2), 70% training and 30% test (7:3) and 50% train and 50% test (5:5). The respective datasets contained 167, 149, 131 and 94 observations for the train set and 19, 37, 55 and 92 observations for the test set.

9.3.3 Predictive analysis

A detailed description of the pipeline here implemented can be found in [19]. Briefly, 2500 runs of the Random Forest were performed changing the following hyperparameters: number of trees, number of features to possibly split in each node and minimal node size after each run. In each run, the mean decrease in the Gini index was used for determining feature importance, which was averaged across all runs, and finally sorted [20]. A classifier with the topmost important features was created, and their predictive performance was accessed using a Super Learner. A construct with the two topmost important feature was initially included in the classifier, and one at a time, features were added, each time performing the Spearman test to remove highly correlated features (Spearman correlation coefficient >|0.8|). Finally, the construct with the least number of features while simultaneously reaching a predictive performance of at least 85% in both the train and test sets was considered the optimal solution.

Finally, the performance of each construct was obtained using the accuracy measure with the probability cut-off estimated for the cut-point that minimized the distance between the Receiving Operator Characteristic (ROC) plot and the point where sensitivity equals 1 and 1-specificity 0 (perfect classification) [21]

9.3.4 Power Analysis

The statistical power of a test is the probability of rejecting a null hypothesis when it is not true [22, 23]. This can be translated as the probability that a significance test will yield statistically significant results, or in other words, to detects a truly existing effect. The power is given by 1 – β, where β, the probability of committing a type II error, meaning

accepting the null hypothesis when it is false [22, 23, 24]. The power is determined as a function of three parameters: significance criterion (α), the sample size (n) and the effect size (ES) [22, 23, 24]. The α value, also known as the type I error, indicates the probability of rejecting the null hypothesis when the null hypothesis is in fact true. Although unofficial, an $\alpha = 0.05$ has become a convention for a minimum basis for rejecting the null hypothesis in most scientific areas. Therefore, here we have used the same for conducting our power analysis. The sample size refers to the number of samples (individuals) in the data. In our case, this corresponds to the number of samples in either the train or test subsets after data partitioning. Finally, the ES, the degree to which the effect under study is manifested, must be determined [25]. The ES can either express the difference between two population parameters or the departure of a population parameter from a constant. In either way, ES can be treated as a parameter which takes the value zero when the null hypothesis is true and some other specific nonzero value when the null hypothesis is false [25]. Thus, the ES may serve as an index of the degree of departure from the null hypothesis. Here the population parameter we sought to determine was the predictive performance (accuracy), which in itself is a proportion that ranges from 0 to 1 [25]. Thus, ES for comparing proportions with different sample sizes was calculated. In such case, if proportion 1 (P_1) and proportion 2 (P_2) represent two accuracies, the effect size (h) is represented by the difference between the arcsine root, or the angular transformation of each proportion:

$$h = 2arcsine(\sqrt{P_1}) - 2arcsine(\sqrt{P_2})[26]$$

Determination of the parameters α, n and ES allow us to elucidate the power. The null hypothesis (H_0) of the power test states that two proportions are no different[27]. The alternative hypothesis (H_1), on the other hand, states that there is a difference between the two proportions. A one-tailed test was implemented were we sought evidence for P_1 being greater or equal to P_2 and a minimum power of 0.8 was establish for evidence of an effect [28, 29]

9.3.4.1 Statistical Software

The statistical analyses were performed in the R software [30] version 4.3.0 with core functions and the following packages: Caret for multicollinearity removal [31], "caTools" for data splitting [32], "OptimalCutpoints" to obtain the optimal cut-points for each predictive model [21], "pwr" to conduct the statistical test of proportions [33], "ranger" to perform the Random-Forest [34] and "SuperLearner" for predictive analysis [16]. The same starting parameter (seed) was used for all analysis conducted to assure findings were biological and not computationally derived. Our analyses were last conducted on the 10th August 2023. All the full reproducible codes are freely available from AF upon request.

9.4 Results

9.4.1 Data partition's effect on variable selection

As a first step in our analysis, we aimed at understanding how different data splitting proportions could impact variable selection. For this, we obtained the average antibody importance values provided by the Gini index given when performing the Random Forest. The average antibody importance distribution for each split is presented in Figure 1A as a density plot. According to our results, the overall antibody importance decreased with the split ratio, with the 9:1, 8:2, 7:3 and 5:5 splits reaching an average antibody importance of 0.031, 0.027, 0.024 and 0.017 respectively (Figure 2A).

Regarding the 10 topmost important antibodies, several were consistently found across the distinct proportions. Indeed, 5 antibodies, against the *MAL6P1.252-1*, *PFC0210c*, *PF10_022_4*, *PF08_010* and *PF10_036.2* antigens, could be found across all four split ratios. From these, the antibody against *MAL6P1.252-1* was the most relevant since it appeared as the most important overall in all scenarios. Meanwhile, the antibody importance against *PFC0 210c*, seemed to decrease as the proportion of data used for the training set decreased. Contrarily, the antibody importance against the *PF10_036.2* antigen increased with a decrease in the split ratio. Lastly, the antibodies against *PF10_022_4* and *PF08_010*, however, did not depict any clear pattern across the different splits. Two antibodies, against *PF14_067* and *PF07_0128* were also found across all splitting procedures except for the one where the train set corresponded to 50% of all the data.

As for the remaining antibodies, these were scattered between the different split ratio, with some of the antibodies within the highest splits not being found in the lower ones, being replaced by other (Figure 1B and C). This is especially noticeable when comparing the topmost antibodies between the 9:1 and 5:5 partitions, where 5 (half) of the antibodies found in the first, against the *PF14_067*, *PFI1520w*, *MAL13P1.133*, *PF07_0128* and *MAL7P1.12.1* antigens, couldn't be found in the latter. Instead, these antibodies were replaced by the antibodies against *PF08_008.4*, *PF11_023*, *PF08_104.3*, *PF10_107.4* and *PF08_1 012.2*, with the last three being exclusively found in this data partitioning (Figure 1A and B). Meanwhile, comparing the 9:1 with the 8:2 and 7:3 splits revealed 2 antibodies that differed between the first and the last two. (Figure 1B). Similarly, the 8:2 and 7:3 partitions also shared 8 antibodies, Finally, a comparison between the 8:2 and 7:3 split ratios against the 5:5 revealed 6 and 7 antibodies shared between them, respectively (Figure 1B and C).

Overall, these results suggest that opting for different proportions of data to perform data partition has the ability to affect variable selection, with more dissimilar proportions showing a greater difference between the variables selected.

9.4.2 Data splitting's effect on predictive accuracy

Then we proceed to assess how the splitting proportion impacted the predictive accuracy of our model. For this, we obtained the accuracy of the train and test set's classifiers with an increasing number of antibodies for each data partition (Figure 2A). Due to the different antibodies elected, one would expect the resulting classifiers' accuracy and antibody composition to vary in respect to the partition used.

According to our results, only classifiers in the 9:1 and 8:2 split ratios were able to reach the target 0.85 in both the train and test subset (Figure 2A). Using the 9:1 partition, the target value was readily attained by a 2-antibody classifier in the train set (Accuracy = 0.852) (Figure 2A). This result is explained by the almost perfect accuracy reached by the RF included in the Super Learner (Figure 2B). In the test subset, however, the target accuracy was only reached by a 35-antibody classifier (accuracy = 0.895) which provided a sensitivity and specificity of 0.917 and 0.860 respectively (Figure 2C and Table I). The full list of antibodies that compose this classifier can be found in Figure 2D. It is worth noting that from the 10 topmost important antibodies using this split ratio, 9 composed this classifier. Concerning the 8:2 split ratio, 4 antibodies were necessary to reach the 0.85 accuracy in the train set accuracy, whilst 75 antibodies were required to reach the target accuracy in the test set (Figure 2A). This 75-antibody classifier 's sensitivity and specificity were 1 and 0.6154 respectively, providing an overall accuracy of 0.87 (Table I). Interestingly, of the 35 antibodies that made up the 9:2 classifier 28 (80%) also composed this 75-antibody classifier.

Among the antibodies found in these classifiers, it is worth highlighting the ones against PF07_0128, PFB0305c and PFC0210c antigens, located in the Erythrocyte Binding Antigen-175 (EBA-175), the Merozoite Surface Protein 5(MSP5) and Circumsporozoite proteins respectively, against which, the reported antibodies have been described to provide protection against clinical malaria [35, 36, 37, 38, 39, 40, 41, 42].

As for the 7:3 and 5:5 splits, once more, the train set's accuracies reached accuracies above 85% and very close to a perfect accuracy with just a few antibodies. Nevertheless, the maximum test set accuracies obtained were 0.8 and 0.78 respectively, with the corresponding classifiers having a total of 26 and 28 antibodies (Figure 2A and 3D). The suboptimal performance of these classifiers is largely related by the model's incapacity to correctly ascertain the susceptible individuals (Figure 2E and 3F) which significantly deteriorated with the decrease of the split ratio.

Finally, our results denote an important aspect concerning the overall predictive ability of the model. As the split ratio decreased, fewer antibodies were necessary for classifiers to reach an almost perfect prediction (Accuracy = 100%) in the train set (Figure 2A). While in 9:1 partition, a perfect accuracy was only reached by classifier with more than 81 antibodies, for the remaining splits only a few antibodies (between 5 to 7) were needed to reach the same accuracy (Figure 2A). Nonetheless, despite the model's extreme accurate predictions

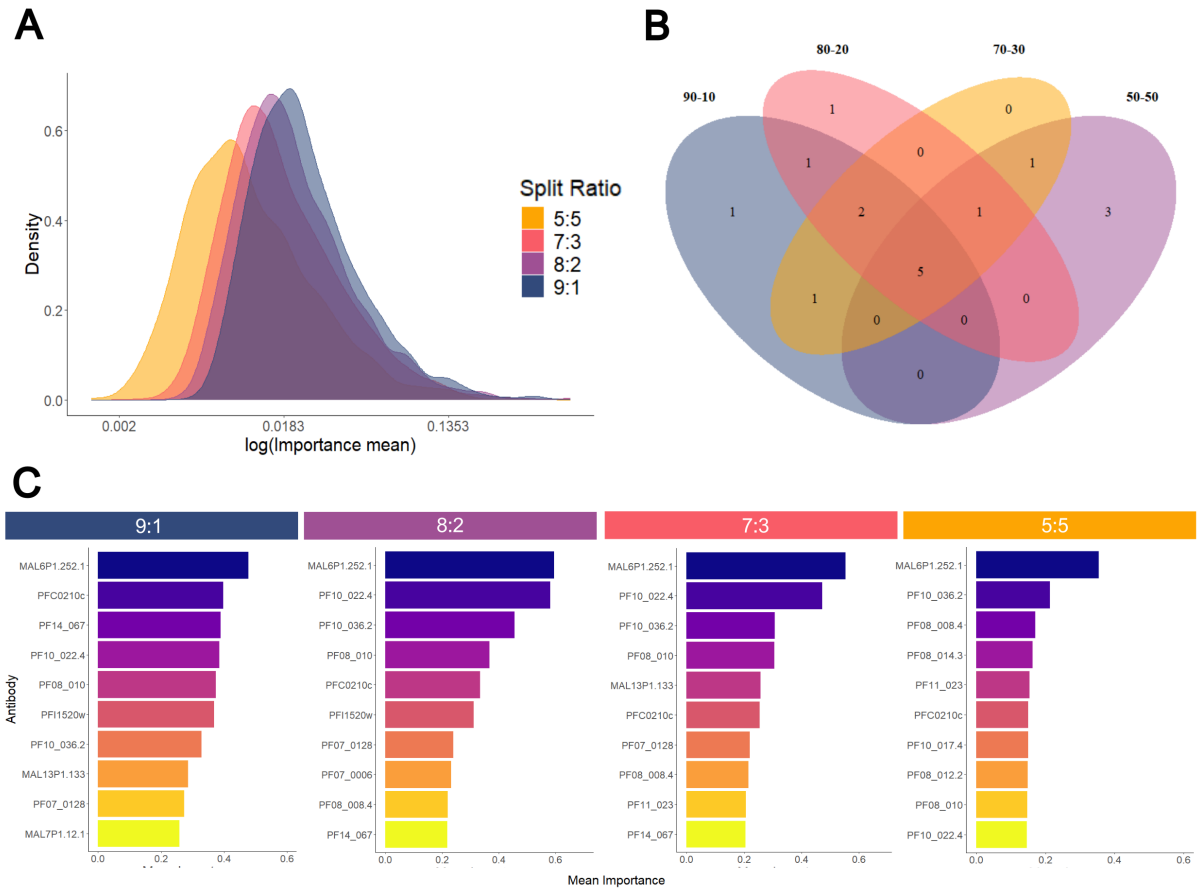


Figure 35: **Split ratio impact on feature selection.** A) Density plot of the antibody's average importance distribution obtained by the RF for each of the split ratios. Veen Diagram. B) Veen Diagram showing the overlap between the 10 most important antibodies for each split ratio and C) the respective bar plot of such antibodies for each split ratio.

in the train set, these were not translated to the test set, where the predictive abilities of the classifiers diminished with the split ratio (Figure 2A). This strongly suggests that as the number of observation available in the train set decreased the model became more prone to overfit the same, leading to a poorer generalization to the test set.

All around, the different data split ratios significantly impacted the overall performance of the models. As the sample size of the train set decreased, less data was available to train the model resulting in a lower generalization capability, rendered by lower performances on new unseen data (test set).

9.4.3 Data partition's effect on power

We then proceeded to estimate the statistical power of the performances obtained by each model. As previously illustrated by our results, while the train sets' performances got increasingly higher, the test sets got worst. This led to a larger disparity between the train and test sets' performances. Thus, initially, we set to understand whether these differences were significantly different.

