

NOTTINGHAM   
TRENT UNIVERSITY

Elucidating new molecular drivers and pathways  
involved in Alzheimer's disease using systems  
biology approaches

Dimitrios Zafeiris

A thesis submitted in partial fulfilment of the requirements of  
Nottingham Trent University for the degree of Doctor of Philosophy

January 2019

## Copyright Statement

This work is the intellectual property of the author and may be owned by the research sponsor(s) and Nottingham Trent University. You may copy up to 5% of this work for private study, or personal, non-commercial research. Any re-use of the information contained within this document should be fully referenced, quoting the author, title, university, degree level and pagination. Queries or requests for any other use, or if a more substantial copy is required, should be directed in the owners(s) of the Intellectual Property Rights.

## Acknowledgements

First and foremost, I would like to thank Professor Graham R. Ball for giving me this excellent opportunity to study such a fascinating subject. Without his tremendous support during brainstorming, learning new techniques and maintaining a healthy state of mind none of this could have happened.

I would also like to thank Dr Alan Hargreaves for his assistance in the more biology heavy aspects of the project.

A great thank you goes to Devika Agarwal for sharing with me her knowledge and expertise in our common field and encouraging me to expand and learn more than I would have otherwise, and of course to the staff and colleagues in the John van Geest cancer research centre.

A personal thanks is owed to my friends who, knowing very little about my work, had the infinite patience to try to understand and encouraged me to go ever onward and especially to Richard Scott for entertaining my rambling at random hours of the day. They shared and lightened the load, allowing me to truly enjoy these years.

Of course, I owe a debt of gratitude to my family for supporting me in this endeavour which started from entering college to the completion of this PhD through everything.

And finally, my eternal thanks to my mother, Chrysa, for always encouraging me to achieve greater heights regardless of the cost to herself. Thank you.

## Abbreviations

A $\beta$	Amyloid beta
AD	Alzheimer's Disease
ADAD	Autosomal Dominant Alzheimer's Disease
ADNI	Alzheimer's Disease Neuroimaging Initiative
ANN	Artificial neural Network
ANNI	Artificial Neural Network Inference
APOE	Apolipoprotein E
APP	Amyloid Precursor Protein
ATP	Adenosine Triphosphate
BP	Back Propagation
CDF	Cerebrospinal Fluid
DNA	Deoxyribonucleic Acid
FDG PET	Fluorodeoxyglucose Positron Emission Tomography
FDR	False Discovery Rate
FNN	Feedforward Neural Network
GPGPU	General-Purpose Graphics Processing Unit
GPU	Graphics Processing Unit
GTP	Guanosine-5'-triphosphate
KEGG	Kyoto Encyclopedia of Genes and Genomes
MAP	Microtubule Associated Protein
MAPT	Microtubule Associated Protein Tau
MAS5	MicroArray Suite 5.0
MCCV	Monte Carlo Cross Validation
MCI	Mild Cognitive Impairment
miRNA	Micro Ribonucleic acid
MLP	Multi-Layer Perceptron
MRI	Magnetic resonance imaging
mRNA	Messenger Ribonucleic acid
MSE	Mean Squared Error
NCBI	National Center for Biotechnology Information
NFT	Neurofibrillary Tangle
PCA	Principal Component Analysis
PET	Positron Emission Tomography
PSEN	Presenilin
p-tau	phosphorylated tau



RMA	Robust Multi-Array Average
RNA	Ribonucleic Acid
RNAseq	Ribonucleic Acid Sequencing
SVM	Support Vector Machines
TAC	Transcriptome Analysis Console
TCGA	The Cancer Genome Atlas
THU	Threshold Processing Unit
t-tau	Total tau
WHO	World Health Organisation

## Abstract

Alzheimer's Disease is the most common form of dementia worldwide with 40 million patients in the USA alone. This neurodegenerative disease is commonly characterised by the presence of amyloid plaques and neurofibrillary tangles in the brain, which result from the deposition of extracellular  $\beta$ -amyloid protein fragments and abnormal tau protein respectively. Over the years, research and medical efforts to control the disease by targeting these proteins have been largely unsuccessful, originally due to the difficulty in detection and targeting, but even with advanced technology, the effects of approaches targeting these proteins have been minimal. Further research is required to fully understand the causes of the disease, how it progresses, which systems are affected and how it can be treated efficiently and effectively.

With the advent of high throughput sequencing technologies such as transcription microarrays, methylation arrays and RNA sequencing, a wealth of high quality data is being generated allowing for the tracking of changes at the genetic level over the course of the disease. This information was analysed using machine learning methods including the in-house Stepwise Artificial Neural Network algorithm as well as the Network Inference algorithm developed by Graham Ball and his research group to elucidate new molecular markers and drivers of the disease and also to evaluate existing ones. The results were analysed using a non-parametric systems biology approach to determine the impact of these markers on the systems involved in the disease and new techniques including the driver analysis were developed to reduce bias and increase clarity.

In order to achieve the most comprehensive set of results and reduce the risk of error and false discovery, the E-GEOD-48350 dataset was selected for its comprehensive and high-quality data and was used to test both old and new methods and obtain a preliminary set of results. These results were validated using other transcription datasets as well as an RNA sequencing dataset, leading to the identification of dysregulated genes related to microtubule stabilisation and immune system regulation in Alzheimer's disease, providing a foundation for further expansion and research.

# Table of Contents

Copyright Statement.....	1
Acknowledgements.....	2
Abbreviations.....	3
Abstract.....	5
Chapter 1: Introduction.....	10
1.1 Alzheimer’s Disease.....	10
1.1.1 Description and Impact.....	10
1.1.2 Risk Factors.....	12
1.2 Characterising Alzheimer’s Disease.....	13
1.2.1 Physiology.....	13
1.2.2 Molecular characterisation.....	14
1.3 Biomarkers.....	15
1.4 Challenges in Alzheimer’s Disease Research.....	19
1.4.1 Diagnostics.....	19
1.4.1 Clinical Trials.....	21
1.5 AD Hypotheses.....	23
1.5.1 Amyloid Cascade.....	23
1.5.2 Inflammation.....	24
Chapter 2: Machine Learning and Data Mining.....	26
2.1 Supervised Approaches.....	26
2.1.1 Artificial Neural Networks.....	26
2.1.2 Support Vector Machines.....	28
2.1.3 Genetic Algorithms.....	29
2.1.4 Decision Trees.....	30
2.1.5 Bayesian Networks.....	31
2.2 Unsupervised approaches.....	31
2.2.1 Hierarchical Clustering.....	32
2.2.2 K-means Clustering.....	33
2.2.3 Principle Component Analysis.....	33
2.2 Biomarker Discovery.....	34
2.3 Systems Biology.....	35
2.3.1 Top down approach.....	37
2.3.2 Bottom up approach.....	38

2.3.3 Middle out approach .....	39
2.4 Study Aims.....	39
Chapter 3: Artificial Neural Networks.....	41
3.1 Introduction.....	41
3.2 Historical Background .....	41
3.3 Architecture.....	44
3.3.1 Perceptron .....	44
3.3.2 Feedforward Neural Networks.....	45
3.3.3 Multilayer Perceptron.....	46
3.4 Learning Rules .....	47
3.4.1 Supervised and Unsupervised Learning.....	48
3.4.2 Bias-Variance Trade-off.....	48
3.4.3 Back-Propagation Algorithm .....	51
3.4.4 Gradient Descent.....	52
3.4.5 Generalisation and Overfitting .....	55
3.5 Optimisation.....	57
3.5.1 Randomisation of Weights.....	58
3.5.2 Learning Rate and Momentum.....	58
3.5.3 Hidden Layer Parameters.....	59
3.6 Advantages and Disadvantages.....	61
3.7 Stepwise Analysis .....	62
3.8 Network Inference.....	66
3.8.1 Model development.....	68
3.8.2 Workflow .....	69
3.8.3 Visualisation.....	71
Chapter 4: Non-Systematic Hypothesis-Free Approach for AD Biomarker Discovery .....	72
4.1 Dataset Selection.....	72
4.2 Data Normalisation .....	74
4.3 Stepwise ANN.....	77
4.3.1 Single Marker Analysis.....	77
4.3.2 Multistep Stepwise Analysis.....	78
4.3.3 Gene Ontology .....	80
4.3.4 Pathway Analysis.....	84
4.3.5 Conclusion.....	85
4.4 Network Inference.....	85
4.4.1 Interactomes .....	85

4.4.2 Hive Plots.....	93
4.5 Driver Analysis .....	94
4.5.1 Master Driver Analysis .....	96
4.5.2 AD Driver Analysis.....	100
4.5.3 Healthy Driver Analysis.....	104
4.6 Commonality Analysis.....	107
Commonalities between all three datasets for the top 50 source genes .....	109
Commonalities between all three datasets for the top 50 target genes.....	110
Pairwise Commonalities Source .....	111
Pairwise Commonalities Target .....	112
Chapter 5: Systems Biology Expansion and Integration.....	113
5.1 Interaction Matrix.....	113
5.2 Disparate Brain Region Variance.....	119
5.2.1 Hippocampus.....	121
5.2.2 Entorhinal Cortex .....	131
5.2.3 Postcentral Gyrus .....	141
5.2.4 Superior Frontal Gyrus.....	146
5.3 Comparison Against Known Markers.....	149
5.3.1 APOE4 - Apolipoprotein E .....	151
5.3.2 APP - Amyloid Beta Precursor Protein.....	152
5.3.3 MAPT - Microtubule Associated Protein Tau .....	154
5.4 Inter-comparison and Cross-Comparison with other datasets and technologies.....	156
5.4.1 Microarrays and RNA-seq .....	156
5.4.2 RNA-seq Driver Analysis .....	157
5.4.3 Comparison against known markers .....	162
5.5 Conclusion .....	170
Chapter 6: Conclusions .....	172
6.1 Novel Methodology .....	172
6.1.2 Predicted biomarkers.....	172
6.2 Quality of Results.....	173
6.3 Hypothesis Free Approach Evaluation.....	175
6.4 Further expansion.....	177
6.4.1 Single cell sequencing.....	177
6.4.2 Top 10 approach.....	178
6.4.3 Commonalities between predictors in a panel.....	179
6.4.4 Epigenetics .....	179

6.4.5 Complete gene analysis.....	180
6.5 Cracking the Algorithm.....	180
References.....	182
Publications.....	198
Appendix.....	226

# Chapter 1: Introduction

## 1.1 Alzheimer's Disease

### 1.1.1 Description and Impact

Alzheimer's disease (AD) is recognised as the most common form of dementia worldwide. This chronic neurodegenerative disease usually starts slowly, often up to 20 years before the first symptoms become visible. The most common early symptom is difficulty in remembering short-term events which gets progressively worse, although the speed of memory degradation appears to vary between individuals (Braak and Tredici, 2012, Mattson, 2008, Gross *et al*, 2010). This is compounded by severe degeneration of multiple brain regions including the hippocampus, entorhinal cortex, neocortex, nucleus basalis, locus coeruleus and raphe nuclei, leading to disruption in mental functions such as comprehension, judgement, language and calculation (Carlson and Birkett, 2017) as shown in Figure 1. Due to the slow progression that is characteristic of the disease, as well as various popular misconceptions surrounding dementia, it is common for patients and their families to assume that this degeneration of a person's mental faculties is a normal part of ageing, thus delaying early diagnosis. It is crucial to emphasise that AD is the abnormal degeneration of mental faculties and while age is indeed the biggest risk factor, it is far from the only one.

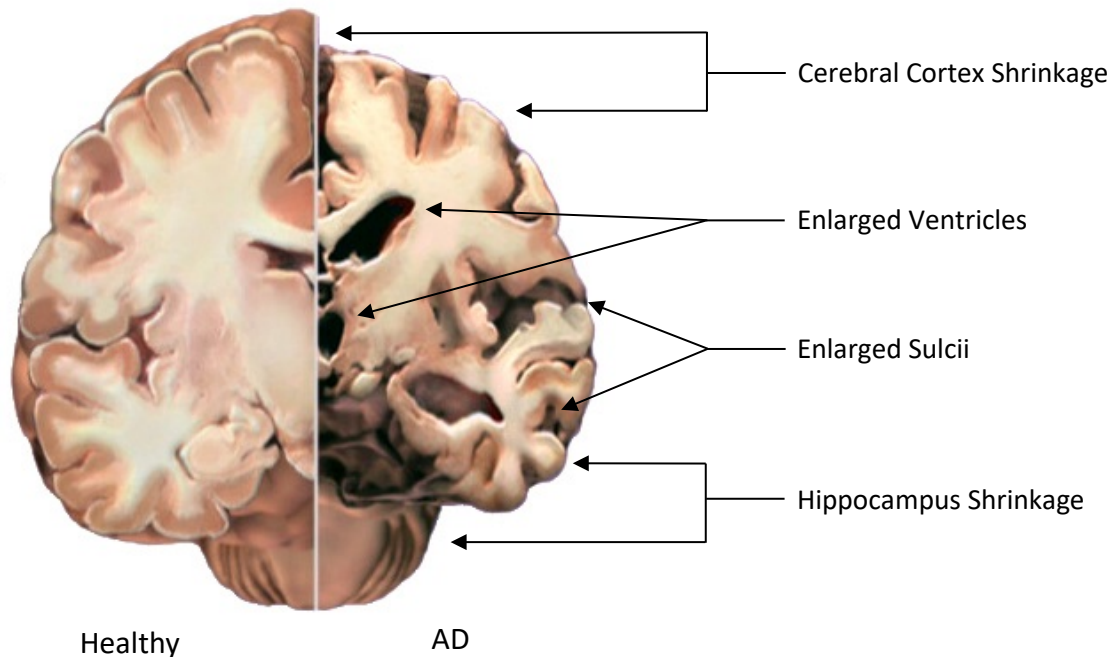
In Duthey's report (2013) it is suggested that AD accounts for as much as 70% of all dementia cases, and the stages were outlined in the WHO report (WHO, 2015) on dementia as follows:

Early stage – Often overlooked as it develops gradually and thought of as a normal aspect of old age, it is the hardest stage to identify early. Forgetfulness, difficulties in communication, timekeeping and decision making, combined with a loss of motivation and reduced activity are common.

Middle stage – Lasts 2-5 years and is distinct due to the degree of degeneration present. Forgetfulness extends to very recent events, speech and comprehension suffers, assistance is required for personal care and behaviour can become erratic and inappropriate.

Late stage – At 5+ years, the late stage of dementia is characterised by total dependence and inactivity. The disease now affects the patient physically as well as mentally and while previous symptoms worsen significantly, they are compounded by inability to eat without

assistance and difficulty swallowing, bladder and bowel incontinence, as well as significant reduction in mobility, often leaving the patient confined in a wheelchair or bed.



**Figure 1:** Physiological differences between a healthy and AD brain section, demonstrating white matter shrinkage in the hippocampus and cerebral cortex. Source: [www.alz.org](http://www.alz.org)

In addition to the enormous emotional cost the disease exerts on patients and their families, it has become a major public concern due to the high healthcare costs associated with it, which, in combination with the overall rise in the elderly population has caused AD to be classified as a priority condition (Duthey, 2013). According to the World Health Organisation, in 2015 there were over 40 million people with dementia in the USA, 15 million of whom suffered from Alzheimer’s disease. Healthcare costs have spiralled to over 900 billion USD, whereas in Europe the costs have risen to nearly 250 billion euros, a rise of almost 40% from 2008. Moreover, it is projected that by 2050, 22% of the world’s population will be over the age of 60, and therefore at increased risk, with patients in third world countries accounting for 80% of the total (WHO, 2015).



### 1.1.2 Risk Factors

Although these statistics are alarming, they fail to show the greatest concern faced when trying to control the disease; the cause is largely unknown. There has been a multitude of studies into understanding AD which have managed to produce a series of risk factors that can be used to further explore it. The greatest of these factors is age. The vast majority of AD patients are over the age of 65 (Ziegler-Graham, 2008) and the risk of developing the disease over the age of 85 rises to 50%. It appears that the risk of developing the disease doubles every six years, leading to an exponential risk increase and as the average human lifespan has increased and more people are able to survive for longer. Curiously, even though incidence rates differ between regions, the risk appears to be consistent regardless of geographic location.

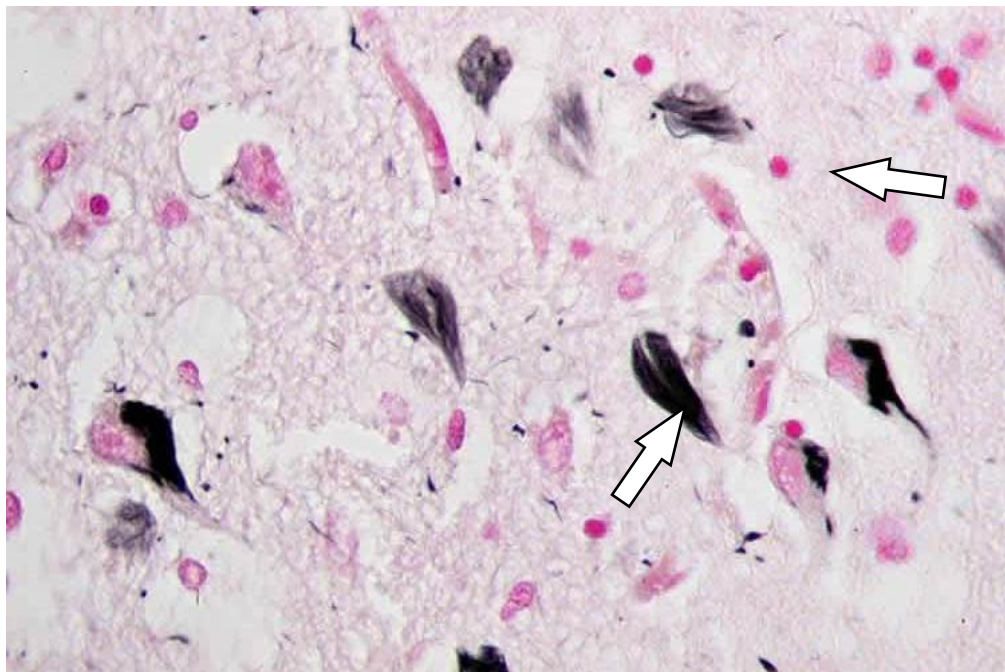
The second largest risk factor is genetics, even though AD is not a genetically inherited condition. In the rare cases where familial AD, also known as autosomal dominant AD (ADAD) caused by mutation in the amyloid precursor protein (APP) and presenilin 1 (PSEN1) and 2 (PSEN2) genes, the disease develops early, between 30 and 40 years and only affects around 0.1% of the population (Veugelen, 2016). However, there is evidence that there are other genetic risk factors resulting from mutations, such as those in the APOE4 allele, that could potentially be used to understand the causes of the disease. These inherent risk factors however, are proving to be insufficient in explaining the underlying molecular mechanisms of the disease, which is the main focus of this thesis, although their influence will also be examined in greater detail in the following chapters in regard to those mechanisms.

Finally, there is increasing evidence that environmental factors, including lifestyle choices and interactions with other disease, have a direct effect on the risk of developing AD. These include obesity and smoking, cardiovascular diseases and type II diabetes, which increase the risk of AD but also increase physical activity, education and better diet which decrease it (Mayeux and Stern, 2012). Oddly enough, the risk of AD appears to be inversely correlated with cancer risk, especially prostate cancer (Behrens *et al*, 2009), likely due to the contradictory nature of the two diseases – uncontrolled growth and abnormal degeneration – leading to depletion in the finite amount of energy used in homeostasis.

## 1.2 Characterising Alzheimer's Disease

### 1.2.1 Physiology

As mentioned earlier, AD is a neurodegenerative disorder characterised by progressive decline in mental ability and, as shown in Figure 1, it also exhibits neuronal loss in multiple brain regions. This has made early detection of the disease inaccurate and impractical, which affects our ability to study and characterise its physiology. Historically, identification of AD could only be performed post-mortem upon examination of the brain tissue. As a result, the physiological hallmarks of AD (Figure 2) have been widely considered to be the presence of amyloid plaques, extracellular deposits of insoluble beta-amyloid ( $A\beta$ ) in the parenchyma of the brain as well as neurofibrillary tangles (NFT), intracellular deposits of hyper-phosphorylated tau protein which fill the neuron and take its shape, preventing it from functioning correctly (Carlson and Birkett, 2017). These features have been considered hallmarks of AD and all current theories stem from them, although there is an ever-increasing body of evidence that indicates this as a small fraction of the whole picture.



**Figure 2:** Amyloid plaques (pink) and neurofibrillary tangles (black) in Alzheimer's disease brain tissue. Source: [www.alzheimers.org.uk](http://www.alzheimers.org.uk)

Amyloid plaques consist of a solid core of defective  $A\beta$  and are surrounded by degenerate axons and dendrites, activated microglia and astrocytes. This defective protein is a result

of the cleaving of the amyloid precursor protein (APP) by beta ( $\beta$ ) and gamma ( $\gamma$ ) secretases. The site at which APP is cleaved by  $\gamma$ -secretase determines whether A $\beta$  will be the long or short form. The short form is the most common (~90%) but the long form is found as often as 40% in the brains of AD patients (Carlson and Birkett, 2017), and while small amounts can be cleared easily, the higher rate of production in AD leads to the clearance system being unable to cope. Moreover, soluble forms of the protein have been shown to be neurotoxic and synaptotoxic (Mucke and Selkoe, 2012). A $\beta$  plaques often form in different but overlapping topological regions to neurofibrillary tangles (Jack *et al*, 2010) but appear to have a smaller effect on neurodegeneration and synaptic loss than the latter (Gomez-Isla *et al*, 1997).

Neurofibrillary tangles are formed as a result of the hyperphosphorylation of tau, a microtubule associated protein (MAP) whose role is to bind to tubulin and stabilise the structure of neurons to maintain their function. When hyperphosphorylated, due to excessive amounts of phosphate ions, it changes from its normal soluble form to oligomeric and fibrillar forms, does not bind to tubulin and impairs axonal microtubule structure and assembly which has been shown to have a neurotoxic effect (Iqbal, 2011). This can persist beyond neuron death (Blair, 2013) which is why AD is the most common tauopathy as well as dementia. It should be noted that although the progression of AD correlates with the number of NFTs in the brain, the formation of such tangles is thought to be a protective mechanism used to sequester toxic, soluble intermediates until they can be converted to a less harmful form (de Calignon *et al*, 2010). It remains a question however whether the formation of NFTs is related to mutations in the MAPT gene, leading to an increase of tau production or misfolded tau protein, or even a spread of the toxic soluble species, similarly to prion diseases (Liu *et al*, 2012).

### 1.2.2 Molecular characterisation

The compartmentalised model for AD described above, was accepted due to the physiology of the disease. It states that the physiological changes mentioned above are directly related to the progression of the disease, i.e. the presence of A $\beta$  plaques and NFTs lead directly to dementia, while absence is indicative of normal cognitive function. Although this has been the accepted model until relatively recently, there is an increasing body of evidence that indicates dementia as the end state of a gradual decline in cognitive

functions that begins decades before the first symptoms become evident, and is characterised by an increase in AD pathology and clinical decline (Jack *et al*, 2010)

These physiological changes however, have led to questions regarding the molecular mechanisms of the disease. The majority of research regarding the molecular mechanisms of AD has been focused on four genes: APP, PSEN1, PSEN2 and APOE. This is partly due to the clear connections these genes have with the development of amyloid plaques and partly due to the certainty of their involvement, as they have been proven to be directly causal to familial AD. However, not all AD patients carry the APOE e4 allele which is linked to familial AD, the incidence rate of which can be as low as 50% (Corder *et al*, 1993), necessitating further research on the subject. Other factors have been considered, including a2-Macroglobulins, which mediate clearance of A $\beta$ , lipoprotein receptor proteins, as they are found in amyloid plaques, and transforming growth factors such as TGF- $\beta$ 1 which are known to be overexpressed in AD. Very few of these factors are considered causative of the disease and are more likely to increase the risk of AD and enhance its pathology by interacting with environmental, pathologic factors or pre-existing conditions (Rocchi *et al*, 2003)

A crucial shortcoming of current research into the molecular pathology of AD is the push to link all possible avenues of thought into the formation of amyloid plaques and NFTs. As discussed shortly, the effect of other molecular factors on neuroinflammation, oxidative stress and synaptic plasticity is considered secondary to A $\beta$  which drastically increases the bias inherent in such studies.

### 1.3 Biomarkers

Biomarkers, or biological markers, are a broad category of medical signs that can be measured accurately and reproducibly and provide objective information about the state of a patient or condition (Strimbu and Tavel, 2011). This is, however, one of many definitions for biomarkers and they all have significant overlap without necessarily contradicting each other. Biomarkers are characteristics that can be measured as indicators for biological processes, which can include body temperature, by-products of metabolism or even genetic mutations. As a result, they are incredibly useful tools to assess the progress of a disease, patient response to therapy and provide researchers starting points

to further analyse and understand a given condition, with good biomarkers reducing the risk of misdiagnosis, which in turn allows for the development of more successful treatments. AD however, is critically short on established, accurate and easy to use biomarkers.

Currently there are very few markers for AD and they pose significant challenges in identifying and making good use of them, with elevated levels of combined A $\beta$ , tau and phosphorylated tau in the cerebrospinal fluid (CSF) used for AD confirmation (Sharma and Singh, 2016). These markers fall within the criteria for ideal AD biomarkers as detailed by a large number of researchers (Gu *et al*, 2012, Blennow, 2014) as they reflect the effects of ageing in the AD brain while describing its pathophysiology, they are highly specific and sensitive, reproducible with clear cut-off values between at least two-fold changes and relatively easy and inexpensive to test for. Furthermore, these markers are identified in the CSF, which makes them highly representative as they come in contact with the central nervous system.

Beta amyloid is used due to the presence of amyloid plaques in the brain being considered a hallmark of the disease, and A $\beta$ <sub>42</sub> specifically, being a hydrophobic and fibrillogenic species, makes up the majority of the deposited protein in the cerebral cortex and hippocampus in the early stages of the disease (Huynh and Mohan, 2017). While mechanisms related to the accumulation of A $\beta$ <sub>42</sub> are unclear, it leads to a decrease in the protein levels in the CSF, although there is little consensus between research groups as to the exact values that can be considered significant. The first incidence of the levels of A $\beta$ <sub>42</sub> being used as a marker is in 2003 by Kapaki *et al* (2003), who set the threshold to a 0.5-fold change compared to normal ageing and was followed up by de Jong *et al* (2006) who confirmed these finding with a higher specificity. Finally, Mulder *et al* (2010) used an intermediate cut-off value in a similar series of experiments which lead to sensitivity and specificity between the previous studies, cementing A $\beta$ <sub>42</sub> levels in the CSF as an AD biomarker. Further studies, such as Vos *et al* (2013) have attempted to use CSF A $\beta$ <sub>42</sub> to differentiate between AD and mild cognitive impairment (MCI) with only moderate success, while others have been attempting to use other A $\beta$  species, such as A $\beta$ <sub>40</sub>, for similar purposes.

Naturally, in-depth analysis of the levels of tau and the degree of its phosphorylation soon followed, as NFTs are the other major hallmark of AD, and much like amyloid, NFTs are found in the hippocampus and cerebral cortex. They are composed of filaments of hyperphosphorylated tau protein and are commonly found to accumulate in AD. The study by Kapaki *et al* (2003) mentioned earlier found a 3.5-fold increase in total CSF tau protein levels (t-tau) between AD patients and cognitively normal controls and the findings were validated by the de Jong *et al* (2006) and Mulder *et al* (2010) studies. Moreover, further research has been performed on the levels of phosphorylated tau (p-tau) levels in the CSF with similar results. In fact, it appears that the best results are only achievable by a combination of these approaches (Vos *et al*, 2013).

The downside of these studies however, regardless of the quality of the results, is that they can only detect markers of the disease after they have started accumulating, which may be too late since the disease likely begins years before these symptoms become evident. As such, there is a need for reliable biomarkers that precede A $\beta$  deposition. Reiman *et al* (2004) suggested that fluorodeoxyglucose positron emission tomography (FDG PET) hypermetabolism occurs in individuals carrying the APOE e4 allele, which produces symptoms similar to AD, and precedes A $\beta$  deposition by affecting glucose metabolism, although it should be noted that the APOE e4 allele appears to simply lower the age at which A $\beta$  deposition begins. Moreover, in recent years technology has allowed for more avenues of thought when looking for AD biomarkers. Task-free functional MRI is capable of measuring functional connectivity and network dynamics that could be used a tool to explore the pathological and physiological processes in AD while being minimally invasive (Jack *et al*, 2013).

As mentioned earlier the, APOE e4 allele is considered an ideal biomarker for detection of familial AD as it is highly sensitive and specific, which has resulted in multiple tests to attempt to establish its significance in sporadic AD. ApoE is produced in astrocytes and regulates lipid homeostasis and metabolism and has three alleles in humans, with the frequency of the e4 allele being approximately 3-fold higher in AD patients than other groups according to recent studies (Liu *et al*, 2013). Its use as a biomarker in sporadic AD however, remains questionable. A study by Elias-Sonnenschein *et al* (Elias-Sonnenschein *et al*, 2010) performed a meta-analysis to evaluate APOE e4 as a biomarker for progression from MCI to AD showed that while APOE e4 was a moderately strong

predictor, it was only usable for highly specialised cases, only being applicable to homozygotes. Further studies to establish APOE  $\epsilon$ 4 as an AD biomarker have been met with only moderate success and inconclusive results, likely due to the nature of the presence of other, stronger but unidentified predictors (Morris *et al*, 2017, Ba *et al*, 2016).

It is however, quite encouraging that, thanks to the increasing interest in the field, there are many studies that are expanding our current understanding of AD, which has led to a plethora of potential markers. Genes such as BIN1, a complex gene playing a crucial role in cytoskeleton dynamics and modulation of endocytosis, has been shown to strongly interact with tau (Chapuis *et al*, 2013), have resulted from genome wide association studies. Clusterin, also known as apolipoprotein J (Lambert *et al*, 2013) is highly expressed in neurons and microglia and could play a crucial role in preventing A $\beta$  fibrillization, while genes such as TREM2, triggering receptor expressed on myeloid cells 2, and genes in the HLA-DRB5-DRB1 region (Huynh and Mohan, 2017) represent other potential candidates. All these potential markers are being discovered in genome wide association studies using modern technologies such as RNA sequencing and multiple systems biology applications which has drastically increased the number of possibilities and the variance within them and thus, is far more likely to explain the cause of the disease, its progression and how to combat it.

Furthermore, blood-based markers related to A $\beta$  proteins, enzymes related to tau pathology and inflammatory markers are being studied extensively. Although plasma levels of A $\beta$  are unstable, as it can become trapped by interacting with other proteins, and platelets contain a high amount of A $\beta$  regardless (Humpel, 2011) it might be possible to detect in the future. While tau can be quite a challenge to detect in the blood, enzymes related to its phosphorylation, such as GSK-3, are less so, and could become biomarkers themselves if the correct ones are identified. The challenge with all peripheral factors however, is that they interact with a large number of other genes/proteins and their levels fluctuate significantly, necessitating larger, interaction-based analyses for them to act as predictors. Finally, personalised medicine could provide the necessary answers as it is highly likely that slight genetic differences between individuals are significant predictors of the disease and will require highly sensitive and specific treatment to overcome.

## 1.4 Challenges in Alzheimer's Disease Research

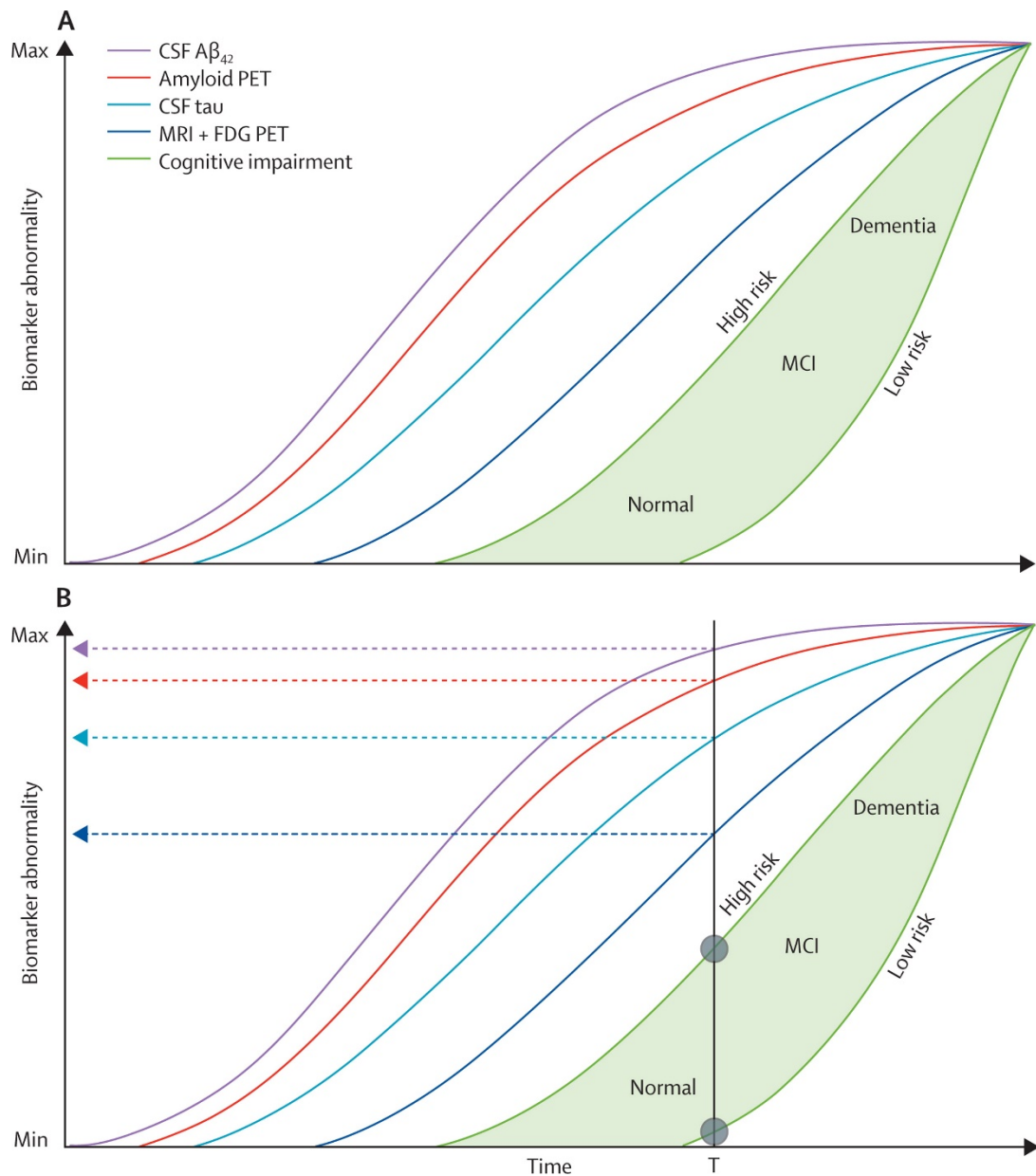
### 1.4.1 Diagnostics

In the previous section it was established that the current acceptable standard for identifying AD biomarkers involves examining the CSF for multiple A $\beta$  peptide species as well as tau. Starting with CSF A $\beta$ 42 and amyloid PET which are most differentiated in the early stages of the disease, followed by testing for CSF tau and FDG PET and finally, structural MRI before moving on to examining clinical symptoms (Jack *et al*, 2013). While the approach of sampling CSF for AD markers by lumbar puncture is proving reasonably effective it is still limited by being rather invasive, although less so than the more accurate but significantly more invasive brain biopsy, being painful for the patient and producing results that are challenging to reproduce. Lumbar punctures are known for causing nausea, weakness and severe backache in elderly patients and are hard to maintain for regular diagnosis over a long timeframe (Sharma and Singh, 2016). Additionally, there is a small risk of bleeding and in some cases, brain herniation which is potentially fatal, thus necessitating the need for more safely and easily obtainable markers present in the blood, serum or other products such as urine.

Potential circulatory biomarkers include molecules such as circulatory miRNA, dysregulations in the expression of which could be linked to AD (Geekiyananage *et al*, 2012) and blood based amyloid markers, as discussed earlier. However, far more likely candidates have presented themselves in the form of neuroinflammatory and oxidative stress markers. The AD inflammation hypothesis is gaining traction and oxidative stress, although common in many diseases, still plays an important role in the development of AD and should not be ignored. The hypotheses will be examined in detail shortly.

Even with all these markers available to us and the research on the topic, the main challenge remains being able to characterise the disease successfully in a clinical setting, as the causes and mechanisms of AD are poorly understood. Jack *et al* suggested a model in 2010, that has since been updated, to combine the biomarkers into a highly sensitive and specific panel for use in diagnostics as follows.





**Figure 3:** Revised model of dynamic biomarkers of the Alzheimer's disease pathological cascade by Jack *et al* (2013). (A and B) Neurodegeneration is measured by FDG PET and structural MRI, which are drawn concordantly (dark blue). By definition, all curves converge at the top right-hand corner of the plot, the point of maximum abnormality. Cognitive impairment is illustrated as a zone (light green-filled area) with low-risk and high-risk borders. (B) Operational use of the model. The vertical black line denotes a given time ( $T$ ). Projection of the intersection of time  $T$  with the biomarker curves to the left vertical axis (horizontal dashed arrows) gives values of each biomarker at time  $T$ , with the lead biomarker (CSF  $A\beta_{42}$ ) being most abnormal at any given time in the progression of the disease. People who are at high risk of cognitive impairment due to Alzheimer's disease pathophysiology are shown with a cognitive impairment curve that is shifted to the left. By contrast, the cognitive impairment curve is shifted to the right in people with a protective genetic profile, high cognitive reserve, and the absence of comorbid pathological changes in the brain, showing that two patients with the same biomarker profile (at time  $T$ ) can have different cognitive outcomes (denoted by grey circles at the intersection of time  $T$ ). Source: Jack *et al*, 2013

According to this model (Figure 3), there is evidence that  $A\beta$  deposition follows a sigmoid pattern, accumulating slowly at first, speeding up and finally reaching a plateau. Thus, it becomes possible to use the current known biomarkers for AD to predict the disease,

although this model exclusively focuses on confirmed AD, since the end state of the disease exists in a state of multiple pathophysiological changes that make it hard to fully characterise. It is worth noting that in this model, while all biomarker curves are sigmoidal in nature, they are not identical, as was the case with the original proposed model, due to the different impact each marker has on the disease, such as the impact of cognitive impairment in patients with a protective profile. Even this model is insufficient though, as it is mostly hypothetical. There are simply not enough high quality, easily accessible markers to translate this model for use in the clinic, and it completely lacks any markers related to the pathophysiological processes of AD. Additionally Jack *et al* also draw attention to the limitations presented by the lack of middle aged patient samples preventing early capture of potentially crucial information, the lack of individuals with end state dementia, reducing the dynamic range of biomarker abnormalities, increasing bias and most importantly, the lack of long-term longitudinal data that have resulted from most AD biomarker studies relying on short-term follow-up, leading to an increase in bias. Furthermore, the model will need to be validated on real data from long-term biomarker studies to combat the need for additional sampling points.

#### 1.4.1 Clinical Trials

Naturally, the push to identify the most sensitive and accurate biomarkers for the disease, especially during the pre-clinical stage is not simply for identification, but because biomarkers are invaluable therapeutic tools due to their ability to aid in prognosis, therapy and evaluate a patient's response to it. Since prevention is preferential to treatment, any potential cure or treatment is most likely to be at its most effective during that pre-clinical stage, necessitating the discovery of a new series of biomarkers, a trend which is reflected in recent research (Riter and Cummings, 2015, Schneider *et al*, 2014, Mattson *et al*, 2015). However, there is another, highly significant trend in AD clinical trials of potential disease modifying therapies; none have managed to go through stage III. Between 1998 and 2014, the Pharmaceutical Research and Manufacturers of America have identified 101 failures of potential AD modifying treatments (Schneider *et al*, 2014) while other studies put the number closer to 170 (Linder *et al*, 2008). It is clear that a change in strategy is required.

Over the last 30 years of clinical trials there has been significant progress in how the disease is approached. It is only recently that AD is understood as a complex multi-stage

disease with a long, asymptomatic preclinical phase, and the trials are changing to reflect this fact. The standard biomarkers for AD (CSF A $\beta$ , CSF p-tau/t-tau) may be highly specific to the disease, unlike most preclinical markers, although there is mounting criticism related to their inability to differentiate AD from other dementias (Engelborghs *et al*, 2008). Moreover, all fluid biomarkers share the limitation of lacking anatomical precision (Rosén *et al*, 2013).

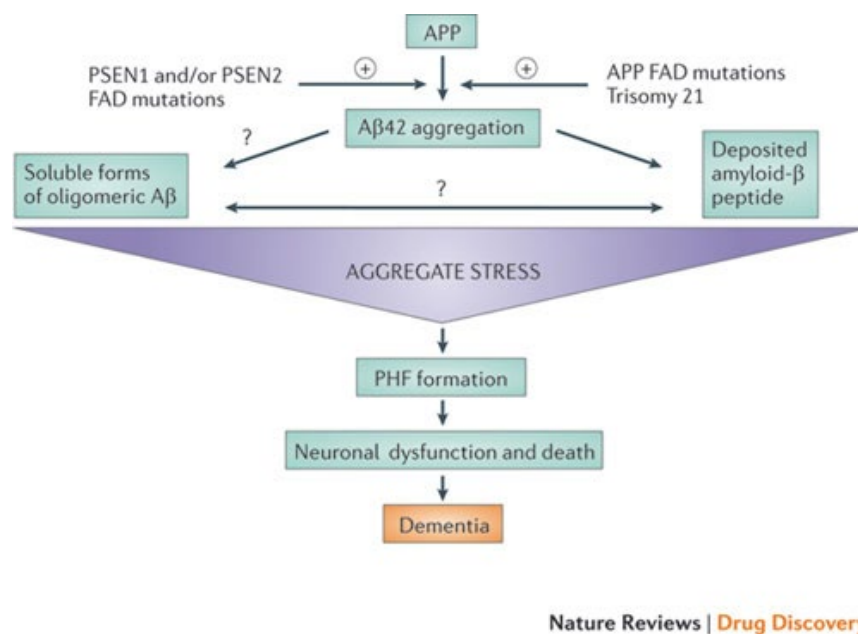
These facts have led to clinical trials attempting to modify the disease in alternative ways. A popular approach is the active amyloid immunisation strategy, which aims to immunise the patient against A $\beta$ 42. Since there is significant evidence that amyloid changes from its normal form to its pathogenic in the preclinical stage, this strategy was most effective when administered before the start of amyloid deposition (Das *et al*, 2001). However, although safe, no studies have managed to produce a positive clinical effect (Ritter and Cummings, 2015). Other approaches have included Bapineuzumab, a monoclonal antibody targeting the N terminus of A $\beta$  with the goal of stopping plaque formation which failed at phase III as although it reduced the levels of p-tau, it failed to significantly alter the A $\beta$  or t-tau levels (Dubois *et al*, 2014). A similar monoclonal antibody, Solanezumab, which targets the middle amino acid section of A $\beta$  is the only partial success to date with two large phase III trials including 2000 patients followed over the course of 72 weeks with the endpoints being delayed cognitive and functional deterioration (NCT00905372, NCT00904683). These trials, once again, showed patients without AD pathology based on PET scans and the study failed to show any reduction in A $\beta$  levels, although there is an indication that Solanezumab prevented A $\beta$  deposition in the preclinical stage (Savoneno *et al*, 2015, Crespi *et al*, 2015).

In summary, there is a need for a wide variety of more representative AD biomarkers that can be used to more accurately guide and gauge clinical trials. However, there is a potential problem stemming from the reliance of so many AD studies on A $\beta$ , its deposition and the resulting plaques. While not incorrect, it is limiting research into other, potentially crucial, factors and biomarkers that could have a significant positive effect on therapy.

## 1.5 AD Hypotheses

### 1.5.1 Amyloid Cascade

The leading theory for the cause of Alzheimer's disease is the amyloid cascade hypothesis, first proposed in 1992 and its influence on AD research cannot be understated. The hypothesis posits that mutations in the APP and presenilin genes PSEN1 and PSEN2 leads to the deposition of A $\beta$  in the brain, which subsequently leads to the formation of NFTs, cell death and dementia. Experiments in animal models have shown that chemically or damage induced lesions lead to an increase in APP levels and accelerate the development of AD (Yar *et al*, 1992, Wallace *et al* 1991). Unfortunately, all approaches based on the amyloid cascade have failed at Phase III clinical trials - tramiprosate, tarenflurbil and semagacestat - and research has not been able to conclusively link the build-up of A $\beta$  to the formation of NFTs (Reitz *et al*, 2012).



**Figure 4:** Diagram of the amyloid cascade hypothesis showing the theorised links between the aggregation of A $\beta$  to cell death and dementia. Source: Karran *et al*, 2011

While it has been made clear that the amyloid cascade hypothesis is not enough to sufficiently explain the development of AD or aid in its detection and, consequently, is currently under heavy scrutiny, it is also not possible to accept the null hypothesis, as autosomal dominant mutations in the aforementioned APP, PSEN1 and PSEN2 genes along with the apolipoprotein E4 (APOE4) allele have been proven to be the key components in familial, or early onset, Alzheimer's disease. Instead, the amyloid cascade hypothesis has to be modified to account for the rate of A $\beta$  deposition and clearance, the

connection with the development of NFTs and the effect of inflammation in the development of AD. Karran *et al* (2011) have attempted to update the hypothesis (Figure 4) for use in therapeutics by presenting four distinct scenarios describing the role of A $\beta$  in AD. These scenarios are:

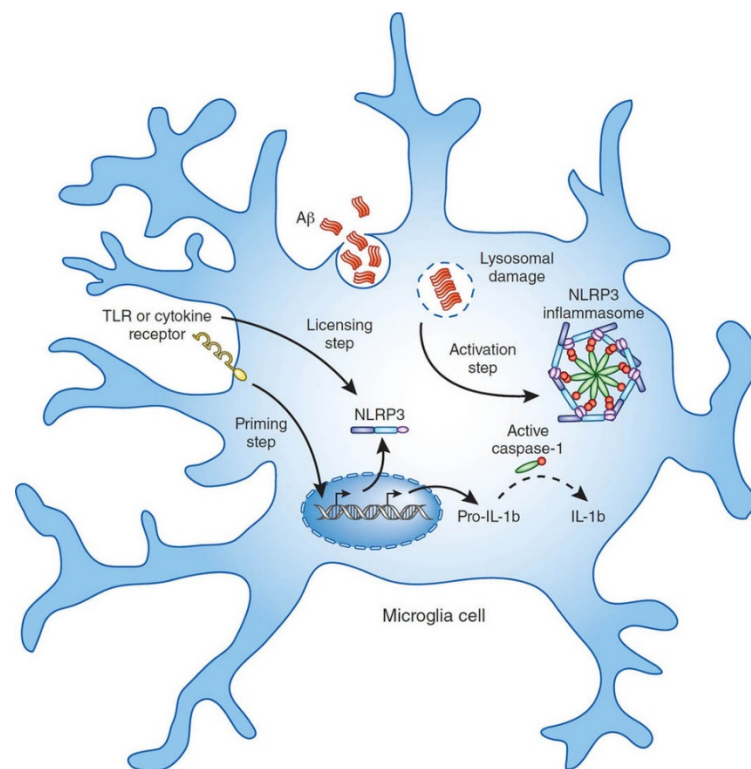
1. A $\beta$  could trigger development of the disease and further accumulation has little to no effect
2. development starts once A $\beta$  reaches a certain, as yet unknown, threshold
3. A $\beta$  is a key driver of AD and its continued deposition accelerates the effect
4. A $\beta$  is irrelevant and the presence of plaques and increased levels of A $\beta$  are a side effect of a different cause.