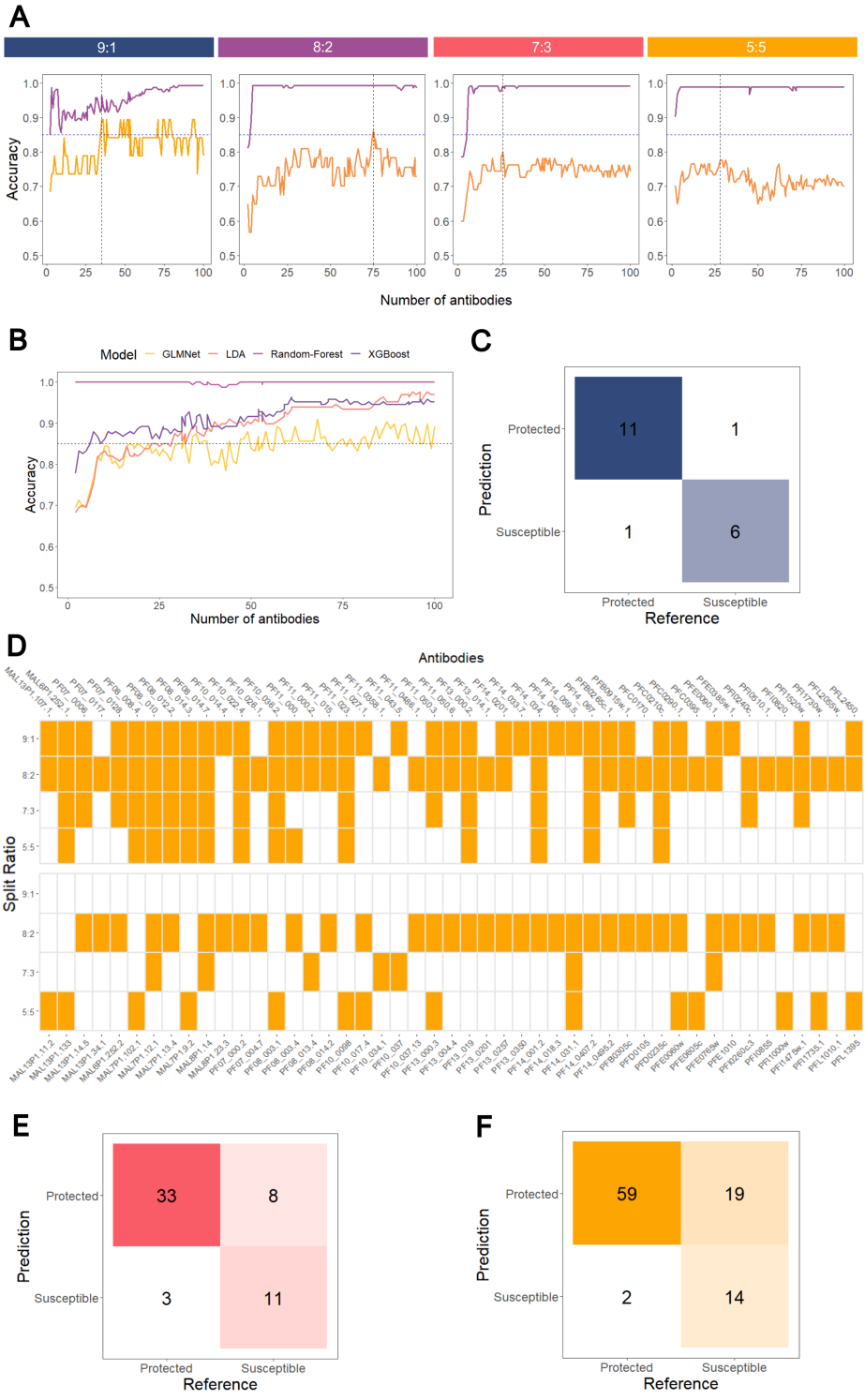


Figure 36: **Split ratio impact on Predictive performance.** A) Accuracy of the SL classifier in the train (purple) and test (orange) datasets as a function of the number of antibodies included for each split ratio. The black horizontal line indicates the 85% (target accuracy), while the vertical line highlights the best classifier. B) Accuracy of the different classifiers assembled by the SL in the train dataset. C) Confusion matrix concerning the best classifier performance on the test set of the 9:1 split. D) Heatmap of the antibodies that compose the best classifier for each split ratio. Orange indicates the presence of such antibody, while white, absence. E) Confusion matrix concerning the optimal classifier performance on the test set of the 7:3 split and F) for the 5:5 split.

In this context, we started by estimating the ES parameter for each splitting ratio as described in the Power Analysis section under the *Materials and Methods*, where P_1 was the train's set accuracy and P_2, the test's set accuracy. From the highest to the lowest split ratios, the respective effect sizes were 0.278, 0.585, 0.748 and 0.773. Once the ES was estimated, the power between the train and test's accuracies was calculated. The power for the 9:1, 8:2, 7:3 and 5:5 splits was 0.31, 0.94, 0.999 and 0.999, respectively. These results indicate that there is enough evidence (power >0.8) to claim that the accuracies between the train and test were not equal for the least three split ratios. However, for the 9:1 split, there was not enough power to claim that the train and test accuracies were dissimilar.

Then, we proceeded to determine whether the train set accuracies for the two lowest splits, in which the train set accuracy didn't reach the target 85%, were significantly different from this value. Thus, the power was once again calculated, although this time comparing the target accuracy of 85 (P_1) to the test sets' accuracies (P_2). The power was estimate at 0.170 and 0.356 for the 7:3 and 5:5 split ratios respectively, indicating the test set's accuracies in both cases didn't significantly depart from the target 0.85. Put it differently, there was not enough power to claim that both the obtained test accuracies differed from 0.85 target. This, however, raised the question of how much should the predicted accuracy be in order for us to identify a significative difference. Our results suggest that the test set accuracy would need to drop all the way down to 70 for us to have enough statistical power to declare a difference between the test set and the target accuracy (Figure 3A). Notwithstanding, this example refers to the split ratio for which the number of observations in the train set is the largest. If we were to make the 80% claim with the different split ratios that consider a small proportion of the test set, it would be necessary for the test set values to reach down to 0.65, 60 and 0.45 (worst than random guessing) for the 7:3, 8:2 and 9:1 split, respectively to claim a difference towards the 0.85 target. These results suggest that as the sample size gets especially low, the reliability of our results become seriously compromised, highlighting the low statistical power of our analysis.

The best way to increase the statistical power of an analysis is by increasing the sample size. Thus, we decided to obtain a statistical difference with a reasonable disparity between the predicted and the 85% target accuracy. To ensure a high reliability of our results we decided to calculate how large would be necessary for the test set here we arbitrarily proposed a 5% difference between the target value and the accuracy. In this sense, we decide to calculate the minimum required sample size to obtain significant statistical power in our analysis when the obtained accuracy equaled 80%. According to our results, a total of 710 observations would be necessary to admit this with 95% certainty. This value however refers to one of the subsets, namely the test set, and thus depending on the split ratio used to achieve such accuracy, the total number of observations is a dataset can range from 1420 up to 7100. The required number of observations to obtain enough statistical power (power = 80%) to claim a difference between the target accuracy of 0.85 and any other ac-

curacy value is given in (Figure 3B).

Finally, given that the accuracies of the train set always surpassed the target 85% accuracy, there was no applicability in calculating the power of the same in regard to the 85% accuracy. Nonetheless, we calculated the uncertainty of such values by assessing their confidence interval. The train set accuracy values for the 9:1, 8:2, 7:3 and 5:5 was 0.96 (95% CI = [0.92, 0.99]), 0.99 (95% CI = [0.96, 1.00]), 0.99 (95% CI = [0.96, 1.00]) and 0.99 (95% CI = [0.94, 1.00]), respectively.

9.5 Discussion

Machine Learning techniques are increasingly being used in biomarker discovery [43, 44, 45, 46, 47, 48, 49]. In the field of malaria, these tools have been widely implemented with the aim of identifying antibodies that confer natural immunity against clinical malaria [1, 2, 3, 4, 14, 50, 51]. However, despite their tremendous performance, to this date, the exact identity of the antibodies such antibodies remain largely elusive. Here we defend that this may be partly attributed to the use of different proportions to split the data into train and testing subset, or even the absence of such procedure when implementing such techniques which may significantly impact the robustness and generalizability of such models [10]. Data partitioning into a train and test subset is an indispensable approach when implementing ML-based models, however depending on the proportion of the data (split ratio) used to create these subsets models may yield differing results [10, 52, 53]. The delicate balance in the allocation of data for training and testing however is very often overlooked in many malaria studies aiming at identifying biomarkers against malaria posing a major problem to the reliability of the machine learning models and the consequent findings reported using such models [54]. In the absence of designated methods to assess the reliability of the results provided by these models, assessing their trustworthiness becomes inherently more challenging. Therefore, here we explored the intricate relationship between the train-test split proportions and the models' performances while assessing the reliability of such models, a quality metric often neglected in machine learning, especially in the malaria field. Based on recent research advocating a specific target accuracy of 85% as a benchmark for model learning [15], here we have provided a pragmatic framework to overcome the challenge of assessing the trustworthiness of machine learning models. By advocating for a specific target accuracy, we were able to estimate the statistical power for proportions being thus able to assess the reliability of the models implemented which may help practitioners to establish more reliable and reproducible machine learning practices ultimately aiding in the identification of the antibodies that confer protection against malaria.

Machine learning refers to algorithms that rely on the application of mathematical and statistical approaches to train a model to learn from data for a particular task [11]. Once

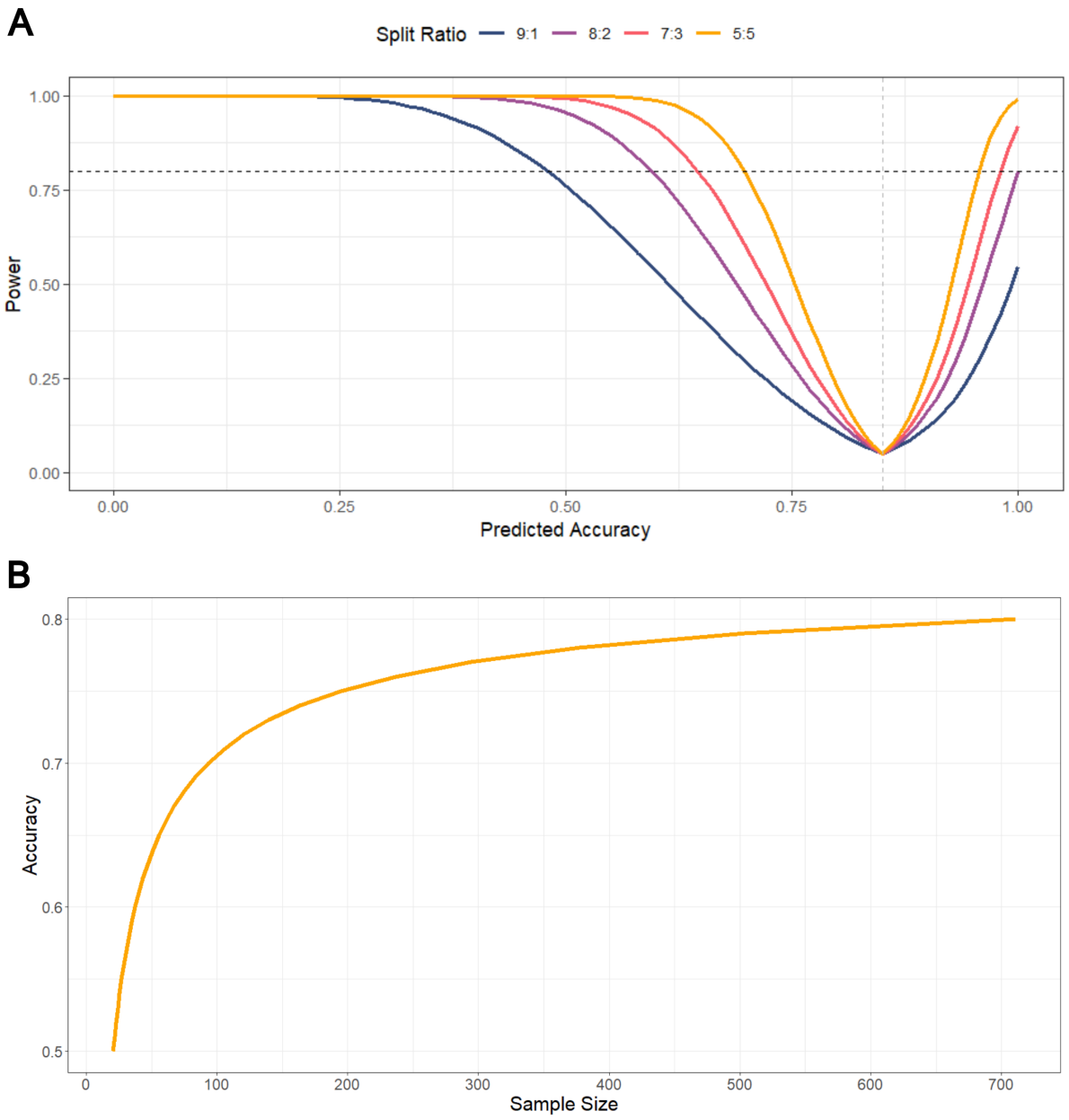


Figure 37: **Power Analysis.** A) Statistical power as function of comparing the predicted accuracy with the target accuracy of 85% for each split ratio. B) Number of samples required to claim enough statistical power (power = 80%) between the predictive accuracy and the target 85%.

the model has been able to learn from the data, it can then be used to make future predictions on new data [10, 11]. This is often done because the aim of implementing such approaches is to correctly predict new observation that will be feed into the model, rather than assertively classifying the observations for which we already know the true outcome [11]. However most often then not practitioners will not have a separate set of data on which they can evaluate the model's performance. To overcome this problem, data splitting into disjoint sets: train ant test sets is often implemented [9]. In such case, the train set is used to train (fit) the model, while the test set, will serve as new, unseen data on which we will assess the model's performance and consequent robustness and generalizability [10]. By, leaving a portion of the data aside that the model hasn't previously seen during training, we can thus understand how it will perform with future data [9]. In this sense, failure to perform data partition raises some key problems. he first concern is the risk of overfitting on the training data, which occurs when a model learns the training data too well, including noise and outliers, consequently rendering these models to perform poorly on new, unseen data [11, 55]. As a result, often, the training performances do not accurately reflect the model's performance on new data. Indeed, it is well known that the performance of the model on the data used to learn the model (training set) is an overly optimistic estimate of the performance on unseen data [9].Finally, in the absence of a separate test set, the model's generalization ability cannot be assessed, and therefore, one can't assess how it will perform on new data. As such, partitioning the data into train/test subsets, is highly advocated to obtain a more realistic view of the model's robustness and its generalizability [54]. Nevertheless, depending on the proportion of data (split ratio) used to compose each subset, models may yield differing results. This issue arises, because ML models are sensitive to the ratios used to divide datasets for training and testing [10, 53].