It should be noted that a major limitation of this hypothesis is that it fails to account for AD patients with little to no AD pathology (Salloway *et al*, 2014) and thus amyloid plaques as identified by PET scan. In recent years, mouse studies have shown that A $\beta$  deposition is a potential driver for tau hyperphosphorylation, fixing one the major limitations of the amyloid hypothesis. Crossing APP transgenic mice with tau knockout mice, resulted in offspring with significantly fewer behavioural deficits (Selkoe and Hardy, 2016) while other studies have shown that soluble oligomers of A $\beta$  can lead to alterations in tau, potentially cascading to AD (Shankar *et al*, 2008) although the mechanisms are still unclear. Strooper and Karran (2015) attempted to provide alternatives including proteostatic stress during the biochemical phase when A $\beta$  aggregates at an abnormally fast pace, defects in the amyloid and tau clearance mechanisms and a decrease in synaptic plasticity. As Selkoe and Hardy (2016) suggest, the amyloid hypothesis, for all its limitations, is essential for therapeutics due to the fact that the complexity of the disease increases drastically after initiation due to the rise in complexity of downstream pathogenic processes, the most likely point of the disease where treatment will be at its most successful.

### 1.5.2 Inflammation

Recent research has also been focused on investigating the role of inflammation in AD, in an attempt to explain the development of the disease. The inflammation hypothesis (Figure 5) posits that deposition of A $\beta$  causes chronic activation of the immune system and disrupts microglial clearance functions. Microglia are immune cells located in the

parenchyma of the brain, making up 20% of the total glial population. Their functions include phagocytosis, induction of inflammation, and antigen presentation to lymphocytes (Aloisi, 2011). However, their roles also include clearance of extracellular deposits of A $\beta$ , and microglial receptors TLR2, TLR4, TLR6 and co-receptors CD36, CD14 and CD47 are activated upon detection of the protein. These receptors can also sense pathogen-associated molecular patterns such as bacterial lipopolysaccharides and viral surface proteins and thus are instrumental for mediating the immune response. Certain bacteria have similar surface amyloids, such as curli fibres, which resemble A $\beta$  aggregates and thus activate toll-like receptors (TLR) and CD36, which in turn triggers the formation of a TLR4-TLR6 heterodimer and results in signalling activation via the transcription factor NF- $\kappa$ B. This leads to a cytokine cascade which further attracts immune cells to the site of the perceived infection.



**Figure 5:** Microglial cell diagram showing the formation of the NLRP3 inflammasome and cytokine cascade as a result of A $\beta$  detection. Source: Heneka *et al*, 2015

Moreover, certain cytokines such as IL-1 $\beta$ , damage the synaptic plasticity by disrupting the formation of dendritic spines, with high cytokine expression being able to disrupt normal hippocampus function. This lead to the hypothesis that chronic activation of the immune systems leads to chronic inflammation and microglial cell death, resulting in increased proliferation and accelerated senescence (Heneka *et al*, 2015, Avdic *et al*, 2014).

## Chapter 2: Machine Learning and Data Mining

Advances in bioinformatics have resulted in a vast amount of data being generated at an accelerated pace. Next-generation RNA and DNA sequencing methods are providing access to incredibly detailed information on entire genomes and allowing us to interrogate more potential biomarkers with an increased level of accuracy. This massive volume of data creates a problem of complexity, which makes it impossible for such information to be utilised using traditional methodologies. Machine learning is an interdisciplinary field of bioinformatics which employs a data-driven class of algorithms to find solutions to a given problem by studying specific aspects of data, such as gene expression patterns across many cases/patients. Although widely and successfully used in a multitude of biological and biomarker discovery studies, the use of these approaches to further our understanding of AD have, to date, been extremely limited. Many such approaches have been developed, each of which will be explained in terms of their utility here and can be broadly characterized in two distinct groups; supervised and unsupervised machine learning.

### 2.1 Supervised Approaches

Supervised learning approaches, the mechanisms of which are discussed in greater detail in chapter 3, are widely applied and use source features to predict a target class (Miotto *et al*, 2017). The supervised approach allows the algorithm to train itself by detecting patterns in large data sets that are predictive of the target class. An example would be highlighting the variance at the genetic level between AD and cognitively normal individuals. We can also make use of previous studies and adjust the algorithm parameters so that it accounts for this information, which allows the power of this approach to increase over time and produce more accurate and robust results. One major advantage of supervised learning is that such approaches are tolerant of the highly complex, nonlinear and noisy data that are often found in biological systems.

#### 2.1.1 Artificial Neural Networks

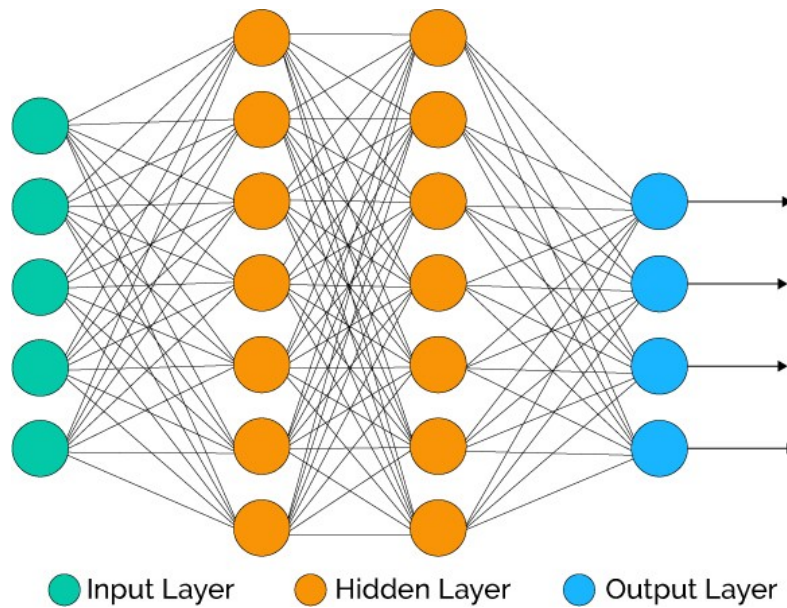
Artificial Neural Networks (ANN) are computational models emulating the function of a network of human neurones for the purposes of encapsulating information in order to analyse large, complex datasets. The learning process is based on the mathematical interconnections between the processing elements that constitute the network architecture

(Chatzimichail *et al*, 2014). This allows them to classify cases based on data by assigning a numerical weight value to each input and adjusting them as they sample the data, effectively learning the optimal solution. The main advantages of ANNs include their high fault and failure tolerance, scalability and consistent generalisation ability, all of which allow them to effectively predict or classify new, fuzzy and unlearned data (Chatzimichail *et al*, 2014, Bertolaccini *et al*, 2017). Additionally, they have been recently used to create panels of biomarkers that can, when used in conjunction with each other, predict diseases such as breast cancer (Abdel-Fatah *et al*, 2016). A basic schematic is represented in Figure 6.

The original ANN architecture, as proposed by Rosenblatt in 1958, was based on the concept of a single artificial processing neuron with an activation threshold, adjustable weights and bias. However, this could only be used for the classification of linearly separable patterns, as it only learns when an error occurs during testing. This is rarely the case with complex problems such as cancer, as patients do not typically fall into a standard distribution and variance in the data is often significant. Typically, ANNs make use of a Multi-Layer Perceptron (MLP) which is made up of multiple perceptrons arranged in layers of three or more, consisting of input, hidden and output layers. These consider the predictor variables, perform feature detection through an activation function and output the results of the algorithm respectively.

ANNs have been successfully used to predict and classify data in different contexts, such as early detection (Mehdy *et al*, 2017), prediction of long-term survival (Huang *et al*, 2017) and biomarker discovery in breast cancer (Abdel-Fatah *et al*, 2016), classification of colorectal cancer tissues (Haj-Hassan *et al*, 2017) and discrimination between benign and malignant endothelial lesions (Makris *et al*, 2017). One of the major disadvantages of ANNs is their liability to overfit when the parameters have not been optimised. Moreover, they often receive criticism for their “black box” approach which allows for little to no interpretation of the results and process. As they are the machine learning method selected for this study, ANNs will be explored in greater detail in the following chapter.

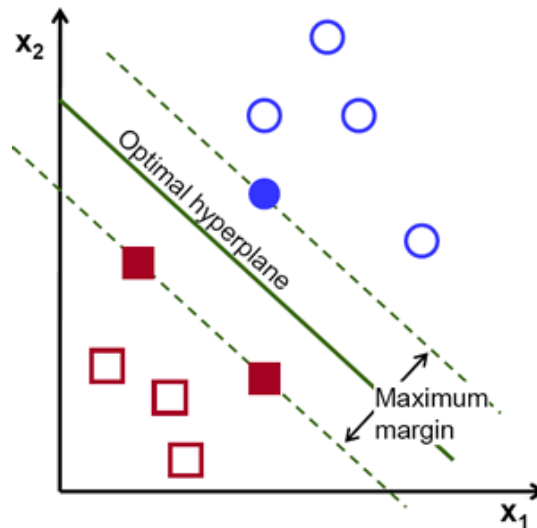




**Figure 6:** Artificial neural network schematic showing the input layer, two hidden layers and the output layer.

### 2.1.2 Support Vector Machines

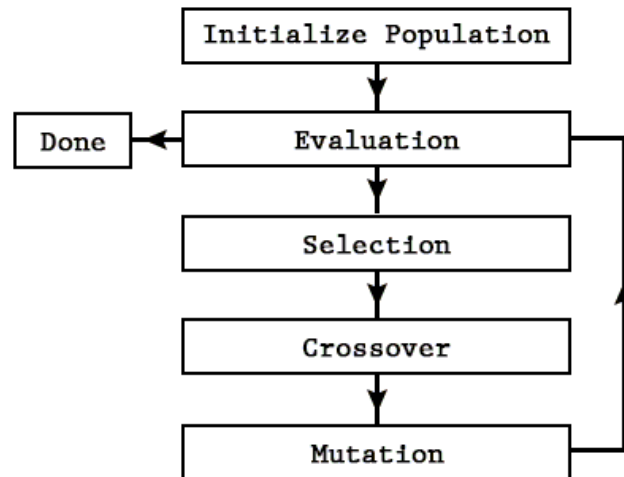
Support Vector Machines (SVM) are supervised learning models that are primarily designed to solve binary problems as shown in Figure 7. They are focussed on finding a hyperplane which separates two classes (van Belle *et al*, 2016) and have been successfully used in pattern recognition and classification. The popularity of SVMs is a result of the availability of a large variety of kernels (functions that separate data) which can be broadly split into linear, polynomial, sigmoid and radial basis function categories. The greatest advantage of SVMs when compared to similar machine learning methods, is that selecting the correct kernel function enables the analysis of non-linear data and overcomes the curse of dimensionality. However, the introduction of more features increases the complexity, and therefore the computing power required. Notwithstanding the practical issues, SVMs have been used for analysing high density data, such as RNA, miRNA and proteomics, and they remain one of the most popular classification methods, especially for cancer prediction and prognosis (Powell *et al*, 2017, Araujo *et al*, 2017, Huang *et al*, 2017).



**Figure 7:** Support vector machine schematic showing the optimal classification between two data series. The hyperplane determines at which point the data is separated in a 2D space.

### 2.1.3 Genetic Algorithms

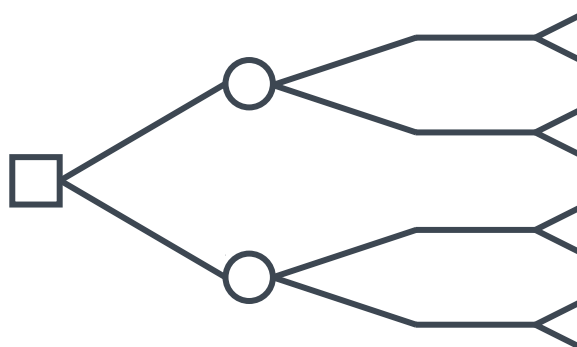
Genetic algorithms (Figure 8) operate on a concept similar to how genetics influences survival of the fittest in nature. Instead of adjusting the weights of the algorithm to train it so that it learns the optimal solutions after a number of iterations, genetic algorithm function by creating a number of solutions encoded as binary strings, each with its own properties and referred to as chromosomes (Srinivas and Patnaik, 1994). From these chromosomes, the algorithm keeps creating random solutions and assigns a fitness score to each of them, with the highest score indicating the optimal solution for the presented problem. The top solutions identified this way, however, instead of being discarded are randomly selected and modified to create a new generation of solutions by a process known as mutation and crossover. This process is repeated until convergence is achieved after the optimal number of generations (McCall, 2005), in an iterative manner. The downside of genetic algorithms is their tendency to converge on the local instead of the global minima, explained in the following chapter, and can be both time consuming and computationally expensive which makes them a non-ideal solution to biological problems.



**Figure 8:** Genetic algorithm flowchart. The schematic is similar to the ANN shown in Figure 6 but with a single output

### 2.1.4 Decision Trees

Tree-based methods involve stratifying a data set into multiple categories (similar to hierarchical clustering) that can then be used to predict possible outcomes based on the values of the input variables. These methods can be used for both classification and regression problems. Decision tree classification algorithms pose a series of questions based on the features of the data set and train to split those features into separate categories, thereby resulting in a dendrogram (Figure 9). Although the advantages of these methods are that they are computationally efficient, have good predictive values, and their results are easy to interpret, their predictive accuracy tends to be lower than their counterparts. To mitigate this issue, methods such as random forests, bagging and boosting are used to construct multiple trees in parallel. These can then be combined to provide a significant boost to their prediction accuracy at the cost of some of their interpretability (James *et al*, 2015).



**Figure 9:** Decision tree schematic showing four clusters originating from the original population.

### 2.1.5 Bayesian Networks

A more recent development in machine learning is the application of Bayes' theorem to create probabilistic graphical models, where the association between a set of variables or nodes can be determined through joint conditional probability distributions (Jiang *et al*, 2010). Bayes' theorem states that the conditional probability of A given B is the conditional probability of B given A scaled by the relative probability of A compared to B. Using Bayesian networks, the association between a set of variables or nodes can be determined through joint conditional probability distributions (Zeng *et al*, 2016). Static Bayesian networks are directed acyclic graphs where each node represents a stochastic variable and arcs represent the probabilistic relationship between a node and its parents, but cannot infer the direction of any given interaction, a feature that is essential in biological networks. However, dynamic Bayesian networks can be cyclic graphs by representing all variables at multiple points in time and drawing edges from variables at an earlier point to a later one. This allows them to infer direction of causality as well as process temporal data, features that are common in biological data.

Although such approaches have been used for multiple biological applications such as inferring cellular networks, modelling protein signalling pathways, data integration, genetic data analysis, and classification (Zhu *et al*, 2017, Field *et al*, 2015, Luo *et al*, 2017), they are limited by the fact that they need larger than average data sets to obtain sufficient prior probabilities to produce an accurate outcome. This in turn makes them extremely computationally expensive. Moreover, they tend to perform poorly on high-dimensional data and their output tends to be complex and as such, can be hard to interpret for non-specialists. Finally, it should be noted that Bayesian networks are not truly Bayesian in nature. They simply adhere to the basic rules of Bayesian statistics on probabilistic inference. It would be more accurate to say that Bayesian networks are directed graphical models with Bayesian elements.

## 2.2 Unsupervised approaches

Unsupervised machine learning approaches are used when the desirable or predefined output is not available. The goal of unsupervised learning problems is to discover the structure of the data and define groups of similar examples, commonly called clustering (Bishop, 2006). Clustering is one of the main unsupervised approaches and it functions

by assigning data points to natural categorical classes or groups, based on similarity or difference of patterns without prior training (Sommer and Gerlich, 2013). Unsupervised learning approaches are best used when the subject is a very large data set with few known variables. This allows the user to find natural patterns in the data and discover novel groups that have not been previously established and using which training can be undertaken. They have been most commonly used to distinguish patterns in microarray data by clustering genes based on their expression levels (Stadler *et al*, 2017, Athreya *et al*, 2017, Vural *et al*, 2016).

However, even though unsupervised approaches tend to be unbiased, biological data tend to show a lot of variance, which in turn leads to less robust results. Moreover, the time required to analyse the results presented by these algorithms is disproportionately large compared to supervised learning approaches, as experimenting with the algorithm parameters and comparing multiple results is required to achieve the desired outcome.

### 2.2.1 Hierarchical Clustering

Hierarchical clustering, the most common unsupervised learning technique, has been widely used for the analysis of microarray data. It is based on measuring distances between data points and defining the first instance of each point as a single cluster, followed by merging the clusters according to distance, with smaller distances between clusters indicating greater similarity. The process continues in an iterative manner until all samples have been used to produce a phylogenetic tree-like structure of the clusters (dendrogram), with individual samples at the bottom, and a cluster containing every element in the data set at the top (Sommer and Gerlich, 2013). Some of the most popular methods to determine cluster hierarchy include Single-linkage, Complete-linkage, Average-linkage, and Centroid distance.

Hierarchical clustering can be implemented using aggressive or divisive approaches. Aggressive methods start with the assumption that each object belongs to a unique cluster, followed by measuring the distance between them and merging in an iterative manner. Divisive methods on the other hand, start by grouping all samples into one cluster, followed by randomly generating clusters and assigning them to vectors, and then sorted by similarity. The process is repeated by redefining the vectors and attempting to reach

convergence. Due to the inefficiency of this approach, it is rarely used in biological tests where the size and complexity of data are major factors.

The major limitation of the hierarchical clustering approach is that as the clusters grow, they might not be representative of the objects within, and it is hard to rectify mistakes that occur early in the clustering process. Furthermore, hierarchical clustering is especially vulnerable to the curse of dimensionality, with datasets with over 50 dimensions, well below the average for biological datasets, having such small distances between the points that the feature space becomes uniform and it is impossible to meaningfully separate the data.

### 2.2.2 K-means Clustering

Much like hierarchical clustering, K-means clustering is a partition algorithm which works by arbitrarily grouping objects into a predetermined number of clusters in an iterative manner. The centroid-average expression of each cluster is assigned randomly, based on the Euclidean distance between each object and the closest cluster average. The algorithm then recalculates the average centroid expression, based on the mean of all objects assigned to it, and repeats the process until convergence is reached, where the average expression of each cluster does not change significantly (Sommer and Gerlich, 2013). Unlike hierarchical clustering, this method has the advantage of being able to deal with large data sets and as a result has been applied to more complex problems. However, the major drawback of this method is that repeating the test can produce significantly different results, as the final assignment of clusters is dependent on the initial random assignment of objects (Rodriguez *et al*, 2014).

### 2.2.3 Principle Component Analysis

Reduction in dimensionality is often necessary for a visual inspection of high-dimensional data, as the number of variables being investigated often exceed the number of samples. This leads to data points being scarcely distributed in a high dimensional feature space (Xanthopoulos *et al*, 2013). The aim of principle component analysis (PCA) is to map the original data into its principle components by linearly transforming the data to reduce dimensionality. These principle components are orthogonally arranged, mutually uncorrelated linear combinations of the original variables, and are often ranked by the

amount of variance they can explain in the data. The highest ranked components contain most of the relevant information, whereas low ranked principle components can be removed if they are not required. This approach is often used as a visualization tool and pre-processing step for classification and clustering (Sommer and Gerlich, 2013). PCA belongs to the linear procedures family and uses a Pearson correlation matrix to seek linear combinations for the highest variance. Essentially, it uses non-linear dimensionality reduction methods.

## 2.2 Biomarker Discovery

These approaches are excellent tools for biomarker discovery and validation, especially when applied to novel questions, but care should be taken to select the appropriate methodology. In recent years, the fallacy of attempting to discover a single best marker that can be used to attempt to predict and treat a disease, regardless of the personal circumstances of the patient has been made apparent due to the failure of such approaches in conditions like AD, cancer, AIDS or diabetes (Barabasi and Oltvai, 2004). While treatments for some of them are available and the quality of life of the patients keeps increasing, it is due to the rise in complexity of available research avenues. The tools and goals used to explore these possibilities are biomarkers. A biomarker is defined as a characteristic that is objectively measured and evaluated as an indicator of normal biological, pathogenic or pharmacological processes to therapeutic intervention (Kennedy *et al*, 2016). The types of biomarkers are rather diverse and usually closely linked to the processes of the disease studied, although they can be broadly categorised as prognostic and predictive.

Prognostic biomarkers are used to estimate disease outcome for the patient, usually in terms of survival. The presence, absence or levels of this marker can be used to determine a differential outcome which reflects the disease's underlying biology, history and ongoing status (Ballman, 2015). Predictive biomarkers tend to be used to determine patient response to therapy and are commonly used in drug development to select the patients most likely to benefit from a new drug or treatment. Quality predictive markers require at least two highly differential groups to make this comparison possible. As a result, the genes, proteins or RNA selected as potential biomarkers have to be able to be used in the aforementioned situations. However, when combined with the failure of the

“one size fits all” approach that was common in the recent past, it is all but essential to generate a panel of such markers. Where a single marker is not enough to differentiate between two closely linked by distinct conditions, which is very common in AD, a combination of such factors would provide both the required sensitivity and specificity to the question presented.

When discussing biomarker discovery, consideration must be given to the false discovery rate, as with enough testing there will be a number of false positive results. This has resulted in the need for false discovery rate correction methods such as the Benjamini and Hochberg FDR-controlling procedure, which functions on the basis of testing a number of hypotheses based on their respective p-values. Considering that a fraction of these discoveries will be false, by sorting their p-values in ascending order, and assigning each hypothesis to a corresponding p-value, the smallest acceptable value can be determined. It then becomes possible to reject all other hypotheses, thus controlling the rate of false discovery.

### 2.3 Systems Biology

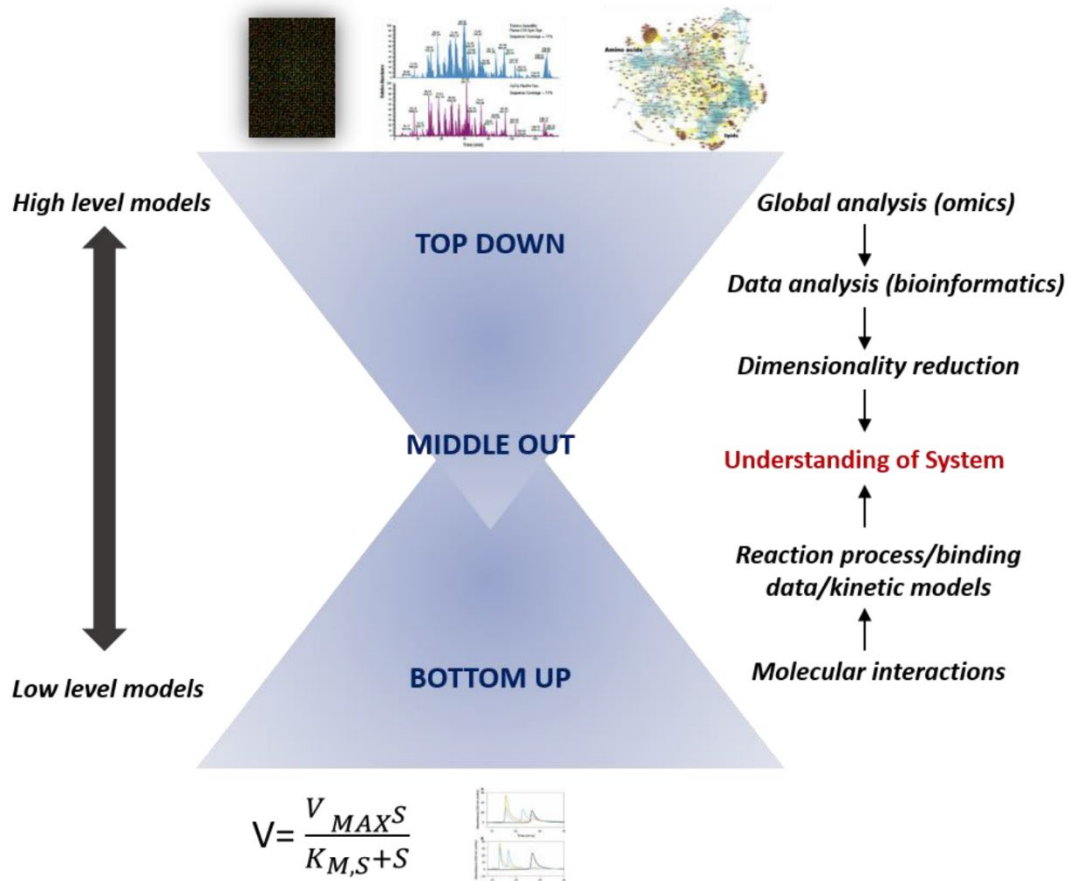
Of course, manually testing and analysing every single factor present in a condition as complex as AD would be costly and time consuming at best and impossible at worst, necessitating understanding of the key dysregulated systems and their functions to allow for accurate pre-processing of the available information before focusing the research on a specific question. To achieve that, the need for systems biology is becoming stronger. Systems biology is a holistic approach to biological systems using mathematical and computational modelling with the goal of understanding not simply the nature and function of individual components of a given system, but how their interconnectivity affects the whole system.

The greatest advantage system biology allows for, is the use of non-parametric, non-reductionist approaches which are capable of considering all possible parameters in a given question and reaching an unbiased, novel conclusion. Reductionist approaches rely on breaking down a system in its base components, understanding them and piecing together the answer from these parts, and although historically commonly used and successful, such approaches suffer from an increase in bias. Additionally, it is becoming



evident that biological systems are more complex than first realised and tend to be more than the sum of their individual components (Barabasi and Oltvai, 2004), the discrediting of the “one gene to one protein” theory being a prime example.

Systems biology is a holistic approach to the study of biology processed in a quantitative manner using computational and mathematical modelling to simulate complex biological systems (Kohl and Noble, 2009). It benefits from advances in the fields of machine learning, omics-based technologies such as genomics and proteomics, and other computational tools and is able to analyse complex molecular interactions in a timely and cost-effective manner. Moreover, the ability to designate the desired system(s), such as the genome in a disease or a specific process like A $\beta$  clearance, rather than relying on predefined ones, and observing the interactions within and between them allows for a more varied, less biased approach to biomarker discovery. There are two major approaches to systems biology, related to the starting point; the top down and bottom up approaches (Figure 10).



**Figure 10:** Schematic representation of the different approaches to Systems Biology adapted from Nielsen and Jewett (2008). There can be three different starting points to studying biological systems, these are: top down, bottoms up and middle out systems biology approaches. These modelling approaches provide a quantitative description of a system. The top down approach uses genome-wide experimental data based on clinical phenotypes, which is produced and analysed to identify the molecular mechanisms, networks and structures within the pathway. In contrast, a bottom-up approach, starts from the kinetic and enzymatic interactions between the different parts of a system to better infer the clinical properties of the system. The middle out approach is the combination of the two and can start at any level where information is available. Source: Agarwal (2017)

### 2.3.1 Top down approach

Anthony *et al* (2012) have defined the top down approach to systems biology as a classic physiology-based approach, beginning with modelling the clinical signs to the molecular processes, which is the current dominant approach due to its compatibility with -omics technologies. The data collected from such technologies can be in the form of DNA microarrays, RNA sequencing, methylation arrays and other technologies (Schena *et al*, 1995, Wang *et al*, 2009, Shahzad and Loor, 2012). This approach can be used to discover novel molecular drivers, markers and pathways and due to the genome-wide transcriptomic information present in such data it can produce results that can be further analysed in wet lab experiments (Khol *et al*, 2010). The advantages of this approach include its ability to use real data of multiple types, such as metabolomic, proteomic or

transcriptomic data, and allow for an unbiased, but focused discovery methodology. However, it should be noted that the complexity of the data obtained this way tends to be high, resulting in the need for equally complex, advanced, and often non-linear and expensive machine learning methods. Additionally, Kohl and Noble (2009) also argue that it can be challenging using this approach to analyse specific phenotype aberrations, and, while the results are of high quality, it can be hard to determine the details of the mechanisms in the pathways/drivers discovered this way. The top down approach however, remains the most widely used systems biology approach due to its robust nature, high quality of results and inherent advantage as a discovery method for novel markers.

### 2.3.2 Bottom up approach

Conversely, the bottom up approach relies on an integrative view of all biological interactions, obtained via a complete genome model for a specific organism (Shahzad and Loor, 2012) in an effort to determine all possible interactions taking place in a living system. Also known as the forward approach, it has been the historical standard for systems biology until recently (Kohl *et al*, 2010) and has been used to infer both functional and clinical properties as obtained via molecular methods in specific subsystems (Bruggeman and Westerhoff, 2007). The mathematical models used in this approach are constructed based on the following four steps, the first being draft reconstruction, where data are collected using bioinformatics methods from specific databases such as NCBI for genomic data, UniProt for proteins and KEGG for pathways. In this step selection of specific organisms or subsystems also takes place and is followed by manual curation of the collected data. This leads to the second step where unnecessary information is removed, gaps resulting from missing data are filled and the new cohort is validated to ensure quality. This cohort can be used in the third step, where mathematical software tools, such as Matlab, SBML (Hucka *et al*, 2003) or custom programs based on linear or quadratic languages, are used to analyse it. Finally, in the last step, the network is validated, by checking for inconsistencies in the results using defined objective functions, and the draft in the second step is readjusted in case of failure (Shahzad and Loor, 2012). This approach has been dominant in mechanistic studies where the expected parameters are more limited but has proven to be unable to cope with the more complex kinetic parameters common in most eukaryotic organisms (Kohl *et al*, 2010).

### 2.3.3 Middle out approach

More recently, in an effort to combine the advantages of the aforementioned approaches whilst minimising their disadvantages, the middle out approach has been gaining traction. Since the two most common approaches currently can be explained as a topology driven approach, stipulating that the relative strength of the main driver is unrelated to the presence or absence of a link, and the transient functional approach which insists on the relevance of kinetics, the middle out approach advocates starting from the relations between the parts in the system (Giuliani *et al*, 2014). By focusing on that relationship as a starting point, it becomes possible to interrogate the data further using methodologies of higher or lower complexity as required, making it ideal for models containing data of multiple scales, such as incorporating transcription, translation and proteomic data in a single analysis, or expanding to multiple organs or tissues and analysing the interactions between them.

## 2.4 Study Aims

As discussed in Chapter 1, there are multiple concerns in the area of Alzheimer's disease research, including but not limited to a lack of understanding of the underlying causes of the disease, lack of clarity on the systems involved in its progression and missing links in the amyloid cascade hypothesis regarding neurofibrillary tangles. Additionally, the current biomarkers for AD have consistently proven to be inadequate due to their lack of sensitivity, specificity and usability in a clinical setting. These facts necessitate a comprehensive study of the disease in an unbiased, robust and efficient manner with the goal of identifying novel markers, drivers and achieving a greater understanding of the mechanisms involved in the disease. Therefore, the aims of the current study are as follows:

1. Use an ANN based integrative data mining and systems biology approach, utilising non-linear, non-reductionist classification and interaction algorithms that can provide statistically significant results and analyse high-dimensional data in a timely and cost-effective manner
2. Identify a number of representative, robust and clinically relevant datasets for use with the aforementioned algorithms. They should include multiple data types, prioritising DNA expression to analyse the genetics of AD pre-transcription and

RNA sequencing data to identify differential genes post transcription but pre-translation. Data from both AD and cognitively normal individuals should be included.

3. Analyse the selected cohorts in a hypothesis-free non-parametric manner to minimise the bias and variance. The methods used at this stage are based on the work of previous PhD students Lee Lancashire, Cristophe Lemetre and Devika Agarwal (Lancashire (2006), Lemetre (2010), Agarwal (2017)).
4. Expand the methodology developed by previous students. Develop a new method to analyse interaction results that can eliminate bias, increase readability of results and allow for further dataset deep mining. Extend the limits previously accepted for the interaction algorithm to obtain a greater number of possible biomarkers.
5. Incorporate the newly developed continuous ANN algorithm into the methodology and use it to compare novel results to previously established markers.
6. Evaluate the significance and quality of the results obtained via this methodology. The biological relevance of the predicted markers cannot be fully established without wet lab validation, which is beyond the scope of the study. Results validation will be based on gene ontology, relevance to AD and consistency across analyses, conditions and datasets.
7. Develop and discover tools to improve the speed, efficiency and accuracy of the methodology based on statistical packages. The goal was to minimise and eventually eliminate human error.

# Chapter 3: Artificial Neural Networks

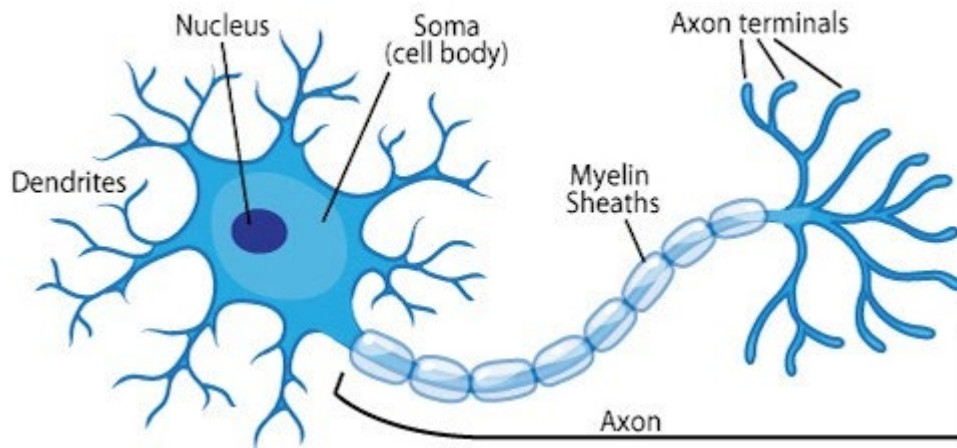
## 3.1 Introduction

As explained previously, ANNs are a form of machine learning, statistical models emulating the function of a neuron, able to identify patterns and linearly separate them by assigning a numerical weight value to each input and adjust them as they sample the data, effectively learning the optimal solution. They can make use of parallel processing in order to predict solutions to complex and non-linear data (Lancashire *et al*, 2009). ANNs are highly fault and failure tolerant, scalable and have consistent generalisation ability, allowing them to predict or classify well for new, unlearned data (Livingstone, 2008, Lancashire *et al*, 2009), which is the focus of this project.

The ANN used for this project is a Multi-Layer Perceptron (MLP) with a back-propagation (BP) algorithm. It is organised in several layers, each with a number of mathematical processing elements depending on the complexity of the problem and the BP algorithm is responsible for feeding the error back through the model, allowing it to adjust the training weights accordingly and stop early if no gains can be made. Although multiple methodologies were considered by my predecessors (Lancashire (2006), Lemetre (2010), Agarwal (2017)), the ANN was determined, after rigorous testing (Lancashire *et al*, 2009) to be the most efficient way to perform hypothesis-free biomarker discovery. It has been shown to have the highest levels of accuracy and predictive power as well as being cost-effective. Furthermore, after years of optimisation it is now possible to focus on expanding the methodology without a constant need to tweak the network parameters.

## 3.2 Historical Background

The ANN and the logic governing it, as well as most biological networks, have their roots in the structure and function of a human neuron (Figure 11). The human brain is effectively a compact, energy efficient parallel processing biological network. This network functions thanks to the interactions between receptors, that transmit the external stimuli received to the brain and effectors which transform these signals into external responses (Haykin, 2009).



**Figure 11:** Schematic diagram of a biological neuron. The understanding of the neuron at the time was very limited. Source: becominghuman.ai

As shown in works by Basheer and Hajmeer (2000) and Nelson and Illingworth (1991), early theories in theoretical neurophysiology as well as neuromathematics and neurocomputing were established between 1890 and 1949 that allowed neuroscientist Warren S. McCulloch and logician Walter Pitts (McCulloch and Pitts, 1943) to introduce computing elements based on the properties of neurons and their synapses into neuronal studies in an effort to understand how the nervous systems worked. This was the beginning on ANN modelling with their observations leading to the description of the first artificial neuron called the Threshold Processing Unit (THU) or McCulloch and Pitts neuron. The theory behind the THU relies on the following assumptions:

1. Neurons are binary and can only be set to one of two states at a time, an “all or none” process.
2. Each neuron has a fixed threshold which does not change over time
3. A neuron can receive inputs from excitatory synapses which all have identical weights
4. A neuron can receive inputs from inhibitory synapses which prevent the neuron from being activated.

If these assumptions are true, this would require the neuron to only be able to process simple logic functions such as **INCLUSIVE OR, OR, OR BOTH** or **AND**. Thus, a neuron can only exist in one of two states  $y$ , which can be activating (1) or inhibiting (0) with a threshold  $\theta$  to define the state. The neuron receives multiple signals ( $n$ ) from inputs ( $x_i$ , where  $i = 1 \dots n$ ) that are weighted by a fixed value  $w$ , which is either +1 or -1 and finally emits an output signal. Based on the “all or none” concept, a neuron could only

fire if the weighted sum of the input vector exceeds the predefined threshold, which results in the following rule:

$$y = \begin{cases} 1, & \text{if } \sum_{i=1}^n w_i x_i \geq \theta \\ 0, & \text{if } \sum_{i=1}^n w_i x_i < \theta \end{cases}$$

This concept was evolved by Rosenblatt (Rosenblatt, 1958) into a fully-fledged neural network by connecting single layer neurons in parallel, in an attempt to explain perception using the retina as a model. This led to the development of the perceptron, which possessed the ability to learn by adjusting the network weights, initially in a stochastic manner, and then by altering the connections between neurons. Due to the limitations of the perceptron (Basheer and Hajmeer, 2000; Hecht-Nielsen, 1988), its use declined with research becoming more focused on artificial intelligence. The perceptron is examined in greater detail in section 3.3.1.

The seminal moment in the decline of ANNs in research was a book published by Minsky and Papert (Minsky and Papert, 1960) which made the limitations of the perceptron clear. Although these limitations were known to the scientific community, Minsky and Papert stress tested the algorithm, clarifying the magnitude and importance of these issues, especially when tested on non-linearly separable data. It is worth noting that although they are quoted as saying that “our intuitive judgment that the extension to multilayer networks is sterile”, they failed to fully understand the capabilities of multi-layered networks, believing that the limitations of the perceptron would extend to these as well. However, when research by John Hopfield in 1982 and 1984 (Hopfield, 1982, Hopfield 1984) led to the introduction of the Hopfield network, followed by Rumelhart, Hinton and Williams (Rumelhart *et al*, 1986) publishing their Back-propagation algorithm based on the work of Widrow and Hoff’s on the Delta rule (Widrow and Hoff, 1960), interest in the field of ANNs was rekindled. The use of the Back-propagation algorithm in conjunction with multi-layer networks has allowed ANNs to become some of the most popular and successful machine learning algorithms for scientific research.



### 3.3 Architecture

#### 3.3.1 Perceptron

The perceptron (Figure 13) was first devised by Rosenblatt in 1958 and is the simplest ANN architecture available. It can only be used to classify linearly separable problems and consists of a single processing neuron with an activation threshold, adjustable weight and bias (Rosenblatt, 1958). Learning transpires during training when an error occurs and the network parameters are adjusted. It is worth noting that Rosenblatt proved that learning converges after a set and finite number of iterations depending on the complexity of the problem.

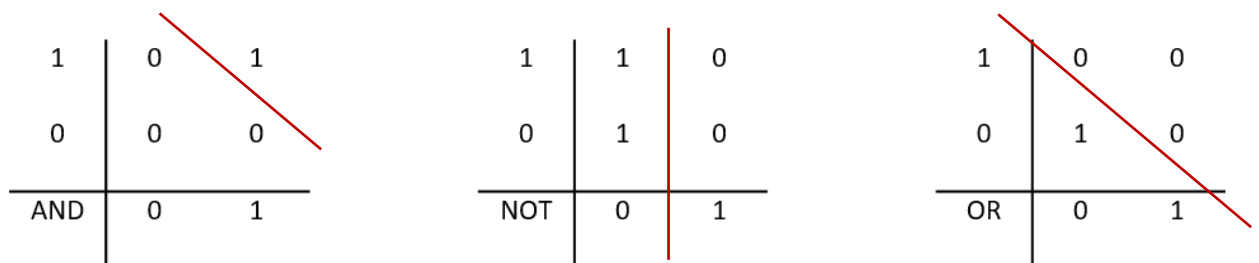
Essentially, the perceptron calculates the weighted sum of all input values and assigns the points in the region of the plane corresponding to whether the value calculated exceeds a predefined threshold. The summed input is given by the following equation.

$$y = f\left(\sum_{i=1}^m w_i x_i + b\right)$$

Where  $w$  is a vector of weights and  $x$  is the input and  $b$  is the external bias. When the value is exceeded as shown in the next equation

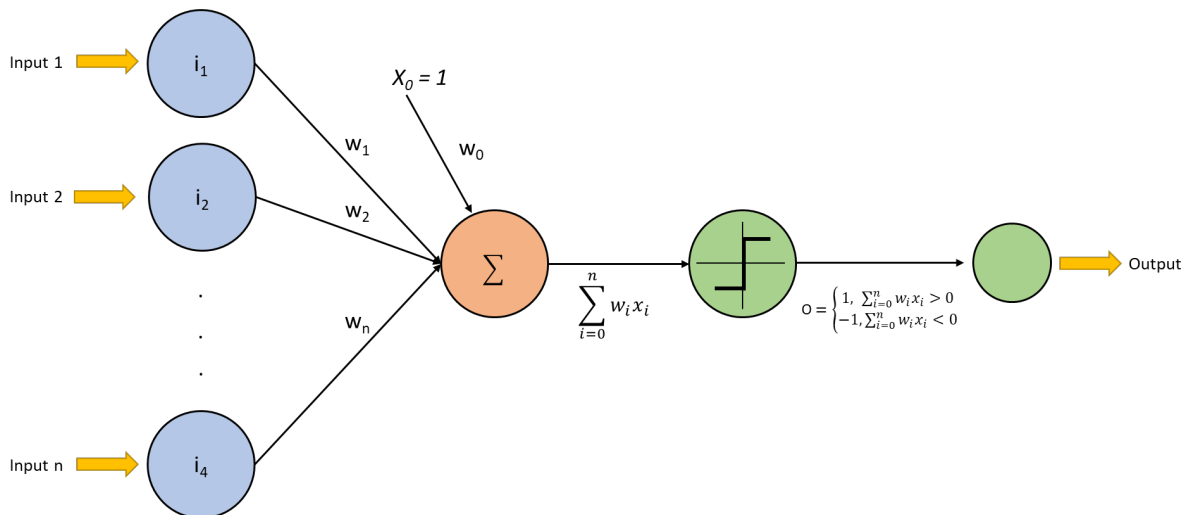
$$f(x) = \begin{cases} 1, & w \cdot x + b > 0 \\ 0, & w \cdot x + b < 0 \end{cases}$$

the neuron activates and classifies that input as either 1 if positive or 0 if negative. This allows the algorithm to classify points based on linearly separable operators such as **AND**, **OR** or **NOT** (Figure 12).



**Figure 12:** Examples of linearly separable problems

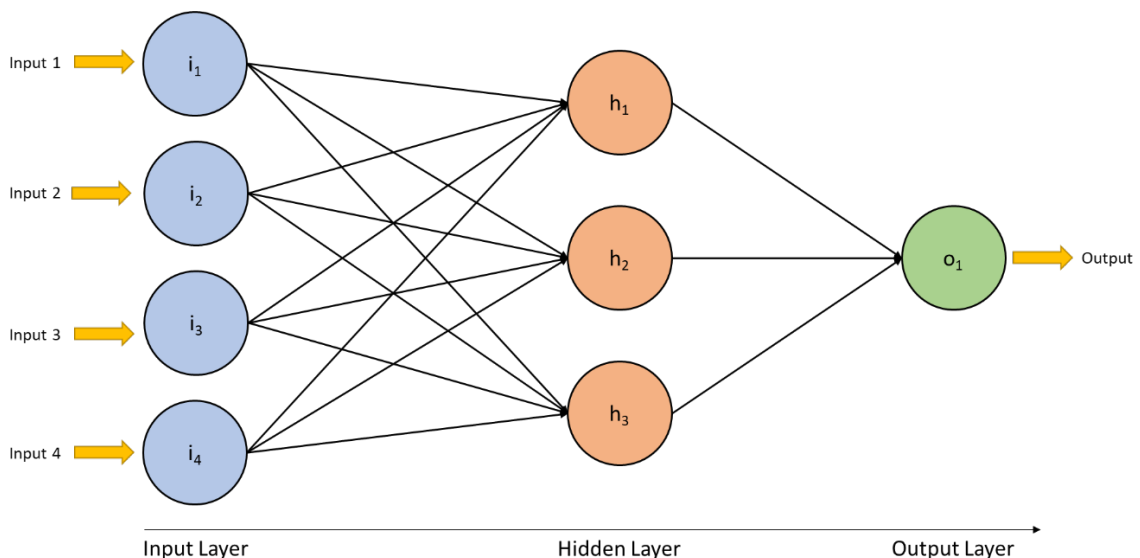
This can only be used for two-class classification problems.



**Figure 13:** Schematic diagram of the Rosenblatt single layer perceptron showing the calculation for each node.

### 3.3.2 Feedforward Neural Networks

Feedforward neural networks (FNN) are one among the first examples of artificial neural networks in practice. They are composed of multiple perceptrons arranged in layers with the first layer taking in the inputs, the middle or hidden layer, which is not connected to any external influence and the output layer producing the results. Each perceptron in each layer is connected to each perceptron in the next layer but not with any perceptrons in their own, thus feeding the information constantly forward.



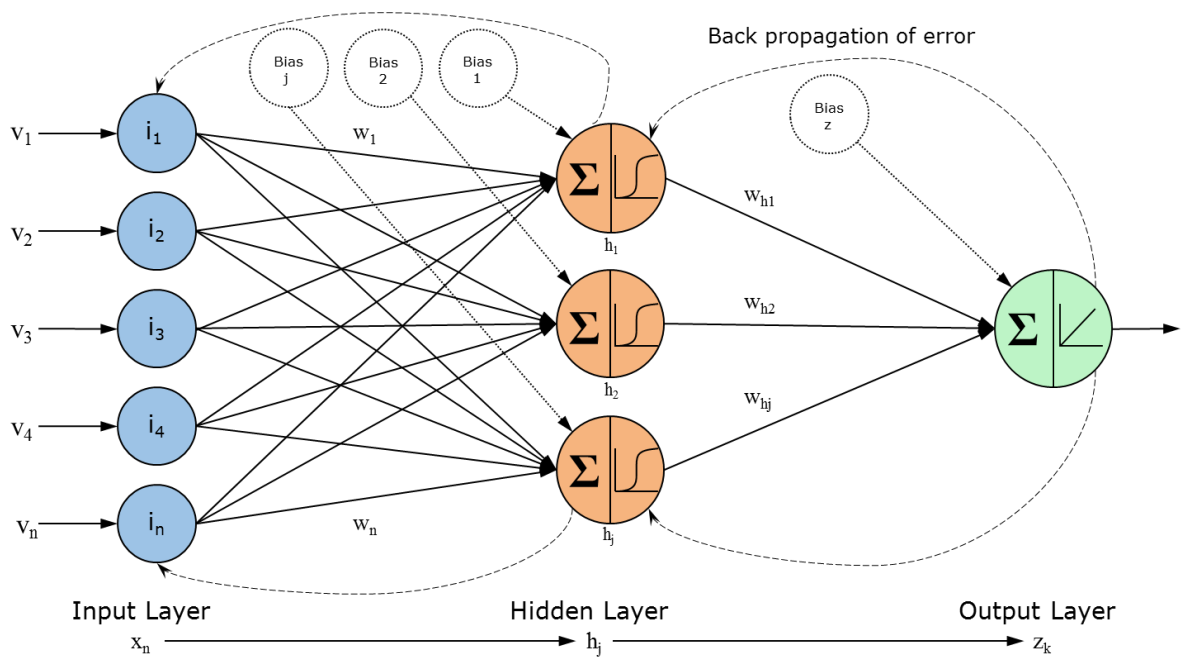
**Figure 14:** Simplified example of a feedforward neural network with a single hidden layer

FNNs (Figure 14) are able to classify data in an arbitrary number of dimensions and groups by simply varying the number of inputs and number and size of the hidden layers,

and as a result have been used to solve classification problems. However, as shown in studies by Minsky and Papert (1998), they are unable to model nonlinear classification problems, an issue they share with the single layer perceptron. To overcome this problem, the multilayer perceptron was developed.