While the literature is filled with studies exploring how variations in the train-test split can influence the final results and how this poses a challenge to the reliability of machine learning models [10, 52, 53], very little guidance is given on how much data should be used to perform such split [17, 54]. Randomly splitting, using an arbitrary proportion for data splitting is the most used approach [17]. A commonly used ratio is 80:20, which means 80% of the data is for training and 20% for testing. However, other common ratios such as 90:10 70:30, and even 50:50 are also used in practice [17]. Nevertheless, no clear guidance on what ratio is best or optimal for a given dataset seems to exist. While deterministic methods for splitting have also been proposed in the literature to aid in this dilemma, these can only be implemented once we specify a splitting ratio, so the challenge for the optimal split ratio remains [17, 56, 57, 58]. Despite the theoretical and numerical investigations on the optimality of data splitting ratio, so far these have not led to any consensus. In their paper, Picard and Berk [59] have recommend 25%–50% of the data to compose the testing set, whereas Afendras and Markatou [60] recommended 50%. The analysis conducted by Larsen and Goutte [61] and Dubbs [62], on the other hand, show that as the size of the data becomes

large, the proportion of the data used for training should increase. In contrast, extensive numerical studies by Dobbin and Simon [54], Pham et al [63] and Nguyen et al. [52] have indicated that a value of around 30% to be a reasonable choice. Finaly, recently, Roshan [64] have proposed an optimal ratio for data splitting of $\sqrt{p}:1$ where p was the number of parameters (predictors) that explained the data well. However, this last one assumes a prior knowledge on the number of parameters that will be included in the model [64]. Our approach however doesn't allow us to know this *a priori*, as the number of predictors in our model is dependent on the model's performance on the test set, and so determination of this parameter is intrinsically associated with the way the data is partitioned. In this sense an alternative must be sought of.

Although using a data splitting that maximizes the number of samples available for training (such as a 9:1 split ratio) may appear a rational selection, this is not always the case. Given that small sample sizes produce an imprecise estimate of accuracy with a wide confidence interval translating into unreliable results the idea of increasing the available data to train the model in order more trustworthy results is logical [65]. Indeed, according to our results, this split ratio was the one that provided the best test set accuracy, or the best generalization of the model, with the accuracy between the train (0.99) and test set accuracies (0.89) being different by only 10%. Another thing to note is that, among all data partitioning used, the 9:1 split provided the highest accuracy for the train set, further highlighting the need for a large sample for this to be able to perform well. However, there is a problem linked with the increased train set proportion, which is the decrease in the test set proportion. As the test set gets smaller, the higher will be the performance metric's variance on this set. In fact, the confidence interval for test accuracy using the 9:1 ratio can be seen to be quite large, ranging between 0.67 and 0.99 a difference in magnitude of 0.32, which is almost twice as high as the test's set accuracy the range of CI interval of for the 5:5 split: $0.86 - 0.68 = 0.18$ (Table I). As a result, a small change in the test set can result in significant fluctuations in the accuracy, making it challenging to draw robust conclusions about the model's generalization performance. Furthermore, if the test set is not representative of the overall data distribution, the model might be evaluated on a biased sample. This can give a misleading perception of the model's performance, especially if the distribution of the test set differs significantly from the data the model is likely to encounter. On the other hand, as we have mentioned, decreasing the available data to train the model may significantly hinder its ability to generalize, especially in the case of datasets with a small sample size. Looking at our results, one can clearly see this pattern. A decrease in a number of observations, available to train the model 7:3 and 5:5 splits, led to incapacity of to reach the target 85% accuracy on the test set, which can be translated as a decrease in the generalizability of the models. Such a decrease in the performance is highlighted by the Super Learner's inability to correctly ascertain the susceptible group of individuals, the minority group, which as almost half of the observations of the protected group (Figures 2E

and 3F). This is yet another setback of ML-based approaches where an imbalance between the two classes often leads the models to predict better the major class in deterioration of the minor class, leading to biased models [66, 67].

These problems are far from trivial and highlight the pitfalls of implementing ML-based models in datasets with a lower number of samples. While these algorithms thrive when implemented in large datasets with thousands of variables, in smaller datasets with just a few hundred of observations they are not so capable. In the latter, these models are much more prone to either overfit or to produce unreliable results as we have highlighted [68]. As a result, it is important to consider these factors carefully when implementing this technique in such datasets, to avoid compromising the credibility of the results. In this context, here we set to assess the reliability of our results in regard to the split ratio used to perform data partition through the determination of statistical power. Although power analysis should be conducted before starting a study, helping to determine the appropriate sample size to achieve the desired level of power to assuring the robustness of the results obtained, often it can be used to retrospectively analyze the likelihood of a study to detect a significant effect, shedding light on the reliability and validity of the study's findings [22]. The concept of statistical power is only relevant in the context of hypothesis testing since the very definition of power is the probability of rejecting the null hypothesis in favor of an alternative hypothesis when the alternative hypothesis is true [23, 26]. A high power (Power $\geq$ 80%) indicates that there is convincing evidence to reject the null hypothesis [29]. In a context of hypothesis free models, such as the ML models, we propose a methodology to evaluate the reliability of their predictive performances based on the power of such analysis. This approach anchored on recent research advocating a target accuracy of 85% as a benchmark for model evaluation [15]. Aiming for this value of accuracy it is possible to construct a general study design to evaluate the power, where the null hypothesis defends that he accuracy of our model equals the target accuracy and the alternative hypothesis states otherwise. Then by assessing the departure of our models' accuracy to this target accuracy (ES) through a proportion analysis, stipulating the significance level ($\alpha = 0.05$) and the sample size (n) we are inherently able to calculate the statistical power of our model's performances regarding the target 85%. Here, given that only the 7:3 (accuracy = 0.8) and 5:5 (accuracy = 0.78)) splits were unable to achieve a test set accuracy of 85% our analysis was restricted to the same. The power values obtained were 0.170 and 0.356 for the 7:3 and 5:5 split ratios respectively, indicating the test set's accuracies in both cases didn't significantly depart from the target 0.85. This results, however, highlighted the low statistical power of our analysis caused primarily by the low sample size, which was further denoted by calculated the necessary sample size to have a significant departure from the target 0.85, which came out as 0.70. Such disparity from the target accuracy represents an immense difference in accuracy, thus one should stipulate a smaller difference between the target and the obtained value to stipulate a significant difference. Other way these results become

non-representative. One way to this is by increasing the sample size. Larger sample sizes constrict the distribution of the test statistic, leading to a decrease in the standard error of the distribution and a reduction of the acceptance region, which, in turn, increases the power [15]. Furthermore, to obtain more reliable results we stipulated a 5% difference between the target value and the accuracy obtain. In this sense we decide to calculate the required sample size to obtain significant statistical power in our analysis when the obtained accuracy equalled 80%. According to such minimum accuracy value, our results demonstrate that a total of 710 samples for the test set would be required for us to have convincing evidence that model with accuracies lower that that threshold would not be generalized as having 85% with a 95% certainty. Consequently, this denotes that in the best-case scenario, a dataset with total of 1420 samples would be needed to have such power (5:5 split). In the worst-case scenario, 9:1, a total of 7100 samples would be necessary to achieve statistical significance. These numbers however, greatly surpass the common sample size found in traditional malaria studies aiming at identifying antibodies against the disease, highlighting the limited reliability of such approaches [1, 2, 3, 4, 14]. Exceptions to this, however, can be seen in the literature. Two primary examples are the papers by Chotirat et al. [51] and Longley et al.50 [50]. In the first the authors analyzed a total of 4,255 plasma samples from a cross-sectional survey conducted in 2012 on endemic areas in the Kanchanaburi and Ratchaburi to assess the ability of P.*vivax* serological exposure markers to detect residual transmission "hot spots" in Western Thailand. Although such paper doesn't introduce a train/test splitting procedure during the predictive stage, one might be less defensive and accept that with this many observations the models' predictions are highly generalizable [51]. Nonetheless as a note of recommendation, we would propose the implementation of a data partitioning procedure to obtain more reliable and realistic results. Longley on the other hand, while aiming to identify P.*vivax* proteins with the ability to detect recent infections and validated their use in the malaria-endemic regions of Thailand, Brazil and the Solomon Islands, has split the analysis into 2 phases: an antigen detection phase a and a validation phase [50]. For the first phase a relatively low number of samples ($\sim$ 30 samples) were used for each region. Meanwhile, during validation, the number of samples ranged from over 700 to above 900, depending on the region [50]. While these types of examples should prevail, allowing for sturdy results to be obtained, this however it is not the case, and thus one may speculate that most of the results obtained using ML-based models in lower datasets may be questionable. This, in turn, may be another factor that has hampered scientists to identify the leading antigens protecting against clinical malaria.

Linking the statistical power of our results with the selection of the the best or optimal split ratio, one can see that the lower the split ratio used, which leads to poorer generalization and lower test set performances, the higher the power concerning of such performance. In this sense it seems that there is a trade-off between generalizability and the models' test performance. In this sense finding an optimal balance between maximizing

performance and the reliability of the results seems to be, in our opinion, the best way to select the optimal split ratio. Nevertheless, as highlighted by our results in lower sample size settings, regardless of the split ratio used, results will hold vary little value. As a comment, we think that proper sample planning should be conducted to ensure the robustness of the results and we strongly advocate that future studies that wish to implement this ML-based approach should start using significantly larger samples than the ones that have traditionally been used to date [1, 2, 3, 4, 14]. Although this would certainly increase the cost of such studies, it would be a step in the right direction to increase the chance of conducting reliable analysis and reliable analysis ultimately improving the chance of identifying the antibodies that confer protection against clinical malaria.

Finally, concerning the antibodies that comprised the classifiers, here we were able to identify a few antibodies that have been described in the literature to be associated with protection against malaria. The first example is the antibody against the *PFB0305c* antigen located in the MSP5 protein. The MSP proteins are expressed at the parasite surface when the parasite is in the bloodstream, thus, providing promising targets for malaria immunity as they are repeatedly and directly exposed to the host humoral immune system [69]. The ability of antibodies against the *MSP5* protein has been widely described in the literature [40, 41, 42, 70]. Besides the antibodies against *MSP5*, Mederos also reported the protective ability of high antibody levels against the *EBA-175* to protect from clinical malaria [70]. Moreover, *EBA-175* is associated with protection from symptomatic malaria elsewhere [1, 39, 71]. The *EBA-175* is the best-characterized member of the EBL family of proteins, a group of ligands involved in invasion of the parasite into the red blood cells (RBCs) and that constitute important vaccine candidates [38]. Finally, the ability of antibodies against the *CSP* protein to induce protection against symptomatic malaria is well established in the literature, with the only available vaccine against malaria, the RTS'S/ AS01 being constructed using *CSP* [35, 36, 37, 72]These findings agree with the literature suggesting the reliability of our methodology to identify relevant antibodies associated with protection to malaria. Nonetheless, most of the antibodies identified in these classifiers had not been previously described in the literature. This evidence suggests that there are antibodies associated with protection against clinical malaria that have not yet been identified, which has been already previously highlighted elsewhere [19]. Nevertheless, further studies are necessary to validate our findings.

9.6 Conclusion

Overall, this paper offers a comprehensive exploration of the intricate relationship between train-test split proportions and the models' performances, while introducing a practical framework for assessing the reliability of such model through a target accuracy benchmark. Here we delved on the search for the optimal split ratio that could not only maxi-

mize the performance but also the reliability of the results obtained. We then introduced the concept of statistical power to assess the trustworthiness of machine learning models by comparing their performance with the target accuracy of 85%. Our results highlight the limitations of implementing ML-based models in low sample contexts and exploit the need for much higher sample sizes than the ones traditionally used in studies of this kind. Finally, we suggest that the optimal split ratio should not only maximize the performance but also the statistical power of the obtained performances. All around, we hope this pilot study may help researchers to establish more reliable and reproducible machine learning practices paving the way for more transparent and comparable results in the field, ultimately aiding in the identification of the antibodies that confer protection against malaria.

References

[1] John Joseph Valletta and Mario Recker. Identification of immune signatures predictive of clinical protection from malaria. *PLoS computational biology*, 13(10):e1005812, 2017.

[2] Carla Proietti, Lutz Krause, Angela Trieu, Daniel Dodoo, Ben Gyan, Kwadwo A Koram, William O Rogers, Thomas L Richie, Peter D Crompton, Philip L Felgner, et al. Immune signature against plasmodium falciparum antigens predicts clinical immunity in distinct malaria endemic communities. *Molecular & Cellular Proteomics*, 19(1):101–113, 2020.

[3] Danica A Helb, Kevin KA Tetteh, Philip L Felgner, Jeff Skinner, Alan Hubbard, Emmanuel Arinaitwe, Harriet Mayanja-Kizza, Isaac Ssewanyana, Moses R Kamya, James G Beeson, et al. Novel serologic biomarkers provide accurate estimates of recent plasmodium falciparum exposure for individuals and communities. *Proceedings of the National Academy of Sciences*, 112(32):E4438–E4447, 2015.

[4] Ralf Krumkamp, Nicole Sunaina Struck, Eva Lorenz, Marlow Zimmermann, Kennedy Gyau Boahen, Nimako Sarpong, Ellis Owusu-Dabo, Gi Deok Pak, Hyon Jin Jeon, Florian Marks, et al. Classification of invasive bloodstream infections and plasmodium falciparum malaria using autoantibodies as biomarkers. *Scientific reports*, 10(1):21168, 2020.

[5] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[6] Animesh Acharjee, Joseph Larkman, Yuanwei Xu, Victor Roth Cardoso, and Georgios V Gkoutos. A random forest based biomarker discovery and power analysis framework for diagnostics research. *BMC medical genomics*, 13(1):1–14, 2020.

[7] Elliot Mbunge, Richard C Millham, Maureen Nokuthula Sibiya, and Sam Takavarasha. Application of machine learning models to predict malaria using malaria cases and environmental risk factors. In *2022 Conference on Information Communications Technology and Society (ICTAS)*, pages 1–5. IEEE, 2022.

[8] Borislava Vrigazova. The proportion for splitting data into training and test set for the bootstrap in classification problems. *Business Systems Research: International Journal of the Society for Advancing Innovation and Research in Economy*, 12(1):228–242, 2021.

[9] Vikas C Raykar and Amrita Saha. Data split strategiesfor evolving predictive models. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part I 15*, pages 3–19. Springer, 2015.

[10] Vikash Singh, Michael Pencina, Andrew J Einstein, Joanna X Liang, Daniel S Berman, and Piotr Slomka. Impact of train/test sample regimen on performance estimate stability of machine learning in cardiovascular imaging. *Scientific reports*, 11(1):14490, 2021.

[11] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 2013.

[12] Faith HA Osier, Gregory Fegan, Spencer D Polley, Linda Murungi, Federica Verra, Kevin KA Tetteh, Brett Lowe, Tabitha Mwangi, Peter C Bull, Alan W Thomas, et al. Breadth and magnitude of antibody responses to multiple plasmodium falciparum merozoite antigens are associated with protection from clinical malaria. *Infection and immunity*, 76(5):2240–2248, 2008.

[13] Faith H Osier, Margaret J Mackinnon, Cécile Crosnier, Gregory Fegan, Gathoni Kamuyu, Madushi Wanaguru, Edna Ogada, Brian McDade, Julian C Rayner, Gavin J Wright, et al. New antigens for a multicomponent blood-stage malaria vaccine. *Science translational medicine*, 6(247):247ra102–247ra102, 2014.

[14] Camila Tenorio França, Michael T White, Wen-Qiang He, Jessica B Hostetler, Jessica Brewster, Gabriel Frato, Indu Malhotra, Jakub Gruszczyk, Christele Huon, Enmoore Lin, et al. Identification of highly-protective combinations of plasmodium vivax recombinant proteins for vaccine development. *Elife*, 6:e28673, 2017.

[15] Robert C Wilson, Amitai Shenhav, Mark Straccia, and Jonathan D Cohen. The eighty five percent rule for optimal learning. *Nature communications*, 10(1):4646, 2019.

[16] Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

[17] V Roshan Joseph. Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 15(4):531–538, 2022.

[18] Peter D Crompton, Matthew A Kayala, Boubacar Traore, Kassoum Kayentao, Aissata Ongoiba, Greta E Weiss, Douglas M Molina, Chad R Burk, Michael Waisberg, Algis Jasinskas, et al. A prospective analysis of the ab response to plasmodium falciparum before and after a malaria season by protein microarray. *Proceedings of the National Academy of Sciences*, 107(15):6958–6963, 2010.

[19] André Fonseca, Mikolaj Spytek, Przemyslaw Biecek, Clara Cordeiro, and Nuno Sepúlveda. Antibody selection strategies and their impact in the analysis of malaria multi-sera data. *medRxiv*, pages 2022–10, 2022.

[20] Bjoern H Menze, B Michael Kelm, Ralf Masuch, Uwe Himmelreich, Peter Bachert, Wolfgang Petrich, and Fred A Hamprecht. A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC bioinformatics*, 10:1–16, 2009.

[21] Mónica López-Ratón, María Xosé Rodríguez-Álvarez, Carmen Cadarso-Suárez, and Francisco Gude-Sampedro. Optimalcutpoints: an r package for selecting optimal cutpoints in diagnostic tests. *Journal of statistical software*, 61:1–36, 2014.

[22] Scott E Maxwell, Ken Kelley, and Joseph R Rausch. Sample size planning for statistical power and accuracy in parameter estimation. *Annu. Rev. Psychol.*, 59:537–563, 2008.

[23] Alexander Grundler, Martin Dazer, and Thomas Herzig. Statistical power analysis in reliability demonstration testing: the probability of test success. *Applied Sciences*, 12(12):6190, 2022.

[24] Karimollah Hajian-Tilaki. Sample size estimation in diagnostic test studies of biomedical informatics. *Journal of biomedical informatics*, 48:193–204, 2014.

[25] Hae-Young Kim. Statistical notes for clinical researchers: effect size. *Restorative dentistry & endodontics*, 40(4):328–331, 2015.

[26] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

[27] Luzia Gonçalves, M Rosário de Oliveira, Cláudia Pascoal, and Ana Pires. Sample size for estimating a binomial proportion: comparison of different methods. *Journal of Applied Statistics*, 39(11):2453–2473, 2012.

[28] KP Suresh and S Chandrashekara. Sample size estimation and power analysis for clinical research studies. *Journal of human reproductive sciences*, 5(1):7, 2012.

[29] Ceyhan Ceran Serdar, Murat Cihan, Doğan Yücel, and Muhittin A Serdar. Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. *Biochemia medica*, 31(1):27–53, 2021.

[30] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[31] Max Kuhn. Caret: classification and regression training. *Astrophysics Source Code Library*, pages ascl–1505, 2015.

[32] Jarek Tuszynski. catools: Tools: moving window statistics, gif, base64, roc auc, etc. *R package version*, 1(2), 2008.

[33] Stephane Champely, Claus Ekstrom, Peter Dalgaard, Jeffrey Gill, Stephan Weibelzahl, Aditya Anandkumar, Clay Ford, Robert Volcic, Helios De Rosario, and Maintainer Helios De Rosario. Package 'pwr'. *R package version*, 1(2), 2018.

[34] Marvin N Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in c++ and r. *arXiv preprint arXiv:1508.04409*, 2015.

[35] Maria Edilene Martins de Almeida, Maria Gabriella Santos de Vasconcelos, Andréa Monteiro Tarragô, and Luís André Morais Mariúba. Circumsporozoite surface protein-based malaria vaccines: a review. *Revista do Instituto de Medicina Tropical de São Paulo*, 63, 2021.

[36] Cristina Fernández-Arias, Sara Mashoof, Jing Huang, and Moriya Tsuji. Circumsporozoite protein as a potential target for antimalarials. *Expert review of anti-infective therapy*, 13(8):923–926, 2015.

[37] Ilka Wahl and Hedda Wardemann. How to induce protective humoral immunity against plasmodium falciparum circumsporozoite protein. *Journal of Experimental Medicine*, 219(2):e20201313, 2022.

[38] Chris Y Chiu, Michael T White, Julie Healer, Jenny K Thompson, Peter M Siba, Ivo Mueller, Alan F Cowman, and Diana S Hansen. Different regions of plasmodium falciparum erythrocyte-binding antigen 175 induce antibody responses to infection of varied efficacy. *The Journal of Infectious Diseases*, 214(1):96–104, 2016.

[39] Matthew B McCarra, George Ayodo, Peter O Sumba, James W Kazura, Ann M Moormann, David L Narum, and Chandy C John. Antibodies to plasmodium falciparum erythrocyte binding antigen-175 are associated with protection from clinical malaria. *The Pediatric infectious disease journal*, 30(12):1037, 2011.

[40] Ronald Perraut, Charlotte Joos, Cheikh Sokhna, Hannah EJ Polson, Jean-François Trape, Adama Tall, Laurence Marrama, Odile Mercereau-Puijalon, Vincent Richard, and Shirley Longacre. Association of antibody responses to the conserved plasmodium falciparum merozoite surface protein 5 with protection against clinical malaria. *PLoS One*, 9(7):e101737, 2014.

[41] Chittibabu Gottimukkala, Charles Ma, Hans J Netter, Santosh B Noronha, and Ross L Coppel. Immunogenicity of malaria vaccine candidate-plasmodium falciparum merozoite surface protein 5 (pfmsp5) expressed in bacillus subtilis. *APCBEE procedia*, 9:113–119, 2014.

[42] Matthew Wayne Goschnick, Casilda Gabrielle Black, Lukasz Kedzierski, Anthony A Holder, and Ross Leon Coppel. Merozoite surface protein 4/5 provides protection against lethal challenge with a heterologous malaria parasite strain. *Infection and immunity*, 72(10):5840–5849, 2004.

[43] Furkan M Torun, Sebastian Virreira Winter, Sophia Doll, Felix M Riese, Artem Vorobyev, Johannes B Mueller-Reif, Philipp E Geyer, and Maximilian T Strauss. Transparent exploration of machine learning for biomarker discovery from proteomics and omics data. *Journal of Proteome Research*, 22(2):359–367, 2022.

[44] Yuqiao Ji, Zhengjun Lin, Guoqing Li, Xinyu Tian, Yanlin Wu, Jia Wan, Tang Liu, and Min Xu. Identification and validation of novel biomarkers associated with immune infiltration for the diagnosis of osteosarcoma based on machine learning. *Frontiers in Genetics*, 14, 2023.

[45] Gokuldas Vedant Sarvesh Raikar, Amisha Sarvesh Raikar, and Sandesh Narayan Somnache. Advancements in artificial intelligence and machine learning in revolutionising biomarker discovery. *Brazilian Journal of Pharmaceutical Sciences*, 59:e23146, 2023.

[46] Qi Hou, Zhi-Tong Bing, Cheng Hu, Mao-Yin Li, Ke-Hu Yang, Zu Mo, Xiang-Wei Xie, Ji-Lin Liao, Yan Lu, Shigeo Horie, et al. Rankprod combined with genetic algorithm optimized artificial neural network establishes a diagnostic and prognostic prediction model that revealed c1qtnf3 as a biomarker for prostate cancer. *EBioMedicine*, 32:234–244, 2018.

[47] Polina Mamoshina, Marina Volosnikova, Ivan V Ozerov, Evgeny Putin, Ekaterina Skibina, Franco Cortese, and Alex Zhavoronkov. Machine learning on human muscle transcriptomic data for biomarker discovery and tissue-specific drug target identification. *Frontiers in genetics*, 9:242, 2018.

[48] Ying Xie, Wei-Yu Meng, Run-Ze Li, Yu-Wei Wang, Xin Qian, Chang Chan, Zhi-Fang Yu, Xing-Xing Fan, Hu-Dan Pan, Chun Xie, et al. Early lung cancer diagnostic biomarker discovery by machine learning methods. *Translational oncology*, 14(1):100907, 2021.

[49] Xiaokang Zhang, Inge Jonassen, and Anders Goksøyr. Machine learning approaches for biomarker discovery using gene expression data. *Bioinformatics*, 2021.

[50] Rhea J Longley, Michael T White, Eizo Takashima, Jessica Brewster, Masayuki Morita, Matthias Harbers, Thomas Obadia, Leanne J Robinson, Fumie Matsuura, Zoe SJ Liu, et al. Development and validation of serological markers for detecting recent plasmodium vivax infection. *Nature medicine*, 26(5):741–749, 2020.

[51] Sadudee Chotirat, Narimane Nekkab, Chalermpon Kumpitak, Jenni Hietanen, Michael T White, Kirakorn Kiattibutr, Patiwat Sa-Angchai, Jessica Brewster, Kael Schoffer, Eizo Takashima, et al. Application of 23 novel serological markers for identifying recent exposure to plasmodium vivax parasites in an endemic population of western thailand. *Frontiers in Microbiology*, page 1727, 2021.

[52] Quang Hung Nguyen, Hai-Bang Ly, Lanh Si Ho, Nadhir Al-Ansari, Hiep Van Le, Van Quan Tran, Indra Prakash, and Binh Thai Pham. Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Mathematical Problems in Engineering*, 2021:1–15, 2021.

[53] Chenkai Liang. Study on the application of big data in business operations as the core strategy. In *2022 3rd International Conference on Big Data and Social Sciences (ICBDSS 2022)*, pages 958–964. Atlantis Press, 2022.

[54] Kevin K Dobbin and Richard M Simon. Optimally splitting cases for training and testing high dimensional classifiers. *BMC medical genomics*, 4(1):1–8, 2011.

[55] Herbert Pang and Sin-Ho Jung. Sample size considerations of prediction-validation methods in high-dimensional data for survival outcomes. *Genetic epidemiology*, 37(3):276–282, 2013.

[56] Roberto Kawakami Harrop Galvao, Mário César Ugulino Araujo, Gledson Emídio José, Marcio José Coelho Pontes, Edvan Cirino Silva, and Teresa Cristina Bezerra Saldanha. A method for calibration and validation subset partitioning. *Talanta*, 67(4):736–740, 2005.

[57] Ronald D Snee. Validation of regression models: methods and examples. *Technometrics*, 19(4):415–428, 1977.

[58] Ronald W Kennard and Larry A Stone. Computer aided design of experiments. *Technometrics*, 11(1):137–148, 1969.

[59] Richard R Picard and Kenneth N Berk. Data splitting. *The American Statistician*, 44(2):140–147, 1990.