### 3.3.3 Multilayer Perceptron

Much like the FNN discussed above, the Multilayer Perceptron (MLP) is a neural network composed of a series of perceptrons arranged in at least three layers linked by weighted connections. Their input, hidden and output layers perform the same function, by taking in the vector of predictor variable as input nodes, a hidden layer to act as a feature detector and an output layer to generate the output signals. The added bias is treated as an additional weight with the purpose of adjusting the activation function. It is important to note that the hidden layer does not interact with any external factors and the number of nodes present can be increased in order to solve more complex problems (Lancashire, 2009). A schematic of the MLP used for this project is shown in Figure 15.



**Figure 15:** Workflow diagram of the artificial neural network algorithm developed by Lancashire *et al* (2008). Notice the sigmoid function in the hidden layer nodes and the linear separation in the output layer node. This allows the algorithm to create accurate but easily readable results.

A significant advantage conferred by the MLP is its ability to process complex, non-linear data. This significant advantage over a single perceptron lies in the fact that single layer

perceptrons are not able to adjust their weights in response to training. The MLP relies on a series of functions, represented in figure 15 inside the nodes, which perform the necessary tasks to run the algorithm. These activation functions, usually found in the hidden layer nodes, are used to convert input signals to output signals by turning themselves on and off depending on whether a signal exceeds a predefined threshold as explained in section 3.3.1. They can be used to normalise the range of values used for output and to readjust the network weights. The two most common activation functions are the **Logistic Sigmoidal** function and the **Hyperbolic Tangent**.

The **logistic sigmoidal** function calculates the linear combinations of all inputs which results in a non-linear output defined as

$$f(o) = \frac{1}{1 + e^{-x}}$$

which normalises the range of the output between 0 and 1. Because of this, it is also known as a squashing function and is commonly used due to its ability to maintain a good balance between linear and non-linear information, allowing it to be used for both single and multi-layered networks (Bishop,1995).

The **hyperbolic tangent** function is very similar to the logistic sigmoidal function but the range of the output is set to between -1 and 1 and is defined as

$$f(o) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

In both these equations  $e$  is Euler's number and  $x$  is the input data.

### 3.4 Learning Rules

Learning rules as applied to ANNs, are algorithms designed to improve the network's performance. Considering that neural networks are trained to classify data by adjusting the weights in an iterative manner, the goal of the learning rule selected is to minimise the error over the training cycles (Agatonovic-Kustrin, 2000). Most learning rules belong in one of two categories; supervised or unsupervised learning, with the most common learning rule currently in use being the **Delta rule**, or **Back Propagation** algorithm.

### 3.4.1 Supervised and Unsupervised Learning

As discussed in chapter 2, supervised learning approaches use source features to predict a target class, whereas unsupervised learning approaches are preferred when a predefined output is not available which also results in them using similar but distinct sets of learning rules.

In supervised learning, since the input and output for all cases is already provided and the algorithm attempts to find the optimal solution to the question given by adjusting the network weights during training (Lancashire *et al*, 2009) it lends itself to learning rules related to weight modification. Such rules include the **Hebbian rule**, first described by Donal Hebb in 1949 which aims to identify how the neurons interact with each other and uses this information to adjust weights (Hebb, 1949). A common function of this rule is to decrease neuron weights when two neurons are activated simultaneously, decrease the signal when they have opposing activations and maintain it when there is no correlation. Other rules include the **Perceptron rule**, which randomly assigns a value to each weight, calculates the resulting error, and repeats the process, adjusting network weights depending on whether the error increases or decreases. While this method can be used for unsupervised learning, allowing the user to set the parameters makes it well suited to supervised learning. Finally, the last major learning method is the **Delta rule** developed by Widrow and Hoff, used for this project and fully analysed in section 3.4.3.

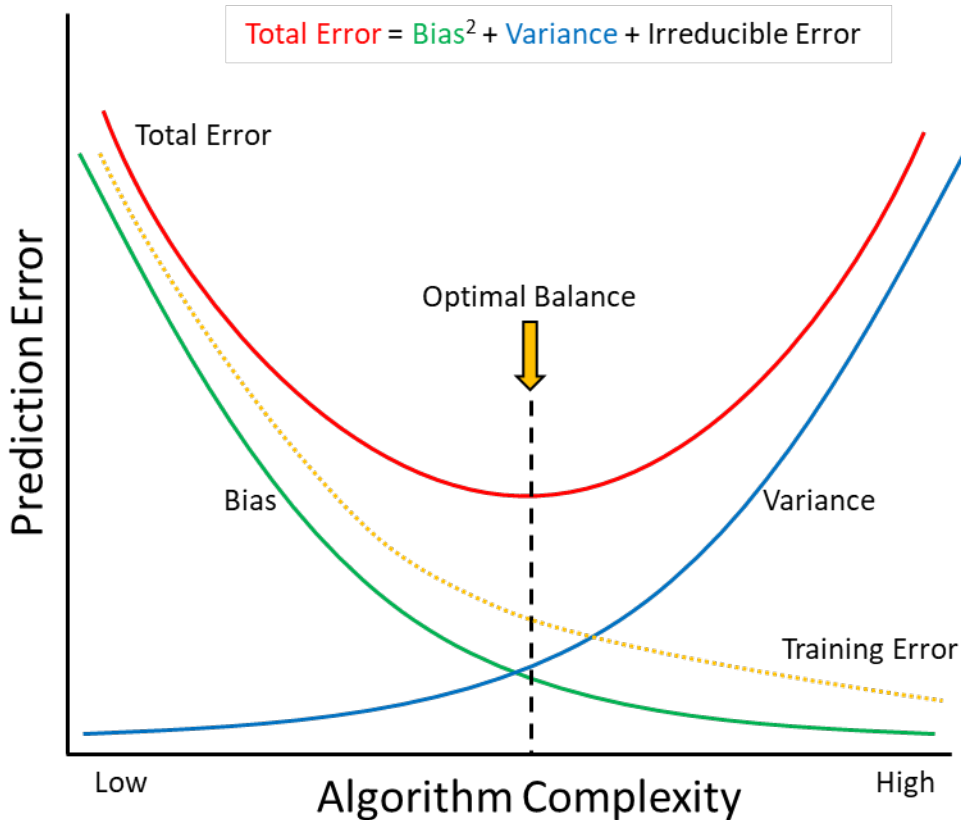
Unsupervised learning conversely uses significantly different rules due to its nature. Since unsupervised approaches are based on the fact that the user provides very little to no guidance to the algorithm, relying on the data patterns to provide an answer, classification occurs based almost entirely on the input data. Of course, unsupervised approaches can still make use of the Perceptron and Delta learning rules modified to accept no information based on expected output.

### 3.4.2 Bias-Variance Trade-off

Before proceeding to the next section, it is crucial to understand one of the core underlying principles of machine learning; the bias-variance trade-off. The most common problem encountered with all forms of supervised learning, and ANNs are no exception, is that of optimisation. The performance of an algorithm and the quality of the results is measured by the error and the goal is to identify key features within the feature space based on a

predefined predictor as measured by the deviation from the training set. The ideal model generalises well on unseen data rather than training data, and good performance on the training set does not imply equivalent performance on the test and validation sets.

This problem is solved by adjusting the complexity of the algorithm. In essence, we have to make a choice and find the balance between simplicity and complexity. In practice however, the choice is not obvious. Too simple a model will lead to a significant increase in bias, causing the model to perform badly on the training set and even worse on unseen data rendering the results invalid, whereas too complex a model, such as an ANN with a hidden layer composed of hundreds of neurons, leads to a drastic increase in variance. This leads to an almost perfect performance on the training set and near complete inability to predict unseen data. This occurs because the algorithm matches the training data perfectly, which includes the irreducible error representing the noise, which is significant in biological data, and the complexity of the model itself will lead to noise, matching the training data so well that every other model is suboptimal by comparison, resulting in a significant increase in error (Geurts, 2010). Since complexity increases exponentially by the addition of non-linear classifiers, the bias-variance trade-off explains the lack of a universally optimal learning method.



**Figure 16:** Graphical representation of the bias-variance trade-off. The optimal balance in the system is at the point where the error is at its lowest in order to achieve the lowest possible bias and variance without compromising the quality of the results. This is not simply the point when bias and variance intersect.

It should be noted that, unlike common parlance, in the context of machine learning bias is defined as the squared difference between true conditional probability of a feature being correctly identified and the prediction of the classifier averaged over the training set. As a result, bias increases if the classifiers are consistently wrong and falls if they are consistently right. Variance on the other hand, is the variation between the prediction of the learned classifiers and measures the consistency of the classification of each feature (Hastie *et al*, 2009).

To calculate that error, it is essential to calculate the bias and variance of each component. Considering we are assuming a relationship between predictor  $Y$  and covariates  $X$ , the relationship can be represented as  $Y=f(X) + \epsilon$  where  $\epsilon$  is the irreducible error represented noise and is normally distributed with a mean of zero. If this model is estimated using linear regression, the expected error at point  $x$  is

$$Err(x) = E[(Y - \hat{f}(x))^2]$$

which can be further decomposed as

$$Err(x) = (E[\hat{f}(x)] - f(x))^2 + E[(\hat{f}(x) - E[\hat{f}(x)])^2] + \sigma_e^2$$

which in turn corresponds to the relationship between error, bias and variance shown in Figure 16.

$$Error = Bias^2 + Variance + Irreducible Error$$

While this relationship seems to indicate that it is impossible to create the perfect model given the presence of error that is irreducible by any model, and the inverse relationship between bias and variance, it is worth noting that the purpose of statistical modelling is to generate a model that can analyse larger datasets than any human, while considering an enormous number of possibilities with a good enough accuracy to provide predictors that can be applied in a real setting. The ANN used for the current project is a low bias, high variance learning method and can be adjusted by altering the number of hidden nodes. Increasing them leads to lower bias and higher variance and, as seen later in this chapter, is the reason why for the majority of the tests in this project the number of hidden nodes was kept low.

### 3.4.3 Back-Propagation Algorithm

As previously established, the MLP uses error correction learning, which belongs to the supervised learning category of rules. During error correction, the algorithm uses a **Back-Propagation** (BP) algorithm in order to adjust the error at the end of each training cycle by comparing the true and predicted outputs. Rojas (1996) breaks down the flow of the algorithm in four steps:

1. Feed-forward computation
2. Back-propagation to the output layer
3. Back-propagation to the hidden layer
4. Update of the weights

The back-propagation step to the output layer uses the following formula

$$\frac{\partial E}{\partial w_{ij}^{(2)}} = \delta_j^{(2)} o_i$$

Where  $\partial E / \partial w_{ij}$  is the partial derivative we need for the next step;  $w_{ij}^{(2)}$  is the weight to the output layer, which is variable as it needs to be updated at each iteration;  $\delta_j^{(2)}$  is the back propagated error from the output; and  $o_i$  is the input, which is a constant. The subscript  $j$



is the cost. Using this formula, it is possible to compute the back propagated error to the hidden layer

$$\frac{\partial E}{\partial w_{ij}^{(1)}} = \delta_j^{(1)} o_i$$

where  $w_{ij}^{(1)}$  is the weight to the hidden layer and  $\delta_j^{(1)}$  is the back propagated error to the hidden layer. This allows the network to update the weights in an iterative manner and stop when it achieves the lowest possible error score. This update occurs according to the following equation known as the **Delta Rule** or **Least Mean Square**:

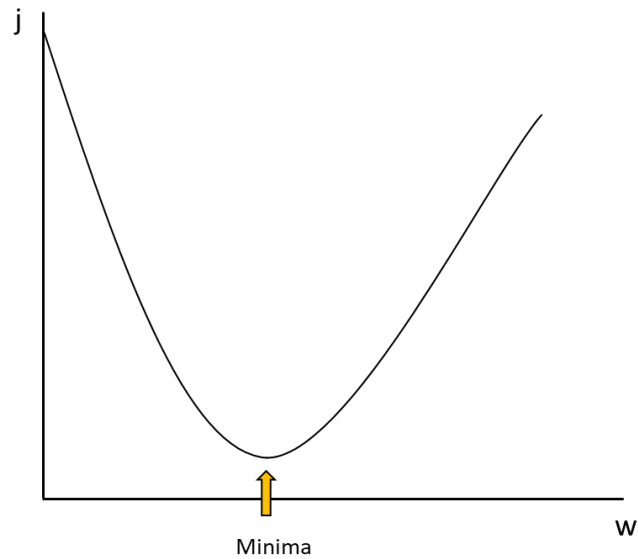
$$\Delta w_{ij}(n) = \eta(t_j - y_j)x_i$$

Where  $\Delta w$  is the weight difference in the  $n^{\text{th}}$  cycle,  $t_j$  is the error of the actual output,  $y_j$  the predicted error and  $\eta$  is the learning rate (discussed in section 3.4.4).  $\Delta w$  will, as a result, be updated in proportion with  $x_i$ , which corresponds to the weight input value.

The BP algorithm was developed by Williams and Hinton (1986) who were able to apply it and use neural networks to solve previously unsolvable problems. In essence, the BP algorithm informs the user of changes in the cost function, which measures the performance of the network, when the weights and bias are altered. However, in order to avoid overfitting and reduce the training time of the algorithm, it is essential to combine the BP algorithm with **Gradient descent**.

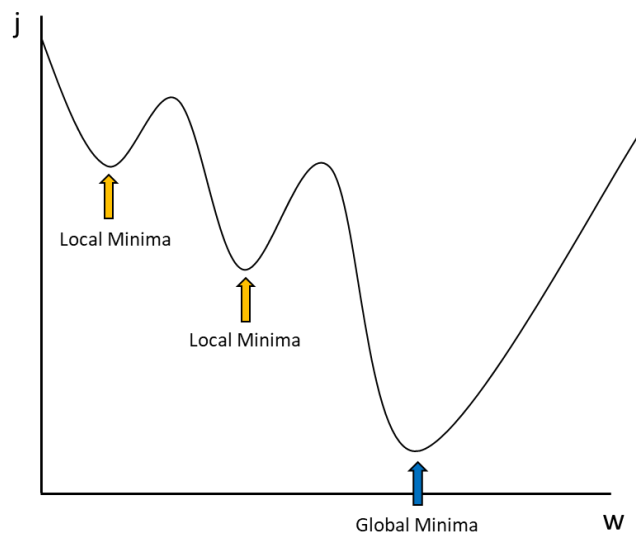
### 3.4.4 Gradient Descent

Gradient descent, also known as **Steepest descent**, is a class of algorithms designed to minimise functions. In the context of neural networks, gradient descent algorithms are used to establish the local minima, the point at which the error is at its lowest by minimising the cost function. As established previously, the cost function establishes how “costly” or wrong the model is and the goal is to make it as small as possible (Figure 17), which leads to the network training well and allowing it to find the weights that best fit the training cases.



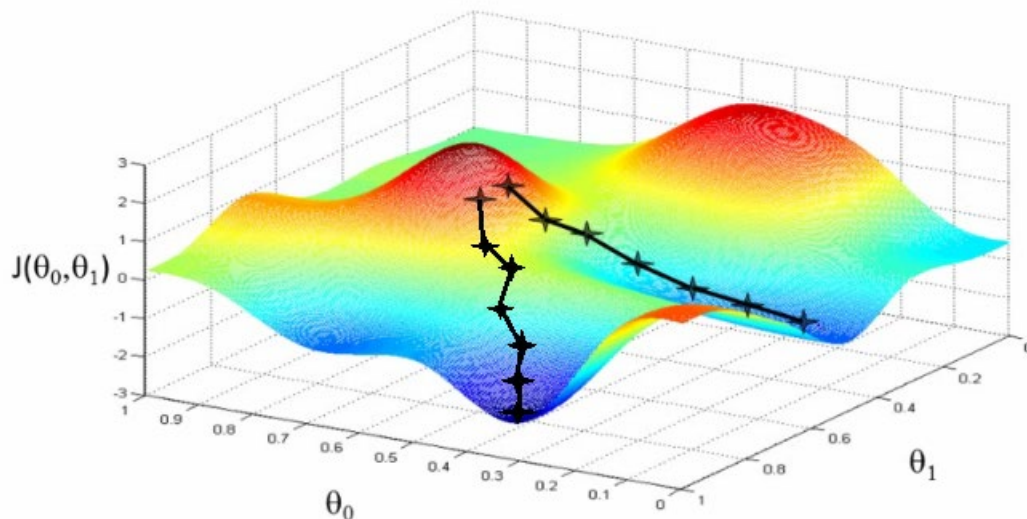
**Figure 17:** Demonstration of a two-dimensional error surface indicating the point where the error is at its lowest, the minima. The  $j$  axis indicates the error while the  $w$  axis corresponds to the weight.

Moreover, gradient descent is an essential addition to a neural network due to the non-convex nature of the cost function often found in biological data. As the complexity of the data increases it becomes possible for the algorithm to identify a point as the lowest possible error (local minima) when there is a lower available error that has not been reached yet (global minima), as shown in Figure 18.



**Figure 18:** Representation of a more realistic two-dimensional error surface. The error falls in irregular increments to a low point, known as the local minima before reaching the lowest possible point for the problem, known as the global minima at which point it increases towards infinity. The  $j$  axis is the error and the  $w$  axis represents the weight.

This is a result of the curse of dimensionality and, as dimensions are added, the potential to converge on the wrong error increases (Figure 19).



**Figure 19:** Illustrative schematic of a three-dimensional error surface with multiple points of local minima. The black line indicates the gradient descent with each point in each line representing a step during the descent. Note that there are two lines as there are two possible global minima points the algorithm is testing before deciding on the global minima. Source: <http://blog.datumbbox.com/tuning-the-learning-rate-in-gradient-descent>)

Additionally, as dimensions increase applying the gradient descent algorithm causes the time required to compute the error to scale exponentially with the number of dimensions, necessitating a method to decrease the required time and power. To combat this problem multiple versions of the gradient descent algorithm were developed, with the two most common being the **Batch** and **Sequential** or **Stochastic Gradient Descent** (Bishop, 1995).

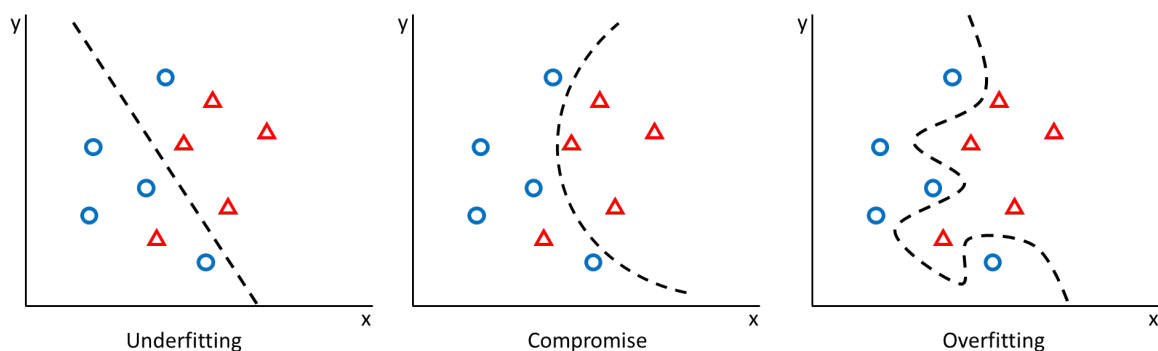
**Batch Gradient Descent** computes the gradient using the whole dataset. The weights are randomised, evaluated and updated as they move in the direction of the negative gradient and, as a result, the largest gradient descent. This method works well with smooth error manifolds and convex datasets and is quite likely to converge on the global minima if parameterised well.

**Stochastic Gradient Descent** on the other hand, evaluates and updates the weights for each training case in random order. In high-dimensional data, where there are a lot of local minima and maxima points, it is superior to the batch version as it deals with the increased noise by sampling small subsets, thus drastically reducing it. Additionally, it is very computationally efficient, significantly lowering the time required to analyse large, complex datasets.

### 3.4.5 Generalisation and Overfitting

One of the most crucial qualities a neural network should exhibit, is the ability to generalise. **Generalisation** is the algorithm's ability to build a statistical model that can learn from the training data and use this information to successfully classify future, unseen data (Haykin, 2009). As this process is reliant on the bias-variance trade-off explained earlier, the right balance between the two must be achieved in order to avoid overfitting.

**Overfitting** (Figure 20) refers to a phenomenon where the network trains itself on the training data too well. This leads to the algorithm "memorising" the training data, which in turn inhibits its ability to apply itself to future unseen data, and potentially incorporate unnecessary features, commonly known as noise, into its framework.

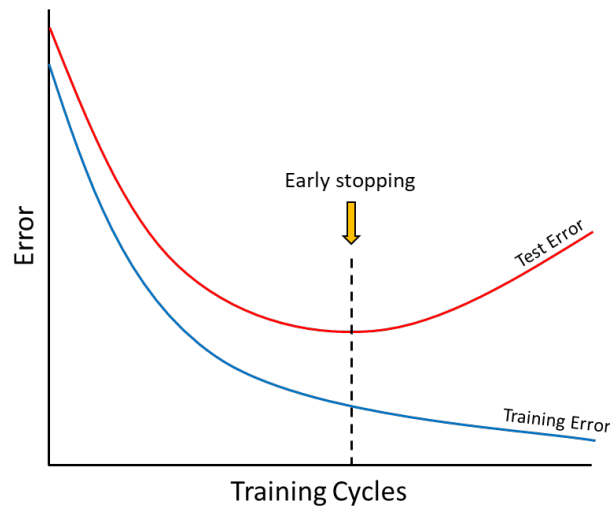


**Figure 20:** Schematic representations of algorithm fitting on a two-dimensional surface. The first graph on the left is a classic underfitting example. It has simply split the data into two categories with little regard to details. While the majority of the points are correctly classified, it is not accurate enough to provide an answer. The last graph on the right represents overfitting. All the points are classified correctly, but the shape of the curve is not usable for any other dataset apart from this one and so, has poor generalisation abilities. The middle schematic has classified most points correctly while being general enough to be equally accurate on unseen data and thus, the best compromise.

Since overfitting is a common issue with neural networks there exist multiple methodologies to avoid it. These include techniques such as **early stopping**, **resampling** with Monte Carlo cross validation (MCCV) and **weight decay regularisation** (Hastie *et al*, 2009).

Early stopping is one of the simplest ways to prevent overfitting and can be one of the most effective when applied to simpler problems. When early stopping is applied, training ends when the test error increases with respect to the training error as shown in Figure 21. The point at which early stopping occurs is commonly decided by a predetermined threshold related to the size of the Mean Square Error (MSE) of the **training set**, used for training the network and minimising predictive error, compared to that of the **test set**,

used to monitor the training set, or the **validation set**, used for further independent validation to achieve an unbiased estimate of network performance. The weights of the trained model with minimum predictive error on the test and validation subset can then be used for further validation on the blind set to check generalisation ability (Lancashire *et al*, 2009). Alternatively, the early stopping point is established after a set number of epochs.



**Figure 21:** Schematic representation of the early stopping algorithm. The training error decreases towards zero over time as the algorithm starts perfectly fitting the data (overfitting) while the test error reaches a point where it increases as overfitting occurs. The early stopping allows the algorithm to stop at the optimal point where the test error is lowest and the algorithm has the highest generalisation ability.

The biggest disadvantage of this method is that the point itself tends to be arbitrarily selected. On one hand this allows it to be easily changed in an ad-hoc fashion (Prechelt, 1998) and has been used successfully (Pouliakis, 2016) but it can end up becoming arbitrary without correct parameterisation, and lead to a lack of consistency for large scale studies (Bishop, 1995).

Resampling is a similar process, where network training is stopped once the optimal error has been achieved. The data are split into the training, test and validation sets assigned randomly according to preset subset sizes and the weights from the training set that produced the lowest error for the test set can then be used for the final mode (Lancashire *et al*, 2009). Since the data needs to be split into three subsets with a predefined percentage in each of these subsets, in order to achieve a truly random and representative sample in each of them, MCCV is commonly used. As shown by Xu and Liang (2001) MCCV is superior to competing methodologies such as leave-one-out cross validation, which tend

to cause overfitting by selecting overly large subsets for prediction. Its advantages include the ability to sample data in smaller subsets and thus having a higher chance of selecting features that can predict correctly, relative difficulty to underfit the data and ability to overcome the influence of collinearity which has the potential to cause random errors in larger datasets. In this project, the most common MCCV split used was 60:20:20, where 60% of the samples in a dataset were used for training with 20% each for testing and validation. This is a fairly novel approach and until recently the majority of experiments using ANNs used a 60:40 split for training and testing, eschewing validation. By adding a validation component however, the predictive power of the algorithm is increased significantly (Lancashire *et al*, 2006).

Weight decay regularisation aims to solve the problem of overfitting by the addition of a penalty term on the network parameters and reducing the freedom of the model, thus making the model unlikely to fit the noise present within the training set and improving generalisation; the ability to predict unseen cases (Lancashire *et al*, 2009). The most common weight regularisation approaches are the L1, L2 and L1/2 also known as lasso, ridge and elastic net respectively (Hastie, 2009). L1 regularisation shrinks certain parameters to zero, effectively removing them from the model which reduces overall test error until enough parameters are removed that there is not enough data to allow the model to learn and the error rises again. L2 regularisation adds a penalty equal to the sum of the squared values of coefficients such as bias and weight, forcing the parameters to remain relatively small making the coefficients more accurate and robust the larger the penalisation. Error tends to follow a sigmoid pattern and allows the user to select between the L1 and L2 approaches depending on the data studied. The last option, L1/2 regularisation combines these approaches by applying a penalty equal to the sum of absolute values to the sum of squared values of the coefficients resulting in less abrupt removal of coefficients. Weight regularisation in general, focuses on penalising larger weights to maintain lower values than what the algorithm would otherwise converge to (Bishop, 1995).

### 3.5 Optimisation

One of the greater challenges faced before putting the algorithm into action is deciding on the parameters of a given analysis. Historically, this has been a limiting factor when

attempting to use machine learning to analyse large and highly complex datasets, although recent advances have allowed for both higher storage capacity and increased processing speed which have not only mitigated the issue, but in time can turn it into an advantage, allowing the user to analyse multiple datasets using a multitude of parameter sets and gain further insight.

### 3.5.1 Randomisation of Weights

The selection of initial synaptic weights is crucial to achieving faster convergence within the network. If the values assigned are very small, approaching zero, they can cause the BP algorithm to become stuck on a flat error surface and generate a linear output, while overly large weights cause the algorithm to converge to the local minima and slow down the learning process. It is crucial to note that the initial network weights are randomised asymmetrically, and while this sounds counterproductive as it prevents every node from receiving the same signal, breaking that symmetry allows the units within the hidden layer to get a wide signal selection preventing it from outputting the same signal every time and leading to a slower learning rate. Naturally, increasing the variance in the original randomisation too much results in the sigmoid function derivative being very small which results in weight updates to be close to zero (Haykin, 2009) necessitating the initial randomisation of weights to achieve a balance between low and high variance.

### 3.5.2 Learning Rate and Momentum

As explained during the section on gradient descend, there can exist multiple points in non-linear data where the error decreases but is not at the lowest possible minimum error for that error surface. As the main goal of the algorithm is to find the point on that surface with the lowest possible error, the **global minima**, it is crucial to have a mechanism in place to ensure that it is not stuck to a point with a low but not lowest possible error, the **local minima**. As the point of gradient descend is to incrementally move across the error surface to the lowest possible point, it is paramount for that descent to be fast enough to escape the local minima but slow enough to not climb out of the global minima. To achieve this result, a **momentum** and **learning rate** terms were added to the algorithm.

**Learning rate** ( $\eta$ ) represents the incremental steps taken by the algorithm when adjusting the weights to find the global minima. If it is too low, the weights are not changed significantly from one iteration to the next leading the algorithm to believe that the lowest possible error has been achieved, while a high learning rate will lead to the algorithm missing the global minima and failing to achieve convergence and entering oscillation (Lancashire, 2006).

**Momentum** ( $\alpha$ ) is intended to prevent the network becoming trapped in local minima of flat regions in the error surface by gradually increasing the size of the steps taken by the network when adjusting weights. The rate and update of the weights of the network when momentum is included is represented by the following equation

$$\Delta w_{ji(n)} = \eta \delta_j x_{ji} + \alpha \Delta w_{ji(n-1)}$$

Where  $\Delta w_{ji}$  is the weight difference,  $\alpha$  is the momentum constant,  $\eta$  is the learning rate,  $\delta_j$  is the error term at the output unit and  $x_{ji}$  is the input value to which the weight is applied (Lancashire, 2006). As seen, the momentum amplifies the effect of the learning rate which leads to a significant decrease in convergence time as well as smoother oscillation. The momentum's optimal range has been shown to lie between 0 and 1 (Mitchell, 1997), although alterations to both learning rate and momentum are to be considered and adjusted depending on the size and complexity of a dataset.

The momentum and learning rate parameters that were used for the current project are 0.5 and 0.1, respectively (Lancashire, 2009)

### 3.5.3 Hidden Layer Parameters

The goal of the nodes in the hidden layer is to act as feature detectors and allow the ANN to classify non-linear input data, by transforming the weighted sum of inputs on the forward pass with a non-linear activation function. The greatest consideration when optimising the hidden layer parameters are the number of layers and their size. It was shown by Basheer and Hajmerr (2000) that if the network architecture is overly constrained by being too small, it will result in the masking of non-linear components and the output of a linear estimate of the desired solution. Conversely, a needlessly large



hidden layer will lead to increased training times as well as overfitting and poor generalisation.

Over the years, multiple solutions to finding the correct size of a hidden layer have been presented, since the optimal hidden layer size can vary due to the size and complexity of the data being analysed. However, very few problems have been proven to require more than a single hidden layer (Heaton, 2008), and thus, the challenge becomes selecting an appropriate number of hidden layer nodes. The generally accepted rules on this topic are that the number of nodes in the hidden layer should not exceed the number of input or output nodes, ideally being two thirds the size of the input and output layers combined and less than twice the size of the input layer.

Ultimately however, the number of nodes in a hidden layer have been decided largely through trial and error, depending on the data analysed. This constructive approach is shown in research by Srećnik *et al* (2002) where the size of the hidden layer was determined by incrementally increasing the number of hidden nodes, beginning from a small number and continuing until a predetermined minimum error was reached. Alternative methods include pruning, where the number of hidden nodes is determined by incrementally decreasing the size of a larger network until the optimal error is reached (Kavzoglu and Mather, 1998), and correlated activity pruning (Roadknight, 2001) which monitors the activating strengths of each hidden node, and calculates the correlation coefficient of their activation energies for paired nodes. This process is repeated until the network fails to achieve generalisation and the correct number of nodes is chosen.

The ideal number of hidden nodes for most of the data used in the current project was determined to be between two and five, which was determined to provide the best network performance and efficiency. This number was selected via correlated activity pruning (Roadknight *et al*, 2005), an approach based on removing hidden layer nodes not actively participating in the solution as determined by their constant output over all training cycles. While for most of the tests performed only 2 hidden nodes were used, for some of the most complex questions asked, the need for higher quality results superseded the need for more efficiency and 5 hidden nodes were used.

### 3.6 Advantages and Disadvantages

Statistician George Box remarked in 1976 that all models are wrong, but some are useful. This aphorism has been considered a core aspect in understanding the problems and limitations of applied statistics, as no matter how good a statistical model, it only serves as an abstraction that allows us to understand an aspect of the real world. ANNs are no exception to that rule and have received considerable criticism since their inception. First and foremost, of these problems is the computational power and time required to fully analyse larger datasets. These parameters scale exponentially with the size and complexity of a dataset. Moreover, overfitting can lead to poor model performance as a result of incorrect parameterisation. Finally, the very nature of ANNs has been under criticism for their “black box” approach to problem solving. Since the workings of the algorithm are not immediately apparent, many argue that the quality of the results is not guaranteed.

The advantages of ANNs on the other hand, are significant enough to confirm that as a statistical modelling method, they belong firmly in the useful category. Although rise in complexity leads to an exponential increase in time and computational power required, ANNs remain one of, if not the, most cost-effective way to handle large datasets. Moreover, their ability to also process highly complex and non-linear data as well as tolerate incomplete and fuzzy data make them ideal for analysing real patient and disease data. With the recent rise in the size and quality of publicly available datasets (ADNI, TCGA, METABRIC), this ability has solidified ANNs as an incredibly valuable tool for analysing patient and disease data, which is commonly complex, large, fuzzy and possibly incomplete. This is compounded by the ability of the ANNs to generalise by considering all of the individual possibilities and split them into similar groups that can be targeted to achieve further insight.

It is worth noting that the criticism ANNs have received over the years has led to a constant desire to improve this technology. As technology advances at an ever-faster pace, most historical criticisms are becoming obsolete. Advances in GPU computing, computing on graphical processing units, as well as parallel computing platforms such as OpenCL and CUDA have massively increased computational power for a fraction of the cost. Before switching to GPGPU computing, an average dataset of 80 samples and 50000 genes needed a week to be fully analysed by the algorithm. Currently a similar dataset can

be analysed in under an hour. Moreover, as GPU power increases, so do the capabilities of the algorithm. Using earlier versions, the author found that it was impossible to exceed 200 samples and 60000 genes without having to split the dataset, whereas currently it is possible to analyse datasets with 1000 samples and 60000 genes, as well as increased algorithm parameters, such as more steps, loops and hidden nodes in an acceptable amount of time. Finally, as understanding of machine learning and statistics increases, it becomes possible to “open” the black box as shown in a later chapter.

### 3.7 Stepwise Analysis

While the methods described above can be used to solve a wide range of problems and are constantly improving, when applied to the field of biomedical science, the challenges faced remain significant. The sheer depth and complexity of the data generated by modern techniques such as next-generation sequencing, whole genome sequencing, RNA sequencing, DNA methylation and even something as comparatively simple as microarray gene expression, have necessitated the development of methods better suited for this kind of data analysis. The main cause of this is the previously mentioned **curse of dimensionality**, which is “*the exponential growth of the input space as a function of dimensionality*” as described by Bellman (1961). This can be equated to an increase in complexity as more parameters are introduced during testing. Using AD as an example, complexity rises as we move from attempting to discover a list of genes that explain the genetic variance between AD and healthy patients, to splitting them across gender, then age, then racial profiles, followed by subcategorising the different regions of the brain and accounting for specific mutations. It is quite possible to construct a scenario where the number of variables exceeds the number of cases, which is the most common situation that is affected by the curse of dimensionality, leading to poor network generalisation. (Bishop, 1995).

Even though the size and quality of the datasets are increasing as more researchers become familiar with machine learning methods, it is still paramount to employ dimensionality reduction methods such as feature selection to decrease the parameters the algorithm must compute. Alternatively, using the **Stepwise ANN** approach it is possible to test each predictor independently for each question and keep testing the resulting combinations with the best predictive performance.

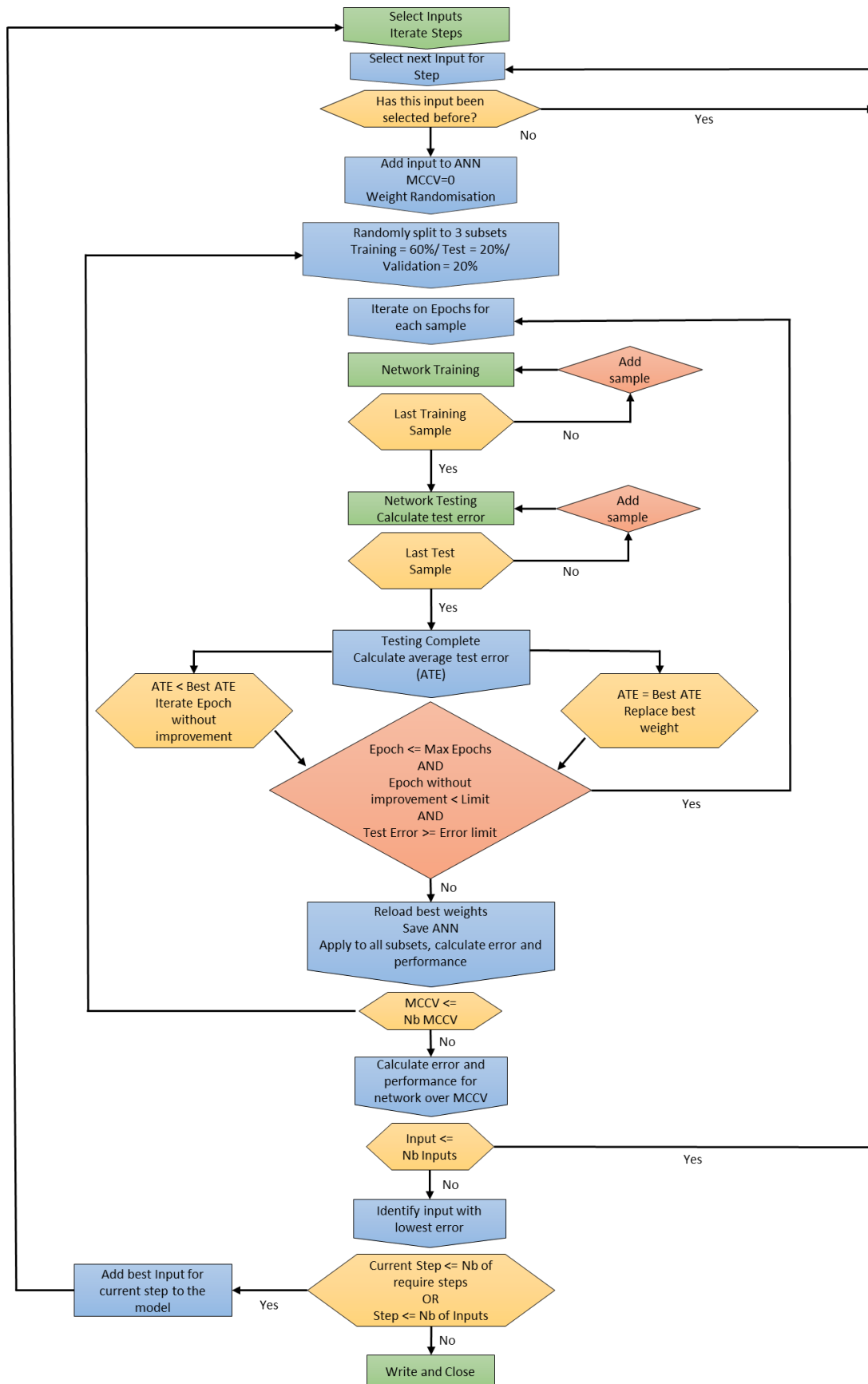
The Stepwise ANN approach used for the current project was first published by Lancashire *et al* (2008) and it has been proven to be capable of identifying patterns in data by identifying the best inputs for classifying a given task based on the predictive performance of each variable. The error, and thus learning, is calculated and backpropagated through the algorithm in an iterative manner leading to quick and efficient identification of the best performing variables are added to a panel that can then be used to predict a given question. This is especially useful in biological datasets, as the algorithm is able to not only identify the genes most likely to explain the variance in a dataset, i.e. the genes most likely to contribute to AD in the male Caucasian population, but generate multi-gene panels with high predictive power and accuracy.

In practice, using a microarray gene expression dataset as an example, this approach uses each gene from the microarray as an individual input, generating  $n$  models corresponding to the number of genes in the experiment, which are then sorted by their predicted performance for unseen cases as selected by the MCCV algorithm. After the genes have been ranked in this manner, the best input is selected and used as a predictor for the models with the remaining genes until  $n-1$  models have been generated. This approach continues in an iterative manner until the algorithm has reached optimal performance or no further improvement can be achieved (Lancashire, 2008).

The algorithm follows the following process as outlined by Lancashire (2006).

1. Each of the variables is used as a single input in a one-input model, creating  $n$  single models.
2. Each model is then trained over 50 events of MCCV, meaning that all the samples were randomly reshuffled to ensure that all are considered blind for a number of models, in order to improve the ability of the network to generalise well for unseen cases. 50 MCCV folds were found to be the number for which the models started to reach consistency (Lancashire, 2006).
3. The predictions and MSE across the 50 sub-models for test subset are monitored and recorded for each single-input model, and these inputs are then ranked based on their MSE.

4. The input within the model predicting the best (i.e. with the lowest error) is then selected for the second step.
5. At the following step, the input that performed the best in the previous one is used as the basis for two-input models.
6. The remaining inputs ( $n - 1$ ) are then sequentially added to create ( $n - 1$ ) two-input models.
7. 50 sub-models are then trained for each of these two-input models, and their performance is monitored as explained earlier.
8. The performances allow us to rank the best two-input model and select the combination of two inputs for the third step.
9. The process is repeated until no improvement in network performance is observed, or if any early-termination condition is met.



**Figure 22:** Schematic representation of the ANN algorithm used in this project. Adapted from Lemetre (2010)

This approach has been successfully used in a wide range of studies (Abdel-Fatah *et al*, 2016, Vafadar-Isfahani *et al*, 2012, Elsheikh *et al*, 2009)

Originally the algorithm was developed for use with Neural Network package of ©Statistica using a Visual Basic program for the purposes of assessing the validity of the approach and using an ad-hoc approach for changing the setting and thus optimising the algorithm. It was then moved to C as the language itself is closer to machine code and thus more efficient. Currently the algorithm runs in a dedicated interface making use of GPGPU computing via OpenCL as this allows for a significant increase in processing speed and efficiency.

The Stepwise algorithm uses a single hidden layer MLP with two hidden nodes and a backpropagation algorithm, using early stopping to avoid overfitting. The maximum number of epochs is 3000 with a window of 1000 epochs, allowing for training to be stopped if 1000 epochs transpire with no improvement in model performance, with a mean square error of 0.01. For the BP algorithm, the learning rate is set to 0.1 and the momentum to 0.5. The initial network weights are normalised between 1 and -1 and randomised. This is followed by the application of MCCV, using a split of 60:20:20 for training, testing and validation respectively, as mentioned earlier. The MCCV approach is repeated 50 times and combined with resampling all the cases to minimise random errors. These parameters were used in the majority of the experiments conducted during the current project, with further deviations disclosed when appropriate.

### 3.8 Network Inference

When applied to real data, the Stepwise approach described is an excellent way of discovering novel biomarkers that can be used as targets for therapy, predictors for early prognosis and a way to process the large amount of data produced at an ever-increasing pace. It should be noted however, that very few of these markers make it to clinical testing, which has resulted in the validity of the results being questioned. As efforts were made towards solving this issue, it became clear that part of the reason for the perceived unreliability of these markers was that it was almost never the case that any disease was caused, or could be targeted, by using a single, or even small selection of biomarkers. Moreover, the relationships of the established markers with other genes in their resulting

networks could have unforeseen consequences that were poorly understood. In response, new methods, such as gene expression analyses focusing on the biology of entire systems were established to examine the causes and effects these complex networks of interacting genes on a given disease (Barabasi *et al*, 2011). This also established the importance of other molecular pathways, particularly regulatory ones, whose effect on the network would be “masked” in more conventional approaches. Christophe Lemetre described it in his PhD thesis thusly:

*If any of the markers (e.g. genes) contained in an expression array of individuals have some influence on the expression of other markers (either positive or negative), we might be able to observe and monitor significantly correlating expression profiles between these interacting markers through the population of individuals. In other words, the influence that one input has upon the prediction of any other given input is proportional to the relationship between the two.*

This logic led to the development of a companion algorithm to the Stepwise ANN with the goal of performing network inference to extract further information by performing iterative calculations to examine the effect that multiple variables can have on a single one.

While there exists a variety of similar approaches including but not limited to Gene Set Enrichment Analysis (Subramanian *et al*, 2005) and DAVID (Huang *et al*, 2009) which analyse entire pathways and the effect genes have on them, they tend to ignore potential gene-gene interactions and other correlations between genes, which is all too common in complex diseases such as AD or cancer. There are other methodologies such as Bayesian approaches (Hartemink *et al*, 2002), likelihood approaches (Liu *et al*, 2005), dynamic ordinary differential equations (Christley *et al*, 2009) and recurrent neural network models (Xu *et al*, 2004) which aim to solve these issues but they are still limited by their inability to consider the entire available pool of variable, focusing instead of a few genes of interest (Lemetre *et al*, 2009) leading them to identify very limited information subsets. While this can be highly beneficial when attempting to establish interactions in a highly controlled and tightly focused environment it is crucial to have a methodology for



unbiased, non-systematic biomarker discovery. Moreover, these methods tend to be further limited by the fact that the resulting networks are non-directional, unidirectional or acyclic. The methodology, proposed by Tong *et al* (2014), is a novel ANN designed to infer directed gene-gene interactions in a pairwise manner, allowing the user to observe how changes in a given gene leads to changes in other genes and the network as a whole.

### 3.8.1 Model development

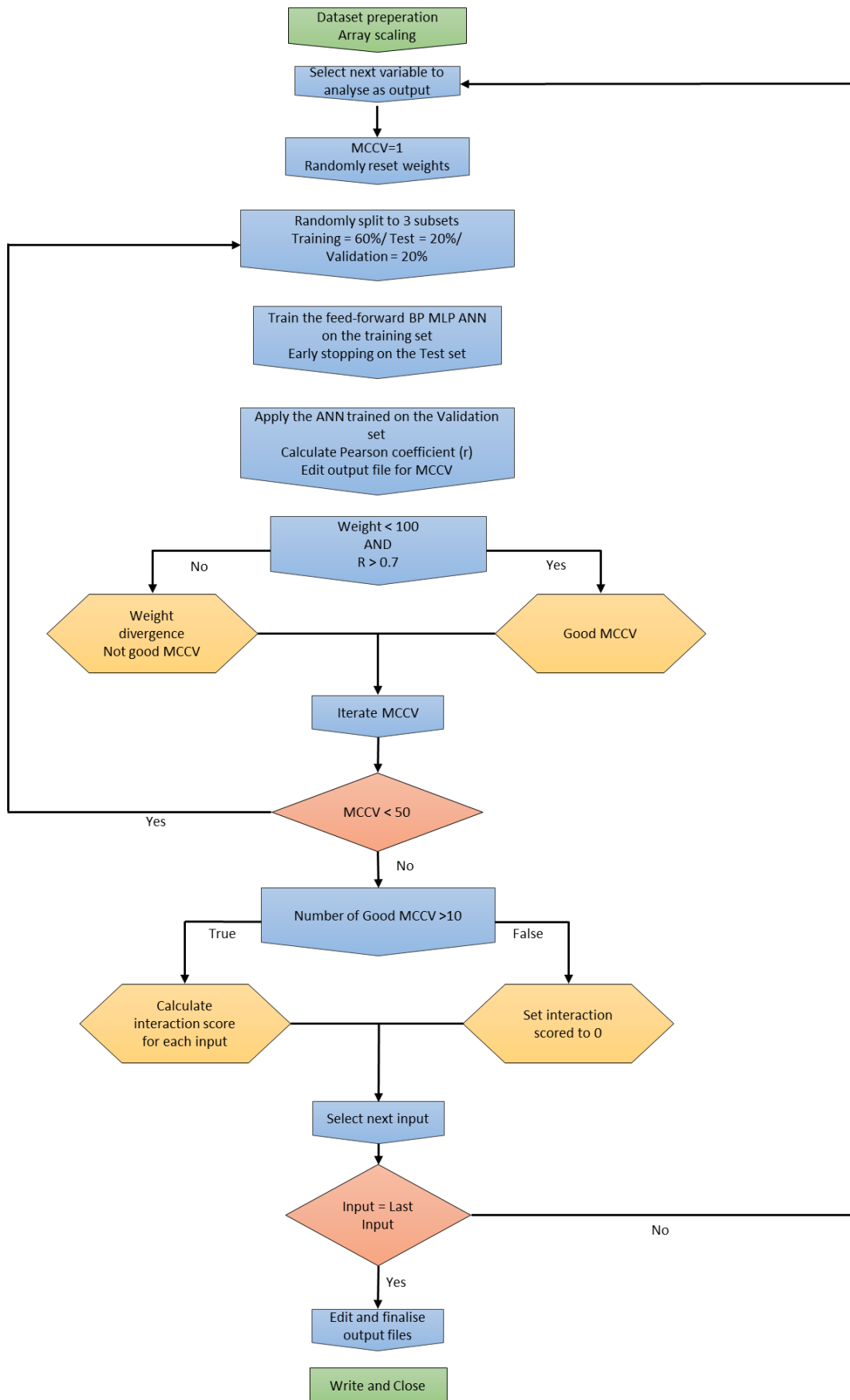
The goal of the interaction algorithm is to quantify the effect of multiple variables on a single one, by iterative calculation of their influence on each other. While the methodology is similar to that of the stepwise algorithm, it is important to note that the interaction algorithm is meant to function in conjunction with it. Thus, rather than identifying the variables with the highest predictive value for a given outcome, the selected variables are used to determine their influence on the whole network rather than individually. This process essentially selects a variable, compares its expression level to every other variable selected as an input, assigns a weight to the predicted influence, which is directly proportional to the intensity of linkage between two variables (Tong *et al*, 2014), and repeats the process in an iterative manner. As mentioned earlier, the advantage of this method is that it can generate directional networks. The magnitude and directionality of the interactions between inputs (sources) and outputs (targets) is determined based on the weights assigned, and determines whether the predicted interaction is uni- or bi-directional and inhibitory or stimulatory.