[60] Georgios Afendras and Marianthi Markatou. Optimality of training/test size and resampling effectiveness in cross-validation. *Journal of Statistical Planning and Inference*, 199:286–301, 2019.

[61] Jan Larsen and Cyril Goutte. On optimal data split for generalization estimation and model selection. In *Neural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop (Cat. No. 98TH8468)*, pages 225–234. IEEE, 1999.

[62] Alexander Dubbs. Test set sizing via random matrix theory. *arXiv preprint arXiv:2112.05977*, 2021.

[63] Binh Thai Pham, Indra Prakash, Abolfazl Jaafari, and Dieu Tien Bui. Spatial prediction of rainfall-induced landslides using aggregating one-dependence estimators classifier. *Journal of the Indian Society of Remote Sensing*, 46:1457–1470, 2018.

[64] V Roshan Joseph and Akhil Vakayil. Split: An optimal method for data splitting. *Technometrics*, 64(2):166–176, 2022.

[65] S Jones, S Carley, and M Harrison. An introduction to power and sample size estimation. *Emergency medicine journal: EMJ*, 20(5):453, 2003.

[66] Adeoti Babajide Ebenezer, O Boyinbode, and Oladunjoye Michael Idowu. A comprehensive analysis of handling imbalanced dataset. *International Journal*, 10(2), 2021.

[67] Yanmin Sun, Andrew KC Wong, and Mohamed S Kamel. Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, 23(04):687–719, 2009.

[68] Daniyal Rajput, Wei-Jen Wang, and Chun-Chuan Chen. Evaluation of a decided sample size in machine learning applications. *BMC bioinformatics*, 24(1):48, 2023.

[69] James G Beeson, Damien R Drew, Michelle J Boyle, Gaoqian Feng, Freya JI Fowkes, and Jack S Richards. Merozoite surface proteins in red blood cell invasion, immunity and vaccines against malaria. *FEMS microbiology reviews*, 40(3):343–372, 2016.

[70] Márcia M Medeiros, Wesley L Fotoran, Rosimeire C dalla Martha, Tony H Katsuragawa, Luiz Hildebrando Pereira da Silva, and Gerhard Wunderlich. Natural antibody response to plasmodium falciparum merozoite antigens msp5, msp9 and eba175 is associated to clinical protection in the brazilian amazon. *BMC Infectious Diseases*, 13(1):1–19, 2013.

[71] Raymond B Nyasa, Helen K Kimbi, Denis Zofou, Jeremy D DeBarry, Jessica C Kissinger, and Vincent PK Titanji. An evolutionary approach to identify potentially protective b cell epitopes involved in naturally acquired immunity to malaria and the role of eba-175 in protection amongst denizens of bolifamba, cameroon. *Malaria journal*, 15(1):1–13, 2016.

[72] Kelly E Seaton, Rachel L Spreng, Milite Abraha, Matthew Reichartz, Michelle Rojas, Frederick Feely, Richard HC Huntwork, Sheetij Dutta, Sarah V Mudrak, S Munir Alam, et al. Subclass and avidity of circumsporozoite protein specific antibodies associate with protection status against malaria infection. *npj Vaccines*, 6(1):110, 2021.

Part IV
General Discussion

Chapter 10 - General Discussion

In this thesis, I have developed new pipelines for the analysis of antibody data aiming at identifying biomarkers for outcomes related to Malaria and CFS. Initially, I set to analyze low-dimensional antibody data with the aim of laying the groundwork for such statistical pipelines. At this point, three statistical pipelines were developed for the identification of IgG antibody data against the 36 P. *faclciparum* parasite aimed identifying biomarkers for clinical protection against disease. I then moved to the analysis of high-dimensional data. Driven by the SARS-CoV-2 breakout which left a trail of struggling suffering from with long-COVID, whose symptoms akin to those of ME/CFS, we set to study the latter. In this context two distinct methodologies were implemented of the analysis of a datasets with more than 3,000 antibodies with the goal of identifying antibody biomarkers against the EBV in ME/CFS patients using public data. Finally we set to understand the influence of sample size on the performance and the generalization ability of the ML-based, addressing some corners raised by the implementation of such approaches on limited samples sizes, which is usually the case for high-throughput antibody studies.

Here we will recapitulate and discuss the main results of the thesis. Furthermore we extend the discussion on some topics that weren't fully addressed throughout the thesis content. Some issues will be excluded because they appear to us small details of the overall picture. The interested reader in such detail should read the specific discussions in each chapter.

10.1 Data dichotomizarion: issues and concerns

One of the main drawbacks when dealing with high-dimensional data is the struggle of identifying the relevant features, since most of the screened antibodies will be irrelevant or redundant for the task at hand [1]. To overcome these limitations, feature selection strategies have been proposed [2]. Thus, in chapter 2 and later extended in chapter 3, we introduced the first developed pipeline, in which the feature selection strategy step relied on dichotomization of the antibody data according to the cut-off value that mazimized the separability between the seropositive and the seronegative populations. Altough data dichotomization is a very common practice due to its simplicity, with some of the most advanced ML techniques employng this stragetgy (i.e, Random forest, Extreme Gradient Boosting) [3, 4, 5], great discussion around the negative consequences on dichotizing continuous predictor variables can be found in the literature [6, 7, 8]. Here we will explore what concerns are raised by some of the articles that address this issue. Although most, if not all of the conclusions drawn from these articles were produced upon regression analysis, such findings can be translated to the classification realm. While one of these concerns was already briefly addressed in section 3, we now open the floor to a broader discussion.

One of the main issues recurrently raised in the literature is the loss of information, caused by dichotomizing data. The conversion of a continuous variable into just two categories, in effect, discards information about individual differences as individuals within a subgroup will be treated as if they were identical with respect to the attribute in question [7]. Furthermore if there are threshold effects in the continuous variable (where the relationship with the outcome changes abruptly at a certain point), dichotomizing may not capture this nuance and can result in the loss of valuable information. Oppositely, dichotomization may also create artificial boundaries that do not accurately represent the underlying nature of the variable. This can oversimplify complex relationships and distort the true patterns in the data as brought up by Robert MacCallum and colleagues [4]. Adding to this, data dichotomization entails another problem, in which, data dichotomization reduces the chance of finding statistical significant results by underestimaing the magnitude of bivariate relationships. This is clearly highlighted by Peters and Van Voorhis [9] which showed that under a bivariate normal distribution, the dichotomization of one of the variables at its mean reduces the population correlation coefficient from p to 0.798 of p. In the same study, dichotomizing of both variables at their mean decreased the correlation even further. A similar finding is illustrated by Julie Irwinn and Gary McLelland [6], where dichotomization of a single variable at its mean, within the scope of bivariate analysis, decreased the squared correlation between both variables to approximately 0.63 of the original squared correlation, reducing the relationship between both variables from a significant value (*p-value=0.02*) to a non-significant value(*p-value=0.07*). Similar findings, in which dichotomiziation of one or both variables significantly undermines the magnitude of bivariate correlations can be found in several other studies [4, 10, 11, 12]. As a result of the reduced ability to detect effects or relationships, the statistical power of an analysis is also compromised, since true existing effects may no longer be detected and thus type II errors (failing to find a significant result even when the effect is present) are more likely to occur [6, 13, 14]. However, if in bivariate analysis lower estimates of effect size and lower power are observed, in multivariate cases, data dichotomization may in fact lead to overestimates of strength of relationship accompanied by an increase in Type I errors, that is, to results that are spuriously statistically significant [7]. Vargha and colleagues [14] also raised this issue, showing that after dichotomization, predictor variables appeared significant when they are not in the original data. This findings may partially explain why this approach excelled across all techniques implemented in section 3.

Patrick Royston and colleagues, [8] whilst addressing the issues of dichotomization in their paper, specifically raised concerns regarding the strategy we used to identify the "optimal" cutpoint. The authors state that because of the multiple testing the overall type I error rate will be very high, being around 25–50%, rather than the nominal 5%. Aligned with this statement, our viewpoint on the topic led us to include the Benjamini-Yekutieli test on the improved iteration of the pipelines shown in section 3, aimed at preventing the

propagation of the FDR rate. Besides, the authors further argue that the cut-off (cut-point) chosen will have a wide confidence interval and will not be clinically meaningful [8]. Although this was observed in our data, where the uncertainty varied substantially among the different antibodies, going as far a saying that such optimal values will not be clinically meaningful might be an overstatement. Nonetheless we reckon that these cut-offs should be used with caution. Finally, the authors finish by highlighting that all studies using optimal outpoints derive such point using univariate analysis and then use the resulting binary variable in multivariable analysis which can lead to results being seriously misleading, unless adjusted, advocating for these data-dependent approach to analysis to be avoided. To overcome this issue, however techniques such as the ones proposed by Mazumdar and colleagues [15] can be used, which allow to search the optimal cutpoint for a specic predictor by adjusting in a multivariable model for other predictors known to be important. Although not implemented here, we leave this reference as recommendation for future work.

Despite all of the these evidence refuting the use of dichotomized data, Robert MacCallum and colleagues [4] advocate that in very rare ocasions data dichotomization may be justified. One of such occasion is when where the distribution of a count variable is extremely highly skewed, to the extent that there is a large number of observations at the most extreme score on the distribution. As an example, the authors provide the following illustration "suppose research participants were asked how many cigarettes they smoked per day. A large number of people would give a response of zero, and the remainder of the sample would show a distribution of nonzero values. Such a distribution indicates the presence of two groups of people, smokers and nonsmokers. Corresponding dichotomization of the measured variable would yield a dichotomous indicator of smoking status, which may be useful for subsequent analyses". Contextualizing this example to our analysis, one can establish a parallel, where individuals not exposed to a pathogen, that will show basal antibody levels (noise) are classified as seronegative. On the other hand, exposed individuals with higher levels of antibody due to a genuine antibody response to a given antigen can be classified as seropositive. In such case, once may conceive that distribution entails the presence of two subpopulations, a seronegative and a seropositive and therefore data dichotomization represents a natural way to eliminate the effect of noise during data analysis. One should also highlight that the above cited findings that suggest refraining from data dichotomization, were not drawn on antibody data, which has its own subtleties. Furthermore, the ideia of stratifying patients, into seronegative and seropositive responders according to their antibody levels, is not new. In fact, this idea is the backbone for the mixture models applied in chapter 3, which were introduced in section 4.1.

Overall, we defend the implementation of this type of approach, nonetheless we reckon that caution should be taken, specially when it comes to selecting and reporting the cut-offs used for data dichotomization.

10.2 Predictive analysis: the logic behind using multiple algorithms

Following the selection of the antibodies that demonstrated a meaningful correlation with the outcome variable, we implemented distinct algorithms to conduct our predictive analyzes. The underlying logic for this routine lays on the fact that there is no prior knowledge on what model will perform best for a given dataset [16]. Although experience and expertise might help narrowing down the selection of models to use in each context, until different models are implemented and compared, there is no concrete way of knowning the best performing one. This stems from the fact that each model is sustained on different statistical assumptions as we have detailed in the *Predictive Algortims section* of the Introduction chapter. As such, each model may perform better on specific type of data or under certain circumstances and worst upon others [17, 18]. Several characteristics of the data may contribute to algorithms producing varying results. Here, we will briefly discuss some of them:

Multicolinearity: Multicollinearity is a condition where there is an approximately linear relationship between two or more independent variables [19]. As a result of multicollinearity, several issues arise such as increase estimates of the standard error of the regression coefficients, causing wider confidence intervals [20], imprecise estimates of regression coefficients with wrong signs and an implausible magnitude of some regressors an an effect of these variables being mixed together [21]. Furthermore, the marginal impact of a variable is hard to measure eroding model interpretability [20]. Due to this, models will have poor generalization ability and overfit the data, performing poorly on new data it has never seen [19]. Evidently, linear models (i.e.,logistic regression), are the ones that suffer most with collinearity among variables. Alternatively, however, sepwise selection and LASSO are some of the models that offer a way to circumvent these issues[20].

Noisy data: An important characteristic of real-world problems is that the data frequently contains noise [22]. That is, the quality of the data may be decreased by errors, or deviations produced in the data collection phase, as a consequence of either human error in translating information or due to limitations in the tolerance of the measurement equipment [22]. As such noisy data is a general term to describe incomplete, inconsistent, corrupted, wrong or distorted data [23]. In general, noisy data may bias the learning process, increasing the learning time and degrading the performance of learning algorithms [23, 24]. Highly flexible, complex models are generally more susceptible to the negative effects of noise[22]. Because of their flexibility (variance), they will end up fitting the noise present within the train data. As a result, they may overfit, meaning they perform well on the training data but generalize poorly to new, unseen data [22]. However, more bias models (i,e., linear models, may handle better noise in the data.

Outliers: Outliers are defined as data points that significantly deviate from the overall pat-

tern or distribution of a dataset, indicating rare events or potentially anomalies/errors [25, 26]. As such, just like noisy data, outliers may deteriorate the performance of algorithms[27, 28]. Nonetheless, linear models are the most sensitive to outliers compared to more flexible models. In models with higher variance, outliers can disproportionately influence the parameter estimates in linear models, leading to biased results and reduced predictive accuracy. Indeed the impact of outliers on such models is well described [29, 30]. In contrast, more flexible models with higher variance, such as decision trees or random forests, may be able to accommodate outliers, to some extent, due to their capacity to capture complex relationships [31].

Non-linearity: Linear algorithms which assume a linear relation between the predictor variables and the outcome, will tend perform poorly on datasets with complex, non-linear relationships[32]. In such cases using more flexible, complex models will therefore lead to better performances. Nonetheless, in cases where data does display a linear relation, linear models will often outperform more complex models[32].