The algorithm, referred to as an ANN based inference algorithm (ANNI) henceforth, consists of a three-layer MLP with a single hidden layer with two hidden nodes, an output layer with a single node, sigmoidal activation function and a backpropagation algorithm to adjust network weights. Inputs are split with an MCCV in proportions of 60:20:20 for training, testing and validation, repeated 50 times per input. A Pearson correlation analysis is used for each repeat to compare the predicted values obtained during training with the actual values obtained after testing, with the resulting correlation coefficient used as a confidence interval for each bootstrap. In order to remove the least significant interactions, a threshold of  $r > 0.7$  for 10 bootstraps minimum was implemented (Lemetre *et al*, 2009). While the algorithm parameters remained the same as those used during the stepwise analysis to achieve the highest level of consistency, the epochs and window were reduced

to 300 and 100, respectively, for the larger analyses. After rigorous testing, it was determined that the algorithm reached convergence at around 300 epochs, the results maintained consistency and the time required decreased by almost 75%.

### 3.8.2 Workflow

The ANNI workflow (Figure 23) used for this project consists of a Stepwise ANN analysis to determine the genes most likely to explain the variance in a question. Between 100 and 200 of the genes with the lowest MSE are selected for use with ANNI, which acts as feature selection, an essential step considering the quadratic relationship for the variables selected and the number of interaction analyses performed. For  $n$  variables selected,  $n-1$  tests take place, leading to  $n(n-1)$  or  $n^2-n$  total interactions, which generates a large enough amount of data to crash the algorithm or at least increase processing time to impractical levels, with individual tests taking weeks. Moreover, even if it were made possible by better hardware there is a high chance that the introduction of so much noise in the dataset would mask significant interactions. There is no gold standard for the ideal number of variables to select, but the number should increase in proportion to the complexity of the question. For the current project, using focused questions where only a small variance is expected, 100 genes were chosen. For more complex problems the number was increase to 200 and for some of the broader questions the number was increased to 500, although the interaction analysis had to be performed as a matrix to account for both time and noise as explained in a later section. This is followed by rescaling the entire dataset between 0 and 1 so that all variables are normalised, and the algorithm is applied as shown in Figure 22.



**Figure 23:** Schematic representation of the interaction algorithm used in the current project. Adapted from Lemetre (2010)

### 3.8.3 Visualisation

The ANNI analysis is followed by visualisation of the results by generating a map of all interactions of interest, an interactome, using third party software such as Cytoscape or in-house application to generate hive plots. This is a crucial step as the results are typically in the form of overly long charts detailing the exact relationships between sources and targets and are hard to read and understand and almost impossible to present and use in a real setting. Cytoscape, developed by Smoot *et al* (2011), is the primary visualisation software used for the current project. Interactomes generated this way show interactions between 100-500 genes and typically between 100-1000 separate interactions. Due to the large number of interactions generated only the ones assigned the largest weights are represented. Each gene is assigned to a **node**, the size of which is proportional to the number of interactions the gene is involved in, with the interaction between nodes represented as a **directed edge** pointed from source to target (source→target). The colour of the edge represents the nature of the interaction with blue for positive and red for negative, with the width of the edge being directly proportional to the strength of the interaction. This type of mapping, as described by Barabási and Oltvai (2004), is a subset of network theory where nodes symbolise markers and edges are the connections between them. The genes with the highest number of connections are designated as **hubs** and are the most likely candidates for therapy or prognosis.

Additional methods to identify potential markers, such as the driver analysis as well as alternative ways to visualise interactomes are explored in the following chapter.

# Chapter 4: Non-Systematic Hypothesis-Free Approach for AD Biomarker Discovery

## 4.1 Dataset Selection

The first step necessary to for biomarker discovery is the selection of a representative dataset that both conforms to the ANN algorithm requirements and can also be used as a control for further development. As explained earlier, the two most crucial parameters are size and complexity. By utilising the software G\*Power 3.1.9 (<http://www.gpower.hhu.de/en.html>), it was possible to perform a power analysis in order to determine that the minimum required sample size. To calculate the required sample size, the parameters chosen were a significance level ( $\alpha$ ) of 0.05, a power ( $1-\beta$ ) of 0.8, an effect size leading to odds ratio of 1.7 and two-tailed test for binary questions or classes (i.e. AD vs Healthy). Based on the assumptions of the power model and using G\*Power 3.1.9 software, the required sample size will be 88 in each class, with sample sizes of  $n=100$  in each group (total sample size  $n=200$ ) sufficient to negate any inter-individual confounding results. The threshold of 0.05 is the minimum threshold to ensure sufficient information for feature detection (Abdel-Fatah *et al*, 2016).

In a power analysis, the  $\beta$  (beta) value is the probability of making a type II error and accepting the null hypothesis even though it is false, when the real difference is equal to the minimum effect size. The power ( $1-\beta$ ) of a test is the probability of rejecting the null hypothesis when the real difference is equal to the minimum effect size. Larger sample sizes provide greater sensitivity than smaller ones by allowing smaller effects to take place and be detected, although too large a sample size risks adding too much noise in the data and/or drastically increasing the required computational power and time. Moreover, as a crucial part of this project has been the consolidation of previously used techniques and the development of new ones, the dataset selected for the primary set of experiments has to be large enough to be representative yet easy to employ and modify, small enough to be analysed in a timely manner and using proven technology to allow for validation of the results and cross comparison with other datasets.

The datasets considered, E-GEOD-48350, E-GEO-5821 and E-GEOD-9770, are publicly available and have been accessed using ArrayExpress (Kolesnikov *et al*, 2014) as well as the Gene Expression Omnibus (GEO) (Barret *et al*, 2012). The parameters required of the potential datasets are:

- Human samples only
- Patient size of >80
- Genes in array >40000
- A minimum of four brain region samples
- Healthy controls between 33% and 66% of the dataset
- Recent publication
- Raw data available in the form of CEL files

It is worth noting that during the first stage of the experiment the size and quality of datasets on AD in these databases was very low on average. Most publications were biased towards specific genes or conditions and could not provide a clear picture. Thus, the selected datasets are all measuring gene expression based on the mRNA levels in the neurons using microarrays. Since then, RNAseq data have become available and it is recommended for future experiments to use such data instead, as they provide more information and the algorithm is now able to handle such data.

The primary dataset selected for the initial series of tests was E-GEOD-48350 (Blair *et al*, 2013). It has 253 patient samples with a ratio of roughly 1:2 AD to cognitively normal using the AFFY-44, Affymetrix GeneChip Human Genome U133 Plus 2.0 array which has information on 54676 gene probes. The dataset includes genetic information about four brain regions; the hippocampus, considered essential due to it being the region most affected by AD, the entorhinal cortex, for which there exists a significant amount of literature linking it to the disease, as well as the postcentral and superior frontal gyrus, neither of which have been shown to be significantly implicated to the development or initiation of the disease. This presents us with a wide range of options and possibilities as well as providing both negative and positive controls as complete datasets, dubbed the Master set for further tests, allowing for the questions to be set to AD against cognitively

normal, comparison between brain regions, comparisons within brain regions and comparisons to other datasets.

The other transcription datasets considered were E-GEO-5821 (Liang *et al*, 2007) and E-GEOD-9770, which were used to study MCI as well as attempt to create a consolidated dataset to increase the power of the techniques used. However, results from these datasets proved inconclusive and were not used after the initial series of tests. The datasets are available in the appendix.

## 4.2 Data Normalisation

In all of the above cases the processed data were available and ready to be used in the algorithm. However, they used different normalisation techniques. As microarrays chips consist of numerous probesets, complementary nucleotide sequences used to measure mRNA levels by binding to them, the level of the mRNA that should directly correlate to the gene expression level in the sample is measured by comparing the intensities of each probe on the chip. This technique, known as relative quantitation, does not produce data that are directly usable due to the large differences in number sizes often leading to an exponential scale and thus, have to be normalised to achieve a linear scale that allows the user to directly compare biological differences. While there are multiple normalisation methods and variants, the two major ones are RMA and MAS5.

RMA (Robust Multi-array Average) is quickly becoming the preferred normalisation technique in biology. It normalises the data across all arrays and compares the various expression levels between them thus leading to comparable distributions. Additionally, it performs background correction on each array leading to better protection against outliers. It functions by performing that background correction first, normalising the data across all arrays, calculating the intensity of each probe and summarising them by performing a median polish, normalising each chip and each gene to its median and repeating until medians converge (Irizarry *et al*, 2003). A popular variant of RMA is GeneChip RMA (GCRMA) which uses the information in probe sequences to estimate probe affinity and use this information to estimate the relationship between non-specific binding and the target sequences. While this is an improvement over the standard RMA methodology due to correcting for the GC content of the oligonucleotides it does not necessarily translate

to better results due to the lack of a control to evaluate normalisation methods (Harr and Schlotterer, 2006).

MAS5 (MicroArray Suite 5.0) normalises the data based on Tukey's biweight, a known and robust statistic, and normalises each value according to the distance from the median by estimating the central location and adjusting for outliers. It has however been criticised for losing information at the probe level, especially in low intensity where too much noise is added (Piccolo *et al*, 2012).

To summarise:

- MAS5 normalises each array independently and sequentially; RMA as the name suggests (robust multi-array) uses a multi-chip model
- MAS5 uses data from mismatch probes to calculate a "robust average", based on subtracting mismatch probe value from match probe value
- RMA does not use the mismatch probes, because their intensities are often higher than the match probes, making them unreliable as indicators of non-specific binding
- RMA values are in log<sub>2</sub> units, MAS5 are not (so values are not directly comparable)

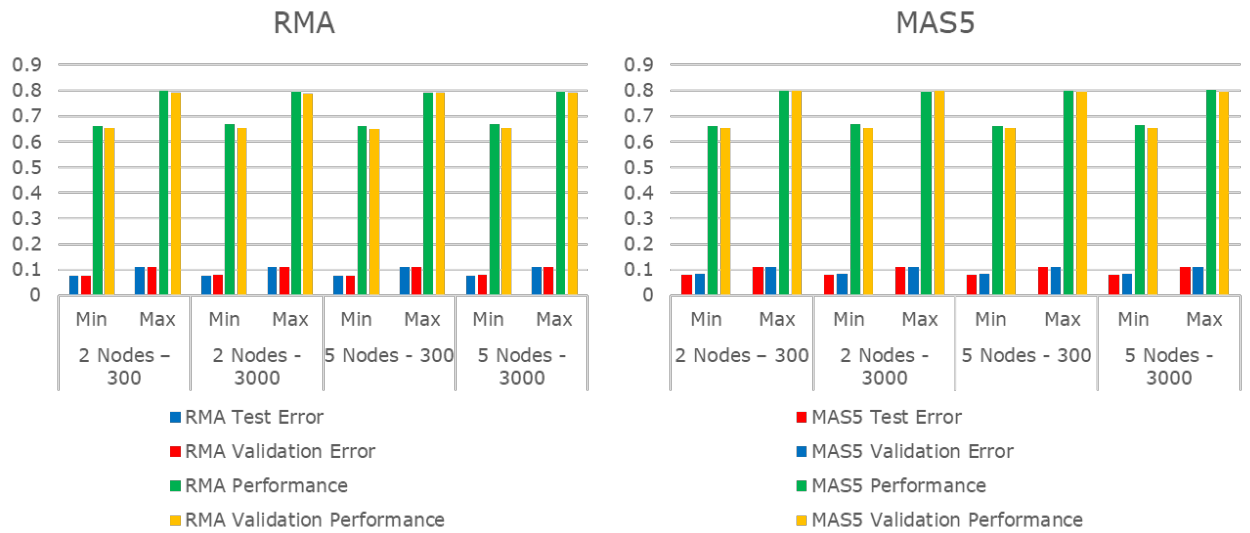
Originally, the normalisation was performed using the Affymetrix Expression Console, now Transcriptome Analysis Console (TAC) software, but has since been moved to the *affy* Bioconductor package for R.

After comparing the two methods on the three selected datasets, it was determined that although RMA normalisation of the data shows higher performance and lower error, it is not statistically significant. However, it does provide significantly more consistent results across multiple tests as show in Figures 24 and 25.



RMA						MAS5					
		Test Error	Validation Error	Performance	Validation Performance			Test Error	Validation Error	Performance	Validation Performance
2 Nodes - 300	Min	0.076014	0.077653	0.660785	0.654	2 Nodes - 300	Min	0.079739	0.082335	0.661765	0.653
	Max	0.111044	0.111738	0.8	0.792		Max	0.110706	0.11137	0.798039	0.798
2 Nodes - 3000	Min	0.075657	0.078575	0.668628	0.655	2 Nodes - 3000	Min	0.080261	0.082595	0.668628	0.655
	Max	0.110697	0.112149	0.795098	0.786		Max	0.110916	0.111687	0.795098	0.799
5 Nodes - 300	Min	0.075853	0.077206	0.661765	0.651	5 Nodes - 300	Min	0.08021	0.082355	0.662745	0.653
	Max	0.111388	0.111904	0.790196	0.791		Max	0.111514	0.111862	0.79902	0.795
5 Nodes - 3000	Min	0.075703	0.078542	0.668628	0.654	5 Nodes - 3000	Min	0.079997	0.082709	0.665687	0.654
	Max	0.111427	0.112054	0.796079	0.791		Max	0.111362	0.112294	0.801961	0.795

**Figure 24:** Table showing the difference in performance between RMA and MAS5 on the E-GEOD-48350 dataset. The tests were repeated for both 2 and 5 hidden nodes as well as 3000 and 300 epochs with 1000 and 100 epochs for early stopping respectively. As shown, the differences in performance are minimal.



**Figure 25:** Graphical representation of Figure 24. There is <0.5% difference between the two normalisation methods.

## 4.3 Stepwise ANN

### 4.3.1 Single Marker Analysis

The stepwise ANN approach (Lancashire *et al*, 2008), as explained in Chapter 3, allows for the identification of a gene or set of genes with the best predictive performance to classify samples based on a certain question by data mining the complete transcriptome. The ANN model functions by modifying the network weights and subsequently adding variables in an iterative manner to find a model with the lowest predictive error. The architecture consists of a single hidden layer, feed forward MLP with a variable number of hidden nodes and a sigmoidal transfer function, using a back-propagation algorithm incorporating supervised learning for updating the network weights. A Monte Carlo Cross Validation (MCCV) strategy was applied to produce a more generalized model with an improved predictive ability for unseen or future cases. The MCCV randomly divides the samples into training, test and validation subsets in 60:20:20 proportion for 50 iterations to provide the most consistent models. The parameters selected for this series of tests are 1 step, 10 loops with a momentum of 0.5, learning rate of 0.1 and threshold of 0.01 (Lemetre *et al*, 2010) as it allows the user to identify a single best predictor for further testing.

However, during the initial phase of the experiment, the Stepwise algorithm was used to identify the dataset to be deep mined as well as the ideal parameters. The parameters that were tested included variations in the number of hidden nodes as well as the epochs and window. Based on the work of Lancashire (2006), Lemetre (2010) and Agarwal (2017) the number of hidden nodes should be between 2 and 5, thus the experiment was repeated using 2,3 and 5 hidden nodes, and the epochs should be between 300 and 3000 with a window of 100 and 300 respectively., so the maximum and minimum were used. As shown in Figure 25, the difference in performance and error is minimal across all parameters, but the time required to compute the results increases exponentially with the addition of hidden nodes and the increase in the number of epochs. Considering a Stepwise ANN analysis on the CPU for a dataset the size of E-GEOD-48350 takes 5-7 days, 2 hidden nodes and 300 epochs with a window of 100 were considered adequate and strike a good balance between power and time.

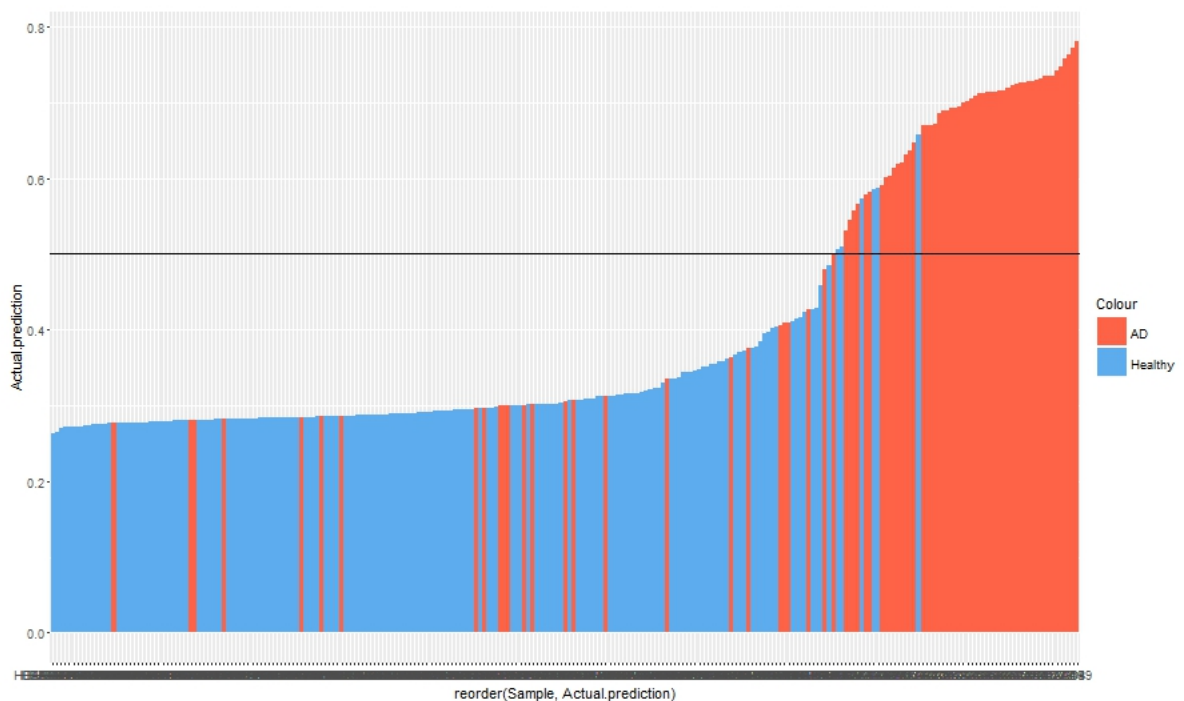
However, for the experiments conducted after August 2016, the algorithm was moved to a GPGPU platform using OpenCL. The time required was reduced drastically; a dataset the size and complexity of E-GEOD-48350 is now able to be analysed by the Stepwise algorithm in 1-2 hours using the same parameters. This has allowed for the increase in the complexity of the parameters, and all experiments conducted after that time use 3000 epochs with a window of 1000. Moreover, in situations where noise in a dataset, as a result of the question being asked, is too focused and therefore biased, or not enough samples are available to define the variance between two questions the number of hidden nodes was increased to 5. Such examples include how the low expression of a particular gene like MAP1LC3 correlates to AD, and the variance between AD and the healthy control in the entorhinal cortex in the E-GEOD-5281 dataset. As this change presents an increased risk of overfitting, the results were monitored carefully to avoid that. Further consideration was given to the momentum (0.5), learning rate (0.1) and threshold (0.01) where slight deviations lead to no noticeable improvement and large ones to under- or overfitting. At the end of a series of tests for any questions the algorithm was perturbed by changing these parameters significantly to cause it to overfit and show the difference between the two states.

The goal of this series of experiments is use categorical variables to identify the genes most likely to explain the variance at the genetic level between AD and healthy individuals as well as identify the best of the three datasets mentioned earlier for use in further deep mining. In order to achieve that goal, after normalisation, the samples in the dataset were classified as “1” if the sample was from a patient diagnosed with AD or “0” if healthy based on the clinical data. In the case of dataset E-GEOD-9770, after being merged and co-normalised with E-GEOD-5281 it was further split into three subsets classified as AD-healthy, AD-MCI and healthy-MCI. This information is used by the stepwise algorithm to train the data in a supervised manner on 60% of the subset and test and validate the rest of the cases. The results are shown in Table 1. In later stages, the classification was expanded to work on a continuous scale as explored in Chapter 5.

### 4.3.2 Multistep Stepwise Analysis

The methodology used in 4.3.1 was developed in order to identify a single best predictive marker by sampling the data for 10 loops. Another method is using a multiple step analysis

to determine the best predictive set of markers. Within each loop, each of the inputs is selected as a model and trained, with the input with the lowest mean squared error being selected for the second step and used as the basis for a two-input model (Lemetre, 2010). This process is repeated until no improvement is gained. The advantage of this technique is that it provides us with a panel of biomarkers and it is possible to determine how well they can be used to predict the condition when used in conjunction with each other. The genes selected by the stepwise ANN were LINC01128, SEMA3A, FCGR2C, AGPAT1, TTTY2B and ANKRD44. Figure 26 shows the classification of AD cases and healthy controls based on this gene panel.



**Figure 26:** Prognostic panel of 253 samples from AD and healthy individuals from array 48350. The cases over the 0.5 threshold should be identified as AD.

The quality of the prognostic panel is determined by the amount of false positive cases over the prediction threshold. Based on the selected genes, this panel shows that only 6 out of 173 healthy individuals were falsely predicted, giving us a false discovery rate of 3.47%.

### 4.3.3 Gene Ontology

For dataset E-GEOD-48350, table 1 shows the top genes identified.

Index	Probeset ID	Gene Symbol	Index	Probeset ID	Gene Symbol
1	215535 s at	AGPAT1	51	223913 s at	MIR7-3HG
2	212117 at	RHOQ	52	212993 at	NACC2
3	32836 at	AGPAT1	53	51158 at	FAM174B
4	224378 x at	MAP1LC3A	54	213558 at	PCLO
5	201938 at	CDK2AP1	55	206382 s at	BDNF
6	212119 at	RHOQ	56	227539 at	GNAI3
7	214770 at	MSR1	57	209515 s at	RAB27A
8	227219 x at	MAP1LC3A	58	241782 at	NEBL
9	212274 at	LPIN1	59	239538 at	ZRANB3
10	1557545 s at	RNF165	60	218614 at	KIAA1551
11	218031 s at	FOXN3	61	219752 at	RASAL1
12	212120 at	RHOQ	62	241385 at	LARP7
13	211630 s at	GSS	63	1559156 at	---
14	223213 s at	ZHX1	64	1559426 at	---
15	214449 s at	RHOQ	65	203961 at	NEBL
16	226996 at	LCLAT1	66	209796 s at	CNPY2
17	223107 s at	ZCCHC17	67	202779 s at	UBE2S
18	204514 at	DPH2	68	223343 at	MS4A7
19	225504 at	HMBOX1	69	210166 at	TLR5
20	224869 s at	MRPS25	70	224478 s at	C7orf50
21	231967 at	PHF20L1	71	204723 at	SCN3B
22	221880 s at	FAM174B	72	213921 at	SST
23	222494 at	FOXN3	73	215267 s at	SLC8A2
24	227890 at	TMEM198	74	231986 at	RIMS1
25	202506 at	SSFA2	75	202767 at	ACP2
26	229917 at	AGAP2	76	202030 at	BCKDK
27	220262 s at	DLK2	77	238697 at	LINC00086
28	205022 s at	FOXN3	78	1555765 a at	GNG4
29	213587 s at	ATP6V0E2	79	204269 at	PIM2
30	235850 at	FAM162A	80	244457 at	---
31	225526 at	MKLN1	81	202737 s at	LSM4
32	203723 at	ITPKB	82	208683 at	CAPN2
33	210951 x at	RAB27A	83	220807 at	HBQ1
34	232011 s at	MAP1LC3A	84	213388 at	PDE4DIP
35	227909 at	LINC00086/7	85	209104 s at	NHP2
36	214306 at	OPA1	86	205184 at	GNG4
37	207842 s at	CASC3	87	218547 at	DHDDS
38	239367 at	BDNF	88	212276 at	LPIN1
39	1553611 s at	KLHL35	89	210992 x at	FCGR2C
40	201253 s at	CDIPT	90	213045 at	MAST3
41	223367 at	DNAJC30	91	235935 at	LRRC73
42	206162 x at	SYT5	92	226326 at	PCGF5
43	212730 at	SYNM	93	225946 at	RASSF8
44	218260 at	DDA1	94	212411 at	IMP4
45	221847 at	LOC100129361	95	213270 at	MPP2
46	243501 at	ATP5F1	96	214665 s at	CHP1
47	236277 at	AF070581	97	203114 at	SSSCA1
48	244261 at	IFNLR1	98	1555889 a at	CRTAP
49	225219 at	SMAD5	99	1555867 at	GNG4
50	209332 s at	MAX	100	204141 at	TUBB2A

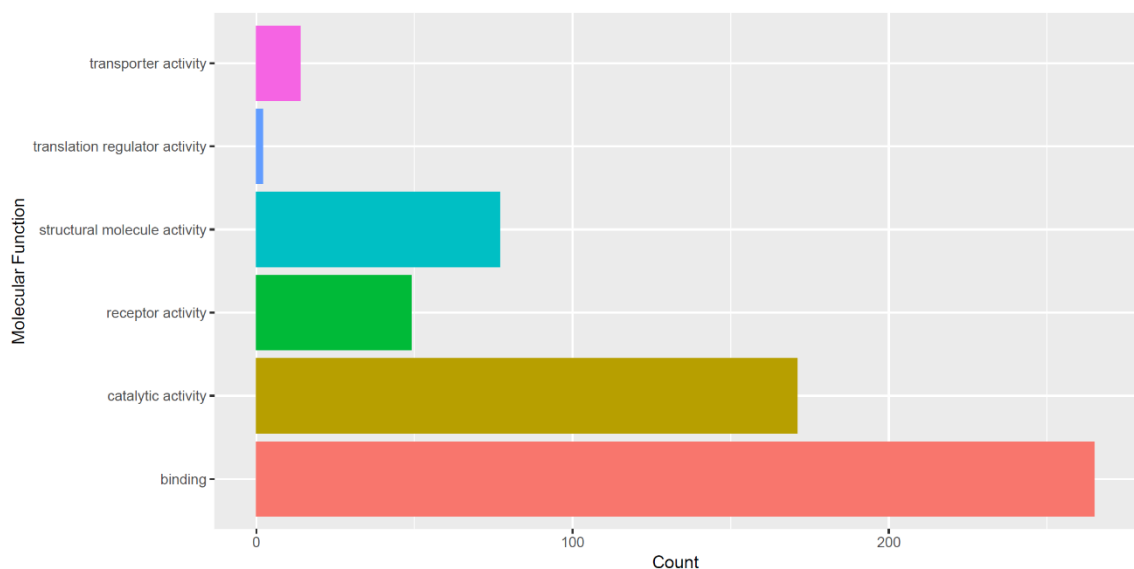
**Table 1:** Table showing the top 100 genes identified by the Stepwise algorithm as the most likely to explain the variance between AD and cognitively normal individuals.

A list of a few genes of interest

- AGPAT1** - 1-Acylglycerol-3-Phosphate O-Acyltransferase 1. This gene encodes an enzyme that converts lysophosphatidic acid (LPA) into phosphatidic acid (PA). LPA and PA are two phospholipids involved in signal transduction and in lipid biosynthesis in cells. This enzyme localizes to the endoplasmic reticulum. This gene is located in the class III region of the human major histocompatibility complex. Alternative splicing results in two transcript variants encoding the same protein.
- FOXN3** - Forkhead Box N3. The protein encoded by this gene acts as a transcriptional repressor. It may be involved in DNA damage-inducible cell cycle arrests (checkpoints).
- CDK2AP1** - Cyclin-Dependent Kinase 2 Associated Protein 1. Encodes for a specific inhibitor of the cell-cycle kinase CDK2
- RHOQ** - Ras Homolog Family Member C. The gene product regulates a signal transduction pathway linking plasma membrane receptors to the assembly of focal adhesions and actin stress fibers. It serves as a microtubule-dependent signal that is required for the myosin contractile ring formation during cell cycle cytokinesis. It also regulates apical junction formation in bronchial epithelial cells.
- MSR1** - Macrophage Scavenger Receptor. This encodes for a macrophage-specific trimeric integral membrane glycoprotein which has been implicated in many macrophage-associated physiological and pathological processes including atherosclerosis, Alzheimer's disease, and host defence.
- MAP1LC3A** - Microtubule Associated Protein 1 Light Chain 3 Alpha. MAP1A and MAP1B are microtubule-associated proteins which mediate the physical interactions between microtubules and components of the cytoskeleton. MAP1A and MAP1B each consist of a heavy chain subunit and multiple light chain subunits. The protein encoded by this gene is one of the light chain subunits and can associate with either MAP1A or MAP1B. Two transcript variants encoding different isoforms have been found for this gene. The expression of variant 1 is suppressed in many tumor cell lines, suggesting that may be involved in carcinogenesis
- TUBB2A** - Tubulin Beta 2A Class IIa, Microtubules, key participants in processes such as mitosis and intracellular transport, are composed of heterodimers

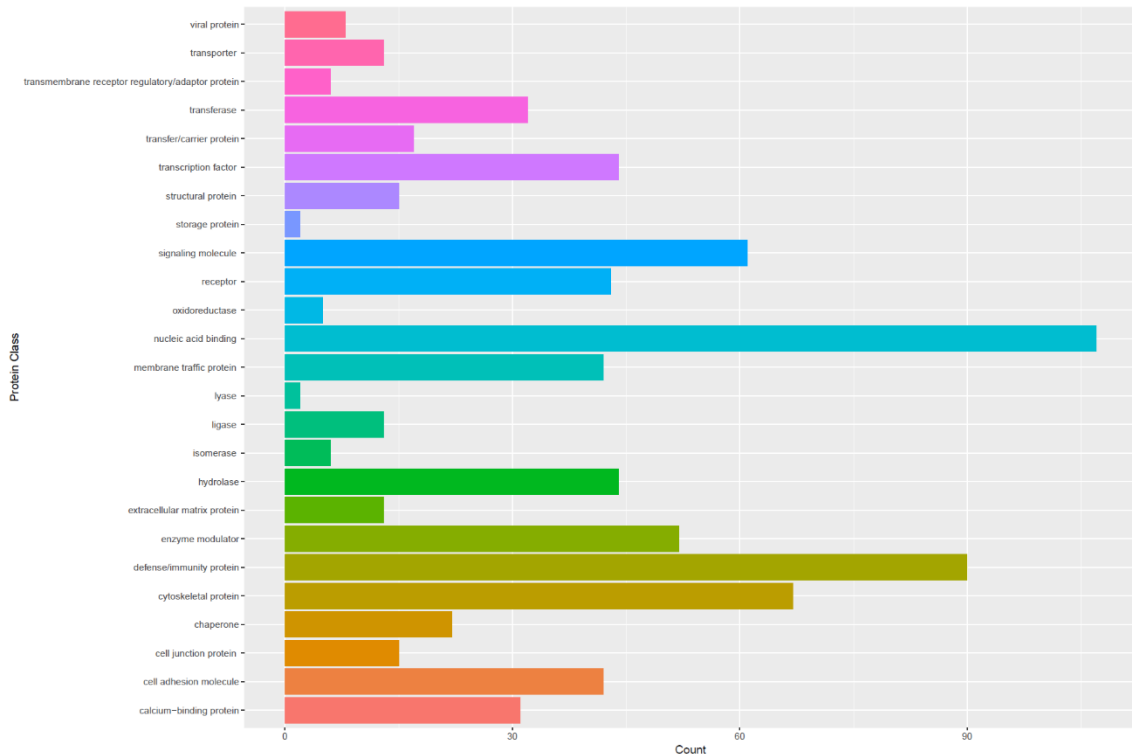
of alpha- and beta-tubulins. The protein encoded by this gene is a beta-tubulin. Defects in this gene are associated with complex cortical dysplasia with other brain malformations-5. Two transcript variants encoding distinct isoforms have been found for this gene. Although it is at the 100<sup>th</sup> position, it persists across multiple analyses as one of the top genes explaining variance between AD and healthy patients.

It is evident by these results that the sheer number of genes involved and identified as significant that it is impractical to attempt to define AD biomarkers at this stage. To combat this problem, the top 500 genes were selected and using resources such as GO, Panther and Bioconductor, were classified according to their molecular function and resulting protein products.



**Figure 27:** Molecular function ontology chart based on the Stepwise ANN results for E-GEOD-48350. Illustrated are the data for the molecular functions of the top 500 genes selected by the ANN. Results obtained via PANTHER on December 2017.

The molecular function, shown in in Figure 27, of the selected genes mostly falls in the binding and catalytic activity categories. Binding is a result of the prevalence of genes like FOXN3 and multiple ZNF variants, whereas the presence of genes that encode MAP kinases is responsible for the catalytic activity.



**Figure 28:** Protein class ontology chart based on the Stepwise results for E-GEOD-48350 showing the protein class of the top 500 genes selected by the ANN. Results obtained via PANTHER on December 2017.

The resultant proteins (Figure 28) on the other hand are significantly more diverse and evenly distributed. The four largest categories are nucleic acid binding proteins, transcription factors, enzyme modulators and hydrolases, with transferases and transporters making up a significant portion of the remaining ones.

However, knowing the molecular functions, protein products and other biological processes, while crucial to our understanding of the disease, are hardly sufficient to reach a conclusion that allows for the discovery of new and validation of previously identified biomarkers. This is especially true in situations where the bias or variance of the data is too high. Highly biased data resulting from the examination of specific cells in tightly controlled environments will only allow for the expression of genes that are directly related to the predefined conditions leading to confirmation bias, whereas high variance will lead to a significantly more even distribution of protein classes and molecular functions, closely mirroring their distribution in a real environment, and avoids the issue of noise masking small but crucial distinctions between such environments. So far, this experiment falls in the second category as the data have not been classified according to

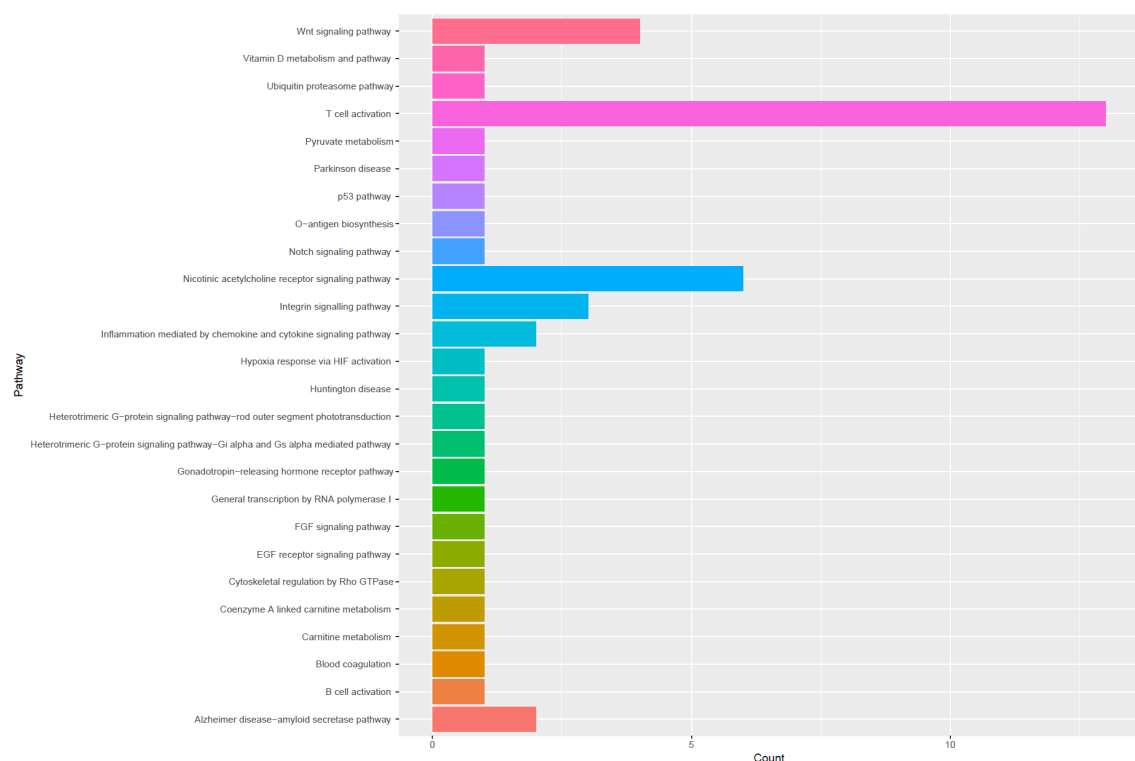


population subclasses. In order to reduce the variance and gain further insight into AD, a pathway analysis was carried out

#### 4.3.4 Pathway Analysis

In order to evaluate the role of the genes selected by the interaction matrix and determine their significance in the development of AD, these 500 genes were used in a PANTHER ontology search in order to determine which pathways their protein products belong to and which diseases these pathways have been associated with. PANTHER is a curated database designed to classify proteins and their genes in order to facilitate high-throughput analysis (Thomas *et al*, 2003)

As shown in Figure 29, some of the major pathways include the T cell activation, p53 and inflammation pathways, which are related to chronic microglial activation and immune response, the microtubule and beta tubulin, axon guidance and cytoskeletal regulation pathways are associated with neuron structure and the tau protein and the Alzheimer's disease presenilin pathway associated with A $\beta$  production.



**Figure 29:** Pathway ontology chart based on the Stepwise results for E-GEOD-48350 showing the pathways of the top 500 genes selected by the ANN. Results obtained via PANTHER on December 2017.

### 4.3.5 Conclusion

Based on the results of the stepwise algorithm, E-GEOD-48350 was identified as the superior dataset for this study, which was further confirmed by network inference. This is due to the consistency of results between tests rather than significantly higher performance or lower test error. Successive repeats show that small deviations in E-GEOD-48350 still preserve essential information such as key drivers and potential markers, while the increased complexity but smaller size of E-GEOD-5821 results in the addition of significant amounts of noise as different areas of the brain are affected by the disease in drastically different ways which will be explored in Chapter 5.

## 4.4 Network Inference

In order to make full use of the results obtained from the Stepwise approach, it is paramount to understand not only the function of the genes or the pathways they are involved in, but their current interactions with other genes in a dysregulated and dysfunctioning system. Considering that living systems encompass constant interactions between multiple organs, organelles and molecules as well as a significant number of environmental factors, it is a challenge to identify the few genes that are responsible for a given disease or that can be used to combat it, whether by prevention or treatment.

AD is especially challenging as not only is the cause unknown, but the system is isolated from the external environment due to the blood-brain barrier, limiting but not eliminating external influences. Not only is the way neurons transmit information and form new synaptic connections to learn unclear, but the mechanisms that are used by the brain to fight infection and clear debris are significantly more sensitive than the rest of the human body. This, coupled with the sheer complexity of the human brain from the way it learns to how it stores and accesses information, necessitate a novel systems biology approach to account for all such variables.

### 4.4.1 Interactomes

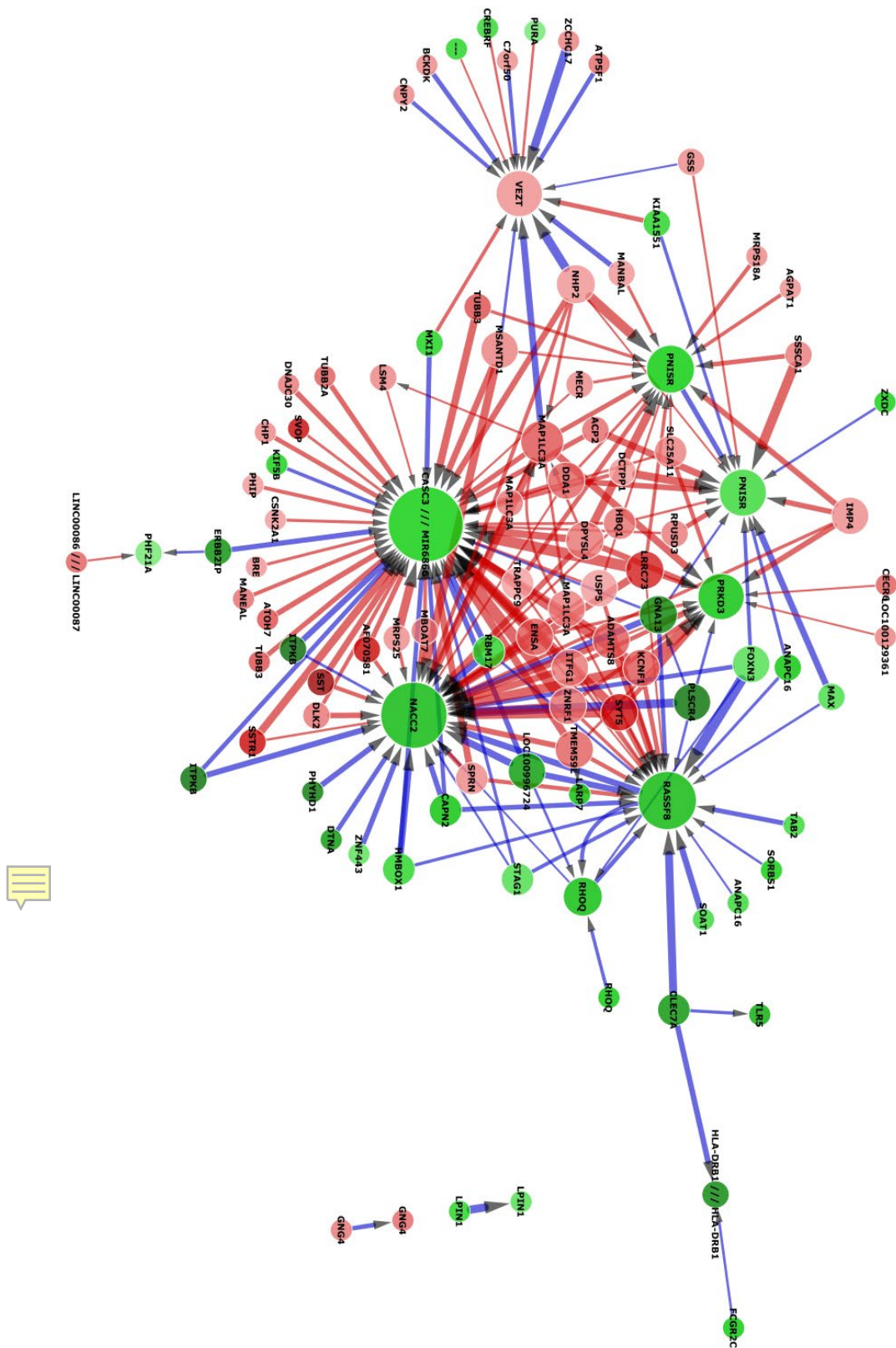
The results obtained from the stepwise ANN approach, shown in Table 1, were further analysed with an interaction algorithm developed by Lemetre *et al* (2010) to perform network inference. The interaction algorithm allows for the iterative quantification of the influence that multiple genes might have on the expression level of a single gene, until all

the genes within the data have been quantified this way, using the same parameter values as those utilized for the ANN stepwise algorithm (Lemetre *et al*, 2010). This allows for the determination of the central role of the most influential genes selected by the stepwise ANN within a system. The interaction algorithm predicts a single probe and assigns a weighted score which is directly proportional to the intensity of linkage between itself and the expression values of all other gene probes (Blair *et al*, 2013), while the intensity and directionality of the interaction between a source and target are determined based on the sum of the weights from an input to an output. The association between gene pairs can be bi- or unidirectional and be either stimulatory or inhibitory. This process was repeated until all gene probes were used as an output iteratively and a large matrix of interaction scores was generated by averaging values across 10 iterations. The results were visualised using Cytoscape.

### Master

The first iteration examined, dubbed the Master set, seeks to explore gene interactions in a non-parametric manner, as described in section 2.3. The goal of this approach is to present and analyse the data when the variance is at its highest while still being statistically significant, while the bias is virtually non-existent. Arnold *et al* (1999) described this approach in their *Kendall's Advanced Theory of Statistics* where they posit that a parametric hypothesis should have a normal distribution with a specified mean and statistically significant variance or have a given mean but unspecified variance. Alternatively, non-parametric approaches either have a normal distribution of normal form with unspecified mean and variance, or two unspecified continuous distributions that are identical. In the latter case the approach is also distribution-free which allows the user to increase the parameters of the training data in an iterative manner and make no assumption about population distributions (Murphy, 2012).

For the Master set, the only parameters that were provided to the interaction algorithm, were the patient status for the Stepwise analysis, and the 200 most differentially expressed genes identified by said analysis. The results are shown in Figure 30.



**Figure 30:** Interactome of the top 100 interactions and 200 genes selected by Stepwise in the Master Stepwise including AD patients healthy controls. Red edges indicate inhibition, blue indicate stimulation. Red nodes indicate that the gene is under-expressed while green shows that it is overexpressed. The thickness of the line corresponds to the strength of the interaction between the nodes and the arrow indicates the target of the interaction. The size of the node corresponds to the number of connections with other genes. Major hubs include NACC family member 2, cancer susceptibility candidate 3, Ras association domain family member 8 and NACC family member 2.

The interaction map, or interactome, generated provides a general view of predicted gene-gene interactions in a set of brain samples in a population of both healthy and AD patients. It is expected, based on our understanding of the disease, that genes that control amyloid clearance, cytoskeleton regulation and microtubule formation to be both highly differentiated and have a significant impact on the network. It is also expected that genes likely to be responsible for the disease are significantly overexpressed and upregulated by factors that promote the disease while being downregulated by genes that are responsible for maintaining the correct function of the brain. The strongest interactions should be between closely associated genes and target dysregulation factors and could themselves be an indication of which pathways are responsible for the progression of the disease. Genes responsible for upregulation factors promoting the disease are potential drivers of the disease and the genes they affect could be potential targets for therapy. Biomarkers however, are expected to be found be, or directly connected to, the largest hubs since they are the easiest to detect.

The largest hubs in this network include:

**CASC3** - Cancer Susceptibility 3, a core component of the exon junction complex (EJC), a protein complex that is deposited on spliced mRNAs at exon-exon junctions and functions in nonsense-mediated mRNA decay. The EJC marks the position of exon junctions in the mature mRNA and the core components remain bound to spliced mRNAs throughout all stages of mRNA metabolism. Additionally, it stimulates the ATPase and RNA-helicase activities of EIF4A3 and plays a role in the stress response by participating in cytoplasmic stress granules assembly and favouring cell recovery following stress. Component of the dendritic ribonucleoprotein particles (RNPs) in hippocampal neurons. May play a role in mRNA transport (Mao *et al*, 2017). Its role in stress response and crucial function as part of hippocampal neurons makes it a likely candidate of dysregulation leading AD. Proper neuron structure and function are crucial to the maintenance of mental health and disruptions are directly correlated with the hyperphosphorylation of Tau and the formation of neurofibrillary tangles. Additionally, it appears to be overexpressed and downregulated by genes such as MAP1LC3A and TUBB3, both essential to the formation of microtubules, ZNRF1, a zinc and ring finger 1, E3 ubiquitin protein ligase which is associated with the immune system as well as mediating the ubiquitination of AKT1 and GLUL, thereby playing a role in neuron cells

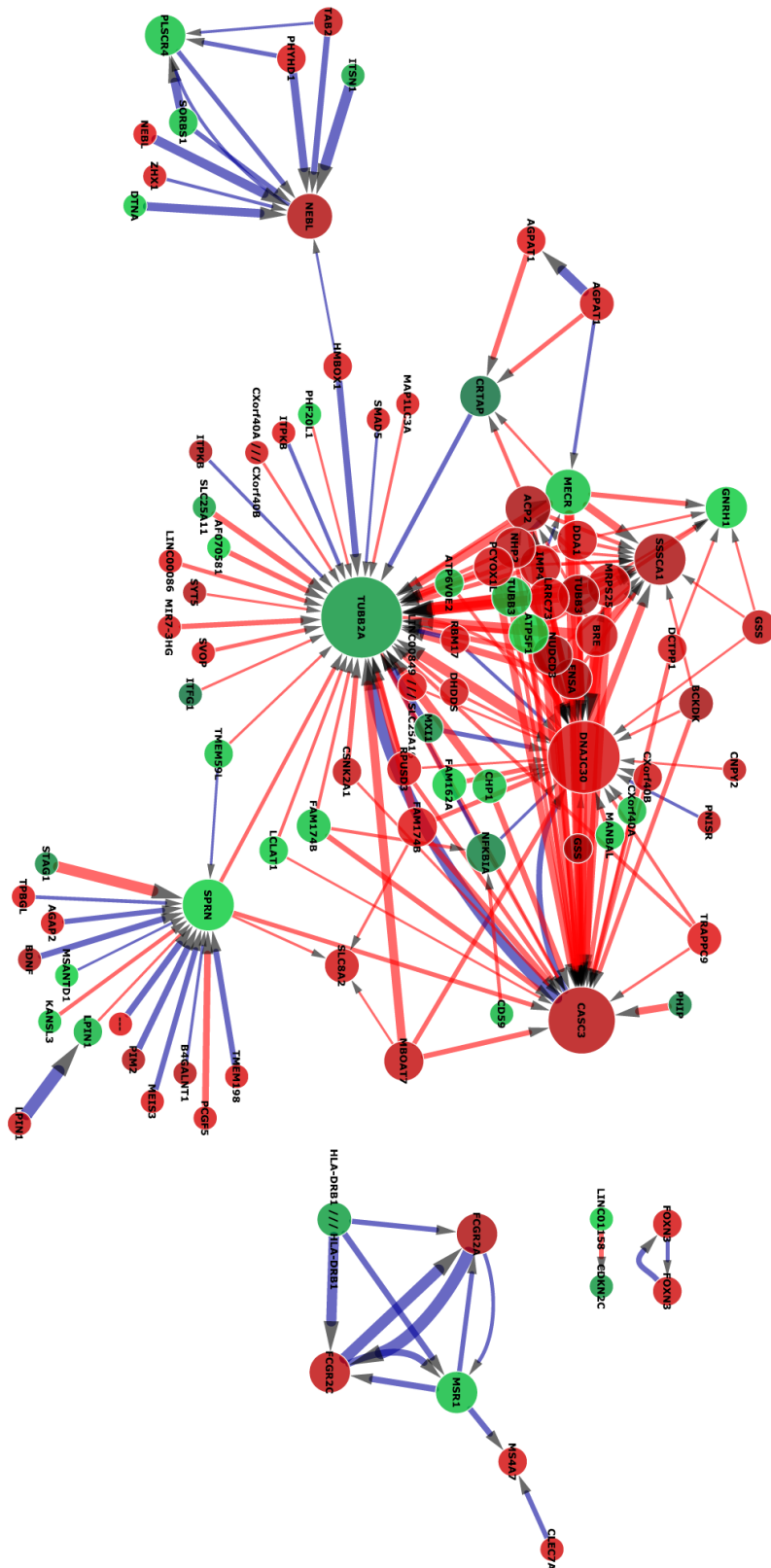
differentiation. Other factors that encode for multiple transmembrane proteins like TMEM59L and KCNF1, a voltage-gated potassium ion channel are linked to the normal function of neurons. Meanwhile, multiple kinases and kinesins appear to be trying to stimulate further expression of the gene.

**NACC2** – NACC Family Member 2, is highly expressed in the brain. It is involved in the recruitment of the NuRD complex to the promoter of MDM2, leading to the repression of MDM2 transcription and subsequent stability of TP53.

**RASF8** - Ras Association Domain Family Member 8. It is a member of the Ras-association domain family of tumour suppression proteins. Also overexpressed, this gene is crucial to the maintenance of junction function and migration of epithelial cells. Normally present at below average expression in the brain, it positively regulates and is regulated by NACC2 creating a cycle. It is predicted to strongly interact with both FOXN3 and RHOQ.

#### AD Only

After examining the results of the Master interactome for the E-GEOD-48350 dataset, it becomes clear that the variance in the data is simply too high. While there are definitely indicators of both possible markers and drivers of the disease, it is possible to obtain a far clearer picture by taking full advantage of the non-parametric distribution-free approach and add significant parameters to the algorithm that are directly correlated to the results obtained previously. By only selecting the most differentially expressed genes for AD patients exclusively, it is possible to create an interactome that creates a “snapshot” of the disease, an indication of the gene interactions in aggregate at the time of the examination. Expected genes include some, but not all of the key genes identified by the Master set and possibly different levels of expression and interaction.



**Figure 31:** Interactome of the top 100 interactions and 200 genes selected by Stepwise in AD. Red edges indicate inhibition, blue indicate stimulation. Red nodes indicate that the gene is under-expressed while green that it is overexpressed. The thickness of the line corresponds to the strength of the interaction between the nodes and the arrow indicates the target of the interaction. The size of the node corresponds to the number of connections with other genes. Major hubs include Tubulin 2A, cancer susceptibility candidate 3, DnaJ (Hsp40) homolog.

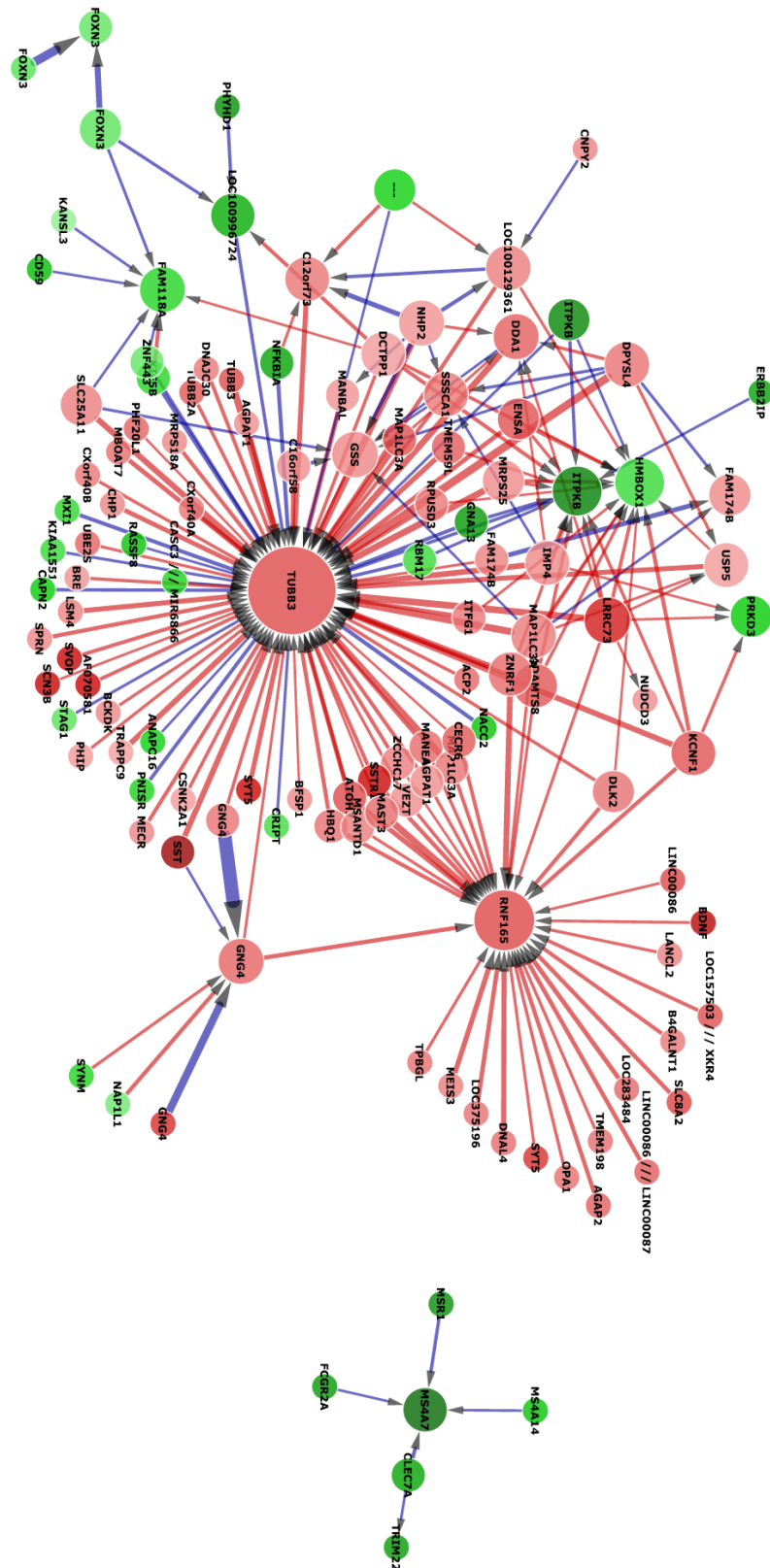
Indeed, it is immediately obvious in Figure 31 that there are significant differences and similarities to the previous interactome. CASC3, identified as a key hub in the combined AD/healthy interactome is still being suppressed by genes such as TUBB3, trafficking proteins and transferases, with one of the largest suppressor genes being BRE, a brain and reproductive organ-expressed tumour necrosis factor (TNFRSF1A) modulator. It is interesting to note that BRE actively suppresses all three major hubs in this network and has been suspected to be significant in homeostasis or cellular differentiation in cells of neural, epithelial and germline origins. Moreover, CASC3 in AD appears to be underexpressed and stimulating the expression of DNAJC30, which is the 2<sup>nd</sup> largest hub as well as the largest hub, TUBB2A which is also overexpressed.

TUBB2A, discussed earlier, appears to be suppressed by TUBB3, hinting at an adversarial relationship between the genes encoding these two tubulin isoforms. Moreover, it is further suppressed by MBOAT7, an acyltransferase and stimulated by highly overexpressed CRTAP, a scaffolding, cartilage associated protein. Unlike the Master set, genes such as AGPAT1, ITFG1 and NFKBIA are present here, with their interactions strong enough to be selected in addition to acting in suppression roles. Finally, the presence of the severely underexpressed NUDCD3 gene, which also suppresses TUBB2A, is known, when depleted at the protein level, to result in the aggregation and degradation of the dynein intermediate chain, mislocalization of the dynein complex from kinetochores, spindle microtubules, and spindle poles, and loss of gamma-tubulin from spindle poles. This affects the conversion of ATP to mechanical energy and could be a catalyst in the creation of NFTs.

### Healthy Controls

The addition of the healthy control as an independent predictor has no particular value by itself. After all, it is almost impossible to define a golden standard for a healthy human brain. This study has mitigated the problem by having a significant cohort of cognitively normal patients, which will allow us, when analysed in aggregate, to provide a counterpoint, a direct comparison not to a healthy brain, but a brain without the neurodegeneration present in AD. Of course, as explained earlier, the controls are all cognitively normal, showing no signs of neurodegeneration due to AD or other conditions, which should present varied enough distribution to reduce the error inherent each patient's individual genetic makeup.





**Figure 32:** Interactome of the top 100 interactions and 200 genes selected by Stepwise in cognitively normal controls. Red edges indicate inhibition, blue indicate stimulation. Red nodes indicate that the gene is under-expressed while green shows that it is overexpressed. The thickness of the line corresponds to the strength of the interaction between the nodes and the arrow indicates the target of the interaction. The size of the node corresponds to the number of connections with other genes. Major hubs include Tubulin 3A, ring finger protein 165.

Immediately it is evident the interactome present in Figure 32 has both crucial similarities and differences with the Master set (Figure 30) as well as the AD interactome (Figure 31). The largest hub is TUBB3 which is underexpressed but regulated both positively and negatively by a wide variety of genes with various expression levels. Similarly to TUBB2A and other genes of the tubulin family, TUBB3 encodes the class III isoform of the beta tubulin family, which is more abundant in neurons than other cells in the nervous system. Mutations in this gene have been shown to cause congenital fibrosis of the extraocular muscles type 3 as well as cortical dysplasia. It appears to be downregulated by factors such as MAP1LC3, ITGF, ZNRF1 as well as TUBB3 itself, possibly as a feedback loop, while being positively regulated by genes such as NFKBIA, CASC3 and NACC2.

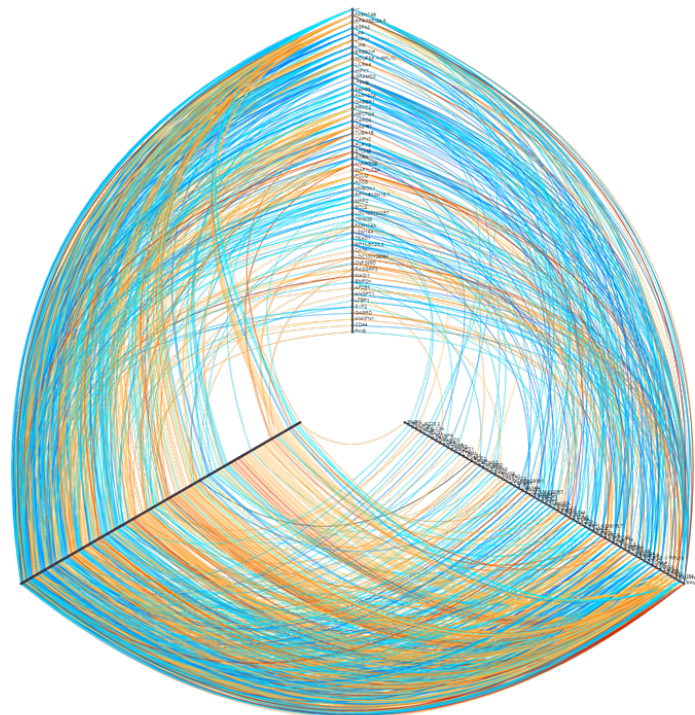
This overlap between the AD and Master sets not only reinforces the previous results by verifying that even after multiple analyses, the results stay consistent, but also allows us to detect the overlap between healthy and AD and cognitively normal individuals. We can use this overlap between interactomes to extrapolate a set of genes that are differentially expressed between these two groups by determining which genes are drivers for the disease, which ones are responsible for maintaining a healthy state, the ones that are essential for the function of the organ and those that can be used for prognosis. This will be explored in the following chapter.

#### 4.4.2 Hive Plots

One of the key challenges in the field of bioinformatics is the issue of visualisation. As shown in figures 30, 31 and 32, the amount of information present can be overwhelming and it is not always clear what is important and what isn't. And while these approaches have expanded the field of biomarker discovery by allowing researchers to consider new possibilities, their use in diagnostics is limited by the fact that the results often require expert specialist to interpret them. If these approaches are to achieve widespread use by clinicians for prognosis, it is paramount to have a clear and easily understandable output.

Developed by Krzywinski *et al* (2012), hive plots offer an alternative network visualisation method to traditional maps. These maps, usually produced by software such

as Cytoscape, Gephi, Netminer and more recently, programming languages such as R, tend to include an overwhelming amount of information, leading to networks that need to be analysed with sorting algorithms to be readable, making them hard to interpret, in addition to that fact that their complexity increases exponentially as more information is included. Hive plots offer a rational visualisation technique, which groups nodes based on specific properties determined by the user as shown in Figure 33. The properties can be inherent network statistics, or information such as features of clinical data.



**Figure 33:** Hive plot variant of figure 31. Top 500 genes and 1000 interactions in AD. The genes on the bottom left axis are all sources and affect the genes in the other two axes. Genes on the bottom right axis are all targets and are regulated. Genes on the top axis are both. They are arranged by number of connections, with genes at the end of the axes being more influential. Blue edges indicate positive and red edges indicate negative regulation.

## 4.5 Driver Analysis

One of the challenges faced when trying to elucidate a marker, driver or therapy target is the selection criteria used. It is crucial to point out that the data used in these experiments presents us with a “snapshot” of the condition investigated, a generalised picture of how each gene is affected by every other gene, while the biological system is in a state of imbalance. As a result, the biggest hubs of most interactomes tend to be either the genes most up- or down-regulated in the network at the time. This has two potential interpretations. The first is that the hub is the source of the imbalance and thus, the most

likely driver of the disease and target for therapy, and the downregulation is a result of the system attempting to restore balance. Alternatively, the hub is the factor preventing the imbalance by working against the disease and is being upregulated in an effort to restore the system to its original state.

The purpose of the driver analysis is to provide a non-biased selection condition based on the sum of the weights each gene exerts on the network, quantifying the amount of influence on a target and the amount of influence of a target. As explained in section 4.4 the interaction algorithm analyses the selected genes in a pairwise manner and assigns each of those pairs a value predicting how strongly their genes interact. Hence, by summing the weight that each source gene exerts on each target and vice versa, it becomes possible to rank them by which ones have the greatest overall effect on the network and which ones are the most affected.

The advantage of this method is the fact that it considers and gives equal importance to non-hubs as it only measures the total effect each gene has on the totality of the network. As such, it is possible to draw attention to genes with a multitude of weak interactions rather than only a few strong ones, which might otherwise not be visible. It is reasonable to assume that such genes may not be the greatest drivers of the disease, but crucial components of the system, and this method allows us to analyse those genes without them being obscured by the hubs and most likely drivers, thus giving a wider and impartial view of the condition. Moreover, the driver analysis is not affected by the complexity of the question, being able to provide comparable results across multiple datasets, in both focused and general conditions.

#### 4.5.1 Master Driver Analysis

Amount of Influence	Gene Symbol	Amount Influenced	Gene Symbol
-344.9320924	KCNF1	-244.2590811	DNAJC30
-332.0028917	MAP1LC3A	-241.0185735	LOC375196
-330.6003391	LRRC73	-240.0149244	SSTR1
-317.5263926	DPYSL4	-239.3294991	AF070581
-315.8452038	TRAPP9	-239.1822618	GSS
-315.0302745	TMEM59L	-239.0658587	RIMS1
-314.3752507	ENSA	-236.8055352	ATOH7
-313.5982365	MAP1LC3A	-233.6844669	MRPS18A
-312.7852517	MAP1LC3A	-231.5888613	PCLO
-311.5632505	ZNRF1	-231.1482094	CDIPT
-306.7346501	SPRN	-231.0994368	SLC8A2
-306.0526841	SYT5	-229.7103817	BRE
-300.3288704	USP5	-228.4404714	NUDCD3
-298.0616905	SST	-227.8733555	C16orf58
-292.4715367	NHP2	-227.3175204	LOC283484
-289.4455	LOC100129361	-226.9368887	GNG4
-288.8005029	MSANTD1	-226.5620476	B4GALNT1
-287.3145565	ADAMTS8	-224.9593742	FAM174B
-286.7106396	SVOP	-223.4391191	CXorf40A
-284.345488	IMP4	-218.8763297	PCYOX1L
-283.9942664	SLC25A11	-216.9777515	DPH2
-281.3528233	AGPAT1	-216.7105047	TUBB2A
-278.9239018	MRPS25	-215.5450909	FAM174B
-278.0725063	PHIP	-212.4694895	UBE2S
-277.0378952	ITFG1	-210.0301289	XKR4
-276.4766275	CECR6	-209.4291068	BCKDK
-274.6556925	C12orf73	-208.911844	ZCCHC17
-270.6691755	AGPAT1	-206.3503681	CXorf40B
-270.1441496	TUBB3	-205.7967278	PHF20L1
-268.5252442	TMEFF1	-205.4487739	GNG4
-268.2006988	HBQ1	-204.7230999	MIR7-3HG
-266.9155943	ACP2	-202.2240239	LINC00849
-266.5628294	MBOAT7	-201.5178359	LINC00086
-264.4619408	CSNK2A1	-198.5046458	VEZT
-264.3780584	MECR	-195.1656867	ATP5F1
-263.4438993	LSM4	-193.8301586	DHDDS
-262.8919387	SYT5	-193.1890934	ATP6V0E2
-262.4929117	CHP1	-192.5004959	LANCL2
-261.2023843	RPUSD3	-192.2768812	AGAP2
-258.1573147	KLHL35	-187.9150484	PIM2
-257.2947927	MANEAL	-187.8787111	CCBL2
-255.8758544	DCTPP1	-186.1024367	C7orf50
-255.5058804	TUBB3	-184.8851746	CNPY2
-255.48536	MAST3	-184.6035574	RASAL1
-255.1961869	MANBAL	-178.2096346	ALKBH6
-253.2531367	BFSP1	-176.5272101	ZRANB3
-253.037765	DLK2	-172.760756	CXorf40A/B
-252.765391	SCN3B	-170.5026804	LINC00086/7
-252.0522753	DDA1	-168.1139266	TPBGL
-247.5769677	SSSCA1	-167.8759895	LCLAT1

**Table 2:** Driver analysis showing the top 100 source genes of the Master set according to their impact on the network.

The Master set driver analysis (Table 2) presents information on the strongest sources, ie genes that target and regulate other genes in the network. While it is impossible to see which genes they affect, a task better handled by the interactome analysis earlier, it is possible to obtain a truly non-biased view of the overall effect each gene has on the network as a whole, preventing the strength of individual interactions to mask consistent but weaker ones. While only 100 genes are shown the analysis was performed on the entire cohort. The full tables are available in the appendix.

It is immediately obvious that all the interactions targeting other genes in the top 100 of them are negative, indicating inhibition. In fact, the first positive interaction is at position 118 from PLSCR4 (see appendix). The interaction values range from -344.93 to a mere 1.26, and while these values are irrelevant by themselves, as the algorithm weights these interactions compared to each other. When using the same dataset and parameters it is possible to understand the general trends present in the most differential genes. It is also interesting to note that while the top genes are present in the interactome, their position and importance are not immediately obvious. In fact, KCFN1 appears to have similar effects to other essential genes, so to see that the gene encoding this voltage-gated potassium channel modifier offers the largest degree of inhibition in the network is information that would otherwise be lost. Similarly, MAP1LC33 is present in most interactomes based on data for AD patients and is most likely a highly mechanistic factor, playing a crucial role in the continued function on neurons. Thus, dysregulations in the gene are potential drivers of the disease and targets for therapy. However, it is unlikely to make a good biomarker as its expression is not significantly differentially expressed between patients.

<b>Amount of Influence</b>	<b>Gene Symbol</b>	<b>Amount Influenced</b>	<b>Gene Symbol</b>
-295.6069097	RNF165	-156.4929965	RHOQ
-295.5138421	MPP2	-155.7561406	FAM118A
-291.9052966	CDKN2C	-155.0436421	NACC2
-284.8066644	LARP7	-152.9363773	MAP1LC3A
-246.9437802	CD59	-151.3816425	LSM4
-243.4545563	NFKBIA	-149.2511146	STAG1
-240.1274623	---	-148.3278591	CAPN2
-235.2026242	RAB27A	-148.0555927	RAB27A
-234.8746866	KIAA1551	-147.5300977	PLSCR4
-218.0155809	CASC3	-147.4155331	---
-217.9567192	PRKD3	-146.2200136	NAP1L1
-216.0890115	MS4A14	-144.6497204	IL13RA1
-214.5850067	ATP5F1	-141.5503578	CXorf40B
-212.4036659	RHOBTB3	-140.7434558	KANSL3
-211.9562747	ZNF443	-140.1021799	MS4A7
-211.6572475	ITPKB	-139.6374288	HMBOX1
-211.1450832	PNISR	-138.8743338	FOXN3
-210.4696329	CRIP1	-137.5025809	SSTR1
-203.0202787	SOAT1	-134.1028393	RHOQ
-202.3836003	CNPY2	-130.9343685	RIMS1
-202.0812448	ERBB2IP	-130.8518394	LCLAT1
-201.592368	ZRANB3	-130.2547258	PURA
-200.4616607	CRTAP	-130.0010477	IFNLR1
-199.4777917	RBM6	-129.9864822	MIR7-3HG
-198.0265765	KLHL35	-129.9352987	ADAMTS8
-196.6397458	TRAPPC9	-129.0403014	PHIP
-196.481397	TM7SF2	-127.8774964	PNISR
-194.4346937	GNRH1	-127.4344562	LOC100996724
-194.3018408	SLC8A2	-126.9756588	RBM17
-192.8069306	---	-126.8301473	TMEM59L
-191.7375446	PCGF5	-126.4250353	LINC01158
-187.1058574	ATP6V0E2	-125.4268825	OPA1
-185.282554	CDIPT	124.7523294	VEZT
-184.396749	ZXDC	-124.4003324	ANAPC16
-183.3796502	NUDCD3	-123.7419584	PCYOX1L
-183.0838539	DCTPP1	-123.1535765	ALKBH6
-182.2584838	SSFA2	-122.6408724	DLK2
-181.7868251	MAX	-121.380003	LANCL2
-180.9677478	CARTPT	-120.3348251	DDA1
-180.1499998	MRPS25	-120.0544629	SMAD5
-178.2302989	B4GALNT1	-118.5565096	CXorf40A
-175.8945416	ITFG1	117.9144891	SCN3B
-172.7780092	PHF20L1	-116.7821859	ENSA
-171.7759594	MAP1LC3A	-116.0499531	USP5
-167.7323661	ACP2	-115.5448579	EGR4
-167.1945548	FAM162A	-115.4805473	TUBB3
-166.5068724	C16orf58	-115.2653275	TNPO1
-161.9070028	KIF5B	-113.2697898	SYT5
-156.8263258	SYNM	-112.7422802	CLEC7A
-156.6992322	LPIN1	-112.2529862	TLR5

**Table 3:** Driver analysis showing the top 100 target genes of the Master set according to their impact on the network

Meanwhile the target data (Table 3) show how the same genes are affected by the sources, and thus potential drivers of the disease, overall. It is quite likely that these results can be used to predict potential markers of the disease as they are predicted to be the genes that are most highly dysregulated. It is also interesting to note that while the source table mostly includes genes that have strong interactions, the target table consists mostly of genes likely to show as hubs in an interactome. This is a flaw of any interactome as it is rather target focused. The trend presented by the weight reduction proceeds at roughly the same rate as the source table, making them directly comparable.

What is immediately clear in this table is the largest hubs of the interactome, the ones with the most connections, being targeted by the largest number of genes, are not necessarily the top genes when considered in aggregate. In fact, the largest hub of the Master interactome (Figure 30), *CASC3*, is in 10<sup>th</sup> position of the driver analysis of the same data. Factors such as *NFKBIA*, which plays a crucial role in neuroinflammation as shown in Chapter 1, are significantly higher in that list. *MPP2* is especially interesting, being a palmitoylated membrane protein 2 and member of the MAGUKs (membrane-associated guanylate kinase homologs family of membrane-associated proteins termed) family, which interact with the cytoskeleton to regulate cell proliferation, signaling pathways, and intracellular junctions. The *MPP2* protein in particular, contains a conserved sequence SH3 (src homology 3), which is found in several other proteins that associate with the cytoskeleton and modulation of signal transduction. Moreover, it has shown as a hub in other tests using different parameters, including when datamining E-GEOD-48350 multiple times to ensure consistency. Finally, the presence of *CD59* near the top of the list is a very interesting find as the gene is involved in lymphocyte signal transduction, activation of T cells and is a potent inhibitor of the complement membrane attack complex, which is essential for the formation of osmolytic pores. Mutations in the gene are associated with cerebral infarctions.

As discussed in the previous section however, the Master set is very useful for providing a general unbiased overview of the brain when both healthy and AD individuals are taken into account, as well as providing a framework to differentiate those two groups. Of course, in order to reduce the variance and therefore noise in the data it is essential to repeat this test on each of the categories present.



#### 4.5.2 AD Driver Analysis

Amount of Influence	Gene Symbol	Amount of Influence	Gene Symbol
-182.1386831	MECR	-110.0662634	GSS
-159.5834977	BRE	-109.0934141	SLC25A11
-158.1099534	MRPS25	-108.9590435	RIMS1
-152.2246763	DCTPP1	-108.5030534	ADAMTS8
-147.4537491	PCYOX1L	-108.3919946	LINC00086
-146.1799208	LRRC73	-108.2252253	C16orf58
-143.9485194	TRAPPC9	-107.8966336	KLHL35
-143.1575419	SPRN	-103.6494094	FAM162A
-142.8118825	CHP1	-103.4715083	LSM4
-141.4160604	NUDCD3	-102.7548017	DHDDS
-141.0775437	FAM174B	-102.5113531	KCNF1
-138.601334	PHIP	-100.3149574	LINC00086/7
-138.5179659	USP5	-100.2204221	ITFG1
-134.4538932	DNAJC30	-99.11777999	MAST3
-133.3060111	ATP5F1	-97.2307461	AF070581
-131.3691153	ACP2	-96.20262049	DDA1
-130.3864472	BCKDK	-95.86815586	EGR4
-129.0010986	TUBB3	-95.80581404	MIR7-3HG
-128.4620616	FAM174B	-95.13247563	MAP1LC3A
-126.5405171	TUBB3	-94.87485074	MRPS18A
-125.3377254	TPBGL	-93.99985858	SSSCA1
-124.8112016	ENSA	-93.29011687	PIM2
-123.74762	MBOAT7	-92.44534428	SYT5
-122.9224799	AGPAT1	-91.3142451	BFSP1
-122.76263	MAP1LC3A	-90.7062403	GNG4
-122.0289962	TMEFF1	-90.65214721	CSNK2A1
-120.1824465	ZRANB3	-89.12010694	CCBL2
-120.0542102	ALKBH6	88.08615246	HMBOX1
-119.9261977	MAP1LC3A	-84.99343162	SLC8A2
-119.1025194	CXorf40A/B	-84.89780465	C12orf73
-118.2783473	GSS	-83.87450702	SST
-118.2513712	CXorf40A	-83.60004752	PCLO
-117.6621415	MANBAL	-83.07529812	LOC157503
-116.8300074	NHP2	-82.16804018	ZNRF1
-116.3334473	LINC00849	-81.7523655	TMEM198
-116.0812768	DLK2	-80.81838571	LCLAT1
-115.8122146	HBQ1	-80.51338384	MANEAL
-115.553884	ATP6V0E2	-79.615437	SCN3B
-115.3596558	TMEM59L	-79.01180814	SYT5
-115.2569221	TUBB2A	-78.65405045	OPA1
-114.824704	TM7SF2	-76.62936544	RASAL1
-114.4498035	LOC100129361	76.37000874	NFKBIA
-114.1397396	CNPY2	74.87758302	NEBL
-113.2421731	CDIPT	74.86398012	CAPN2
-113.0615487	SVOP	-74.19434296	UBE2S
-112.7488907	IMP4	-74.14761085	RNF165
-112.4748854	CECR6	-73.08625641	VEZT
-111.8793248	RPUSD3	72.23933782	DTNA
-111.4082992	AGPAT1	-72.06927054	C7orf50
-110.6449825	CXorf40B	-70.63588881	PHF20L1

**Table 4:** Driver analysis showing the top 100 source genes of the AD set according to their impact on the network.

When compared to the Master table, in AD (Table 4) the trend of the data appears to be reversed. The sources appear to have an overall weaker effect on the network than the targets, the range of the weights being -182.14 to 0.93 for the sources and -243.94 to 0 for the targets, indicating that when the noise is reduced, the impact of the genes dysregulated in the network is greater than the genes causing the dysregulation.

Also, much like the Master set, the order of hubs and drivers is different. While BRE remains an important driver, MECR has significantly higher impact than first thought. BRE is predicted to be important based on information by both the interactome and the driver analysis, while the impact of MECR is only clear in the driver analysis. Although MECR is involved in fatty acid elongation and metabolism and has been shown to be associated with multiple diseases, it has never been linked to neurodegeneration. Moreover, multiple genes linked to the management of phosphate in the body such as ACP2 and DCTPP1 attain prominence through this method. Genes related to energy management such as ATP5F1, an ATP synthase and multiple genes responsible for the transport of molecules are present and prominent. Finally, multiple tubulin variants and other factors related to microtubule formation are present, validating the results obtained through the interactome.

Amount Influenced	Gene Symbol	Amount Influenced	Gene Symbol
-243.9362947	NFKBIA	95.1592248	CHP1
-204.8803941	MPP2	-94.97707761	AGAP2
-203.1684278	CASC3	-94.78535562	MRPS25
-202.1148961	CRTAP	-93.29199133	SOAT1
-189.9328862	LOC283484	-93.02043903	SLC25A11
-183.1340921	RHOQ	92.95818554	NEBL
-168.5932874	LINC01158	-89.19683302	TMEM198
-162.3023047	SSSCA1	-88.20571643	ADAMTS8
-158.0796119	DPH2	-87.75606472	RPUSD3
-156.5903832	NAP1L1	87.24289995	LINC00086/7
-151.2712694	TUBB2A	-87.16643655	KIAA1551
-149.1587904	PNISR	-86.98821533	RHOBTB3
-145.8134477	MXI1	-86.49191671	ATP6V0E2
-144.5393016	CRIP1	-85.70089494	ANAPC16
-141.983114	CDKN2C	-85.04905916	IFNLR1
-141.4163932	TM7SF2	-84.25908194	ANAPC16
-139.1827213	SMAD5	84.20632957	RASAL1
-137.6643741	RBM6	-83.22683976	SSFA2
-136.0585466	RBM17	-82.7406998	DDA1
-131.1232322	PNISR	-82.47893052	SVOP
-130.501559	CAPN2	81.80847792	LPIN1
-127.9621124	DNAJC30	79.68006188	MANBAL
-127.9597284	RAB27A	-79.46229619	IMP4
-126.2716554	B4GALNT1	-78.42937775	SYT5
-123.9710662	CD59	-78.18631572	---
-122.2632773	OPA1	-77.84423807	SST
121.3426363	MECR	-77.60351709	HMBOX1
-118.927459	---	-77.39595343	FAM162A
-118.1313983	ACP2	-77.2765248	STAG1
-115.2617753	DLK2	-76.55492724	TPBGL
-115.2267633	FOXN3	-75.69138322	USP5
-111.778133	KIF5B	-75.64504438	TLR5
-111.3621972	KANSL3	-75.07030272	CDK2AP1
-110.5266025	---	-74.3680467	ITFG1
-109.251917	ITPKB	74.09671932	VEZT
-106.7183667	BFSP1	-72.42907592	KCNF1
106.5614286	BCKDK	-72.09524192	MANEAL
-106.1307335	PPP1CC	-70.87094747	MAX
-105.8565258	RHOQ	-70.15294438	MAP1LC3A
104.2470044	BRE	-69.74324176	GNA13
-102.4273651	LARP7	69.21059613	WDR18
-102.0573963	GNRH1	-67.10926795	PRKD3
-101.6725985	MS4A14	-65.93977793	IL13RA1
-100.8067491	SYT5	-65.09039257	DPYSL4
-99.68302581	SLC8A2	-63.30007291	KLHL35
-99.61550915	LOC100996724	63.18712549	NEBL
-97.85840191	TUBB3	-62.86952041	PCYOX1L
-97.51079723	TMEM59L	62.42801714	PLSCR4
-95.30363152	MAP1LC3A	-61.85914028	DCTPP1
-95.22983168	PURA	61.64292155	FOXN3

**Table 5:** Driver analysis showing the top 100 target genes of the AD set according to their impact on the network.

As shown in Table 5, the effect of the target genes on the network, the potential biomarkers and targets for therapy, is significantly more varied. Positive interactions appear earlier than the rest and the strength of the interactions drops quickly, with an almost 20% decrease from the first to second positions. What is very interesting however, is the presence of factors suspected to play a crucial role in the development of AD very high in the list. NFKBIA, RHOQ, MPP2, CASC3 and tubulin variants are all present and highly dysregulated. In fact, the rapid decrease in interaction strength until position 32 (FOXP3), reaffirms that these genes are significantly more dysregulated than normal and actively being suppressed. Moreover, they all support the hypotheses presented in Chapter 1 related to gaps in the amyloid beta theory and the proposed inflammation hypothesis, without disputing proven factors such as the importance of APP or tau.

To complete this series of experiments, however, it is essential to know how the healthy controls compare to both the AD and Master sets. Overlap between these three should indicate factors that are essential to the function of the brain, regardless of dysregulation as until a patient succumbs to the disease, normal function has not stopped, just being disrupted. The cause and potential markers can be found in how much these two conditions diverge from each other. Moreover, genes present in both the Master set and AD set are most likely essential to the progress of the disease. Conversely genes present in the Master and healthy sets, but absent in the AD set are likely to be factors that can be used for prevention.

### 4.5.3 Healthy Driver Analysis

Amount of Influence	Gene Symbol	Amount of Influence	Gene Symbol
-219.7890081	KCNF1	-154.4252554	DNAJC30
-217.8977369	ENSA	-153.8438645	NHP2
-201.1928644	MAP1LC3A	-153.4988645	DPH2
-200.6455532	TRAPPC9	-153.4976702	SCN3B
-200.0633687	LRRC73	-153.0764731	GSS
-199.6888257	USP5	-152.8861622	NUCD3
-198.8483625	DCTPP1	-152.3825606	LSM4
-198.2033843	MAP1LC3A	-149.2278934	MRPS18A
-197.4110719	TMEM59L	-147.8421792	CDIPT
-196.472449	ADAMTS8	-147.2322794	SLC25A11
-194.1985939	MAP1LC3A	-145.5588682	CHP1
-194.0896538	SYT5	-145.1360367	---
-193.1686487	ZNRF1	-143.3272651	C16orf58
-188.4855853	SPRN	-138.8601386	MANBAL
-185.9048485	MSANTD3- TMEFF1	-138.2850936	BRE
-185.3449112	ACP2	-137.9653693	LANCL2
-184.3247506	SVOP	-137.789322	LINC00849
-182.3766567	PHIP	-137.7619685	AF070581
-181.1850507	MAST3	-137.1631822	PHF20L1
-179.6302113	ITFG1	-136.2795038	ATOH7
-178.8265106	LOC100129361	-135.3447047	TUBB2A
-178.4084616	SYT5	-134.9713193	GNG4
-178.0958956	CSNK2A1	-133.381258	PCLO
-177.4502159	DLK2	-133.3078383	BCKDK
-173.4453284	CECR6	-129.7705703	FAM174B
-173.0089619	AGPAT1	-128.7479423	FAM174B
-172.6739967	DPYSL4	-125.7117172	RNF165
-172.5012996	IMP4	-124.4151796	VEZT
-172.490897	AGPAT1	-123.6192103	LOC375196
-172.4616359	RPUSD3	-122.3607693	C7orf50
-170.5902293	TUBB3	-119.4724442	LOC283484
-168.9139457	RIMS1	-119.3154824	LINC00086/7
-168.7283549	HBQ1	-119.0923605	CXorf40A
-167.0813607	MRPS25	-116.3363936	DHDDS
-165.9981127	TUBB3	-115.2999034	ZCCHC17
-165.8881007	BFSP1	-115.2382008	CCBL2
-165.6247537	SST	-114.7920584	TM7SF2
-161.9981384	MANEAL	-114.5485256	AGAP2
-161.0374528	SSSCA1	114.1790576	CLEC7A
-160.7676669	MBOAT7	-112.008759	XKR4
-160.7132947	PCYOX1L	-107.9512567	LINC00086
-159.8264398	SLC8A2	-107.0458271	MIR7-3HG
-159.3695579	MECR	-106.3001334	RASAL1
-158.6228504	C12orf73	-106.155753	ATP5F1
-158.0791634	GNG4	-105.6383663	CXorf40B
-157.9231056	B4GALNT1	-104.9146807	ZRANB3
-157.1613359	DDA1	-104.8095803	TPBGL
-156.5389736	SSTR1	-104.7689174	ALKBH6
-156.1484712	KLHL35	-101.62153	MEIS3
-155.3216096	MSANTD1	-100.2220291	PURA

**Table 6:** Driver analysis showing the top 100 source genes of the cognitively normal set according to their impact on the network.

Much like the Master set, the sources shown in in Table 6 are very strong and KCNF1 is the top gene identified as a driver. The rate of interaction strength decrease is fairly smooth and the top genes identified are exactly what was expected. Factors such as MAP1LC3A, which appear to be ubiquitous when analyzing brain expression data, mediate interactions between microtubules and as such retain their function even during neurodegeneration, while ENSA (endosulfine alpha) belongs to the highly conserved cAMP-regulated phosphoprotein family and is still found in the AD analysis.

Amount Influenced	Gene Symbol	Amount Influenced	Gene Symbol
-333.1584791	RNF165	-122.7065075	MPP2
-259.8655391	ZRANB3	-122.1831065	SYNM
-234.4622366	CCBL2	-121.4371374	KLHL35
-230.1139033	TUBB3	-121.046184	PCYOX1L
-204.0677344	NFKBIA	-121.0297796	GSS
-204.0415324	KIAA1551	118.7070932	FAM174B
-200.4477734	RHOBTB3	-118.5158837	ENSA
-197.6155715	---	-117.8560196	ANAPC16
-197.5402309	PNISR	-117.424957	MXI1
-193.8665492	CASC3	-116.9206148	RHOQ
-189.4131282	CDKN2C	-116.5051402	FAM162A
-187.409986	MRPS25	-114.312991	BRE
-184.3184892	CRIP1	-113.2417781	RBM17
-182.1055429	RAB27A	-111.9495872	RAB27A
-181.2418003	---	-111.5160173	SYT5
-178.9025556	NUDCD3	-111.135889	SOAT1
-178.4729506	GNRH1	-111.0922798	RASAL1
-177.2665027	LARP7	-110.4090812	---
-174.268136	CD59	-109.729288	SPRN
-173.4248265	TRAPPC9	-108.1170065	SLC8A2
-167.2046632	ZNF443	-107.3700912	TPBGL
-165.0193211	SSFA2	-106.5578195	ITPKB
-163.9127191	LANCL2	-106.4312861	DNAJC30
-162.8919791	PRKD3	-106.2598637	RHOQ
-160.8236023	AF070581	-105.7657074	KANSL3
-160.4849276	ITSN1	-105.6263901	BFSP1
-160.3730733	LINC00086	-105.2374447	RASSF8
-159.0672036	FOXN3	-104.2288672	SST
-154.2983617	ATP5F1	-101.6404965	ANAPC16
-148.4091853	CAPN2	-99.89733923	PHF20L1
-146.8628242	CRTAP	99.34226627	GSS
-146.5937516	USP5	-98.64342083	LOC100996724
-146.5098019	HMBBOX1	-98.50228716	RHOQ
-144.9163019	ZXDC	-97.05957196	PURA
-144.2603504	ITFG1	-96.5092495	PNISR
-143.1525307	ERBB2IP	-95.55664447	CNPY2
-142.5112438	TUBB2A	-94.00998718	TLR5
-140.9546076	NACC2	-93.7118169	TNPO1
-137.7882976	LPIN1	-90.48280289	IFNLR1
-136.3981393	SMAD5	-89.32922227	SVOP
-136.1505093	TUBB3	-87.56893434	PLSCR4
-132.5124926	STAG1	87.41929151	FOXN3
-131.2440752	CDIPT	-86.45165447	KIF5B
-129.8965902	MAX	-85.45435342	TM7SF2
-128.7950734	DDA1	-85.44126783	EGR4
128.7424583	MANBAL	-83.61695397	MEIS3
128.4658022	SSSCA1	-81.30115277	SCN3B
-126.0040924	RBM6	-80.55298325	RHOQ
125.9057641	CHP1	-80.10009736	AGAP2
-124.0414893	IL13RA1	-79.77969666	PCGF5

**Table 7:** Driver analysis showing the top 100 target genes of the cognitively normal set according to their impact on the network.

The target analysis (Table 7) however, breaks the trend and shows a few irregularities in the top genes, with RNF165, ZRANB3, CCBL2 and TUBB3 being inhibited significantly more by the network before the rate of reduction in strengths becomes smoother. Even accounting for this irregularity, the results still follow the expected pattern with the top genes encoding zinc finger proteins, transferases and even RHOBTB3, a conserved member of a conserved subfamily of the Rho GTPases, of which RHOQ, which appears to be highly dysregulated in AD, is a member.

It is now possible to start forming a hypothesis based in this non-systematic, non-parametric approach which relates to the conservation of mechanistic factors in cognitively normal individuals versus the dysregulation of key factors in AD patients. However, it should be noted that the list of potential candidates for biomarker for use in prognosis and therapy, as well as drivers of the disease is still very large at 200 genes. There is still a significant amount of noise present and although the list has been filtered down from almost 60000 genes, a panel of 200 genes would have too much noise and would take too long to examined to be successfully implement in a clinical setting. It needs to be reduced to its most significant components.

#### 4.6 Commonality Analysis

By combining the results of the previous steps in the pipeline examined, it is possible to perform a commonality analysis. This allows us to compare different questions directly and draw conclusions based on the degree of similarity of the most important genes identified across multiple analyses. It becomes possible to determine which genes have mechanistic importance and are crucial to the function of the system by knowing which genes are conserved between varied conditions, such as healthy and diseased, as well as which genes maintain health, and which drive the disease by comparing these conditions between datasets.

The probability that a gene will be common across questions is a function of the number of genes in the dataset multiplied by the number of questions examined.

$$\left( \frac{\text{gene subset by rank}}{\text{number of genes}} \right)^{\text{number of questions}}$$



The goal of this section is to compare broad questions against each other and determine whether there are any common factors across them. As the probability that a given being common across all three questions (AD, Healthy, Master) is very low even considering the similarity of the datasets and the chance is

$$\left(\frac{50}{54675}\right)^3$$

any genes that appear as commonalities should be examined further in order to identify whether their structure, function or pathways they are involved in has any significance and how that can be used to explain the variance seen in AD.

## Commonalities between all three datasets for the top 50 source genes

Gene Symbol	Gene Title
ACP2	acid phosphatase 2 lysosomal
AGPAT1	1-acylglycerol-3-phosphate O-acyltransferase 1
CECR6	cat eye syndrome chromosome region candidate 6
DCTPP1	dCTP pyrophosphatase 1
DLK2	delta-like 2 homolog (Drosophila)
DNAJC30	DnaJ (Hsp40) homolog subfamily C member 30
ENSA	endosulfine alpha
HBQ1	hemoglobin theta 1
IMP4	U3 small nucleolar ribonucleoprotein
LOC100129361	chromosome X open reading frame 69-like
LRRC73	leucine rich repeat containing 73
MAP1LC3A	microtubule-associated protein 1 light chain 3 alpha
MBOAT7	membrane bound O-acyltransferase domain containing 7
MECR	mitochondrial trans-2-enoyl-CoA reductase
MRPS25	mitochondrial ribosomal protein S25
MSANTD3-TMEFF1 /// TMEFF1	MSANTD3-TMEFF1 readthrough /// transmembrane protein with EGF-like and two follistatin-like domains 1
PCYOX1L	prenylcysteine oxidase 1 like
PHIP	pleckstrin homology domain interacting protein
RPUSD3	RNA pseudouridylate synthase domain containing 3
SPRN	shadow of prion protein homolog (zebrafish)
SVOP	SV2 related protein homolog (rat)
TMEM59L	transmembrane protein 59-like
TRAPPC9	trafficking protein particle complex 9
TUBB3	tubulin beta 3 class III
USP5	ubiquitin specific peptidase 5 (isopeptidase T)

**Table 8:** Commonalities between the 50 most influential source genes on each network.

As shown in Table 8, there are 27 probes, and 25 common genes that are included in the 50 genes that were predicted to have the strongest influence on the entire network. Genes such as TUBB3, TMEM59L, MAP1LC3A, and AGPAT1 have been analysed extensively in the preceding sections and their functions are theorised to be crucial in the normal function of the human brain. These genes should be monitored when attempting to find how they affect AD, not by their mere presence, but which genes they interact with and whether they inhibit or stimulate them. While not necessarily drivers of the disease,

changes to their targets and interference in different pathways, as well as irregular expression could cause a knock-on effect that dysregulates a crucial gene immediately or even multiple steps away, leading to development of the disease. Their overlap is shown in Figure 34.

Their molecular function falls in 4 major categories. Binding, catalytic, structural and transporter activity, while the largest predicted protein function categories are cytoskeletal proteins, hydrolases and nucleic acid binding proteins which cytoskeleton regulation being the pathway they are most involved in in AD.