Class-imbalance: Class imbalance refers to a situation where the number of instances in different classes is significantly uneven [33, 34]. In such cases models might be biased toward the majority class, since they will tend to achieve high accuracy by simply predicting the majority class for most instances. Therefore, imbalanced datasets can lead to poor generalization to the minority class, making it challenging for the model to accurately predict instances belonging to these underrepresented class [33, 35, 36]. As a consequence accuracy, a commonly used evaluation metric, may not be a reliable measure of model performance in imbalanced settings, since a high accuracy can be achieved by predicting the majority class, even if the minority class is misclassified frequently [36, 33]. Therefore, often other evaluation metrics such as precision, recall or F1-score are used instead [37]. Some algorithms are more sensitive to class imbalance than others. For example, decision trees and XGBoost tend to handle imbalance better than logistic regression and random forest [36, 38, 39, 40].

High-dimensionality: High-dimensional refers to an instance where the number of features p, are close or larger to the number of observations n [32, 41]. Data with high dimensionality inherently presents numerous challenges, so much that a termed called the *"course of dimensionality"* was coined to illustrate the hurdle involved in working with this data [42, 43]. The term "course" was not chosen arbitrarily, as there are numerous challenges in handling this type of data. Here now open the floor ot discuss a few of them.

One of the main issues associated with high-dimensionality is te increased chance of overfitting. Given that in most cases, the vast majority of the variables in the data provide little or no explanation of the outcome, data will contain high noise [44, 45]. Therefore, high-dimensionallity leads to overffiting due to the increased risk of fitting the noise in

the data, capturing spurious patterns in the training data rather than the true underlying patterns. This leads models to struggle in generalizing to new, unseen data [44]. Flexible models, such as ensemble methods, that have a high capacity to fit intricate patterns and relationships in the data often suffer the highest impact with increased dimensional, as they will more likely overfit, compared to more strict models [22]. As a result the model's generalization ability decreases, specially of more flexible models, which will lead to poor model, performances. As such, high-dimensional data often leads to a decrease in model accuracy [46].

High-dimensional data will also tend to decrease model interpretability. Given that a large amount of variables will hold no association with the outcome variable (generating noise in the data) models will also struggle in identifying the effective information, this is , the key features that explain the outcome, as they will be submerged within the data. Ultimately, this will result in too many variables being included int he final model causing a decrease in model interpretability [47]. Not only that, but managing such data will increase computational complexity. As the number of features increases, the number of parameters or coefficients to be estimated by the model also increases. This results in a larger parameter space, which requires more computational resources for optimization during the training process, affecting its efficiency and increasing the time required to complete the task [45, 46]. Besides, storing and manipulating high-dimensional data requires also more memory. A larger memory requirement can lead to increased storage costs and slower processing times. Training machine learning models on high-dimensional data typically requires more iterations extending the overall training time. Therefore, techniques such as cross-validation, commonly used for model evaluation, becomes more computationally intensive with high-dimensional data. [48]. Not only that, but some algorithms are inherently sensitive to the dimensionality of the input data. High-dimensional data can lead to slower convergence or increased iteration counts for algorithms, amplifying their computational complexity. In some cases, when high-diemntionality can render model's usage unreasonable. An example of such instance was evidenced by Peduzzi and colleagues [49] whom demonstrated that in cases where the ratio of predictors per event (each class of the outcome) is larger than 10, logistic regression coefficients may become highly bias. Although these rule was later relaxed by Charles and colleagues [50] to a ratio of about 5 to 9 variables per event, this would still indicate that for a total of 20 variables a minimum of 100 to 180 cases per event to obtain reliable coefficients. This constraint inherently led to the exclusion of logistic regression from the models used in our predictive analysis when analyzing high-dimensional data such as described in chapters 6 and 7.

Therefore, usually a number of algorithms are applied to a dataset and the quality of the resulting models is evaluated using an appropriate measure, most commonly classification accuracy and the model that provides the most accurate and reliable predictions

for such datasets is chosen [16]. This strategy brings forth several advantages. The first is increased robustness of the analysis conducted. If a single algorithm is implemented and it performs poorly (underfits the data) either because it isn't complex (flexible) enough to cope with the data at hand or its not suitable for that specific data (as we have just seen above), the implementation of alternative models, which may be more suitable to the data at hand, will help to obtain a more reliable result [32]. On the opposing a hand, comparing different models may also help to identify whether a model has overfitted the data. Furthermore, given that each algorithm may uncover different patterns and relationships within the data, by comparing the results from various algorithms, one can gain a more comprehensive understanding of the underlying data structure [32]. Finally, one can also combine predictions from multiple algorithms through ensemble methods, such as bagging or boosting, which often results in improved performance compared to using a single algorithm. This because, combing predictions from multiple algorithms through ensemble methods can mitigate individual algorithm weaknesses and enhance overall predictive power [51]. This fundamental logic was what motivated us to chance from implementing each model separately to using the Super-Learner, to conduct our predictive analyzes [52] which has demonstrated that an ensemble of the algorithms in the collection can outperform a single algorithm.

10.3 Hybrid approach : a comprehensive analysis of the data

Notwithstanding its predictive performance, the accuracy of the previous developed pipeline could be dependent on the uncertainty around the optimal cut-off for each antibody. As demonstrated by our analysis, this uncertainty varied substantially from one antibody to another. Not only that, but this simplistic approach neglect the distinct data distributions that can emerge within the antibody data, which could compromise the results obtained. As such in chapter 3 we aimed at developing a more tailored approach for the analysis of antibody data.

The same relied on an hybrid parametric/non-parametric pipeline that integrated Box-Cox transformation followed by a t-test, together with the use of finite mixture models and the Mann–Whitney-Wilcoxon test as a last resort. In this approach, feature selection relied on data transformation and dichotomization via mixture modeling, thus accommodating different data patterns providing a more thorough analysis to the data. Parametric testing assumes a normal distribution of the data values, or a "bell-shaped curve", often, called a Gaussian distribution [53]. Checking the normality assumption is key when implementing parametric tests, such as t-tests, that assume such a distribution, which, if not met, may lead to unreliable results [54]. However, when the distribution is not normal (i.e., the distribution is skewed), the distribution is not known, or the sample size is too small (<30) to assume a normal distribution, often one must rely on non-parametric tests

[53, 55]. In general, parametric tests, are more powerful, meaning they have a higher ability to detect significant effects or differences when they truly exist in the population, therefore, often requiring a smaller sample size than non-parametric tests [53]. In this sense, parametric testing is strongly advocated over non-parametric testing in cases where the data is normally distributed. As such, non-parametric tests should only be considered when the distribution is highly skewed and proper transformation cannot change it to normal distribution[55]. For this reason, this second pipeline started with the implementing of the Box-Cox transformation for each antibody, which sough the optimal transformation to the data. This technique applies a deterministic power function to the data using the estimate of the power transformation parameter (λ)[56] approximating the data to a normal distribution. After this transformation, the Shapiro-Wilk normality test was then implemented to assess whether the transformed antibody data followed a normal distribution. Here, we specifically relied on the Shapiro-Wilk test because it has been shown to be a powerful method that performs particularly well [57, 58]. In cases where normality was met, a two sample t-test was implemented, as commonly described in th literature, otherwise we resorted to the use of mixing models [59, 60, 61]. Otherwise, we returned to the use of finite mixture models. Mixture models are a tool often used in seroepidemioligc studies analysis in order to help classifying individuals into either antibody-positive or antibody-negative [62, 63, 64, 65, 66]. Thus we decided to invoke this well-established tool in our serological analysis which allowed us to identify distinct sub populations of individuals concerning their antibody levels within the data. Selection of the best fitting model was done prioritizing the simplicity of the model (lowest AIC) while providing a good fit to the data ($p\text{-}value>0.5$). For antibodies whose data provided evidence of two latent serological populations, we divided the individuals into two latent serological groups using the optimal cut-off by maximization of the χ^2 statistic. In situations where the models were not fit by a mixture of two models but there was evidence for a single latent population we implemented two linear regression models using the antibody values as the response variable. The first model comprised only the intercept (i.e., not including any covariate), while the second model comprised the malaria protection status as the single covariate. We then computed the *p-value* of the Wilks likelihood ratio test to compare the two models. This approach allowed us to determine weather there as an association between antibody values and the outcome of the patient (the same technique was used in section 4). Finally as a last resort for antibodies where there was no evidence of a single or two latent populations the χ^2 test was implemented. Lastly, due to the multiple testing nature of this pipeline all *p-values* obtained were adjusted via the Benjamini-Yekutieli test to ensure a global FDR of 5% and the Super Lerner was used to conduct our predictive analysis.

All around this pipeline resorted to several distinct filter methods, specifically implemented to deal with the data distributions at hand, thus providing a tailored approach for antibody selection. Nonetheless, this comprehensive data approach carries a higher com-

plexity cost than the data dichotomization strategy. Indeed, this approach is expected to increase the computational time dramatically as the number of antibodies under analysis increases. Furthermore, due to its statistical complexity, this strategy may be less appealing to the malaria research community where the availability of qualified staff with statistical skills remains scarce. Thus, a simpler alternative such as data dichotomization seems a more viable solution at the moment, especially, when it comes to analyzing data featuring thousands of antibodies as it was not much simpler to implement but also performed better, contradicting our previous expectations. Although overestimation of the strength of the relationship between variables (as explained in section 10.1 may aid to explain the increase performance of the data dichotomization pipeline over this antibody tailored approach, we doubt that such contribution would be so significant as to lead this pipeline to be the best performing, and thus one must recognize the suitability of such an approach. In fact, both pipelines performed fairly similar, reaching a predictive performance of close to 80%, however at the expense of a considerable number of covariantes.

10.4 Random Forest: a sturdy algortihm

Due to the increased interest in ME/CFS because of its significant overlap with the post-COVID syndrome (long COVID or post-acute sequel of COVID) here I transition efforts to the analysis of ME/CFS data in order to understand if ME/CFS patietns were at increased risk of developing COVID-19. Thus, in section 4 analysis of ME/CFS gene expression and methylation data were conducted. Driven by this primary analysis we migrated efforts to implement the previously developed pipelines to high-dimensional antibody data against EBV antigens in ME/CFS patients. The feature selection strategies used in the previous sections however, relied on univariate methods, which often fail to fully account for intricate interactions within the data an thus may lead to suboptimal results [67, 68]. Not only that but the implementation of statistical tests to so many covariates would tremendously increase the chance of committing type I errors. Furthermore the time complexity to he implementation of such approaches, specially of the hybrid approach would become a burden. Thus, embedded feature selection strategies relying on ML methods, on the other hand, may provide a better option, fully exploiting the conditionality and richness of the data in a reliable time frame [32, 69]. In this sense, in section 6 we provided a feature selection strategy relying on a machine learning approach. Among the several machine learning techniques currently available, here we decide to use the random forest. The reasoning for this selection was-folded. First, Random Forest is a highly robust model that is able to perform remarkably well with very little tuning required [32]. This was proved by our preliminary analysis to the data, where other ML-based approaches such as the elasitc-net and XGBoost were implemented to the data (results not shown). Even though boosting methods may outperform Random Forest on most problems, these will require a more strenuous

tuning procedure which will required longer time periods to conduct the analysis, otherwise their performance may take a toll [32]. Secondly, given that in boosting, each tree is grown subsequently, where each tree in particular takes into account the other trees that have already been grown, it becomes very difficult to parallelize the implementation of the algorithm. In Random Forest however, given that each tree is grown individually, this process can be easily parallelized, being much faster to implement than boosting algorithms [32, 70]. R particularly, has the *ranger* package that renders the parallelized implementation of Random Forests to be highly efficient. These two factors were the main reason why we decided to use Random Forest for our feature selection analysis and may be the underlying reason to explain why in section Section 1.3, a great amount of the cited literature has resorted to the use of the Random Forest to conduct their analysis. Overall this approach was able to provide highly accurate prediction, surpassing the target 85% accuracy in both the train and test sets relying on a 26-antibody signature when implemented on the ME/CFS data.

10.5 Machine Leaning: powerful tools on large datasets

As we have mentioned towards this work, ML-based algorithms tend to perform greatly on high-dimensional data. Nonetheless, such performance may come at the cost of overfitting the data, which may be overlooked, when train and validation/test splitting is neglected [32]. This problem becomes even more worrying in the context of microarray data, where usually the number of samples is relatively small as highlithed by the studies portrayed in [71]. Using a sample size that is too small when developing a prediction model leads to imprecise parameter estimates and increases the risk of overfitting, which can yield inaccurate and unstable predictions leading to poor model performance when evaluated in 'new' individuals from the same population, ultimately limiting the generalization ability of the model [72]. Thus, while on the dataset used to develop the analysis, model may render extremely accurate results, if implemented on a slightly different dataset, the model performance may decay significantly. Prediction models are often developed with no sample size calculation and as a consequence many are too small to provide precise estimates [73]. Indeed reviews have found that inadequate sample sizes are a key contributor to high risk of bias (inaccuracy) in prediction model studies [74, 75]. As an example Wynants et al. found that 67% of studies were at high risk of bias due to inadequate sample sizes [74]. This could be a contributing factor for the inconsistencies found in malaria specifically, where antibodies such as the MSP1 and AMA1 have been described as diagnostic biomarkers in some studies, and not in others [76].

Althought the rule of thumb of 10 outcome events per variable (EPV) have typically been used to guide the calculation and justification of the sample size for developing a prediction model, this rule of thumb has been shown to have no rationale, especially in

prediction model research, as its evidence base is mainly informed by simulation studies that investigate the performance of estimating covariate-outcome relationships [74, 77, 78]. Thus, given that sample size is a crucial design consideration for any research study, in our last work, we sought out to analyze the impact of sample size on the performance and the statistical power of the model estimates. For this, we conducted an analysis using different train-test splits and implemented a Machine Learning model to the data. Finally, we determined the sample size necessary to reach a statistical power of 80% certainty of the obtained performances. This analysis revealed that most studies in the literature, have an insufficient sample size to claim the results published with enough statistical power.