### Commonalities between all three datasets for the top 50 target genes

Gene Symbol	Gene Title
<b>CAPN2</b>	calpain 2
<b>CASC3 /// MIR6866</b>	cancer susceptibility candidate 3 /// microRNA 6866
<b>CD59</b>	CD59 molecule
<b>CDKN2C</b>	cyclin-dependent kinase inhibitor 2C (p18
<b>CHP1</b>	calcineurin-like EF-hand protein 1
<b>CRIP1</b>	cysteine-rich PDZ-binding protein
<b>CRTAP</b>	cartilage associated protein
<b>GNRH1</b>	gonadotropin-releasing hormone 1 (luteinizing-releasing hormone)
<b>LARP7</b>	La ribonucleoprotein domain family
<b>MPP2</b>	membrane protein
<b>NFKBIA</b>	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor
<b>PNISR</b>	PNN-interacting serine/arginine-rich protein
<b>RBM6</b>	RNA binding motif protein 6
<b>SMAD5</b>	SMAD family member 5
<b>SSSCA1</b>	Sjogren syndrome/scleroderma autoantigen 1
<b>TUBB2A</b>	tubulin
<b>TUBB3</b>	tubulin

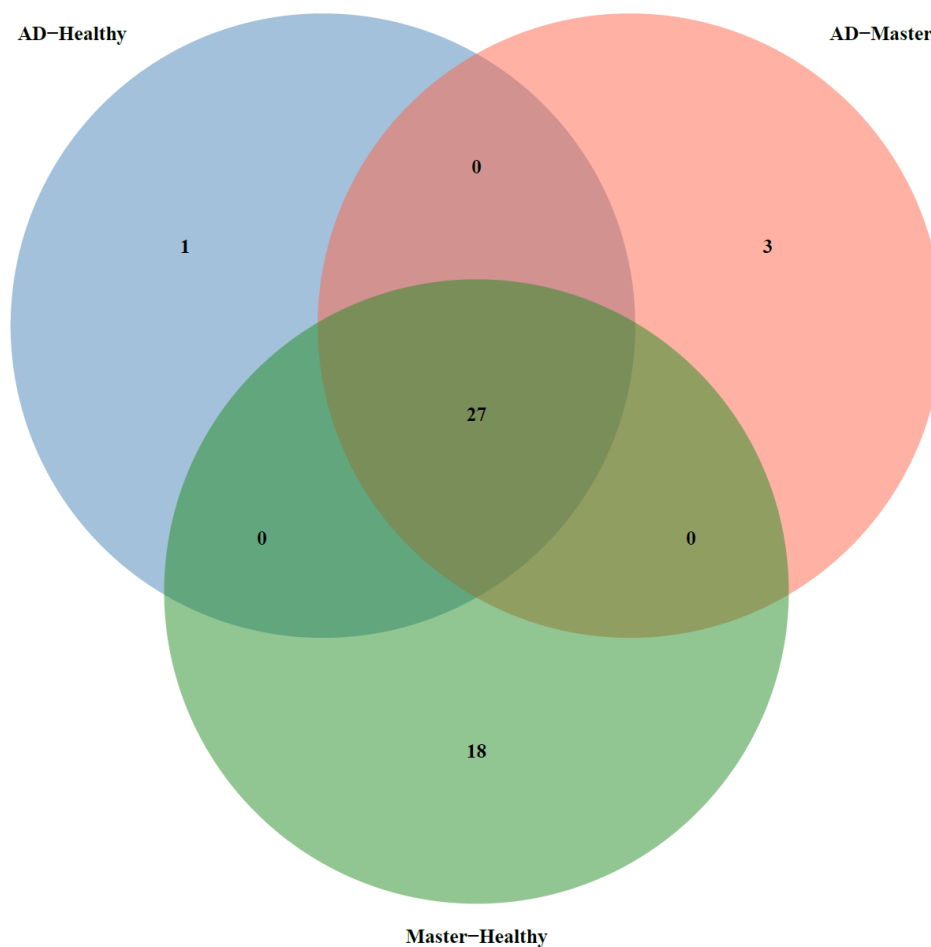
**Table 9:** Commonalities between the 50 most influential target genes on each network.

The target genes on the other hand (Table 9), show a significantly smaller degree of overlap. With 18 gene probes and 17 genes, it seems to indicate that while there is a certain degree of universality to the drivers, perhaps resulting from the fact that a driver does not have to significantly change but can be a factor in causing the disease simply by altering the expression of other genes, sources, being potential biomarkers, are more likely to

diverge between AD and cognitively normal individuals. Curiously, TUBB3 is the only gene to be common across all tests, appearing in both AD, healthy and combined sets as a crucial factor. Their overlap is shown in Figure 35.

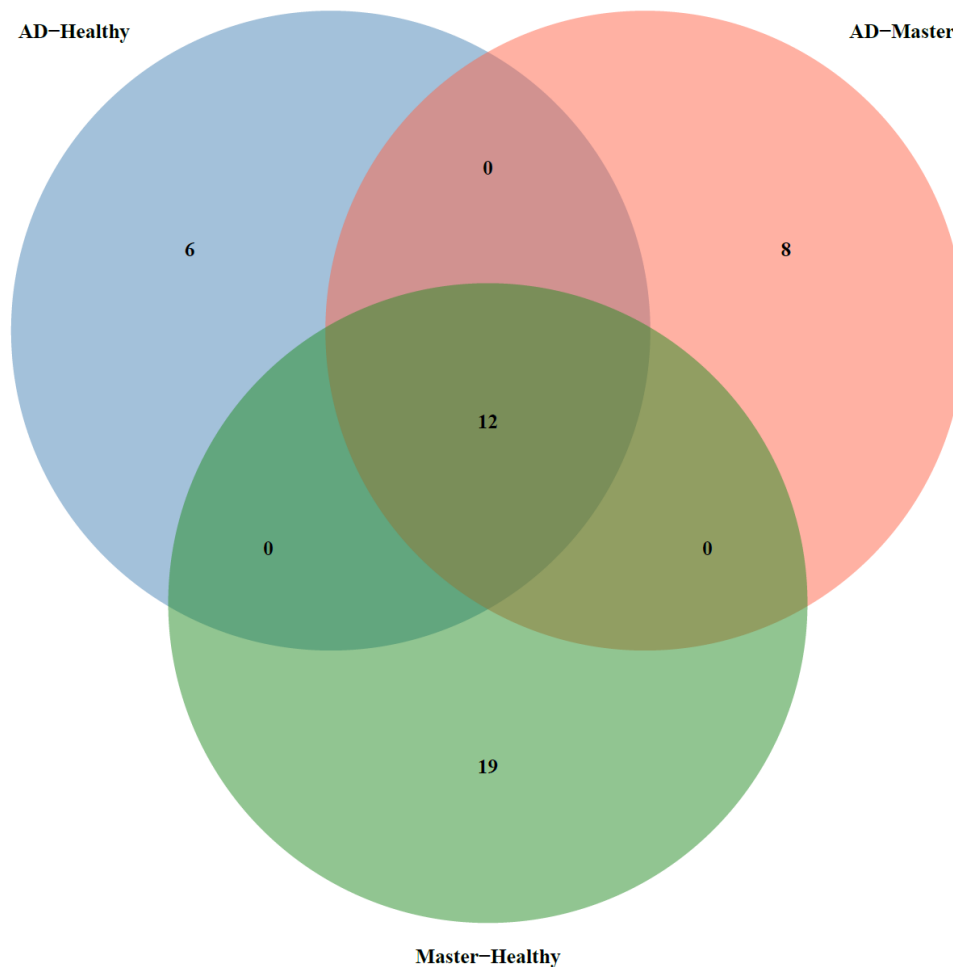
Target genes appear to be mostly composed of structural molecules, but binding and catalytic activity still remain relevant, with cytoskeleton predicted proteins remaining the most dominant class, but with more variety in predicted proteins including calcium binding cell junction proteins. When analysing possible pathways however, there is little consensus, with the three major pathways being the Huntington’s disease, gonadotropin-releasing hormone receptor and cytoskeleton regulation pathways, with inflammation and T cell activation also being relevant.

### Pairwise Commonalities Source



**Figure 34:** Venn diagram showing the pairwise commonalities for the most influential gene between three stepwise analyses AD-Healthy, AD-Master and Master-Healthy. Of note is the high number of commonalities between AD-healthy and AD-Master indicating greater variation between AD and cognitively normal individuals rather than between AD ones. Complete table available in the appendix.

## Pairwise Commonalities Target



**Figure 35:** Venn diagram showing the pairwise commonalities for the most influenced gene between three stepwise analyses AD-Healthy, AD-Master and Master-Healthy. Of note is while the number of commonalities between AD-healthy and AD-Master remains high, similarly to the most influential genes, there is greater variation for the most influenced ones while the variance between AD and cognitively normal controls remains high. Complete table available in the appendix.

## Chapter 5: Systems Biology Expansion and Integration

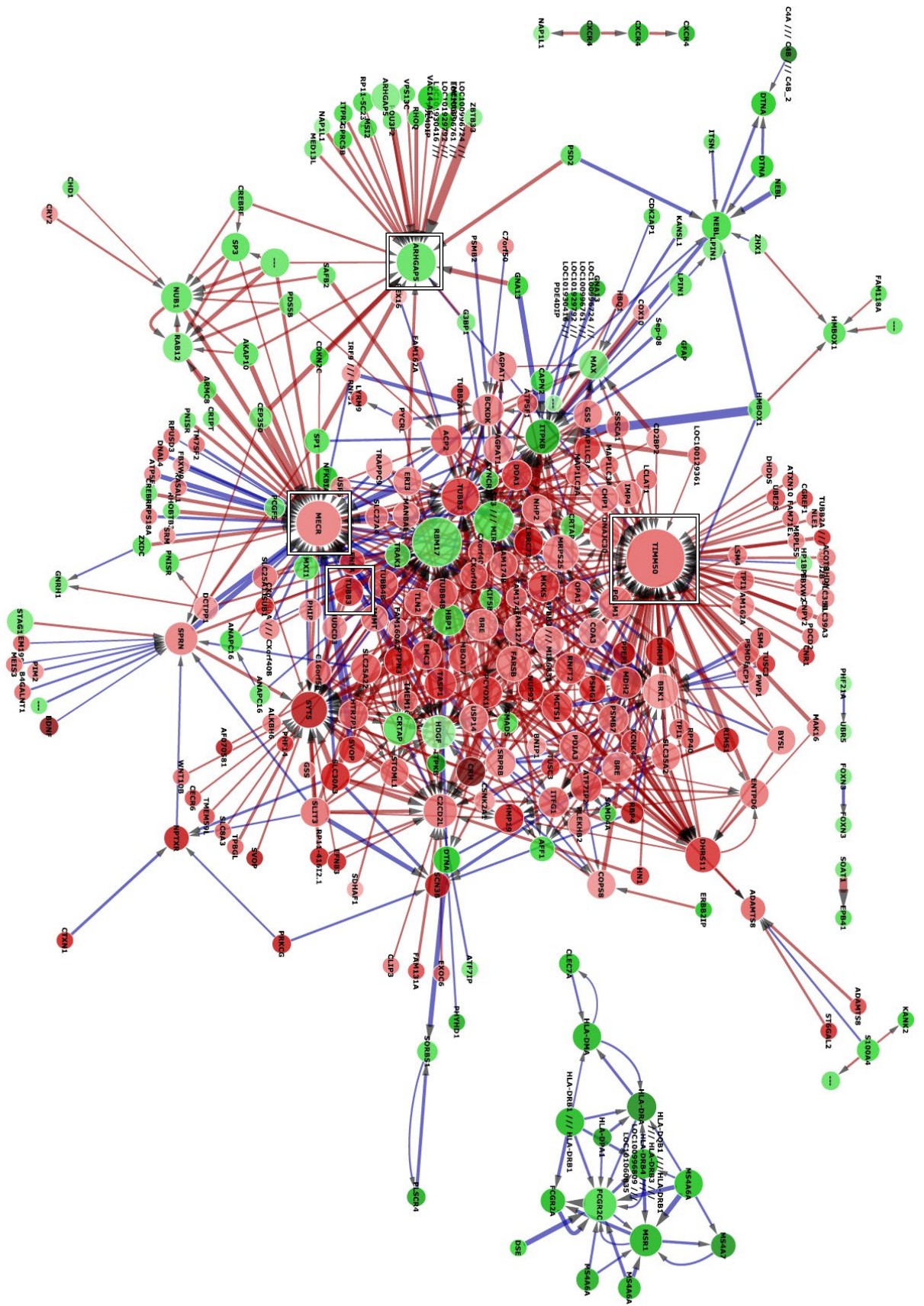
### 5.1 Interaction Matrix

One of the greatest problems encountered during the analysis of the interaction results is that the method used to predict a single best marker during the preceding stepwise analysis, the selection process is stochastic; there is a random probability element and while the results can be statistically significant, it makes the process imprecise. Moreover, this issue can be exacerbated during the network inference step if not taken into account. For the previous tests, this was countered by repeating both the stepwise and network inference analyses multiple times, until a convergence of results was reached. The most representative example was then selected via commonality analysis and used to perform biomarker discovery.

In an effort to not only counter that effect but also increase the overall power of this method, the 500 genes most likely to explain the variance between the conditions selected (AD-Healthy) selected by the stepwise process were split into 5 datasets of 100 genes each. This was followed by merging these datasets into 20 sets of 200 genes each for network inference. While only 10 are technically required to generate a complete matrix, they were analysed twice each to achieve convergence. The number of gene probes selected for this approach was determined by analysing the performance statistics of the stepwise algorithm. The training, test and validation performance starts to plateau after the first 400 genes, indicating that the differentiation between the given conditions was decreasing and was thus liable to introduce noise in the analysis if the number of selected genes was increased. Moreover, due to the increased computational requirements of this analysis, with a full run taking up to 30 days, it was deemed essential to maintain the minimum number of significant genes while also generating data that was feasible to analyse in a shorter timescale.

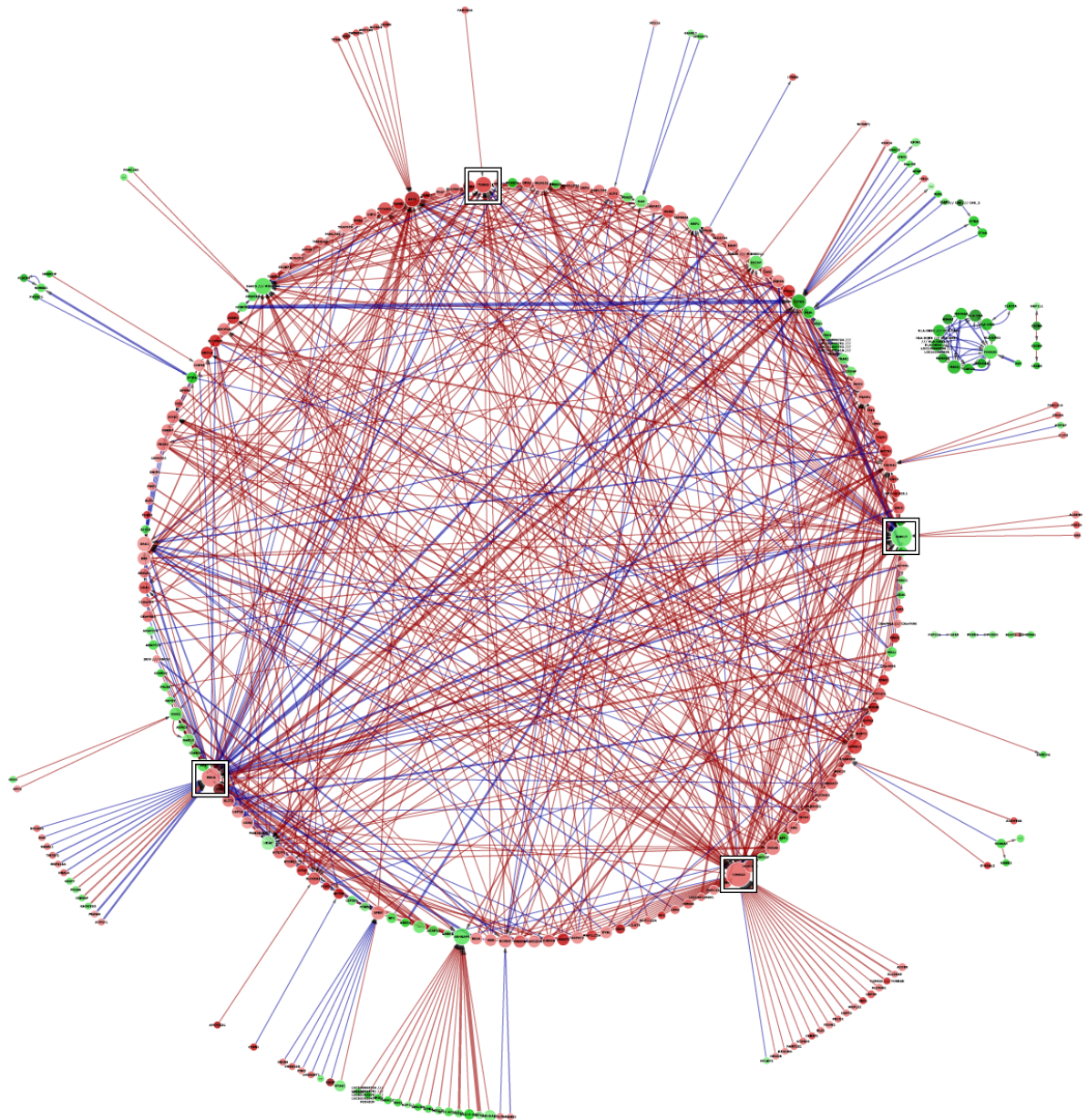
Once the 20 network inference analyses were completed, the data was consolidated and the top 1000 strongest interactions were selected and visualised with Cytoscape. Duplicate interactions resulting from multiple tests were considered and removed to allow for more clarity in the interactomes. Expanding to 1000 interactions between 500 genes over 200 interactions for 200 genes has proven to be essential in identifying a more representative view of the disease. The reasoning behind developing this technique is that the previously

examined single marker approach, focuses only on a small subset (~0.1%) of the genes actively influencing a given condition, in this case, AD. Moreover, by only selecting the 100-200 strongest interactions, it is virtually guaranteed that in the resulting network, the biggest hubs, hence the most likely drivers of the disease and targets for therapy, will be kept to a minimum and will be biased towards the most differentiated genes as seen in Section 4.4.1. It is important to note however, that for a highly focused analysis where the variance in the expression of key factors is small, such as when studying a specific subset of genes in a subset of a disease, ie. proliferation markers in untreated triple negative breast cancer patients, the very nature of the data would result in a network where all the hubs are equally important. Thus, in such cases, identifying key markers and drivers based on interaction strength and hub centrality is still likely the superior choice, until advances in technology allow us to consider a larger number of genes in a shorter amount of time without the addition of noise in the network.



**Figure 36:** Expanded interactome encompassing 500 gene probes and 1000 predicted interactions in the E-GEOD-48350 AD cohort. Noteworthy hubs include TIMM50, ARHGAP5, MECR and TUBB3.





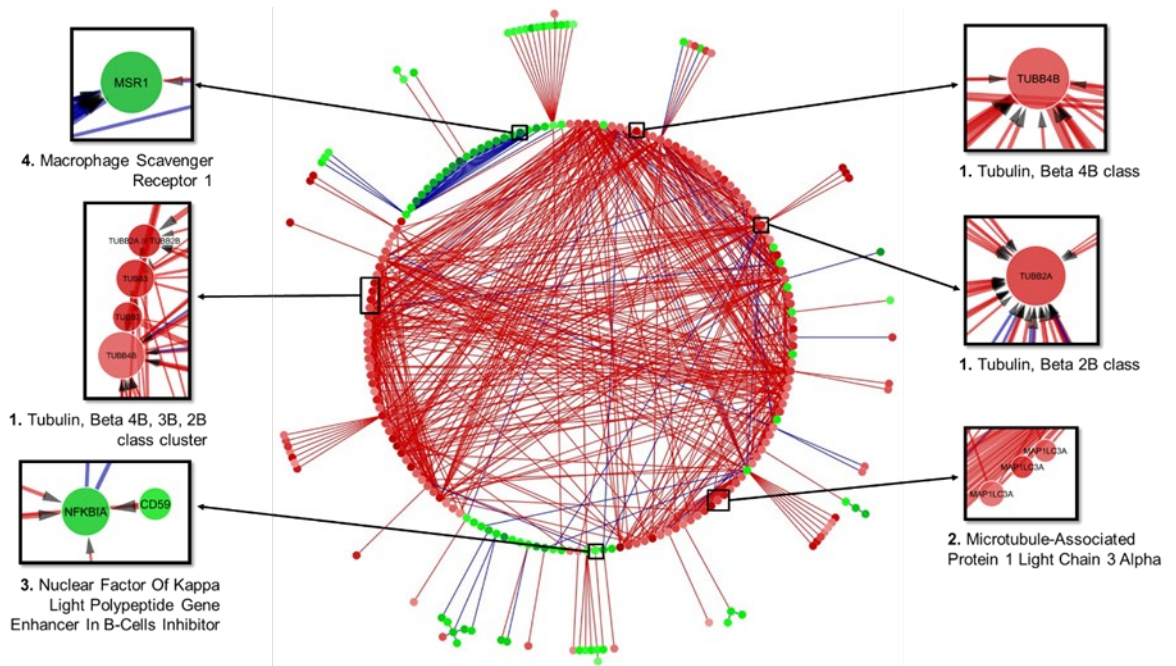
**Figure 37:** Circular interactome of Figure 36, highlighting the hubs and clarifying the ration of positive to negative interactions.

In Figures 36 and 37 it is possible to see the effects of direct effects of more than doubling the number of gene probes provided to the algorithm, especially when compared to the smaller AD interactome as shown in Figure 31. It is worth noting that even though both the number of gene probes and interactions were dramatically increased (150% increase in genes, 500% increase in interactions) the number of major hubs, defined as genes with a combined edge count of 40 or more, has not increased as much as expected. Certain new genes have emerged as hubs, such as TIMM50 with 128 connections, but the majority of hubs remain consistent with the smaller, more focused interactomes. The genes are TIMM50, RBM17, MECR, CASC3 /// MIR6866, ARHGAP5 and TUBB3. There is also

evidence that there is a certain degree of consistency between methodologies, although there are some significant changes due to the stochasticity present in the system, namely, the loss of TUBB2A as a major hub. Conversely the strength of this approach is not present in the hubs, but in the genes affecting said hubs.

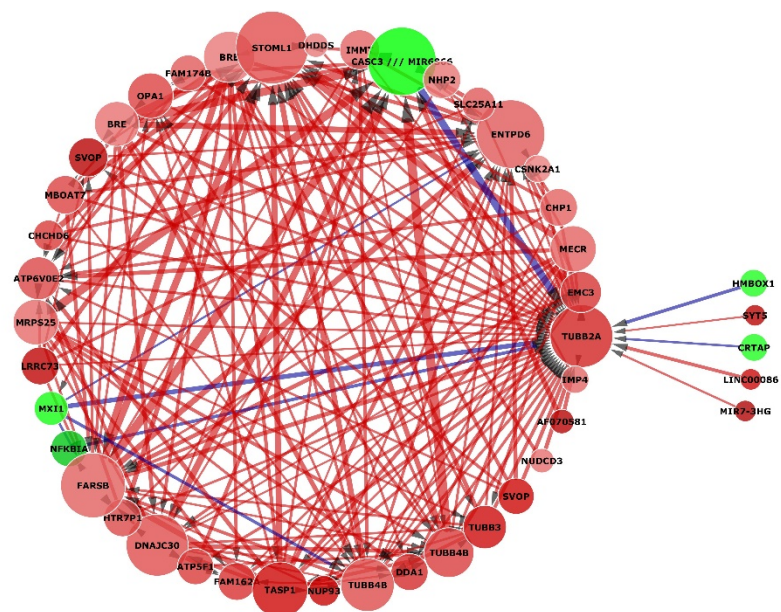
The increase in the number of interactions allows us to observe the ratio of inhibition to stimulation and provide clarity on the issue of the maintenance of balance in biological systems. As expected, a bigger cut-off point for the number of interactions has allowed for greater variety in their types. Biological systems in diseased individuals tend to be in a state of dysregulation. As the disease progresses, the system becomes unable to regulate itself, which leads to an overrepresentation of certain interaction types. This is often the case for cancer patients where the inhibition of tumour suppression factors, among others, is significantly stronger than their counterparts, leading to extremely biased interactomes. While that is a very useful tool to identify biomarkers and targets for therapy, it is harder to determine which factors started driving the disease in the first place and which interactions were crucial in the development of said disease. Moreover, in a disease such as AD, where the cause is largely unknown, and the samples are from patients who did suffer from it, but the degree of severity or how far it had progressed is much harder to quantify, the ability to construct a more inclusive, but still usable interactome is paramount.

Additionally, one of the major advantages of this method is the ability to generate a large and complex interactome that can be focused on a gene or genes of interest and analyse their interactions in greater detail as shown in Figure 38.



**Figure 38:** Circular layout interactome of the 1000 strongest interactions between 500 genes in AD using the E-GEOD-48350 dataset. Based on the overall expression of all brain regions. Novel targets identified.

The hubs in this interactome are not too dissimilar to the ones discussed in the preceding figure 38, but the focus of this technique is to decouple any discovery analysis from being exclusive to the major hubs and attempt to find out how the smaller hubs impact the network.



**Figure 39:** Focused Tubulin interactome based on Figure 38. Tubulin beta 2A interactions in AD when all brain regions are accounted for. Of note is its positive regulation by an NFKB inhibitor, NFKBIA.

In this example tubulin 2 beta (TUBB2A), a structural component of microtubules and a gene closely associated with tau, has consistently been in the top genes identified in AD across multiple tests. By taking advantage of the fact that the previous interactome (Figure 38) has high enough complexity to be able to break into smaller ones that are still biologically relevant, it is possible to analyse all significant predicted interaction TUBB2A has with other significant genes without having to resort to the interaction algorithm for a second time and increase the time requirements for a single analysis. Furthermore, if enough genes are identified as relevant to the question, they can then be used as continuous predictors in Stepwise and those results used for network inference increasing the overall power.

In Figure 39 we can observe that TUBB2A is underexpressed but also downregulated by the vast majority of predicted interactions, including by other tubulin variants such as TUBB3 and TUBB4B as well as BRE which was discussed earlier. It is interesting however that both CASC3 and NFKBIA, both of which are overexpressed in this case, are upregulating TUBB2A, weakly in the case of NFKBIA but relatively strongly in the case of CASC3. CASC3 also appears to be very strongly downregulated by TUBB4B, MRPS25 a mitochondrial ribosomal subunit involved in mitochondrial translation and organelle maintenance and biosynthesis, and FARSB, a Phenylalanyl-TRNA Synthetase Beta Subunit involved in tRNA aminoacylation and has been found to be associated with muscular dystrophy. Thus, it is possible to surmise that the dysregulated state of the TUBB2A gene in the network is directly correlated with mechanistic dysregulations in other genes that in turn affect genes responsible for regulation of TUBB2A itself. CASC3 and NFKBIA are failing to significantly upregulate TUBB2A back to normal levels due to dysregulation within themselves.

## 5.2 Disparate Brain Region Variance

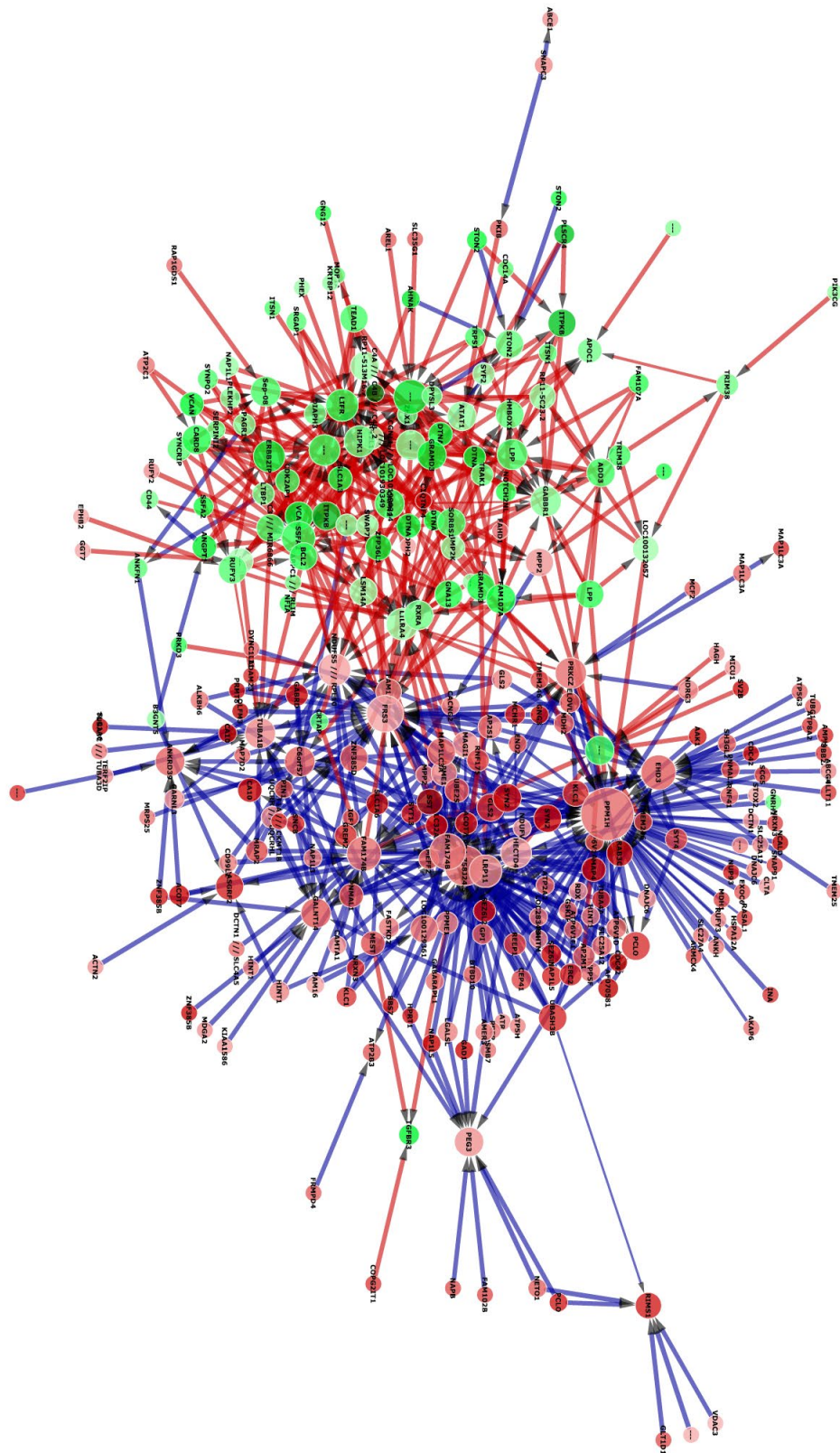
As explained during the introduction to AD in Chapter 1, while it is a neurodegenerative disease that causes loss of brain tissue, and hence, function, different regions of the brain have significantly varied roles and are affected by the disease in different ways. It is a widely accepted fact that the hippocampus is the centre of the brain that is most affected during AD. Moreover, the non-parametric approach followed thus far, has been in aiming to reduce the bias inherent in setting a null hypothesis and therefore assuming significance

of specific genes and brain regions. While biologically relevant, even such information can lead to masking of factors that can be used to predict the presence of AD, the way of develops, the affected pathways and potential therapy targets. Thus, the results of the previous experiments can be used as pruning techniques and provide a framework to inform further deep mining in a given dataset.

The areas analysed in this section include the hippocampus and entorhinal cortex. An interaction matrix analysis was performed on both AD individuals and cognitively normal controls of the E-GEOD-48350 dataset. The stepwise algorithm was used to identify the 500 genes most likely to explain the variance between AD and healthy individuals exclusively in the brain regions specified. The interaction matrix used 500 genes and 1000 interactions to generate an interactome and followed by a driver analysis to identify the resulting differentiation from the analysis performed during section 4.5. The results were visualised using Cytoscape 3.5.1.



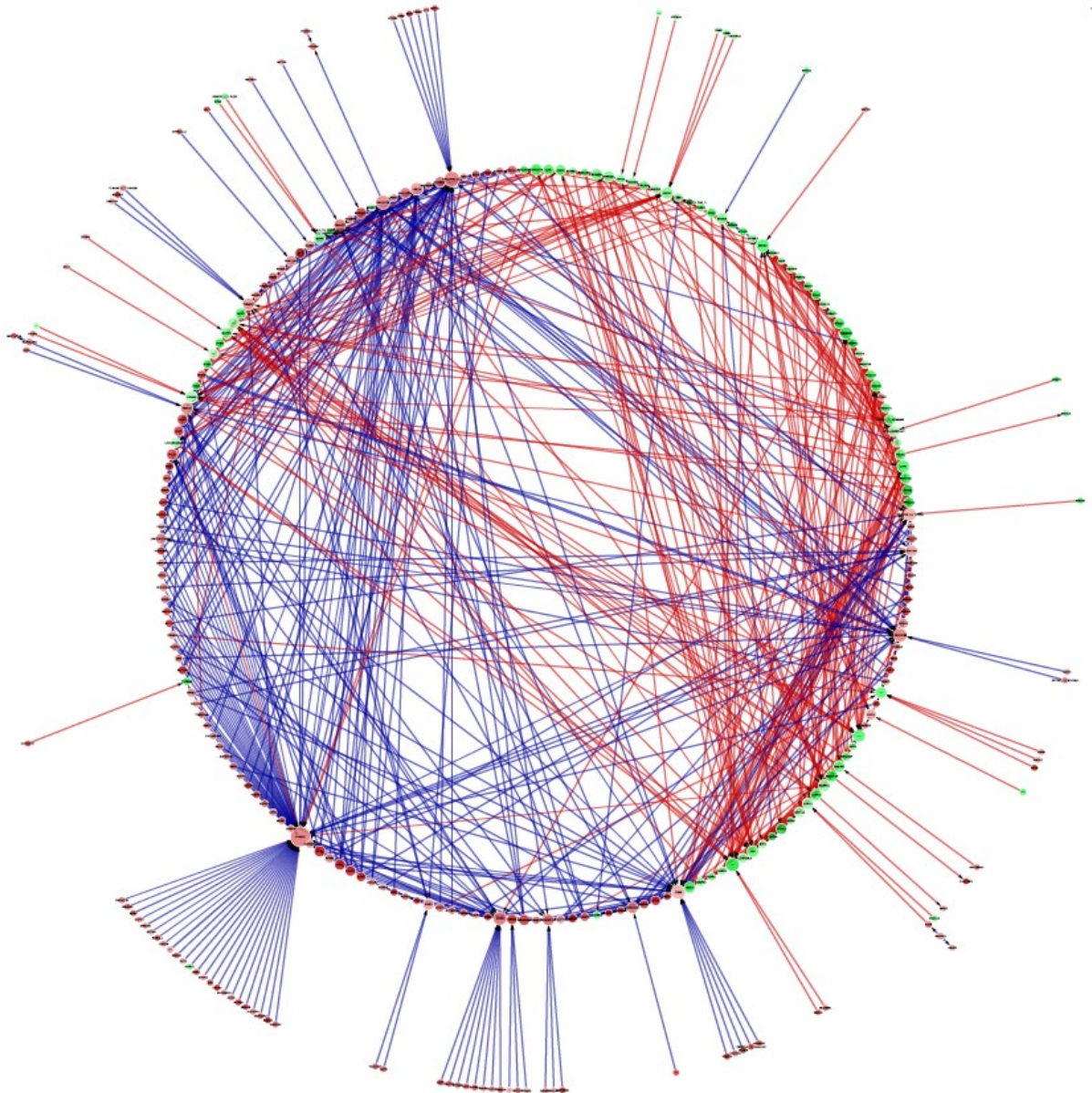
## 5.2.1 Hippocampus *AD Interactome*



**Figure 40:** Force directed interactome of the 1000 strongest interactions between 500 genes in the hippocampus, AD brain. Obtained via analysis of the E-GEOD-48350 dataset.

As seen in Figure 40, upon separating the data to only include gene expression data exclusively from the hippocampus from AD patients only, a rarely seen duality presents itself. In most complex diseases such as cancer, the dysregulation that is represented in the interactomes is a direct result of the mechanisms of the disease. Successful cancers can hijack the body's immune response, avoid detection and proliferate uncontrollably. This in turn, leads to the body mounting a very strong response by attempting to upregulate anti-tumour factors and suppress proliferation factors among others in order to prevent the abnormal cells from disrupting the function of crucial organs. Diabetes is similarly represented, as due to chronically high sugar levels the function of the organs affected get significantly damaged. This leads to interactomes that are either mostly up- or down-regulated.

However, irrespective of the cause, non-familial AD is a direct result of the failure to regenerate damaged cells and clean away debris over a long period of time. Moreover, the isolated nature of the brain, the increased regulation of substances that can cross the blood brain barrier and most importantly the brain's plasticity, are crucial defence factors that other organs lack. Plasticity is especially important as the brain can tolerate extensive damage before showing significant dysregulation, which is why AD is so hard to identify early. As a result, the interactomes of affected regions show both up- and downregulation as it is possible to observe both suppression factors that could potentially be the direct cause of the disease and healing factors that are attempting to restore balance, as the mechanisms for it are still present and functional. In fact, dysregulation in the mechanisms involved in immune response and debris clearance could be used as predictors for early prognosis of AD as they are still functional, but increasingly ineffective.



**Figure 41:** Circular interactome of the 1000 strongest interactions between 500 genes in the hippocampus, AD brain. Obtained via analysis of the E-GEOD-48350 dataset.

This duality in the interactome shown in Figure 41 however, reveals an interesting pattern within the data. Based on a fold change analysis of the original microarray data for AD in E-GEOD-48350, the genes that are overexpressed are downregulated overall. Conversely, underexpressed genes are predicted to be mostly downregulated. It is a fact that the hippocampus is the most dysregulated brain region in AD, so this is possible proof that the system is attempting to restore balance by suppressing the high expression of factors such as HIPK1, a kinase which plays an important role in senescence, ITPKB, a kinase that regulates inositol polyphosphates or BCL2, a protein phosphatase which is a crucial



apoptosis factor. In short, the system is attempting to decrease the effect of genes involved in cell death.

The factors that are underexpressed on the other hand, appear to be upregulated and significantly more dysregulated, with an overall larger number and stronger individual interactions. The largest hub is PPM1H, another protein phosphatase which dephosphorylates CDKN1B, a CD kinase inhibitor involved in diseases such as Type IV Multiple Endocrine Neoplasia and familial Primary Hyperparathyroidism. Another such gene is FRS3, a fibroblast growth factor receptor substrate which is involved in regulation of RAS signalling.

While these genes and others like them seem to indicate that there is a significant effort to re-establish homeostasis, of further interest are the genes that do not fall inside these clearly defined categories. These genes include multiple tubulins such as TUBA1B and TUBB2A which are underexpressed but being simultaneously up- and downregulated, TGFBR3 which encodes for the transforming growth factor beta, type III receptor and plays a crucial role in cell adhesion and is associated with diseases such as familial cerebral saccular aneurysm. TGFB itself activates transcription factors of the SMAD family, which in turn, regulates gene expression. ATP2C1 is an ATPase which catalyses the hydrolysis of ATP and is underexpressed while still attempting to downregulate CARD8. CARD8 itself is caspase recruitment domain containing family of proteins, and is involved in pathways negatively regulating the activation of NF $\kappa$ B, which as explained during the introduction, has a key role in the theory of neuroinflammation, and is quite likely an attempt to slow down or stop the chronic immune response leading to said neuroinflammation. Other irregularities include MAP1LC3A and MPP2 explained earlier and CD44, a cell-surface glycoprotein involved in cell-cell interactions, cell adhesion and migration and interacts with, among other things, matrix metalloproteinases (MMPs). MMPs, and MMP-9 in particular have long been suspected in playing a key role during AD and have been shown neuroprotective capabilities (Fragkouli *et al*, 2014). Finally, one of the most highly underexpressed and downregulated genes is C1QTNF4, a complement-C1q tumour necrosis factor-related protein whose role is not clearly defined but has been suspected of acting like a pro-inflammatory cytokine, leading to the activation of NF $\kappa$ B and upregulate production of IL6.

AD Driver Analysis

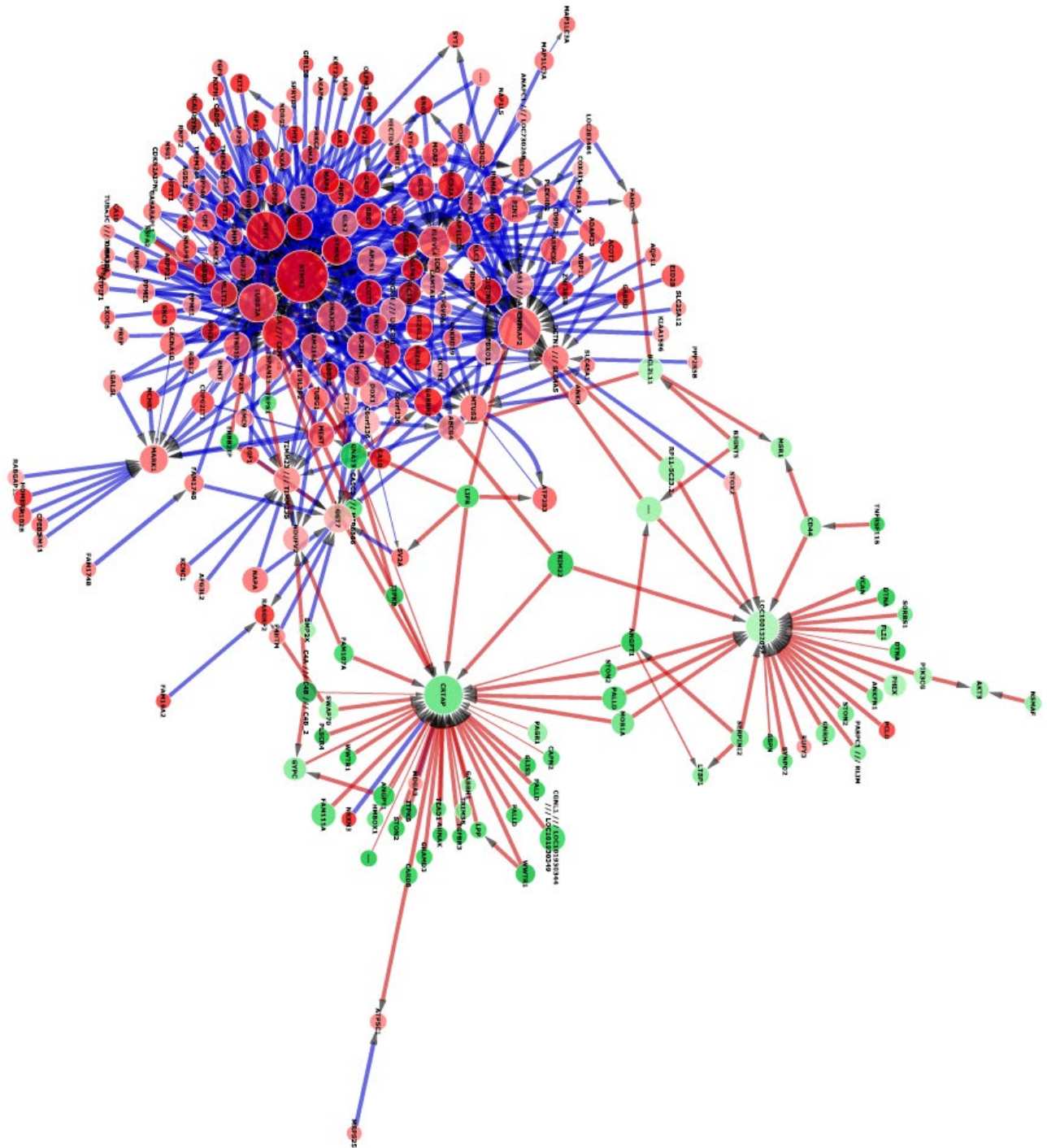
Amount of Influence	Gene Symbol	Amount Influenced	Gene Symbol
-1213.081747	ITPKB	2381.32828	RP4-758J24.5
-1155.143791	GNA13	1814.197348	PPM1H
-1148.207036	RHOBTB3	1792.055655	C6orf57
-1130.228993	VCAN	1754.53677	PRKCZ
-1122.165399	PRKD3	1738.174507	FAM174B
-1119.483983	ITPKB	1733.508537	FAM174B
-1113.48562	TRAK1	1694.647661	LRP11
-1108.416853	CASC3 /// MIR6866	-1686.025743	CAPN2
-1090.674524	SRGAP1	1643.902991	RASGRF2
-1087.335279	LPP	1612.688333	FASTKD2
-1028.750389	LIFR	-1609.099447	1561158_at
-1026.678888	GLIS3	-1592.564443	RXRA
-1025.359853	TEAD1	-1560.277805	HIPK1
-1018.653673	CARD8	-1533.256201	SWAP70
-1018.095788	ERBB2IP	1529.733365	GALNT14
-1017.418527	RUFY3	1523.370265	LOC100129361
-1012.441445	242611_at	1504.669747	PEG3
-1010.030914	CRTAP	-1473.99368	RP11-513M16.7
-992.2756126	PABPC1 /// RLIM	1437.631403	HECTD4
-982.4210022	SORBS1	-1435.335801	SYF2
-979.1729048	233323_at	-1431.019129	1557286_at
-973.5676705	SYNCRIP	-1430.602853	TGFBR3
-971.9687449	SEPT8	-1419.648077	FAM107A
-967.7392151	SSFA2	-1390.736715	244457_at
-967.402376	BCL2	1376.260249	BNIP1
-966.0628739	DTNA	-1366.651015	LTBP1
962.5225317	KLC1	-1352.118161	B3GNT5
-949.0374794	GRAMD3	-1351.336497	CRTAP
-935.7444619	FAM107A	-1320.758301	RP11-5C23.2
-933.9110942	SSFA2	1315.432274	ABCE1
-930.2480972	HMBOX1	1314.897356	FAM174A
-917.4727487	TRPS1	-1312.652293	HMBOX1
-913.4533421	PALLD	1310.768154	AP2S1
-913.3942276	FAM107A	1302.110691	GPS1
-909.7624557	BCL2L11	-1289.970537	MOB1A
-905.7671419	CDK2AP1	1282.458469	ALKBH6
-904.8356429	VCAN	-1270.746624	KRT8P12
-904.3397815	CAPN2	1264.244601	MAGI1
-902.6081661	233323_at	1255.168137	ANKRD39
-899.9210524	NOTCH2NL	1251.151278	DNAJC6
-896.8863383	ZFP36L1	1240.782655	EHD3
893.9583626	ZNF385B	1231.946711	238466_at
-888.5853093	ADD3	1227.9842	AREL1
-880.9829708	WWTR1	-1218.986806	ATAT1
-876.8780354	PALLD	-1211.954338	LILRA4
861.0770622	SYN2	-1207.832696	LIFR
-860.3346604	NFIA	1207.407361	TUBA1B
-859.6319203	228297_at	1204.867524	GABBR2
-851.0770584	DTNA	-1195.223424	ITPKB
849.629429	AP2M1	1194.137548	PLEKHB2

Table 10: Driver analysis showing the top 50 source and target genes according to their impact on the network for AD in the hippocampus. The probe IDs in red have not been mapped as of 2017.

The driver analysis (Table 10) was carried out on the 500 selected genes of the matrix interaction. The most influential source genes showed significant similarities and differences to the results of previous analyses on AD. Genes identified in the interactome such as ITPKB and CASC3 as well as trafficking proteins like TRAK1 and kinases like PRKD3 are expected. Of note is the disproportionate presence of BCL2 when compared to the interactome. However, the sources of interest include RHOBTB3, a member of the highly conserve family of Rho GTPases similar to RHOQ discovered during earlier testing, as well as SRGAP1. SRGAP1 encodes for a GTPase activator and works in conjunction to CDC42, a GTPase of the same family, to negatively regulate neuronal cell migration. Moreover, when combined with receptor ROBO1, it can deactivate CDC42. Its presence so high on the source list as a downregulating factor, indicates that its function is being stronger than expected, resulting in slower cell migration and impediment of the regeneration process. CARD8, discussed earlier, has a strong, negative effect on the network, suppressing the expression of related genes.