10.6 Antibodies: Unraveling potential new biomarkers against disease

The implementation of our pipelines lead to the identification of distinct antibody panels with protective potential for outcome. Although antibody biomarkers have been proposed for both diseases, the literature for malaria is far more vast. Indeed a vast array of studies have been developed for the sole purpose of identifying biomarkers against disease. Examples of such studies include [76, 79, 80, 81, 82, 83] in which several antibody biomarkers have been proposed. Among the antibodies identified in these studies, several were also found as potential biomarkers in our analyzes such as *pf113*,*eba-175* and the antibodies belonging to th MSP family (*msp2*, *msp3*, *msp4* and *msp7*) [84, 85, 86, 87, 88, 89, 90, 91, 92]. Nonetheless, *msp1* and *ama1*, immune responses commonly associated with protection to clinical malaria and often referred to as potential vaccine candidates, were not found among the signatures that conferred protection against clinical malaria [93, 84, 94, 95]. Similar findings have also been reported elsewhere [76, 93, 96, 97], which highlighting the need for sturdier pipelines that may help to increase the reproducibility. among studies. Prevailing consensus however, seems to appear concerning *msp1*and *ama1*'s potential to serve as serological exposure biomarkers [98, 99, 100, 101, 102]. More interestingly however, in this work we have identified novel potential biomarkers against clinical malaria, were not described in the literature. This was particularly the case for the analysis conducted in Chapter 9, where most of the antibodies identified had not even been categorized/defined. This results suggests that there may be potential biomarkers for clinical malaria that have not yet been studied. One of the reason for this may be attributed to the just recent introduction of high-throughput technology which has just now allowed us to screen the full antibody repertoire against he parasite's antigens.

Concerning ME/CFS, the literature on antibody biomarkers is far more scarce. Indeed until recently,ME/CFS flew under the radar for most of the scientific community becoming more renowned after the SARS- coV-2 pandemic due to its symptomatic similarities with long-COVID. However, a few papers have been able to identify some antibody biomarkers against EBV and HHV [103, 104, 105], indicating a considerable pathway forward for

the identification of diagnostic biomarkers against disease. One consideration, however, that could shorten this path, would be the consensus of the diagnostic criteria to define ME/CFS patients, which would certainly reduce the heterogeneity between patients and conceivably increase the replicability of findings across studies.

10.7 Concluding remarks

With the ever growing availability of high-throughput antibody data, novel methodologies better capable to cope with such data are key for advancing the identification of diagnostic and treatment biomarkers against diseases. Due to their incredible performances, ML's use has spread to the most diverse corner of our daily lives and in he immunological filed the story is no different. However ML-base antibody data analysis are still in an infant stage and thus there is considerable work to be done to fully explore the potential of these methods in the identification of antibody signatures against disease. Nonetheless we believe that, integrated within robust pipelines such as the ones here implemented, can seriously drive such research. Indeed, we are convinced that pipelines such as the ones here suggested may help tp increase the reproducibility of the findings among studies, increasing the chance of unveiling new biomarkers against disease. However, the limited cohort size in high dimensional studies poses a significance hindrance to such analyses. Therefore great efforts by the scientific community should be redirected into creating reliable datasets that would provide those analyzing them with sturdy and accuracte result.

References

[1] Gregory Piatetsky-Shapiro and Pablo Tamayo. Microarray data mining: facing the challenges. *ACM SIGKDD Explorations Newsletter*, 5(2):1–5, 2003.

[2] Verónica Bolón-Canedo, Noelia Sánchez-Marono, Amparo Alonso-Betanzos, José Manuel Benítez, and Francisco Herrera. A review of microarray datasets and applied feature selection methods. *Information sciences*, 282:111–135, 2014.

[3] SL Prince Nelson, Viswanathan Ramakrishnan, Paul J Nietert, Diane L Kamen, Paula S Ramos, and Bethany J Wolf. An evaluation of common methods for dichotomization of continuous variables to discriminate disease status. *Communications in Statistics-Theory and Methods*, 46(21):10823–10834, 2017.

[4] Robert C MacCallum, Shaobo Zhang, Kristopher J Preacher, and Derek D Rucker. On the practice of dichotomization of quantitative variables. *Psychological methods*, 7(1):19, 2002.

[5] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[6] Julie R Irwin and Gary H McClelland. Negative consequences of dichotomizing continuous predictor variables. *Journal of Marketing Research*, 40(3):366–371, 2003.

[7] Scott E Maxwell and Harold D Delaney. Bivariate median splits and spurious statistical significance. *Psychological bulletin*, 113(1):181, 1993.

[8] Patrick Royston, Douglas G Altman, and Willi Sauerbrei. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in medicine*, 25(1):127–141, 2006.

[9] Charles Clinton Peters and Walter Roe Van Voorhis. *Statistical procedures and their mathematical bases*. School of education, the Pennsylvania State college, 1935.

[10] Robert F Tate. The theory of correlation between two continuous variables when one is dichotomized. *Biometrika*, 42(1/2):205–216, 1955.

[11] Venkat Srinivasan and Amiya K Basu. The metric quality of ordered categorical data. *Marketing Science*, 8(3):205–230, 1989.

[12] Lloyd G Humphreys and Allen Fleishman. Pseudo-orthogonal and other analysis of variance designs involving individual-differences variables. *Journal of Educational Psychology*, 1974.

[13] Jacob Cohen. Partialed products are interactions; partialed powers are curve components. *Psychological bulletin*, 85(4):858, 1978.

[14] András Vargha, Tamas Rudas, Harold D Delaney, and Scott E Maxwell. Dichotomization, partial correlation, and conditional independence. *Journal of educational and behavioral statistics*, 21(3):264–282, 1996.

[15] Madhu Mazumdar, Alex Smith, and Jennifer Bacik. Methods for categorizing a prognostic variable in a multivariable setting. *Statistics in medicine*, 22(4):559–571, 2003.

[16] Sunil Gupta, Kamal Saluja, Ankur Goyal, Amit Vajpayee, and Vipin Tiwari. Comparing the performance of machine learning algorithms using estimated accuracy. *Measurement: Sensors*, 24:100432, 2022.

[17] Shahadat Uddin, Arif Khan, Md Ekramul Hossain, and Mohammad Ali Moni. Comparing different supervised machine learning algorithms for disease prediction. *BMC medical informatics and decision making*, 19(1):1–16, 2019.

[18] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168, 2006.

[19] Jireh Yi-Le Chan, Steven Mun Hong Leow, Khean Thye Bea, Wai Khuen Cheng, Seuk Wai Phoong, Zeng-Wei Hong, and Yen-Lin Chen. Mitigating the multicollinearity problem and its machine learning approach: a review. *Mathematics*, 10(8):1283, 2022.

[20] Wonsuk Yoo, Robert Mayberry, Sejong Bae, Karan Singh, Qinghua Peter He, and James W Lillard Jr. A study of effects of multicollinearity in the multivariable analysis. *International journal of applied science and technology*, 4(5):9, 2014.

[21] Aylin Alin. Multicollinearity. *Wiley interdisciplinary reviews: computational statistics*, 2(3):370–374, 2010.

[22] David F Nettleton, Albert Orriols-Puig, and Albert Fornells. A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial intelligence review*, 33:275–306, 2010.

[23] Rashida Hasan and Cheehung Chu. Noise in datasets: What are the impacts on classification performance?[noise in datasets: What are the impacts on classification performance?]. In *Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods*, 2022.

[24] Shivani Gupta and Atul Gupta. Dealing with noise problem in machine learning datasets: A systematic review. *Procedia Computer Science*, 161:466–474, 2019.

[25] Sanae Borrohou, Rachida Fissoune, and Hassan Badir. Data cleaning survey and challenges–improving outlier detection algorithm in machine learning. *Journal of Smart Cities and Society*, 2(3):125–140, 2023.

[26] Denis Cousineau and Sylvain Chartier. Outliers detection and treatment: a review. *International Journal of Psychological Research*, 3(1):58–67, 2010.

[27] Sidi Boubacar Ould Estaghvirou, Joseph O Ogutu, and Hans-Peter Piepho. Influence of outliers on accuracy estimation in genomic prediction in plant breeding. *G3: Genes, Genomes, Genetics*, 4(12):2317–2328, 2014.

[28] Azme Khamis, Zuhaimy Ismail, Khalid Haron, and Ahmad Tarmizi Mohamm. The effects of outliers data on neural network performance. *Journal of Applied Sciences*, 5(8):1394–1398, 2005.

[29] R Dennis Cook and Sanford Weisberg. Residuals and influence in regression, 1982.

[30] David A Belsley, Edwin Kuh, and Roy E Welsch. *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons, 2005.

[31] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[32] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 2013.

[33] Yanmin Sun, Andrew KC Wong, and Mohamed S Kamel. Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, 23(04):687–719, 2009.

[34] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[35] Victoria López, Alberto Fernández, Salvador García, Vasile Palade, and Francisco Herrera. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information sciences*, 250:113–141, 2013.

[36] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications*, 73:220–239, 2017.

[37] Xinjian Guo, Yilong Yin, Cailing Dong, Gongping Yang, and Guangtong Zhou. On the class imbalance problem. In *2008 Fourth international conference on natural computation*, volume 4, pages 192–201. IEEE, 2008.

[38] Bartosz Krawczyk, Michał Woźniak, and Gerald Schaefer. Cost-sensitive decision tree ensembles for effective imbalanced classification. *Applied Soft Computing*, 14:554–562, 2014.

[39] Bartosz Krawczyk, Mikel Galar, Łukasz Jeleń, and Francisco Herrera. Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy. *Applied Soft Computing*, 38:714–726, 2016.

[40] Mikel Galar, Alberto Fernández, Edurne Barrenechea, and Francisco Herrera. Eusboost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. *Pattern recognition*, 46(12):3460–3471, 2013.

[41] Naveen Naidu Narisetty. Bayesian model selection for high-dimensional data. In *Handbook of statistics*, volume 43, pages 207–248. Elsevier, 2020.

[42] Richard Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.

[43] Claude Sammut and Geoffrey I Webb. *Encyclopedia of machine learning and data mining*. Springer Publishing Company, Incorporated, 2017.

[44] Mustafa Abdul Salam, Ahmad Taher Azar, Mustafa Samy Elgendy, and Khaled Mohamed Fouad. The effect of different dimensionality reduction techniques on machine learning overfitting problem. *Int. J. Adv. Comput. Sci. Appl*, 12(4):641–655, 2021.

[45] Naveen Venkat. The curse of dimensionality: Inside out. *Pilani (IN): Birla Institute of Technology and Science, Pilani, Department of Computer Science and Information Systems*, 10, 2018.

[46] Weikuan Jia, Meili Sun, Jian Lian, and Sujuan Hou. Feature dimensionality reduction: a review. *Complex & Intelligent Systems*, 8(3):2663–2693, 2022.

[47] Andreas Backhaus and Udo Seiffert. Classification in high-dimensional spectral data: Accuracy vs. interpretability vs. model size. *Neurocomputing*, 131:15–22, 2014.

[48] William Stephenson and Tamara Broderick. Approximate cross-validation in high dimensions with guarantees. In *International Conference on Artificial Intelligence and Statistics*, pages 2424–2434. PMLR, 2020.

[49] Peter Peduzzi, John Concato, Elizabeth Kemper, Theodore R Holford, and Alvan R Feinstein. A simulation study of the number of events per variable in logistic regression analysis. *Journal of clinical epidemiology*, 49(12):1373–1379, 1996.

[50] Eric Vittinghoff and Charles E McCulloch. Relaxing the rule of ten events per variable in logistic and cox regression. *American journal of epidemiology*, 165(6):710–718, 2007.

[51] Giovanni Seni and John Elder. *Ensemble methods in data mining: improving accuracy through combining predictions*. Morgan & Claypool Publishers, 2010.

[52] Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

[53] Richard Chin and Bruce Y Lee. *Principles and practice of clinical trial medicine*. Elsevier, 2008.

[54] Banda Gerald and Tailoka Frank Patson. Parametric and nonparametric tests: A brief review. *Int J Stat Distrib Appl*, 7:78–82, 2021.

[55] Umesh Wadgave. Parametric test for non-normally distributed continuous data: For and against: Array. *Electronic Physician*, 11(2):7468–7470, 2019.

[56] Özgür Asar, Ozlem Ilk, and Osman Dag. Estimating box-cox power transformation parameter via goodness-of-fit tests. *Communications in Statistics-Simulation and Computation*, 46(1):91–105, 2017.

[57] Henry C Thode. *Testing for normality*, volume 164. CRC press, 2002.

[58] Derya Öztuna, Atilla Halil Elhan, and Ersöz Tüccar. Investigation of four different normality tests in terms of type 1 error rate and power under different distributions. *Turkish Journal of Medical Sciences*, 36(3):171–176, 2006.

[59] Nicholas Pudjihartono, Tayaza Fadason, Andreas W Kempa-Liehr, and Justin M O'Sullivan. A review of feature selection methods for machine learning-based disease risk prediction. *Frontiers in Bioinformatics*, 2:927312, 2022.

[60] Richard J Fox and Matthew W Dimmic. A two-sample bayesian t-test for microarray data. *BMC bioinformatics*, 7:1–11, 2006.

[61] Husam Ali Abdulmohsin, Hala Bahjat AbdulWahab, and Abdul Mohssen Jaber Abdul Hossen. A new hybrid feature selection method using t-test and fitness function. *Computers, Materials & Continua*, 68(3), 2021.

[62] Nuno Sepúlveda, Gillian Stresman, Michael T White, Chris J Drakeley, et al. Current mathematical models for analyzing anti-malarial antibody data with an eye to malaria elimination and eradication. *Journal of immunology research*, 2015, 2015.