Meanwhile, the most targeted genes on the network include PPM1H, a protein phosphatase, TGFBR3, multiple kinases, and an alpha-tubulin TUBA1B. More beta tubulins are included in the complete list. Also, although rarely seen, ATAT1, an alpha tubulin acetyltransferase, a neuronal cell component crucial to the microtubule growth appears to be negatively regulate. ATAT1 is involved in coenzyme binding and tubulin N-acetyltransferase activity and only acetylates older microtubules, being unable to act on unstable ones. Genes such as APGAT1 which fulfil similar purposes have been discovered in previous test, suggesting that slower/weaker acetylation of older microtubules could play a key role in the development of AD. Curiously, one of the upregulated factors is AREL1, apoptosis resistant e3 ubiquitin protein ligase 1, which inhibits apoptosis. It is possible that it is being upregulated in an attempt to keep the neurons alive and functioning to prevent further damage. Finally, the presence of ITPKB as both a significant source and target indicate that it is a crucial component of the system regardless of disease state. It will be further analysed when examining the cognitively normal controls for the hippocampus.

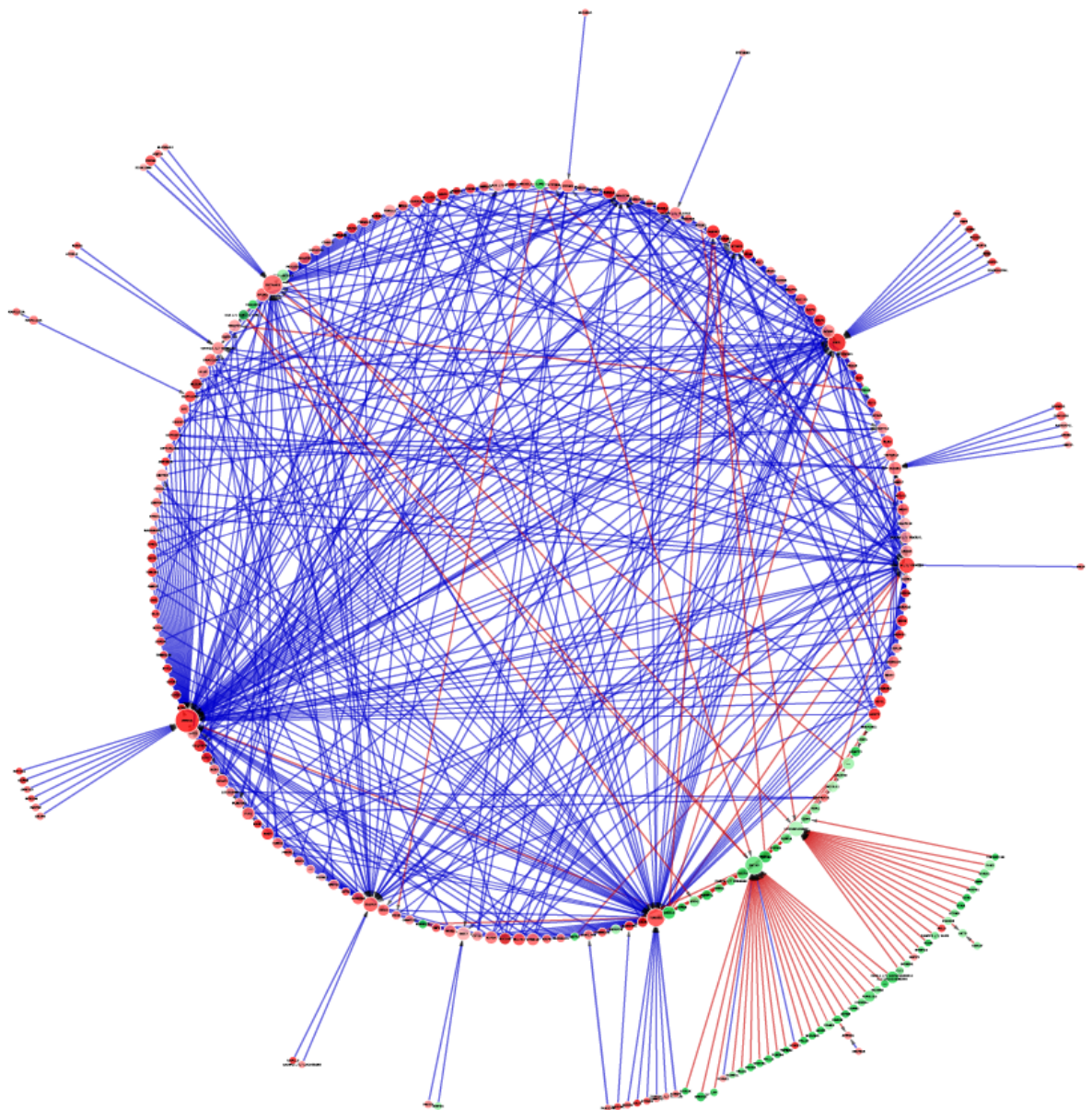
## Healthy Interactome



**Figure 42:** Force directed interactome of the 1000 strongest interactions between 500 genes in the hippocampus, cognitively normal brain tissue. Obtained via analysis of the E-GEOD-48350 dataset.

In order to differentiate AD from cognitively normal individuals, the same methodology was applied to a cohort of exclusively healthy controls from the same dataset and brain region (Figure 42). All test parameters were kept identical to the previous experiment. It is immediately obvious that while the healthy interactome bears similarities to the AD

one, it also has significant differences, especially to the details. There is a far greater degree of positive regulation as well as genes with a negative fold change. The number of genes that break the pattern of negative-upregulate/positive-downregulated is also much smaller than the AD interactome in figure 41. This could be a direct result of the fact that these are post mortem samples. The individuals these samples originally belonged to, died without ever suffering from AD, meaning that any dysregulation that could have potentially led to its development in the future did not have time to develop. The expected results from such a test should show a number of factors involved in the normal function of neurons and maintenance of mental health by a fully functioning immune system.



**Figure 43:** Circular interactome of the 1000 strongest interactions between 500 genes in the hippocampus, cognitively normal brain tissue. Obtained via analysis of the E-GEOD-48350 dataset.



The largest upregulated node in the healthy interactome is STMN2, a gene encoding for a member of the stathmin family of phosphoproteins which plays a key role in the maintenance of stability in microtubules as well as neuronal growth. Dysregulations in this gene make it a prime candidate for being a key driver of AD, and in fact multiple studies have shown links of STMN2 with APP (Li, 2005), indicating that when the levels of STMN2 drop it can lead to a build-up of APP. Moreover, similar proteins have been implicated in NFT formation (Okazaki, 1995). Reduced levels of STMN2 of this gene has been suspected of playing a crucial role in the development of AD as well as Down's syndrome but with no solid proof. More recent studies have found strong links between the drastic reduction of the gene and the development of prion diseases such as Creutzfeldt-Jakob disease (Mead, 2009), although the reduced levels of the resulting protein are most likely a symptom of downregulation upstream by other genes such as SCG10. Finally, STMN2 plays a key role in tubulin binding and stabilises them when phosphorylated by MAPK8, allowing it to control the length of neurons. Naturally, there are strong interactions between STMN2 and TUBB2A, another major hub, in which they form a positive feedback loop, upregulating each other. Additional hubs include NEFL, a neurofilament light polypeptide involved in the maintenance of the neuronal calibre and involved in the intracellular transport for axons and dendrites, CKMT1B, a creatine kinase involved in the transport of high energy phosphate, and CNTNAP2, a contactin associated protein-like that encodes a member of the neurexin family, which function as receptors and cell adhesion molecules in the nervous system. It is worth noting that this protein contains EGFR domains and mediates interactions between neurons and glia during the development of the nervous system. Finally, the only major downregulated, overexpressed hub is CRTAP, which is involved in the degradation of the extracellular matrix and mentioned previously.

Healthy Analysis

Amount of Influence	Gene Symbol	Amount Influenced	Gene Symbol
-2409.38252	GNAI3	4519.617995	STMN2
-2405.895792	LIFR	4058.037015	TUBB2A
-2153.593846	ITPKB	3919.679162	CNTNAP2
-1961.500623	ITPKB	3888.015445	DNAJC30
-1938.311441	GRAMD3	3732.346586	CKMT1A/B
-1924.913063	VCAN	3697.770309	NEFL
-1922.678685	ERBB2IP	3645.044496	STMN2
-1911.597588	1557286_at	3479.563337	TIMM23/B
-1895.432825	CARD8	3350.460166	GAD1
-1868.745806	ADD3	3214.152569	SYT1
1859.310485	AP2M1	3213.11753	ARMC2-AS1 /// ATP5J2
-1828.941544	HMBOX1	2988.745013	ELOVL4
-1808.646486	SSFA2	2839.719429	BEX4
-1778.580875	PABPC1 /// RLIM	-2594.893387	LOC100132057
1744.553734	MOAP1	2580.324251	NAP1L5
1721.578914	ACOT7	2559.681444	PCLO
-1693.086362	GLIS3	2516.49015	EHD3
-1685.903208	PALLD	2484.915441	FAM216A
-1675.529513	VCAN	2466.581447	ATP5H
-1657.416606	TRPS1	2464.248071	GNG3
-1653.117321	CAPN2	2413.053921	C1QTNF4
-1647.983734	FAM111A	2390.111457	C6orf57
-1643.611024	NOTCH2NL	2379.425454	RIMS1
-1635.520913	SEPT8	2314.091955	HIGD1A
-1624.04146	PRKD3	2302.746557	VDAC3
-1617.252873	STON2	2223.283026	MDH1
-1616.707867	NFIA	2136.621832	CEP41
-1613.263754	SWAP70	2122.0799	KRT222
1606.682435	AP2S1	-2116.065213	ANGPT1
-1606.220537	233877_at	2086.459849	DNMIL
1596.751606	MDH2	-2080.683999	PEG3
-1584.30775	SSFA2	2077.126409	NDUFAB1
-1581.243845	ZFP36L1	-2035.795926	242611_at
1577.716054	KIF3A	-2021.675415	LTBP1
-1572.171615	GRAMD3	-2013.726691	AKT3
-1570.886333	WWTR1	2000.280919	MAP1LC3A
1563.171841	FRMPD4	1974.606218	ARMCX4
-1550.159051	SYF2	1968.200746	DCTN1 /// SLC4A5
-1543.888993	FAM107A	1961.044033	CISD1
1542.889975	DDX1	1959.476079	NUP93
-1537.669358	AHNAK	-1945.483864	NSMAF
-1535.919976	CASC3 /// MIR6866	1928.30344	GAP43
1529.933531	BBS7	1914.695147	C14orf2
-1529.267405	LSM14A	-1905.410475	AGK
-1519.091946	TRIM38	1898.734893	PLEKHB2
1510.461285	SLC32A1	1886.580483	UQCRH /// UQCRHL
1506.156909	GOT1	1880.605941	MLLT11
-1500.454172	SSPN	1875.811868	SYT13
1493.098478	ACOT7	1874.327004	LOC101930324 /// NSF
1490.021051	C1QTNF4	-1852.854091	232791_at

Table 11: Driver analysis showing the top 50 source and target genes according to their impact on the network for cognitively normal controls in the hippocampus. The probe IDs in red have not been mapped as of 2017.

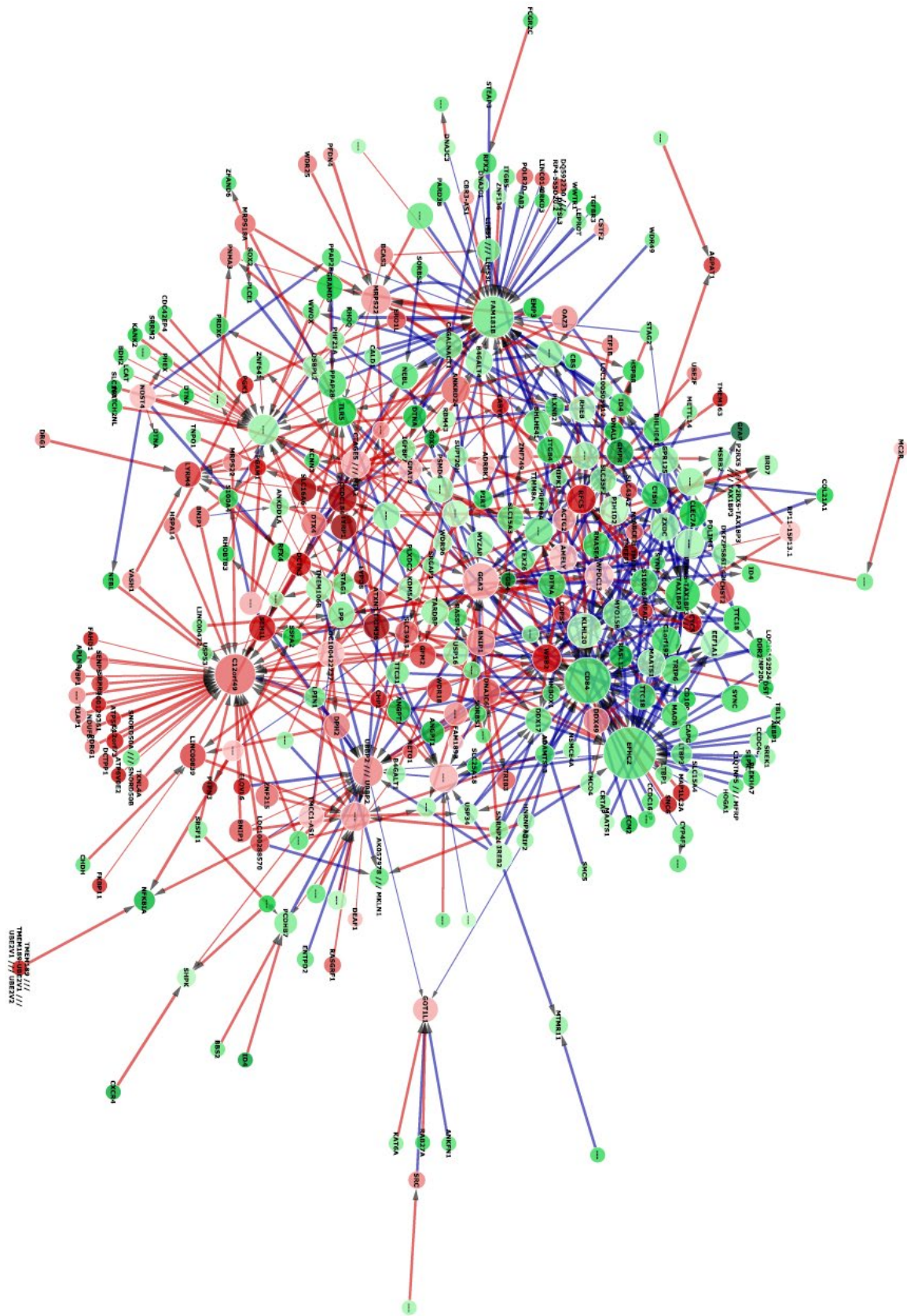
The driver analysis (Table 11) mostly verifies the results of the interactome. Of particular note are a number of genes that are predicted to also be strong sources in the AD driver analysis during the previous section. These include GNA13, LIFR, ITPKB, CARD8 and VCAN. It is thus reasonable to assume that the dysregulation present in AD is not a direct result of these genes. Rather, the drivers of health are present in both healthy and diseased states, it is simply their effects that are lessened, or overshadowed by the significantly stronger drivers of the disease. This is evident when looking at the list of targets which have STMN2 and TUBB2A as the absolute most strongly upregulated ones in cognitively normal individuals while they are not present in the AD drivers. Moreover, genes such as MAP1LC3A and multiple ATP synthases are significantly upregulated in healthy controls when compared to their AD counterparts. This is in addition to the upregulation of apoptosis factors as sources whereas there was evidence of downregulation of factors that drive apoptosis in AD.

Gene ontology analysis of the genes involved in the interactome for the healthy brain showed a balance between a wide variety of pathways and protein products, whereas an equivalent analysis for the AD interactome returned a larger percentage of pathways involved in AD as well as a larger percentage of enzyme modulators and transferases.

### 5.2.2 Entorhinal Cortex

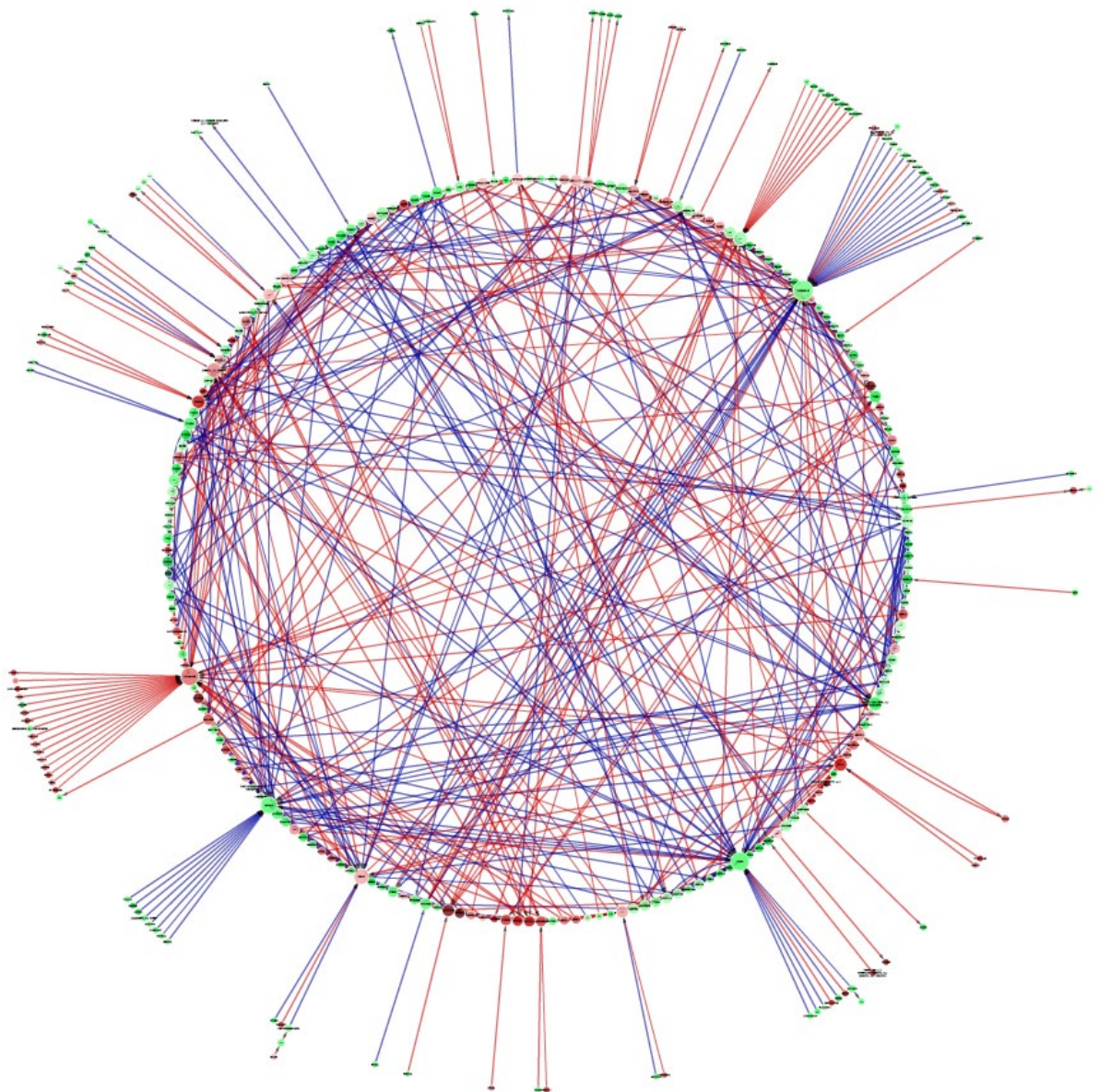
To ensure the validity of the results obtained for the hippocampus, suspected to be the most differentiated region in AD, and as such the region most likely to provide answers to its drivers and targets for therapy, this methodology was also applied to the remaining brain regions available in the E-GEOD-48350 dataset





**Figure 44:** Force directed interactome of the 1000 strongest interactions between 500 genes in the entorhinal cortex, AD brain tissue. Obtained via analysis of the E-GEOD-48350 dataset.

As seen in Figure 44, there is a similar proportion of positive and negatively expressed genes according to their fold change as in the AD hippocampus (3:1). While the entorhinal cortex is not as heavily affected in AD as the hippocampus, studies have shown significant dysfunction in the preclinical stages of AD (Khan *et al*, 2014) although the reasons for this vulnerability are still unknown. However, other sources have reported significant amounts of NFT formation in the entorhinal cortex (Polydoro *et al*, 2013), and as such genes related to microtubules and phosphorylation are expected to play key roles in this interactome.



**Figure 45:** Circular interactome of the 1000 strongest interactions between 500 genes in the entorhinal cortex, AD brain tissue. Obtained via analysis of the E-GEOD-48350 dataset.

The two largest positively regulated hubs identified in the interactome in Figure 44 are EFHC2 and CD84. EFHC2 is implicated in calcium ion binding and has been suspected to play a role in epilepsy but there is little concrete evidence. CD84 on the other hand, is a member of the signaling lymphocyte activation molecule family and has an important role during the immune response by mediating natural killer cell cytotoxicity and increasing proliferative T-cell response. Moreover, the interactome indicates that it is closely linked with EFHC2 and it does respond to cytosolic calcium. Among the non-hubs but still heavily dysregulated genes included in the interactome include MAP1LC3A, which aggressively downregulates EFHC2, and MRPS22, a mitochondrial ribosomal protein that aids in protein synthesis and has been associated with Combine Oxidative Phosphorylation deficiency. When compared to the dysregulation in the hippocampus for the same cohort, there is a clear distinction between the two regions. As all evidence suggests that the entorhinal cortex mostly gets affected during the preclinical stage in AD, by the time the disease has fully developed, the damage has been done and the genetic makeup of the region has stabilised. This provides an extra layer of challenge for achieving reliable early prognosis for AD as the dysregulation is not immediately present at the genetic or RNA level and there is a severe lack of comprehensive protein datasets that could help answer that question.



AD Driver Analysis

Amount of Influence	Gene Symbol	Amount Influenced	Gene Symbol
-2226.88	LOC100422737	-2405.420705	CCDC184
-1953.25	PPAP2B	-2295.738482	NDUFAF1
-1811.37	237448 at	-2157.529211	PTPN3
-1657.15	RHOBTB3	2097.824919	EFHC2
-1599.87	LOC101930112 /// SPG7	-1983.407522	PPFIA1
-1506.29	AMELY	-1973.642031	LYRM4
-1488.07	CHDH	-1935.089836	SLC39A3
-1487.83	CD2AP	-1918.344225	LINC00839
-1481.1	DNAJC3	-1827.502366	TRIM36
-1447.39	237448 at	-1743.727266	TRIB3
-1441.42	GRAMD3	-1694.200669	LMTK2
-1436.27	SRRM2	-1657.038565	DRG1
-1410.46	240262 at	-1626.551919	1567527 at
-1367.98	TNPO1	-1623.440597	MRPS22
-1337.45	AK057978 /// MKLN1	-1581.32752	PGAM1
-1325.36	231528 at	-1564.112782	243788 at
-1290.39	NFKBIA	-1543.430787	USP24
-1252.44	230850 at	-1514.618022	PSMD4
-1245.2	FKBP11	-1491.175367	CHP1
-1242.34	240247 at	-1465.343123	WSB2
-1236.81	RAB12	-1459.546581	AGPAT9
-1222.54	H3F3A /// H3F3AP4 /// H3F3B	-1438.424875	237218 at
-1213.43	DNAJC6	-1437.161334	WFDC13
-1207.24	TMCO4	-1433.600819	MINOS1
-1201.49	ANKRD24	-1432.227992	NDUFB2
-1167.07	RRAGC	-1430.070255	PGK1
-1166.83	PPFIA1	-1426.230107	DCTN3
-1163.37	243528 at	-1419.575789	BNIP1
-1160.84	231063 at	-1395.105705	234838 at
-1153.75	KDM5A	-1382.511927	ADAMTSL3
-1153.67	CNNM3	-1363.085747	MAP1LC3A
-1150.81	UBR5	-1361.318114	TYRP1
-1149.73	USP34	-1361.169992	PLXNA1
-1142.83	WVOX	-1353.908526	MC2R
-1139.01	MTMR11	1334.596438	KLHL20
-1137.19	243014 at	-1329.005161	LOC100288570
-1124.46	234034 at	-1314.000942	C12orf49
-1111.64	ERO1L	-1307.078813	CTAGE5 /// MIA2
-1103.98	1556962 at	-1279.228607	SRSF11
-1100.51	1559332 at	-1275.917971	229859 at
-1093.24	TTC31	-1259.693846	TMEM163
-1092.92	LINC00839	-1251.838109	HMOX2
-1090.68	WSB2	-1236.792542	BRD7
-1086.21	232198 at	-1224.665884	TMCC1-AS1
-1065.45	HSPA14	-1219.318887	PFDN4
-1062.91	HIPK1	-1218.289383	SHPK
-1060.55	PHF21A	-1215.489131	SEH1L
-1057.99	RHOQ	-1201.064901	SRC
-1053	RHOBTB3	-1183.916437	240262 at
-1050.37	239857 at	1175.461243	CD84

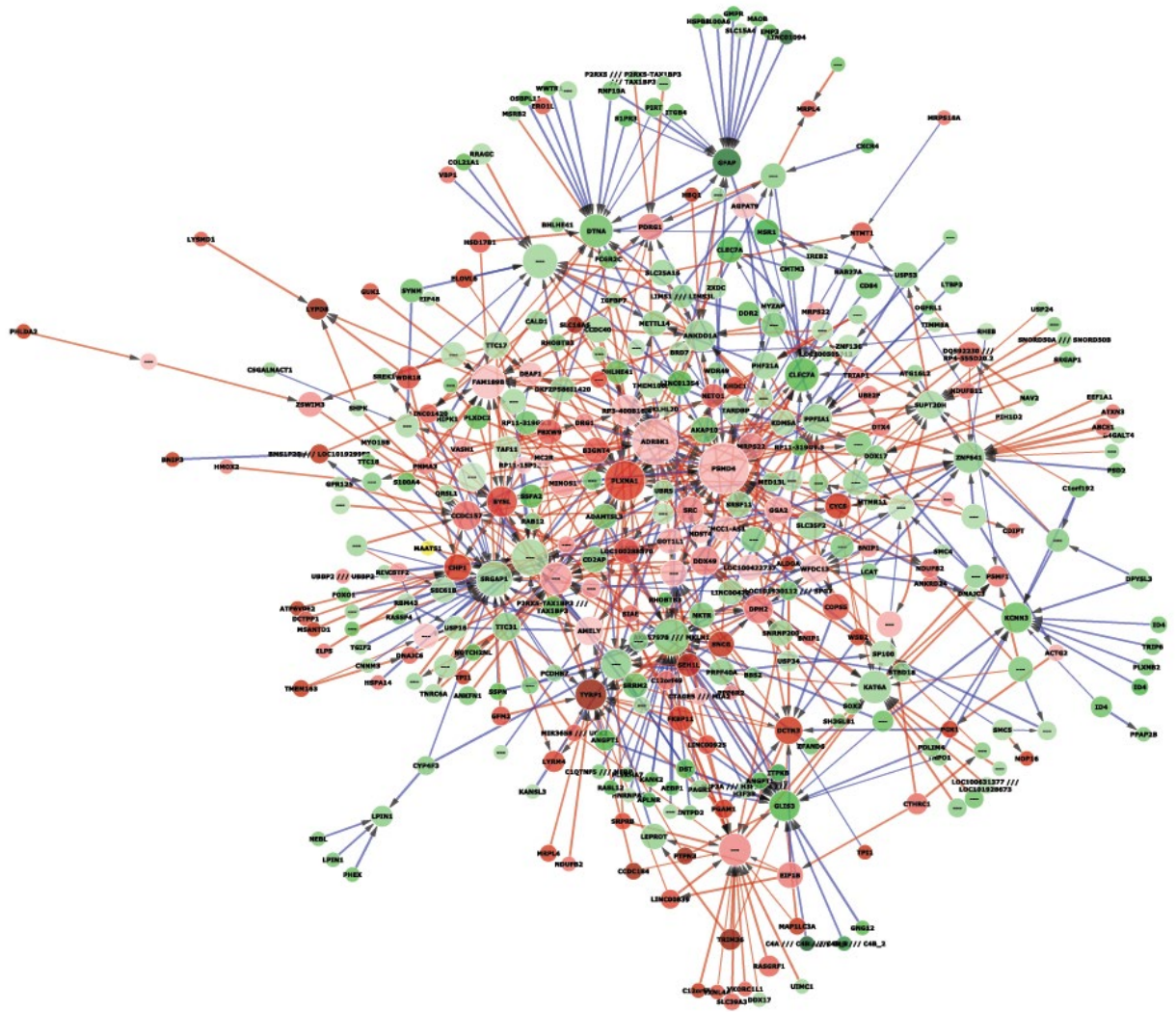
Table 12: Driver analysis showing the top 50 source and target genes according to their impact on the network for AD in the entorhinal cortex. The probe IDs in red have not been mapped as of 2017.

The driver analysis (Table 12) is also slightly weakened by the relatively large percentage of unassigned gene probes. However, it allows us further insight than the previous interactome and provides further information on the differentiation of key genes. Indeed, the top source genes include PPAP2B, also known as PLPP3, a gene encoding for phosphatidic acid phosphatase which converts phosphatidic acid to diacylglycerol and has been shown to hydrolyse phosphatidic acid. This could lead to its relevance to the formation of NFTs as downregulation of this gene, as seen in this case, could impede the brain's ability to hydrolyse and clear phosphate groups, leading to their buildup and resulting phosphorylation of tau. Moreover, downregulation of RHOBTB3, a RHO GTPase could lead to the same problems seen in other regions as explained earlier as well as the downregulation of NFKBIA. The most influential source genes repeat this pattern with multiple other RHO GTPase related genes, phosphor regulating factors and immune system regulators such as NKTR in position 62 which is found on the presence of NK cells and facilitates target binding.

Meanwhile, the targets include multiple downregulated phosphatases and kinases, which most definitely impede the ability to remove excess phosphoric acid and its ions as well as leading to disruption in the ATP cycle restricting the available energy in the system. A prime example of this phenomenon is the severe downregulation of AGPAT9. Genes such as MAP1LC3A, which have been quite reliably influential in previous tests, are still present although related genes such as tubulins and other microtubule regulating factors are not, possibly due to the fact that the damage in the region has already been done by the time the patient has succumbed to the disease, as mentioned.

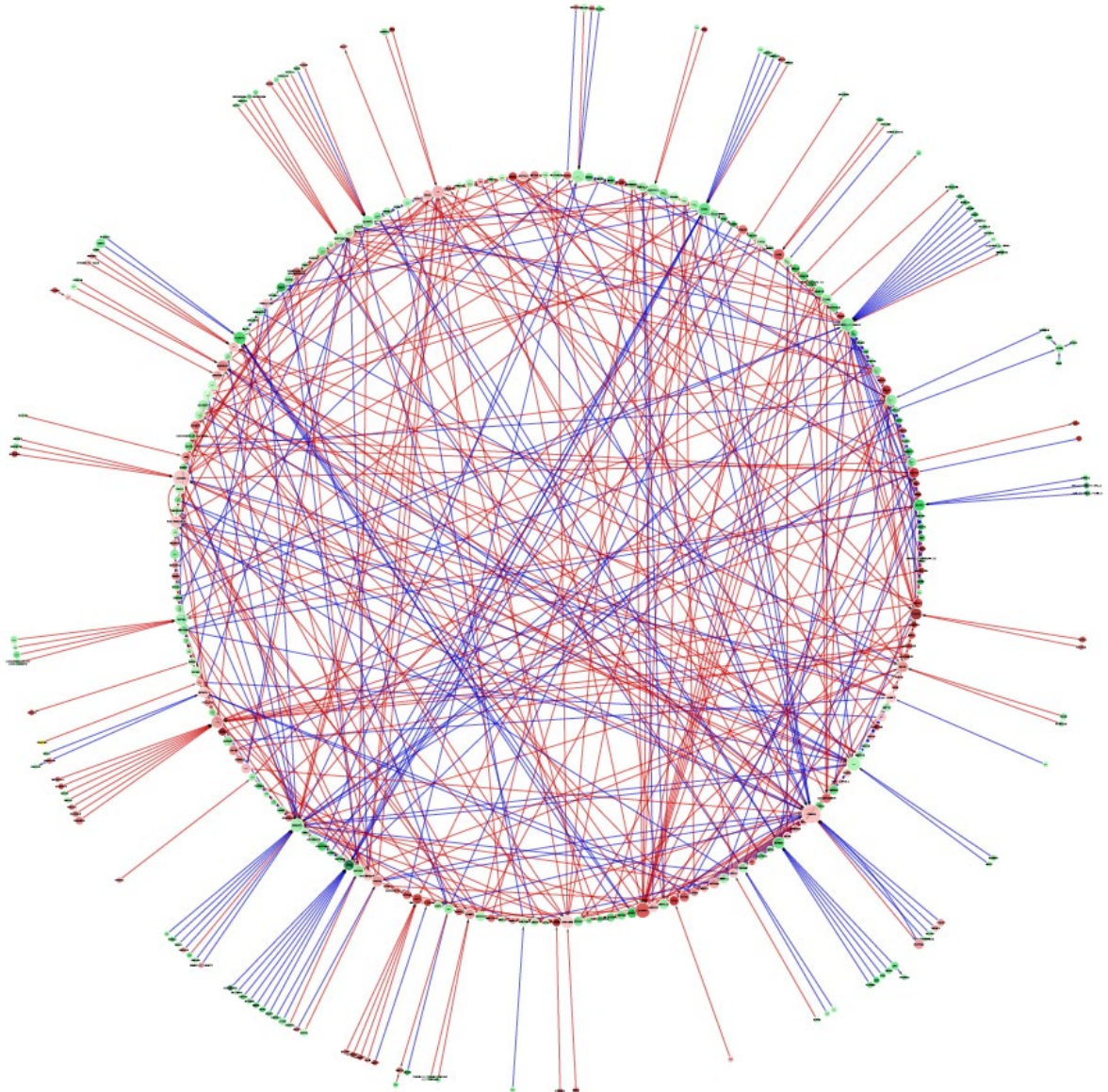
#### *Healthy Interactome*

Although in the AD cohort for the entorhinal cortex the expected genes were directly correlated with factors that lead to the formation of NFTs, in the healthy cohort it is expected to show greater similarity with the hippocampus healthy cohort. If the mechanisms that are involved in clearance of damaged neurons as well as the ones directly correlated with phosphate processing, it can be surmised that the entorhinal cortex is perpetually combating the effects of phosphoryl accumulation.



**Figure 46:** Force directed interactome of the 1000 strongest interactions between 500 genes in the entorhinal cortex, cognitively normal brain tissue. Obtained via analysis of the E-GEOD-48350 dataset.

Much like the previous tests there is a consistent balance of negative to positive fold change and a large enough amount of positive and negative regulation to allows us to collect statistically significant data. It is worth noting that in this case the interactome is quite complex with little clear divide between different regulation types. This is most likely a direct result of the interactome being based on a healthy, fully functioning system.



**Figure 47:** Circular interactome of the 1000 strongest interactions between 500 genes in the hippocampus, cognitively normal brain tissue. Obtained via analysis of the E-GEOD-48350 dataset.

The largest hubs in figures 46 and 47 are involved in the entorhinal cortex of cognitively normal individuals include PSMD4, a subunit of a proteasome complex involved in the maintenance of protein homeostasis by removing misfolded, damaged proteins or ones that are no longer required. Moreover, PSMD4 appears to have an affinity for polyubiquitin chains, which is quite interesting because there are enough ubiquitin genes that are dysregulated, and are evident in the preceding interactomes and driver analyses, but have never been deemed crucial to the condition. The function of this protein however is absolutely essential to the normal function of the human brain, which is further reinforced by the fact that it is not significantly underexpressed and is being up and downregulated by a significant number of genes. Another major hub is the similarly

regulated ADRBK1, also known as GRK2, a G protein-coupled receptor kinase that phosphorylates the activated form of the beta-adrenergic and related G-protein-coupled receptors. Phosphorylation factors, especially central ones being down and upregulated are fully expected to be present in a healthy entorhinal cortex. Another highly expressed, highly upregulated factor is GFAP, encoding for a glial fibrillary acidic protein, which encode for a major filament protein of mature astrocytes. In a similar highly expressed vein is MSR1, which has been discussed preciously, and CLEC7A, a c-type lectin essential in the activation of NFkB and the TLR2-mediated inflammatory response, as well as enhancing cytokine production in dendritic cells.



Healthy Analysis

Amount of Influence	Gene Symbol	Amount Influenced	Gene Symbol
-2055.35269	LOC101930112 /// SPG7	-2926.656486	PLXNA1
-1887.274866	NKTR	-2500.323389	ANKDD1A
-1772.076267	240262_at	-2417.31448	LOC100288570
-1674.912185	1558740_s_at	-2356.129179	LYRM4
-1663.982229	PLXDC2	-2191.086398	PGAM1
-1591.617262	TTC31	-2111.332118	1564192_at
-1489.467654	239857_at	-2057.01101	CYCS
-1476.278986	RAB12	-1983.069152	DRG1
-1447.314771	KDM5A	-1978.898011	SEH1L
-1438.495887	242233_at	-1964.958496	TYRP1
-1407.077744	CCDC40	-1938.007631	CHP1
-1394.240064	231063_at	-1880.841699	242233_at
-1381.757153	AK057978 /// MKLN1	-1873.881528	FAM189B
-1358.615057	242696_at	-1788.892702	BYSL
-1356.715103	AMELY	-1672.796488	CCDC184
-1353.94991	USP24	-1662.242611	228297_at
-1345.502474	BRD7	-1601.963374	229859_at
-1338.763092	DDX49	-1579.632409	LYPD8
-1333.654121	ANKDD1A	-1572.023741	1567527_at
-1325.624031	SLC35F2	-1552.095487	UBFD1
-1263.157738	EFHC2	-1535.838205	1560557_at
-1242.749194	SRRM2	-1507.115485	NOP16
-1231.060993	CTHRC1	-1493.502815	SLC39A3
-1224.731572	TTC17	-1479.526697	DCTN3
-1223.165671	234838_at	-1474.7252	CCDC157
-1221.234455	WFDC13	-1471.145615	TRIB3
-1213.870404	DDX17	-1470.976423	EIF1B
-1213.558752	SMC4	-1454.859265	CLTA
-1202.967069	MYO15B	-1450.641354	KDM5A
-1184.289502	QRSL1	-1431.30231	WDR18
-1171.08059	MSRB2	-1391.994517	1559332_at
-1158.988367	FBXW9	-1390.681502	AGPAT1
-1145.319796	SRSF11	-1384.475985	AGPAT9
-1144.427096	233007_at	-1371.324079	CD84
-1123.715192	239957_at	-1365.551339	TPI1
-1116.307871	TNPO1	-1357.94373	242362_at
-1104.782451	SNORD50A /// SNORD50B	-1352.485004	CDIPT
-1099.585717	ADRBK1	-1347.184406	UBR5
-1092.836168	ACTG2	-1342.367093	PGK1
-1084.326908	TIMM8A	-1332.516406	ADRBK1
-1073.831529	REV1	-1326.819807	237448_at
-1071.838864	SLC25A16	-1316.71338	ABCE1
-1071.221176	231528_at	-1316.709625	SUPT20H
-1059.332188	MTMR11	-1310.819144	SIAE
-1057.547907	FKBP11	1305.685299	1558740_s_at
-1037.20213	SRGAP1	-1288.103742	PFDN4
-1004.080002	233876_at	-1258.666369	SRSF11
-993.553969	HMOX2	-1256.974908	RHOBTB3
-951.0615682	1567527_at	-1247.409671	238714_at
-937.5932135	237218_at	-1240.624871	SNCG

**Table 13:** Driver analysis showing the top 50 source and target genes according to their impact on the network for cognitively normal controls in the entorhinal cortex. The probe IDs in red have not been mapped as of 2017.

Much like most healthy interactomes, the entorhinal cortex interactome shows a significant degree of balance between hubs and their sources (Table 13), with both up and downregulation of and by multiple factors, necessitating a driver analysis to gain further insight. Indeed, the most influential source genes include genes such as NKTR, a natural killer cell triggering receptor, a myotubularin protein coding gene MTMR11 and multiple transport molecules and enzymes. There are no significantly differentiated due to the balance present in a healthy system. The target however, include genes such as PGAM1, a gene that catalyzes the reaction of 3-phosphoglycerate to 2-phosphoglycerate, AGPAT1 and APGAT9, RHOBTB and CD84, discovered in the AD cohort. The largest target however appears to be the heavily downregulated PLXNA1 gene, which encodes for plexin A1, a coreceptor for class 3 semaphorins that aid in axon guidance, cell migration and invasive growth.

It is clear from these results that a healthy system cannot be examined in a vacuum. Only by comparing it to a dysregulated system due to disease is it possible to glean the required information. Of course, it does not require to be considered a control. In fact, when the dysregulated system is used as the control, the healthy system allows for the discovery of factors that are absolutely crucial to the maintenance of homeostasis and health, but are suppressed or simply masked by other genes in a disease system.

### 5.2.3 Postcentral Gyrus

The following sections will be more limited in scope due to the lack of sufficient evidence in literature to link either the postcentral gyrus (Canu *et al*, 2011, Zhang *et al*, 2015) or the superior frontal gyrus (Cinco *et al*, 2015) to AD. The interactome analysis has been successfully carried out but the results provided no further information than that present in the driver analysis, so in the interest of brevity the drivers will be examined exclusively.

AD Driver Analysis

Amount of Influence	Gene Symbol	Amount Influenced	Gene Symbol
-1464.936233	DDX54	-1973.931954	AR
-1179.027568	RAD51D	-1933.572158	NUDT9P1
-1087.578723	ALOX15B	-1851.793089	ABCC12
-1080.622499	C12orf45	1804.165749	DGKZ
-1016.69935	ADRA1D	-1774.456944	LOC286189
-1011.676087	SOS1	-1606.320931	TFIP11
-1011.287052	RPS6	-1586.001369	SESN1
-1009.584267	RAB40C	-1542.802097	DNAL4
-987.9211308	STOX2	-1516.663898	HTR7P1
-963.4832192	CTIF	-1468.332493	ATP5F1
-945.8896223	CD59	-1425.087791	PHF20L1
-908.5333394	AGPAT1	-1423.47152	CA4
-865.2530647	MAP1LC3A	-1400.597669	GUSBP5
-838.310936	TRIM22	-1373.224043	LRCH3
-818.4887235	DUSP4	-1366.049086	DOPEY1
-813.7914104	GNB2	-1361.666576	CA4
-791.0628184	GSS	-1354.892422	DISP2
-789.7588241	FAM83H	-1352.689816	R3HDM4
-786.682945	PTPRC	-1343.537184	AR
-778.2661672	PPP1R3D	-1324.445007	239358_at
-776.4001765	FYB	-1309.36311	GNB2
-775.4624098	FDFT1	-1291.724818	CCDC176
-774.727238	TYROBP	-1288.6565	DOPEY1
-770.4636407	LOC389906	-1273.589748	HSPA6
-753.6084113	SYTL3	-1254.143212	G3BP1
-750.0823808	CARD16	-1249.391637	LINC00461 /// MIR9-2
-746.8725727	LINC00263	-1228.592257	SLC30A3
-745.8424282	1554963_at	1212.433295	LOC101927424
-743.9049148	ARAP1	-1209.97314	C1orf95
-740.7199148	WDR77	-1207.113261	SPAG9
-735.922145	MAP4	-1188.497582	STOX2
-730.9337404	SAMHD1	-1161.200508	TMEM180
-723.7250261	SIRT2	-1145.476043	240248_at
-715.3868645	KCNQ2	-1136.28564	KLHL35
-708.6175104	SOAT1	-1120.587071	PHC3
-707.4666258	1567575_at	-1116.284094	NOMO3
-707.2431256	NAT14	-1113.025718	MS4A6A
-706.8118627	SRMP1 /// SRMP1	-1092.128676	EZH1
-705.1177343	KMO	-1079.603041	236766_at
-692.7333625	RAB27A	-1063.377498	PCGF5
-688.0222904	TMEM8B	-1056.963417	CTSS
-683.7112662	MEIS3	-1048.189126	SNX9
-680.749865	LSM4	-1045.977048	SLX4
-672.67009	RPL14	-1017.048194	AASDH
-666.8094881	GKAP1	-1004.203334	242181_at
-660.9792509	GNG4	-998.2054565	215845_x_at
-656.1201136	AGPAT1	-990.5007932	NEXN
-647.7846893	DGKZ	-978.2134236	LSM4
-647.3039133	ELF2	-961.0000017	MUL1
-639.2112922	RNPS1	-953.9697503	ARID2

Table 14: Driver analysis showing the top 50 source and target genes according to their impact on the network for AD in the postcentral gyrus. The probe IDs in red have not been mapped as of 2017.

As shown in Table 14, while the impact of AD on the postcentral gyrus has not been studied extensively and the results have proven inconclusive, this analysis serves a key role in allowing for further reduction in bias by performing more inclusive and extensive deep mining of the information present in the E-GEOD-48350 dataset. There are no expected outcomes as the region appears to be less affected than other regions by the progression of the disease. Nevertheless, the presence of genes such as AGPAT1 and MAP1LC3A in the most influential genes lends further credence to the theory that these genes are essential for the maintenance of brain health and dysregulations in them can have far-reaching consequences. There are also multiple factors encoding for transmembrane proteins and RNA signaling molecules, translation initiation factors and other genes related to the function of RNA, but they are common across both AD and cognitively normal controls preventing them from being used as reliable markers of the disease. Of note is TRIM22, a member of a tripartite motif including zinc binding domains and involved in interferon signaling which, when combined with the effects of CD59, which has been present in previous analyses, reinforces the role of the immune system to the development of the disease.

Similarly, the most influential factors include the usual suite of ATP synthases, signaling molecules and kinase regulating factors such as SPAG9, which is related to kinase binding and MAP-induced scaffold activity. Overall the results of the postcentral gyrus are inconclusive and do not provide a clear driver for AD, contrary to the information datamined from the hippocampus and entorhinal cortex.

Healthy Driver Analysis

Amount of Influence	Gene Symbol	Amount Influenced	Gene Symbol
-1275.97	C12orf45	-1828.26	PHC3
-1119.04	RGS8	-1612.74	LPIN1
-1059.51	RNF121	-1610.01	KLF12
-1048.84	RAD51D	-1578.27	215845_x_at
-1017.33	RAB40C	-1530.72	SLC11A2
-968.412	DDX54	-1488.99	LOC286189
-919.211	MRPS18A	-1465.15	RBM45
-884.774	KLHL35	-1461.93	215284_at
-854.371	TMEM8B	-1453.96	RP5-1085F17.3
-844.552	SRRD	-1453.2	NEBL
-839.153	KCNQ2	-1451.5	ALS2
-829.745	TM7SF2	-1434.51	1559235_a_at
-827.67	KLHL14	-1426.71	243682_at
-823.97	EXOSC5	-1322.05	G3BP1
-789.81	RASAL1	-1299.17	DNAL1
-787.272	STMN1	-1247.66	RHOQ
-786.855	BFSP1	-1235.52	ZBTB33
-783.146	MAP1LC3A	-1232.17	ICA1L
-781.861	NOL9	1213.012	KIAA2026
-769.47	NUDT9P1	-1192.67	244503_at
-767.787	KMO	-1189.58	CNTLN
-756.708	ZCCHC17	1189.32	WFDC2
-752.284	C1orf95	-1187.18	TIFA
-749.695	B4GALNT1	-1179.44	TNXA /// TNXB
-746.77	NAT14	-1163.6	FRYL
-746.103	HRK /// LOC283454	-1140.48	GUSBP5
-744.963	AGPAT1	1117.584	FOXN3
-731.299	1565579_at	-1087.55	MLLT4
-730.915	LMBR1	-1073.65	FLRT2
-723.051	ADRA1D	-1072.52	BDNF
-721.577	WDR77	-1064.62	CENPVP1 /// CENPVP2
-718.824	DPM2	-1060.37	C3orf62
-705.39	LSM4	-1057.07	ST6GAL2
-703.738	SNRNP48	-1042.75	HSPA6
-702.194	PKIG	1022.383	PGM5
-697.484	MSANTD3-TMEFF1	-1008.65	USP5
-697.297	SLIT3	-1007.13	TNXA /// TNXB
-693.64	FCHSD2	-1002.93	1567575_at
-686.128	DGKZ	-998.348	MAP4
-685.468	C11orf31	-990.666	CTA-254O6.1
-674.217	FAM132B	-982.602	PURA
-673.654	HRK	-970.085	TTC23
-672.88	ANKRD24	967.7535	FOXN3
-671.997	FRMPD2	-962.999	MAP1LC3A
-661.261	AGPAT1	-953.689	HOTAIRM1
-660.994	C11orf31	-952.058	PACSIN2
-659.397	SRM	-950.972	KMO
-655.541	LOC389906	-941.991	LOC101929243
-653.078	SYNCRIP	-941.909	MS4A6A
-650.3	STOX2	-941.026	MKLN1

**Table 15:** Driver analysis showing the top 50 source and target genes according to their impact on the network for cognitively normal controls in the postcentral gyrus. The probe IDs in red have not been mapped as of 2017.