[63] Mohd Amirul Fitri A Rahim, Sriwipa Chuangchaiya, Paisit Chanpum, Laun Palawong, Panuwat Kantee, Nor Diyana Dian, Inke Nadia D Lubis, Paul CS Divis, Akira Kaneko, Kevin KA Tetteh, et al. Seroepidemiological surveillance, community perceptions and associated risk factors of malaria exposure among forest-goers in northeastern thailand. *Frontiers in cellular and infection microbiology*, 12:953585, 2022.

[64] Lindsey Wu, Julia Mwesigwa, Muna Affara, Mamadou Bah, Simon Correa, Tom Hall, Susheel K Singh, James G Beeson, Kevin KA Tetteh, Immo Kleinschmidt, et al. Seroepidemiological evaluation of malaria transmission in the gambia before and after mass drug administration. *BMC medicine*, 18:1–14, 2020.

[65] Andrew J Vyse, NJ Gay, LM Hesketh, Richard Pebody, P Morgan-Capner, and Elizbeth Miller. Interpreting serological surveys using mixture models: the seroepidemiology of measles, mumps and rubella in england and wales at the beginning of the 21st century. *Epidemiology & Infection*, 134(6):1303–1312, 2006.

[66] Karen Kerkhof, Lydie Canier, Saorin Kim, Somony Heng, Tho Sochantha, Siv Sovannaroth, Inès Vigan-Womas, Marc Coosemans, Vincent Sluydts, Didier Ménard, et al.

Implementation and application of a multiplex assay to detect malaria-specific antibodies: a promising tool for assessing malaria transmission in southeast asian pre-elimination areas. *Malaria Journal*, 14:1–14, 2015.

[67] Zena M Hira and Duncan F Gillies. A review of feature selection and feature extraction methods applied on microarray data. *Advances in bioinformatics*, 2015, 2015.

[68] Younes Bouchlaghem, Yassine Akhiat, and Souad Amjad. Feature selection: a review and comparative study. In *E3S web of conferences*, volume 351, page 01046. EDP Sciences, 2022.

[69] Iqbal H Sarker. Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, 2(3):160, 2021.

[70] Ludmila I Kuncheva. *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, 2014.

[71] Jingeun Kim, Yourim Yoon, Hye-Jin Park, and Yong-Hyuk Kim. Comparative study of classification algorithms for various dna microarray data. *Genes*, 13(3):494, 2022.

[72] Richard D Riley and Gary S Collins. Stability of clinical prediction models developed using statistical or machine learning methods. *Biometrical Journal*, 65(8):2200302, 2023.

[73] Paula Dhiman, Jie Ma, Cathy Qi, Garrett Bullock, Jamie C Sergeant, Richard D Riley, and Gary S Collins. Sample size requirements are not being considered in studies developing prediction models for binary outcomes: a systematic review. *BMC Medical Research Methodology*, 23(1):188, 2023.

[74] Laure Wynants, Ben Van Calster, Gary S Collins, Richard D Riley, Georg Heinze, Ewoud Schuit, Elena Albu, Banafsheh Arshi, Vanesa Bellou, Marc MJ Bonten, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *bmj*, 369, 2020.

[75] Constanza L Andaur Navarro, Johanna AA Damen, Toshihiko Takada, Steven WJ Nijman, Paula Dhiman, Jie Ma, Gary S Collins, Ram Bajpai, Richard D Riley, Karel GM Moons, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *bmj*, 375, 2021.

[76] John Joseph Valletta and Mario Recker. Identification of immune signatures predictive of clinical protection from malaria. *PLoS computational biology*, 13(10):e1005812, 2017.

[77] Constanza L Andaur Navarro, Johanna AA Damen, Maarten van Smeden, Toshihiko Takada, Steven WJ Nijman, Paula Dhiman, Jie Ma, Gary S Collins, Ram Bajpai, Richard D Riley, et al. Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. *Journal of clinical epidemiology*, 154:8–22, 2023.

[78] Paula Dhiman, Jie Ma, Constanza L Andaur Navarro, Benjamin Speich, Garrett Bullock, Johanna AA Damen, Lotty Hooft, Shona Kirtley, Richard D Riley, Ben Van Calster, et al. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. *BMC medical research methodology*, 22(1):101, 2022.

[79] Danica A Helb, Kevin KA Tetteh, Philip L Felgner, Jeff Skinner, Alan Hubbard, Emmanuel Arinaitwe, Harriet Mayanja-Kizza, Isaac Ssewanyana, Moses R Kamya, James G Beeson, et al. Novel serologic biomarkers provide accurate estimates of recent plasmodium falciparum exposure for individuals and communities. *Proceedings of the National Academy of Sciences*, 112(32):E4438–E4447, 2015.

[80] Camila Tenorio França, Michael T White, Wen-Qiang He, Jessica B Hostetler, Jessica Brewster, Gabriel Frato, Indu Malhotra, Jakub Gruszczyk, Christele Huon, Enmoore Lin, et al. Identification of highly-protective combinations of plasmodium vivax recombinant proteins for vaccine development. *Elife*, 6:e28673, 2017.

[81] Faith HA Osier, Gregory Fegan, Spencer D Polley, Linda Murungi, Federica Verra, Kevin KA Tetteh, Brett Lowe, Tabitha Mwangi, Peter C Bull, Alan W Thomas, et al. Breadth and magnitude of antibody responses to multiple plasmodium falciparum merozoite antigens are associated with protection from clinical malaria. *Infection and immunity*, 76(5):2240–2248, 2008.

[82] Carla Proietti, Lutz Krause, Angela Trieu, Daniel Dodoo, Ben Gyan, Kwadwo A Koram, William O Rogers, Thomas L Richie, Peter D Crompton, Philip L Felgner, et al. Immune signature against plasmodium falciparum antigens predicts clinical immunity in distinct malaria endemic communities. *Molecular & Cellular Proteomics*, 19(1):101–113, 2020.

[83] Faith H Osier, Margaret J Mackinnon, Cécile Crosnier, Gregory Fegan, Gathoni Kamuyu, Madushi Wanaguru, Edna Ogada, Brian McDade, Julian C Rayner, Gavin J Wright, et al. New antigens for a multicomponent blood-stage malaria vaccine. *Science translational medicine*, 6(247):247ra102–247ra102, 2014.

[84] Roméo-Karl Imboumy-Limoukou, Sandrine Lydie Oyegue-Liabagui, Stella Ndidi, Irène Pegha-Moukandja, Charlene Lady Kouna, Francis Galaway, Isabelle Florent,

and Jean Bernard Lekana-Douki. Comparative antibody responses against three antimalarial vaccine candidate antigens from urban and rural exposed individuals in gabon. *European Journal of Microbiology and Immunology*, 6(4):287–297, 2016.

[85] FHA Osier, LM Murungi, G Fegan, J Tuju, KK Tetteh, PC Bull, DJ Conway, and K Marsh. Allele-specific antibodies to plasmodium falciparum merozoite surface protein-2 and protection against clinical malaria. *Parasite immunology*, 32(3):193–201, 2010.

[86] Spencer D Polley, David J Conway, David R Cavanagh, Jana S McBride, Brett S Lowe, Thomas N Williams, Tabitha W Mwangi, and Kevin Marsh. High levels of serum antibodies to merozoite surface protein 2 of plasmodium falciparum are associated with reduced risk of clinical malaria in coastal kenya. *Vaccine*, 24(19):4233–4246, 2006.

[87] Vashti Irani, Paul A Ramsland, Andrew J Guy, Peter M Siba, Ivo Mueller, Jack S Richards, and James G Beeson. Acquisition of functional antibodies that block the binding of erythrocyte-binding antigen 175 and protection against plasmodium falciparum malaria in children. *Clinical Infectious Diseases*, 61(8):1244–1252, 2015.

[88] Matthew B McCarra, George Ayodo, Peter O Sumba, James W Kazura, Ann M Moormann, David L Narum, and Chandy C John. Antibodies to plasmodium falciparum erythrocyte-binding antigen-175 are associated with protection from clinical malaria. *The Pediatric infectious disease journal*, 30(12):1037–1042, 2011.

[89] Ronald Perraut, Marie-Louise Varela, Charlotte Joos, Babacar Diouf, Cheikh Sokhna, Babacar Mbengue, Adama Tall, Cheikh Loucoubar, Aissatou Touré, and Odile Mercereau-Puijalon. Association of antibodies to plasmodium falciparum merozoite surface protein-4 with protection against clinical malaria. *Vaccine*, 35(48):6720–6726, 2017.

[90] Madhusudan Kadekoppala and Anthony A Holder. Merozoite surface proteins of the malaria parasite: the msp1 complex and the msp7 family. *International journal for parasitology*, 40(10):1155–1161, 2010.

[91] Kerrianne Mello, Thomas M Daly, Carole A Long, James M Burns, and Lawrence W Bergman. Members of the merozoite surface protein 7 family with similar expression patterns differ in ability to protect against plasmodium yoelii malaria. *Infection and immunity*, 72(2):1010–1018, 2004.

[92] Sodiomon B Sirima, Simon Cousens, and Pierre Druilhe. Protection against malaria by msp3 candidate vaccine. *New England Journal of Medicine*, 365(11):1062–1064, 2011.

[93] Richard Thomson-Luque, Thomas C Stabler, Kristin Fürle, Joana C Silva, and Claudia Daubenberger. Plasmodium falciparum merozoite surface protein 1 as asexual blood stage malaria vaccine candidate. *Expert Review of Vaccines*, 23(1):160–173, 2024.

[94] Elissa M Malkin, David J Diemert, Julie H McArthur, John R Perreault, Aaron P Miles, Birgitte K Giersing, Gregory E Mullen, Andrew Orcutt, Olga Muratova, May Awkal, et al. Phase 1 clinical trial of apical membrane antigen 1: an asexual blood-stage vaccine for plasmodium falciparum malaria. *Infection and immunity*, 73(6):3677–3685, 2005.

[95] AA Holder, Guevara Patiño JA, C Uthaipibull, SE Syed, IT Ling, T Scott-Finnigan, and MJ Blackman. Merozoite surface protein 1, immune evasion, and vaccines against asexual blood stage malaria. *Parassitologia*, 41(1-3):409–414, 1999.

[96] Palak N Patel, Thayne H Dickey, Ababacar Diouf, Nichole D Salinas, Holly McAleese, Tarik Ouahes, Carole A Long, Kazutoyo Miura, Lynn E Lambert, and Niraj H Tolia. Structure-based design of a strain transcending ama1-ron2l malaria vaccine. *Nature Communications*, 14(1):5345, 2023.

[97] Danny W Wilson, Freya JI Fowkes, Paul R Gilson, Salenna R Elliott, Livingstone Tavul, Pascal Michon, Elija Dabod, Peter M Siba, Ivo Mueller, Brendan S Crabb, et al. Quantifying the importance of msp1-19 as a target of growth-inhibitory and protective antibodies against plasmodium falciparum in humans. *PloS one*, 6(11):e27705, 2011.

[98] Lou S Herman, Kimberly Fornace, Jody Phelan, Matthew J Grigg, Nicholas M Anstey, Timothy William, Robert W Moon, Michael J Blackman, Chris J Drakeley, and Kevin KA Tetteh. Identification and validation of a novel panel of plasmodium knowlesi biomarkers of serological exposure. *PLoS neglected tropical diseases*, 12(6):e0006457, 2018.

[99] Nguyen Ngoc San, Nguyen Xuan Kien, Nguyen Duc Manh, Nguyen Van Thanh, Marina Chavchich, Nguyen Thi Huong Binh, Tran Khanh Long, Kimberly A Edgel, Eduard Rovira-Vallbona, Michael D Edstein, et al. Cross-sectional study of asymptomatic malaria and seroepidemiological surveillance of seven districts in gia lai province, vietnam. *Malaria Journal*, 21(1):40, 2022.

[100] Zulkarnain Md Idris, Chim W Chan, James Kongere, Tom Hall, John Logedi, Jesse Gitaka, Chris Drakeley, and Akira Kaneko. Naturally acquired antibody response to plasmodium falciparum describes heterogeneity in transmission on islands in lake victoria. *Scientific reports*, 7(1):9123, 2017.

[101] Phubeth Ya-Umphan, Dominique Cerqueira, Gilles Cottrell, Daniel M Parker, Freya JI Fowkes, Francois Nosten, and Vincent Corbel. Anopheles salivary biomarker as

a proxy for estimating plasmodium falciparum malaria exposure on the thailand–myanmar border. *The American journal of tropical medicine and hygiene*, 99(2):350, 2018.

[102] Pedro M Folegatti, André M Siqueira, Wuelton M Monteiro, Marcus Vinícius G Lacerda, Chris J Drakeley, and Érika M Braga. A systematic review on malaria seroepidemiology studies in the brazilian amazon: insights into immunological markers for exposure and protection. *Malaria journal*, 16:1–15, 2017.

[103] Sabine Gravelsina, Anda Vilmane, Simons Svirskis, Santa Rasa-Dzelzkaleja, Zaiga Nora-Krukle, Katrine Vecvagare, Angelika Krumina, Iana Leineman, Yehuda Shoenfeld, and Modra Murovska. Biomarkers in the diagnostic algorithm of myalgic encephalomyelitis/chronic fatigue syndrome. *Frontiers in Immunology*, 13:928945, 2022.

[104] Manuel Ruiz-Pablos, Bruno Paiva, Rosario Montero-Mateo, Nicolas Garcia, and Aintzane Zabaleta. Epstein-barr virus and the origin of myalgic encephalomyelitis or chronic fatigue syndrome. *Frontiers in Immunology*, 12:656797, 2021.

[105] Brandon S Cox, Khaled Alharshawi, Irene Mena-Palomo, William P Lafuse, and Maria Eugenia Ariza. Ebv/hhv-6a dutpases contribute to myalgic encephalomyelitis/chronic fatigue syndrome pathophysiology by enhancing tfh cell differentiation and extrafollicular activities. *JCI insight*, 7(11), 2022.