Curiously, analysis of the cognitively normal interactome as well as analysis of the drivers (Table 15) shows remarkable similarities between the postcentral gyrus of AD patients and healthy controls, with the presence of genes such as *AGPAT1*, *STOX2* a little studied protein possibly related to growth restriction and *KCNQ2*, encoding for a potassium channel protein regulating neuron excitability and inhibited by M1 muscarinic acetylcholine receptors, the role of which in epilepsy has been extensively studied (Wang *et al*, 1998, Rim *et al*, 2018). Moreover, the most influenced genes include common factors such as *MAP1LC3A* and *FOXN3* suggesting a similar genetic makeup with other brain regions in healthy controls in regard to factors related to AD.

## 5.2.4 Superior Frontal Gyrus

### AD Driver Analysis

Amount of Influence	Gene Symbol	Amount Influenced	Gene Symbol
-1213.79	MRPS18A	-1754.53	CKLF
-1090.63	C19orf10	-1684.33	KLHL6
-1058.41	C19orf10	-1551.9	R3HDM4
-1048.87	JMJD4	-1502.52	LEAP2
-993.251	BAZ1A	-1501.61	PNPT1
-964.158	CHID1	-1492.24	242772_x_at
-962.807	RASSF8	-1414.01	1554266_at
-961.634	FAM21EP /// FAM21FP	-1394.47	SASH1
-944.845	TSTA3	-1380.39	DDX54
-939.47	RBM17	-1369.88	CTSS
-934.189	SOAT1	-1298.27	ACTL6A
-918.27	CASP7	-1288.67	PPM1L
-906.806	RAB29	-1282.56	SUPT20H
-902.566	RNF114	-1243.1	MAST3
-858.93	FOXN3	-1200.37	KLHL21
-843.336	MANBAL	-1174.53	PIM2
-837.253	BEND7	-1155.5	RPS27L
-835.29	RNASET2	-1151.12	ZDHHC23
-832.957	ENAH	-1143.25	ENSA
-809.333	PPBP	-1131.71	PER3
-805.221	ITPR2	-1125.65	GNA13
-803.346	TLR7	-1121.14	KIAA1217
-781.731	HAVCR2	-1119.4	PRPF4B
-781.369	TRMT61A	-1118.73	PCDHGC3 /// PCDHGC5
-780.433	PBX3	-1115.36	SEMA5B
-780.231	PRKD3	-1111.9	CYTL1
-767.834	RNF114	-1100.22	PURA
-765.025	DCTPP1	-1094.09	CDIPT
-752.738	IRF9 /// RNF31	-1088.33	DLK2
-747.514	FOXN3	-1079.76	PDLIM4
-745.337	FBXW2	-1079.72	CD37
-740.842	ANAPC16	-1055.16	TBC1D24
-735.363	FOXN3	-1044.97	RC3H2
-733.385	NIPSNAP1	-1042.82	CD84
-732.693	SPR	-1034.88	DHFR
-729.794	DDOST	-1028.84	LCP2
-729.146	HMBOX1	-1023.12	CRY1
-710.391	LPIN1	-1019.95	CDK16
-710.324	PARVG	-1010.3	PTGES3L
-710.021	MYL5	-1002.55	ITGAM
-707.564	RNASET2	998.731	CRIP2
-699.659	ANAPC16	-980.45	GPATCH2
-698.79	TAB2	-972.252	P2RX7
-695.322	RHOQ	-970.605	C3AR1
-695.316	RHOQ	-969.315	PBX3
-692.518	216675_at	-958.59	CREBRF
-691.935	PPM1L	-952.357	SWAP70
-688.825	GKAP1	-943.685	CEP350
-684.884	LPIN1	-942.43	LINC01158
-683.976	ARHGDI2A	-941.928	CXCR4

**Table 16:** Driver analysis showing the top 50 source and target genes according to their impact on the network for AD in the superior frontal gyrus. The probe IDs in red have not been mapped as of 2017.

Similar to the postcentral gyrus, there is very little evidence linking the superior frontal gyrus to AD (Table 16), especially its development, leading to this analysis being used to further reduce bias and study the impact of AD in other brain regions. Of particular note among the most influential genes is CASP7, encoding for a protein of the caspase family involved in apoptosis, which is similar to genes found in the hippocampus of AD patients earlier, as well as RHOQ and FOXN3. The number of genes related to the RAS oncogene family and other GTPases is larger than expected, indicating a greater similarity to the dysregulation seen in the hippocampus of AD patients than the postcentral gyrus, while the high number of transferases such as SOAT are significantly harder to use as AD biomarkers due to their persistence across brain regions in both AD patients and healthy controls.

The most influenced genes further support previous findings with genes such as CKLF, a highly downregulated cytokine which acts as a chemoattractant for neutrophils, monocytes and lymphocytes, KLHL6, encoding for a member of the kelch-like family of proteins and is involved in B-cell receptor signaling or LEAP2 a protein with antimicrobial properties mostly expressed in the liver. This might have been misidentified instead of a protein with similar properties in the brain. Genes such as PNPT1, an RNA binding and degradation protein of the highly conserved polynucleotide phosphorylase family which has been shown to degrade and clear oxidized RNA upon exposure to interferon beta (IFNB) can also be linked to neuroinflammation. In fact, there appears to be a higher than average immune system factors in the most influenced genes of AD patients according to this driver analysis. This could be a response to the spread of the disease which leads to an increased immune response in an attempt to prevent further damage, or a byproduct of the disease itself.



Healthy Driver Analysis

Amount of Influence	Gene Symbol	Amount Influenced	Gene Symbol
-987.77	KCNIP3	-2047.59	DHFR
-952.769	SON	2037.828	DOCK11
-865.787	CKLF	-1906.49	MSL2
-857.041	227503_at	-1772.76	PCDHGC3
-843.541	FOXN3	-1771.93	216675_at
-830.164	KTN1	-1616.08	IRF9 /// RNF31
-814.172	243859_at	-1589.52	ZCCHC17
-807.437	CREB3L1	-1541.89	ARL4D
-802.468	KBTBD2	-1516.76	FAM21EP
-800.791	ACBD5	-1473.31	MEIS3
-763.572	NAP1L1	-1451.45	GEM
-759.172	RBM17	-1373.93	GPATCH2
-737.244	RHOQ	-1348.31	C22orf29 /// GNB1L
-711.927	HES4	-1325.94	SOCS4
-709.766	ARL5A	-1242.75	LINC01094
-702.432	CDK16	-1224.18	ZCCHC17
-697.999	CDK2AP1	-1163.28	LGALS1
-697.512	235422_at	-1156.45	SOX2
-696.048	ITPR2	-1151.7	FAM204A
-685.166	RASSF8	1149.141	LCP2
-683.218	DDX54	-1134.89	DCTPP1
-675.298	RAB40C	-1128.27	KIF5A
-673.808	ITPKB	-1111.44	NUS1P3
-673.275	241258_at	-1082.71	MYL5
-672.605	MCM3AP-AS1	-1081.52	RPS27L
-671.643	239476_at	-1075.66	DHFR
-660.488	LINC01158	-1063.6	CES2
-659.507	RHOQ	1047.412	227503_at
-652.606	229541_at	-1045.43	ST6GAL2
-652.181	ADAMTS8	-1004.48	PNRC1
-648.489	RHOBTB2	-985.555	CHD9
-637.393	MBD2	-980.524	ARHGAP5
-637.345	BAZ1A	-978.525	FBXW9
-622.261	ATP11C	-975.17	FAM66B
-622.198	1554266_at	-966.201	ITGB8
-619.03	FGD5-AS1	-963.636	C5AR1
-607.55	MIR6778 /// SHMT1	-960.113	MYH11
-604.991	HMBOX1	-959.265	ARSG
-602.276	PTPN12	-956.923	BRD9
-600.845	LEPREL1	-952.538	JMJD4
-599.371	KIAA1217	950.1384	SRSF4
-592.976	MXI1	-948.727	FCGR2C
-592.362	TM2D1	-945.457	ZNF664
-588.981	ENAH	-935.734	ANKRD40
-587.092	LARP7	935.083	RP5-1085F17.3
-586.616	HNRNPH3	-934.113	BAZ1A
-575.083	PRKD3	-926.617	KIF5B
-574.925	240180_at	-921.744	STAG1
-574.705	LOC100287387	-901.119	ESYT2
-572.817	SYNCRIP	877.4567	FOXN3

**Table 17:** Driver analysis showing the top 50 source and target genes according to their impact on the network for cognitively normal controls in the superior frontal gyrus. The probe IDs in red have not been mapped as of 2017.

Much like the superior frontal gyrus of AD patients, cognitively normal controls (Table 17) show significant similarities in their most influential genes with FOXP3 and multiple Ras proteins being prominent. There is also a significant number of kinases such as ITPKB and ion channels like KCNIP3, which responds to intracellular calcium in particular and has been shown to interact with presenilin, and genes such as ACBD5, an enzyme involved in lipid and acyl-CoA binding. These results corroborate previous findings of a brain in a healthy state having a significant number of varied and essential genes being among the most influential ones.

This trend continues with the most influenced genes including MSL2, which encodes for a protein component of a histone acetyltransferase complex related to chromatin organization and DHFR, a highly conserved reductase essential for the synthesis of purines, thymidylic acid, and certain amino acids, involved in proliferation and expressed by most organisms. SOCS4 is particularly noteworthy as a suppressor of cytokine signaling involved in the negative feedback system that regulates cytokine signal transduction and inhibits EGF signaling and stands in stark contrast with the significant number of highly influenced factors driving the immune response in AD. This pattern continues with the now common selection of kinases and transcription factors. The differences in the most influence genes between a cognitively normal individual and an AD patient are highly pronounced in the superior frontal gyrus. However, due to the lack of research in this region, there is a significant challenge to compare these results with published literature and other studies and determine their importance, if any, to AD, but could provide a new avenue for research regarding the response of different brain regions to the disease.

### 5.3 Comparison Against Known Markers

Over the last few sections, as well as the previous chapter the E-GEOD-48350 dataset was deep mined in a non-parametric hypothesis free manner. This allowed for the unbiased discovery of dysregulated factors, potential markers and a very large number of genes that can be used in conjunction with each other to explain how AD develops and progresses. However, even through all this datamining there was very little evidence of the major genes that are known for a fact to play a crucial role in AD, even though multiple factors that interact with or regulate them have been considered significant, such as tubulins and

phosphorylation factors for tau and STMN2 for APP. Based on the amyloid cascade hypothesis, there are a few know factors that have been used as biomarkers in AD. As mentioned earlier, mutations in the APOE gene lead to familial AD, the MAPT gene encodes for the tau protein and the APP gene encodes the amyloid precursor protein. In the E-GEOD-48350 dataset there are 3 APOE, 3 APP ad 6 MAPT probes. Instead of averaging these probes to create an approximation of the overall expression of their corresponding genes, it was decided that each probe will be considered as its own predictor and analysed in a continuous manner. The advantage of this technique is that it reduces the variance focusing the entire analysis on a single probe without significantly increasing the bias as every patient sample has its own expression level for each probe and they are all considered valid predictors. As a result, the probability that a given gene will be able to explain the variance in AD patients based on the expression profiles of these three genes is

$$\left(\frac{1}{54675}\right)^{12}$$

as shown in section 4.6. Moreover, an interactome and driver analysis should also provide a clearer picture of the genes involved in the regulation of these three factors, and how they differ from previous tests.

In order to prevent this chapter from getting overloaded with information and getting needlessly large, the results will be restricted to the most crucial ones obtained via driver analysis. Over the previous sections, it was proven that not only can the drier analysis provide further insight into the complexities of gene-gene interactions, but also avoid the bias inherent in a force directed network without contradicting it. Moreover, as there are multiple probes for each of the genes of interest in E-GEOD-48350, the selected genes will include genes common across all iterations of the gene probe. The complete table is available in the appendix.

The driver analyses for all available probes have one specific feature in common; targets, the genes most influence by the network are significantly more heavily regulated. In fact, the range of values for sources genes can be less than a quarter that those of the targets and have a much more even rate of descent. This is supported by the nature of the data

discussed earlier during the section on the non-parametric approach. As for this series of tests, the variance was decreased enormously by focusing the analysis on a single point in the array, and by performing a continuous stepwise analysis based on the genes suspected as being causes of the disease, we have found the genes most likely to explain the variance not in the disease, but in the expression of the specified gene. This results in a highly target focused analysis as the genes most likely to explain the variance of the predictor (APP, MAPT, APOE4) are likely to be their drivers.

Moreover, the probes all have significant degree of dissimilarity. For APOE4, the 50 most influential genes for each probe have no gene that is common between all three probes. The first probe (203381\_s\_at) in particular, only shows a single common gene between itself and one other probe. The most influenced genes on the other hand show 2 genes as common across all probes and the first probe remains visibly dissimilar to the other two as shown in Table 18.

### 5.3.1 APOE4 - Apolipoprotein E

<b>Most Influential genes for APOE4</b>			
<b>Gene</b>	<b>Probe 1 203381_s_at</b>	<b>Probe 2 203382_s_at</b>	<b>Probe 3 212884_x_at</b>
<b>HSDL2</b>	40	34	31
<b>SMYD2</b>	26	5	43
<b>ADD3</b>	14	15	
<b>AGT</b>	42	46	
<b>PDLIM5</b>	28		33
<b>APOE</b>	51		2
<b>THRA</b>		47	8
<b>PVALB</b>		4	3
<b>ANK1</b>		18	38
<b>GPR125</b>		11	5
<b>VAMP1</b>		17	12
<b>PCDHGA1</b>		33	39
<b>NTRK2</b>		49	47
<b>FAM167A</b>		1	40
<b>FIBIN</b>		9	15
<b>FBXO33</b>		8	6
<b>IL17D</b>		43	51
<b>S100A16</b>		24	34
<b>ZNF385B</b>		37	10
<b>NPAS3</b>		44	41
<b>NARF</b>		32	24
<b>GTDC1</b>		7	7

**Table 18:** List of the most influential common genes obtained by driver analysis between the three probes for the APOE4 gene in the AFFY-44 array. The number indicates the position that gene is found in its respective analysis.

The only genes common between the APOE4 analysis and the AD analyses are the ZNF385B, ADD3 and IL17D genes, which are also found in the AD hippocampus driver analysis. ZNF385B is a member of the Zinc Finger Protein, members of which have been found regularly in previous experiments, with this one related to p53 binding. This is interesting as the relationship between AD and p53 has been theorised previously, as explored in the introduction, and this is further proof. ADD3 is an adducin and plays a key role in the spectrin-actin network and has been identified as relevant most likely due to being a cytoskeleton associated protein, and IL17D is a cytokine whose expression is dependent on that of NFkB. In conclusion, APOE4 is significantly distinct compared to previous experiments as expected. As APOE4 has been found to be directly associated with familial AD and none of the patients in the E-GEOD-48350 cohort are subject to it, this is further proof that AD and familial AD are similar but functionally different conditions.

### 5.3.2 APP - Amyloid Beta Precursor Protein

As far as APP is concerned, the probes are completely dissimilar, with no commonalities among the 50 most influential sources or targets (Table 19). As a result, it is impossible to reach a conclusion as to the possible drivers for APP in AD. In order to determine any common features that can be used as predictors in future tests, the number of genes was expanded to 200.

<b>Most Influential genes for APP</b>			
<b>Gene</b>	<b>Probe 1 200602_at</b>	<b>Probe 2 211277_x_at</b>	<b>Probe 3 214953_s_at</b>
<b>MATR3</b>	165		59
<b>ATP1B1</b>	60		77
<b>MORF4L2</b>	186		3
<b>PRKACB</b>	137		36
<b>POLD4</b>	30		163
<b>SNTA1</b>	22		188
<b>SCG5</b>	116		8
<b>TCEAL1</b>	80		47
<b>DNAJC6</b>	176		63
<b>ZBTB11</b>	56		73
<b>DYNC1I1</b>	198		17
<b>ROM1</b>	1		150
<b>EPB41L3</b>	106		83
<b>CALM1</b>	180		43
<b>SORBS3</b>	21		147
<b>TNFSF12-13</b>	35		181
<b>PTS</b>	113		92
<b>RTN4</b>	127		21
<b>PREPL</b>	144		16
<b>PBXIP1</b>	17		191
<b>PPM1H</b>	169		75
<b>UBE3A</b>	173		94
<b>SS18L1</b>	183		88
<b>PBXIP1</b>	20		149
<b>APP</b>	13		139
<b>TTC38</b>	2		143
<b>SEPN1</b>	168		32
<b>NDFIP2</b>	26		157
<b>C10orf54</b>	24		167
<b>TP53I13</b>	87		80
<b>JAZF1</b>	163		76
<b>FAM174A</b>	195		117
<b>TCEAL7</b>	41		200
<b>LGI4</b>	48		72
<b>CHIC1</b>	158		67
<b>SLIT2</b>		8	70

**Table 19:** List of the most influential common genes obtained by driver analysis between the three probes for the APP gene in the AFFY-44 array. The number indicates the position that gene is found in its respective analysis.

After the expansion, it is clear that the second probe is highly dissimilar to the other two, leading to the addition of significant amounts of noise in the analysis. Using the first and third probes, it is possible to obtain a small number of common genes and the information contained is quite relevant. Between the using the AD hippocampus and APP as predictors

DNAJC6, which reliably shows as a major hub in multiple interactomes, appears to be a common factor and its roles in phosphatase activity and ATPase stimulation confirm previous predictions, as is PPM1H which has been discussed in a previous section. Between this test and the entorhinal matrix in AD the only similarity is once again DNAJC6. This is most likely an indication of the importance of that gene to the function of the brain, of not necessarily health maintenance. While it is unlikely that it can be used as a predictor for prognosis or target for therapy, dysregulations in this gene might lead to long term problems.

### 5.3.3 MAPT - Microtubule Associated Protein Tau

Similar to the issues with APP discussed above, the problems with attempting to find common ground based on MAPT with other questions is hampered by the fact that there are 6 probes for it in the AFFY-44 array and there are significant variations between them. Probes 1 and 4 appear similar, as do probes 2 and 5, with probe 3 having commonalities and differences with all of them and probe 6 being completely different. Attempts to consolidate the information contained in these probes by averaging their expression values and using that as a predictor, would result in a highly biased, most likely incorrect conjecture about the nature of the gene. The dissimilarity issue will be considered further in section 5.4.

<b>Most Influential genes for MAPT</b>						
<b>Gene</b>	<b>Probe 1 203928_x_ at</b>	<b>Probe 2 203929_s_ at</b>	<b>Probe 3 203930_s_ at</b>	<b>Probe 4 206401_s_ at</b>	<b>Probe 5 225379_a t</b>	<b>Probe 6 233117_a t</b>
<b>MAPT</b>	42	27			11	
<b>GNAO1</b>	48	4			9	
<b>KCNQ2</b>	51	3			4	
<b>GRIN1</b>		28	1		17	
<b>MAPT</b>		19		1	12	
<b>TMEM130</b>			6	13	27	
<b>CREG2</b>	3			9		
<b>CNTNAP5</b>	5			51		
<b>RBFOX1</b>	6			10		
<b>BASP1</b>	37			24		
<b>ARHGEF9</b>	38			43		
<b>KIF3C</b>	41			5		
<b>MAPT</b>	43			6		
<b>SYT1</b>	45			2		
<b>SCN3B</b>	47			41		
<b>GNAO1</b>		8			6	
<b>CDK5R1</b>		29			1	
<b>KALRN</b>		13			14	
<b>DLGAP2</b>		26			7	
<b>GRIN1</b>		50			26	
<b>MAST3</b>		15			2	
<b>CEP170B</b>		46			29	
<b>DNM1</b>		23			3	
<b>CALY</b>			22	8		
<b>ZNRF1</b>			2		16	
<b>PRKCE</b>			26		50	
<b>SCN2A</b>			5	47		
<b>GRIN2A</b>			7	20		

**Table 20:** List of the most influential common genes obtained by driver analysis between the six probes for the MAPT gene in the AFFY-44 array. The number indicates the position that gene is found in its respective analysis. Note the pattern of commonalities used to cluster the separate probes as well as the complete lack of commonalities for probe 6.

There also appears to be no significant overlap between the genes common across all MAPT probes (Table 20) and the genes identified by previous tests in either AD as a whole, or in specific brain regions. This is most likely an issue stemming from the aforementioned lack of similarity between the different probes. Since none of the probes presented had commonalities exceeding 50% with any other probe, it would be ill advised to consider the common genes are significant to the disease without further testing.



Moreover, this process has revealed a flaw in current AD marker discovery. The dissimilarity of probes present in all three of the genes identified by literature as the most relevant to AD, casts doubt on previous experiments using gene expression arrays as the only tools to predict markers and validate results. Thanks to advances in sequencing technologies and increase of available computing power, it is possible to expand the search to other datasets and validate the significance of previous findings.

## 5.4 Inter-comparison and Cross-Comparison with other datasets and technologies

### 5.4.1 Microarrays and RNA-seq

One of the major disadvantages of the methods and results used in the preceding series of tests is that they are based on the E-GEOD-48350 array. Even though it is a clear and comprehensive dataset, with both positive and negative controls, a large enough number of cases to be able to datamine specific subsets of the dataset, such as the brain regions and based on real patient data, it is still a collection of data obtained via DNA microarray. In the biological process, gene expression is further away than desired from the real outcome. This data does not provide any information that is altered, lost or otherwise not expressed due to mutations, epigenetic factors, lifestyle genetic changes or dysregulation at the transcription and translation steps.

To alleviate this issue, another dataset was selected to validate the previous results in the form of E-GEOD-84890 an RNA expression profiling and DNA methylation dataset for AD patients. The advantages of RNA sequencing include the ability to reduce the bias in detecting changes at the genetic level resulting from single nucleotide polymorphisms, indels, fusions of even the presence of novel transcripts of previously discovered genes as well as increasing the specificity and sensitivity of the methodology. The RNA data in E-GEOD-84890 include 97 RNA samples from the middle temporal gyrus of AD patients and 98 RNA samples from the middle temporal gyrus of age and sex matched cognitively normal controls. While more limited than the previously examined dataset, the increased granularity present in E-GEOD-84890 resulting from the ability of RNA-seq to provide more biologically relevant results, in spite of the decreased variance and increase in bias due to the focus on a single region, should be able to provide a robust platform to validate both the methodology and the results obtained thus far.

#### 5.4.2 RNA-seq Driver Analysis

The initial set of experiments using this dataset were based on a categorical stepwise analysis to determine the genes most likely to explain the variance between AD and cognitively normal controls in the middle temporal gyrus, followed by an interaction analysis to perform network inference on the resulting genes exclusively for AD and control cases. The resulting interactomes (not shown) were used to understand the nature of the interactions between genes before a driver analysis was used to obtain an unbiased list of the effect of genes on the network as a whole. The interaction analysis used the matrix parameters of 500 genes in 20 sets of 200 genes each.

AD Driver Analysis - 84890

Amount of Influence	Gene Symbol	Amount Influenced	Gene Symbol
-1958.219882	WASF1	-2083.646267	CDH23
-1771.708153	C5ORF27	-2073.839527	PPDPF
-1681.90543	UBE2W	-2008.11657	LOC641972
-1569.987205	TNIP1	-1950.653977	GRTP1
-1538.37216	ATXN10	-1857.454956	CXORF45
-1525.982519	KDSR	-1806.232575	C5ORF22
-1456.170588	PPM1B	-1782.58809	ACRV1
-1444.406897	ATP5L	-1762.870243	C18ORF1
-1443.206771	KRT222	-1729.529326	SPHKAP
-1440.776943	CAPRIN1	-1727.605793	LOC143543
-1398.87505	GABRA6	-1716.716444	ILK
-1393.557333	NCALD	-1656.52827	PCSK1
-1387.381056	MAP3K6	-1653.539556	CYP2C8
-1375.322974	PLEKHA9	-1637.153128	KCNJ10
-1362.456705	SEC61G	-1609.53879	GPN1
-1362.251365	CHMP4B	-1593.680342	PPP2R2C
-1353.581317	PRSS16	-1589.530603	BOLA3
-1333.016106	RLBP1L1	-1576.435682	LOC388481
-1324.533163	SEH1L	-1574.295104	MZF1
-1321.421495	DHDDS	-1568.208178	TBL2
-1318.454802	ATXN1L	-1557.275271	ADIPOR2
-1313.044019	PSMD1	-1552.214294	SYN1
-1293.863812	LOC100131541	-1552.112077	BOLA3
-1293.122535	ENO3	-1550.23252	FAM80B
-1292.391918	SYNJ1	-1545.676833	AGAP2
-1285.161931	WBP11	-1538.084895	ADCYAP1
-1278.640019	GTF2H3	-1533.413701	FANCF
-1277.803989	LOC642995	-1533.286163	F8A1
-1273.104627	C17ORF102	-1530.602362	DNALI1
-1271.807839	MUC1	-1510.801632	LOC649095
-1265.945699	KCTD7	-1502.116088	CCNH
-1264.079219	CNNM3	-1485.785739	STS
-1251.704961	MGC12760	-1481.622216	RHOQ
-1250.495855	C6ORF168	-1477.877914	PSD2
-1249.600678	FBXW4	-1474.397924	NSBP1
-1242.045164	RTKN2	-1465.563653	PRKAR1B
-1235.527989	B4GALT4	-1463.123966	NRN1
-1235.024006	SVOP	-1461.872398	VKORC1L1
-1234.320421	VGLL4	-1460.565362	UQCRHL
-1233.785913	DKFZP586I1420	-1459.488359	FOXJ1
-1228.796667	CCKBR	-1440.07732	FAM113A
-1227.661396	PLEKHH3	-1423.203493	UQCC
-1216.135735	LOC730173	-1422.665137	TMEM16C
-1209.316648	ZNF32	-1415.475921	CBLB
-1207.153644	FLJ35258	-1414.377083	DCUN1D5
-1206.517827	C12ORF11	-1397.246752	C6ORF225
-1205.192332	ZNF786	-1391.790025	INPP5D
-1201.566951	DISC1	-1391.034877	ATP5F1
-1201.226621	LOC100128781	-1388.760245	C16ORF53
-1201.122274	CHST6	-1382.625062	SST

**Table 21:** Driver analysis showing the top 50 most influential and influenced genes according to their impact on the network for AD in the E-GEOD-84890 dataset.

Even though the data presented originates from a different brain region than those explored earlier, the middle temporal gyrus has shown signs that it is significantly affected during AD (Galton, 2001). The results (Table 21) are expected to show significant similarities between the genes most dysregulated in other brain regions, but also expected to have key roles in the progression of the disease. Indeed, the most influential gene on the network is the heavily suppressed WASF1, a Wiskott-Aldrich syndrome protein related to the regulation of the actin cytoskeleton via downstream regulation of Rac, a Rho GTPase. Dysregulation of genes interacting with Rho GTPases has been a consistent feature of the results obtained thus far, and the pattern continues to a different dataset, in a separate brain region, at the transcription step all but ensuring the crucial role Rho GTPases play in AD. Meanwhile, UBE2W is a ubiquitin containing enzyme, dysregulation of which appears to be another consistent feature of AD, which has been theorised to play a key role in DNA repair. Moreover, it has been shown to monoubiquitinate the N-terminus of the TAU/MAPT substrate, linking it to AD. Similar to the results seen earlier, TNIP1 is a gene encoding for the TNFAIP3 Interacting Protein 1, which inhibits NFKB activation and TNF-induced NFKB-dependent gene expression. PPM1B has a similar role and belongs to the same family as the PPM1H gene discovered in the hippocampus of AD patients in E-GEOD-48350.

The genes most influenced in the predicted interactome on the other hand, include multiple GTPases such as GPN1 and AGAP2, as well as the RHOQ gene consistently present in previous tests, and ATP5F1 and ATP synthase. There appears to be a pattern for energy regulation genes in the targets of AD interactomes. FOXJ1 is another gene from a similar family as previously discussed hubs, specifically FOXN3, although its exact function is unknown. Finally, there are genes such as GRTP1 whose function is unknown but are suspected to encode for GTPase activating proteins. Attention should also be drawn to CDH23, a cadherin related gene encoding for calcium dependent cell-cell adhesion glycoproteins. While CDH23 has not been shown to impact AD, a recent methylation study by Lord and Cruchaga (2014) indicates that it may be a major epigenetic marker.

Healthy Driver Analysis - 84890

Amount of Influence	Gene Symbol	Amount Influenced	Gene Symbol
-1456.495782	CCNT1	-2063.323668	CYP2C8
-1412.178441	CAB39L	-2015.844058	PPDPF
-1345.960358	PSMD1	-1997.835086	SCG3
-1322.028315	LRP5	-1840.195571	ZNF565
-1311.106611	PSMD5	-1803.535868	CDK2AP1
-1268.627084	KDSR	-1797.388131	C6ORF168
-1247.204153	LOC143543	-1738.701272	CHMP4B
-1242.831606	LOC100128266	-1681.199377	NRN1
-1230.075309	CLPTM1	-1654.916256	SST
-1225.868251	KDSR	-1645.536377	TTC7B
-1225.751547	GPN1	-1635.881228	SPHKAP
-1201.317747	UBE2W	-1632.348351	LATS2
-1173.741212	WBP11	-1613.461585	KIAA0556
-1173.384916	SERF1A	-1611.063644	PCSK1
-1171.462712	CCDC102A	-1591.319621	ADCYAP1
-1166.776063	CHD7	-1588.65639	F8A1
-1161.471531	ITSN1	-1586.502806	PRKCG
-1149.832257	C15ORF24	-1546.674027	C19ORF30
-1120.849593	SLITRK1	-1532.422965	NEDD8
-1117.054661	FBXW4	-1528.639488	XPNPEP2
-1109.797022	NSBP1	-1519.002158	LOC100132839
-1106.243085	ZXDB	-1507.202325	SYN1
-1104.006132	SAMD11	-1501.372545	LOC649095
-1081.116572	MGC12760	-1500.160444	PHF16
-1079.902753	TMEM118	-1500.141354	ATP6V1E1
-1079.083135	ATXN10	-1486.882555	LOC402221
-1076.315685	SOSTDC1	-1470.883204	LOC642921
-1074.333758	SUB1	-1464.905126	NPTX2
-1070.333564	ENAH	-1453.383537	C16ORF53
-1060.611483	C12ORF11	-1453.286793	HS.303060
-1057.802023	GSDMD	-1452.104188	BNIP3
-1055.140557	RNF216	-1447.704631	BRE
-1052.172809	MLLT6	-1444.649884	CHCHD2
-1049.071821	TBC1D20	-1438.146632	SCG3
-1048.995355	LASS1	-1435.942757	LOC641972
-1046.867427	NOTCH1	-1435.870436	CTDSP2
-1044.166788	RAPGEF3	-1424.487757	FXR1
-1044.134222	KHK	-1401.704573	UHRF1
-1039.794465	FAM160A2	-1400.379505	HS.335413
-1037.49264	SCG3	-1399.221307	LOC649270
-1033.418515	CDC14A	-1396.822085	KCNC2
-1020.682617	PTRF	-1393.515347	C15ORF57
-1019.820879	SYNJ2BP	-1392.798912	MEIS3
-1017.476062	GLO1	-1389.247794	OLA1
-1014.910057	TUBB4Q	-1382.43849	SEC61G
-1011.52942	ITGB5	-1374.943869	PRSS16
-1011.184684	SLC35A2	-1374.033653	GPN1
-1004.771574	VIL2	-1349.773658	DCUN1D5
-1004.447735	NFKB1	-1349.15908	FAM127A
-992.3480245	C16ORF5	-1333.988156	PPP4R1

**Table 22:** Driver analysis showing the top 50 most influential and influenced genes according to their impact on the network for cognitively normal controls in the E-GEOD-84890 dataset.

Analysis of the healthy drivers (Table 22) however, presents a different picture to the one for AD. The greatest influence on the network is in the form of CCNT1, a member of the highly conserved cyclin C subfamily, encoding for a cyclin-dependant kinase, overexpression of which leads to tumour growth. PSMD1 encodes for a proteinase complex involved in ATP-dependent degradation of ubiquitinated proteins and is a crucial component in homeostasis by removing damaged or misfolded proteins, a crucial factor in the progression of AD as misfolded APP and dysregulation in tau leading to damaged neurons cannot be effectively cleared. Moreover, PSMD5 which is predicted further down the list, is another key component of the same complex. UBE2W, and likely other similar genes, appear to be conserved as they are found in analysis of AD as well as cognitively normal cohorts, in both tests for 84890, as well as the APP and Master analysis tests for 48350. There is a high probability that it is dysregulation in genes related to ubiquitination but not directly controlling it that are relevant to the progression of the disease. There also appear to be a large number of apoptosis and chromatin regulating genes which certainly converge with previous results relating to cognitively normal brains as clearance of debris, homeostasis, DNA repair and a fully functioning immune response are consistent factors in these samples, but found rarely, or highly dysregulated in AD regardless of brain region.

While most genes identified in this analysis can be used to explain the development and progression of AD by their absence, there are two that have been analysed previously in this thesis; tubulin and NFKB. TUBB4Q is a tubulin pseudogene so not directly relevant but was most likely identified by the algorithm due to its similarity with other tubulin genes such as those found previously. Meanwhile, NFKB1 encode for a subunit of the NFKB protein complex, whereas the dysregulated genes often found in the analysis of the AD cohorts, include genes encoding for NFKB inhibitors and are usually being suppressed, but have been found upregulated relative to tubulin, reinforcing the connection between the neuronal damage and formation of NFTs and the immune system.

The targets of these genes, and most affected genes in the network, are directly related to their respective sources. Genes such as CDK2AP1, a cyclin dependant kinase, NRN1, a neurotin expressed during the development of the nervous system and strongly associated with plasticity, TTC7B, which is crucial for phosphate synthesis and SYN1, a synapsin encoding for neuronal phosphoproteins and acts as a substrate for multiple kinases as well

as phosphorylation are fully expected to be highly regulated in a healthy system. Another gene previously seen in AD as well as cognitively normal studies is BRE, also known as BABAM2, and is thus most likely highly conserved, as well as GPN1 which is a GTPase.

#### 5.4.3 Comparison against known markers

Similar to section 5.3, it is possible to use the expression values of specific genes, such as known or suspected markers, the APOE4, APP and MAPT genes in this case, as predictors for a continuous stepwise analysis, thus obtaining a set of genes most likely to influence the expression of the predictor genes. Moreover, as the data is based on RNA-seq, unlike section 5.3, the results are significantly closer to the resulting proteins that cause the dysregulation in AD. Additionally, there are less transcripts of those genes in the RNA-seq than the microarray, reducing the variance between them.

*APOE4* – 84890

<b>Amount of Influence</b>	<b>Gene Symbol</b>	<b>Amount Influenced</b>	<b>Gene Symbol</b>
-55.3296	ATG5	-162.647	SLC4A2
-55.3286	MAP6	-138.502	PLXNB3
-54.8841	C10ORF88	-122.533	CD4
-54.7469	PDS5B	-116.642	MTSS1L
-54.0002	TOPBP1	-115.098	C9ORF16
-53.1253	LOC100125556	-109.236	TSC22D4
-52.8089	PTTG1	-104.852	PHF1
-51.8272	C6ORF115	-103.608	C1QTNF5
-51.4383	LOC100127952	-103.141	ZNF385A
-48.3982	CST3	-101.169	EMID1
-47.9805	DCLK2	-100.264	BCL7C
-45.8396	SLC3A2	-98.6892	PLEC1
-45.7144	CYB5R3	-98.6693	FLJ10357
-45.6531	AES	-97.2157	PACS2
-45.6307	MLEC	-95.3303	TNKS1BP1
-45.6301	ACADVL	-89.4024	HIP1R
-45.2363	AGXT2L2	-89.3585	JOSD2
-44.7907	CTDNEP1	-88.9531	EPHX1
-44.7278	BAT1	-86.8357	INO80E
-44.6286	AGPAT1	-85.891	S100A1
-44.5971	MED12	-85.0158	PTPN23
-44.4986	MLL4	-84.5998	PTPN6
-44.4604	XKR8	-82.5412	IRF3
-44.1857	HDGF	-80.9215	BAD
-43.6019	NDRG2	-79.7947	TMEM214
-43.4076	CSRP1	-79.6953	TENC1
-43.4023	TMEM179B	-79.5999	KIAA0195
-43.3812	LOC731096	-79.474	GPS2
-43.3712	GTF3C5	-79.4024	UNC5B
-43.2866	PLEKHB1	-78.55	AGTRAP
-43.2653	CDK2AP2	-77.9451	RENBP
-43.0004	PRPF3	-77.8405	MYH14
-42.9846	LOC729495	-77.668	C17ORF62
-42.9554	FAM108A2	-76.8407	G6PD
-42.9505	TMEM132A	-75.0923	BSG
-42.8277	CECR1	-74.9885	ARHGDI2
-42.775	BANF1	-73.3141	FLJ20489
-42.7094	ERCC1	-73.2281	SLC25A1
-42.6907	HS.397465	-72.2711	SH3BGR1
-42.6376	CYTSB	-71.3583	CDC42BPB
-42.5857	SHISA5	-71.2695	FXD1
-42.503	ABCA2	-70.0323	IFT140
-42.4798	TTYH1	-68.5697	ACIN1
-42.3122	GHDC	-68.387	ITPK1
-42.3033	PC	-67.8762	GNAI2
-42.1588	DDR1	-67.7181	OS9
-42.1348	TECR	-67.0586	LOC728908
-42.1182	AKNA	-66.3526	SDF4
-42.082	LOC644988	-66.2642	MAPK8IP1
-42.006	TAOK2	-65.81	SERF2

**Table 23:** Driver analysis showing the top 50 most influential and influenced genes according to their impact on the network based on the expression of the *APOE4* gene for AD patients in the E-GEOD-84890 dataset.



As there is only a single APOE4 probe in the consolidated Ensembl array, it is possible to examine the effect of APOE4 in the middle temporal gyrus in AD without attempting to find commonalities to account for false discovery. As a result, the most influential genes in the interactome (Table 23) include ATG5, which encodes for an autophagy related protein, which conjugates with other proteins such as ATG12, 7 and 10 as a E1-like activating enzyme in a ubiquitin-like conjugating system. This allows it to be involved in multiple cellular processes such as autophagic vesicle formation and mitochondrial quality control after oxidative damage, both of which are key factors in AD. Additional genes include PDS5B, a cohesion factor essential in chromosome segregation during mitosis, MAP6, a microtubule associated protein involved in stabilisation of the microtubules, AGPAT1, discussed previously, and AES, an amino-terminal enhancer of split, which is essential in neurogenesis during embryonic development.

Regarding target genes, they include PLXNB3, variants of which have been previously discovered, CD4, a T-cell surface glycoprotein regulating T-cell activation, multiple zinc finger and transmembrane proteins as well as genes regulating cytoskeleton activity and actin binding, such as MTSS1L. The presence of S100A1 is especially noteworthy; being a member of the S100 family of calcium binding proteins, it is heavily influential in the innate immune response, and its presence in the most downregulated genes in AD could lead to malfunctions in the immune response as well as homeostasis attempting to decrease or stop the chronic immune response supported by the neuroinflammation theory.

As APOE4 is responsible for familial AD, these results are not sufficient to explain why the disease develops in healthy adults but provides crucial information as to the mechanisms involved in the progression of the disease.

Influencer Gene	Influencer Gene	Influenced Gene	Influenced Gene
Probe 1	Probe 2	Probe 1	Probe 2
FAHD2A	ATG5	LUM	HERC3
ARFGAP2	MAP6	LOC653879	LOC100132774
NBAS	C10ORF88	DOCK9	ZNF84
REEP2	PDS5B	ITM2B	SLBP
HS.336643	TOPBP1	ARCN1	SLC25A4
AK2P2	LOC100125556	TACC1	ATRN
YIPF2	PTTG1	TIA1	ATP6AP2
GNAZ	C6ORF115	PIGY	DENND4C
LOC728791	LOC100127952	RNASE4	SPNS1
LOC100131850	CST3	DAZAP2	PAFAH1B1
ING4	DCLK2	C13ORF23	FLJ12078
YAF2	SLC3A2	LRRC33	HERC1
LOC440345	CYB5R3	BTF3	CXORF45
CSDE1	AES	POLR2D	FYTTD1
FAM134C	MLEC	SAMSN1	LOC100131866
B4GALT2	ACADV1	LMBRD1	C5ORF25
SIPA1L1	AGXT2L2	GP1BA	CEP164
GP1BA	CTDNEP1	CD46	KIAA1279
DAZAP2	BAT1	ZAK	RRAGA
CAP1	AGPAT1	FAM134C	BRMS1
RNF115	MED12	MR1	SLC25A4
P2RY12	MLL4	UTRN	GLRB
TACC1	XKR8	CD74	GNAS
ST6GAL1	HDGF	SIPA1L1	UQCC
DDX5	NDRG2	MBD1	AP1G1
KIAA0196	CSRP1	DDX5	PARP10
FAM91A1	TMEM179B	KIAA0196	LOC646786
SSR1	LOC731096	FCER1G	CCNDBP1
MAPK1	GTF3C5	HS.336643	RNF160
DDX3X	PLEKHB1	RAB27A	TMEM14A
POLR2D	CDK2AP2	EIF4G2	OCIAD1
HIF1A	PRPF3	MAPK1	ACTR10
LOC653879	LOC729495	SCAMP1	ITFG1
PICALM	FAM108A2	FAM96A	PRNP
MAT2B	TMEM132A	JAK1	TSPAN13
STT3B	CECR1	C14ORF149	FUK
APEX1	BANF1	CRIM1	EIF4H
GPR177	ERCC1	APP	HINT1
FAM96A	HS.397465	CSNK1A1	CSNK2A1
IREB2	CYTSB	RIOK3	CSRNP2
REV3L	SHISA5	B4GALT2	THYN1
LOC440595	ABCA2	C3	APOO
C14ORF135	TTYH1	LOC440595	EIF1B
KDEL2	GHDC	UEVLD	CAPRIN2
ARPC3	PC	YAF2	STAM
RB1	DDR1	LOC440345	ARAF
RRM2B	TECR	SUMO2	NEK8
ARF4	AKNA	CAV2	C9ORF130
TIA1	LOC644988	CCT7	WDR23
FCER1G	TAOK2	RNF115	HECTD1

**Table 24:** Driver analysis showing the top 50 most influential and influenced genes according to their impact on the network based on the expression of the APP gene for AD patients in the E-GEOD-84890 dataset.

Even though there is almost no overlap between the two APP probes there is a great degree of similarity between the functions of the predicted genes (Table 24). Gene ontology reveals that binding and catalytic activity are among the two most common molecular functions of these most influential genes, with nucleic acid binding and transcription factors being common across their resulting proteins. Moreover, Probe 2 shows significant commonalities with previous studies, including ATG5, AGPAT2, MAP6, CTDNEP1 and AES, which have also been seen in the APOE4 analysis, verifying that they share similar progression paths, even if the source of the disease differs. The most influenced genes also include a selection of biologically relevant results, including receptors and transporters for their sources as well as nucleic acid binding, which is consistent with previous predictions. Curiously, the largest category of resulting protein products is enzyme modulators, specifically G-protein modulators. This complex signalling system is regulated by the binding and hydrolysis of GTP, heavily dysregulated in AD, phosphorylation of which can alter the duration and intensity of signals, which explains the presence of so many kinases.

However, most of the results obtained through this test have not been discovered in previous studies. This may be an indicator that certain regions of the brain, such as the hippocampus, contain a large number of genes that are significantly affected during AD, which coincides with current literature, but other regions may be greater influencers and show little to no signs of neurodegeneration but drive it instead.

MAPT - 84890

Influencer Gene	Influencer Gene	Influencer Gene	Influencer Gene
Probe 1	Probe 2	Probe 3	Probe 4
PKM2	ZFP1	NT5M	ITGB5
UCKL1	LOC652140	LOC643990	SS18
PGD	FAM179B	LOC100101121	FAM129B
ATP5D	CXCL17	ANKRD52	YES1
MLF2	HAAO	LOC440804	LRRC32
TRPC4AP	ADM2	LOC642278	FGR
ZMYM3	EPYC	STK16	LAPTM4A
GPX4	DSP	RNU105B	SMARCA5
SMARCB1	LOC653689	PCID2	OLFM1
MAP1S	C19ORF12	C19ORF24	RUNDC3A
EPB41L1	LOC642740	LOC649864	SYN1
NCDN	RNF213	FAM55A	NPTX1
UBE1	CCDC56	FGFR1	DOC2A
UBA1	LGALS12	SLC2A11	LRRTM1
C19ORF29	LOC653073	LGALS12	SFRS14
APLP2	PROKR2	HSD11B1	LOC286411
DBP	GLIPR1L2	LOC100134359	TAGLN3
SNRNPB	HS.566764	NAPA	RBP4
BAI2	LRP2BP	LOC100128140	SYN1
STX1A	OR1L8	HS.403584	PAK1
GATAD2B	GPR175	LOC100133999	ERCC3
LPPR2	PLA2G4E	LOC728992	ROBO2
ARHGDI3	C1ORF38	LOC100128310	LOC730744
HYOU1	LOC442329	CDC42EP1	CLTB
C2CD2L	LOC646562	CCDC80	CDH13
YPEL3	CHRNE	LOC652837	SCN2B
AP2A1	LOC641772	LOC644079	LOC652900
GAS7	UBE4A	HS.557356	CRSP2
COPE	HS.566008	ATP1A3	HS.553187
DHX30	LYPD3	SCARNA17	KALRN
C12ORF53	ALPPL2	LOC100130358	GLS2
NRGN	ACPT	CRB2	FRMPD4
HDGF2	PLCB1	MTUS2	MARCH11
ARHGEF4	CHD4	LOC149351	RIMBP2
GDI1	VNN1	RAB37	CDH8
ACTB	SLC22A7	HS.290834	SULT4A1
PKD1	HS.537603	LOC440105	DDX24
CAMK1	MAPT	HS.545462	RNF41
TUBG1	KRTAP8-1	PEBP4	SYNGR3
NPDC1	AQP7	MYH16	PRKCE
MLL	FRS2	SERPINA2	CHRM1
DGCR6	TBX6	FCGR3A	C6ORF168
WNK2	GUCY2E	CLK2P	PABPC1L2B
BAT2L	LOC646012	LOC389816	SEPT3
BCL7B	OAS1	LCE2D	UBE2E2
SPTAN1	LOC645183	LOC653270	NPTXR
CEND1	FAM155B	SNORD4B	ARHGEF7
CORO2B	LOC641819	LOC728344	ORC5L
MOGS	C9ORF62	TRIM31	SLC9A6
C6ORF1	PIB5PA	C9ORF16	CREG2

**Table 25:** Driver analysis showing the top 50 most influential genes according to their impact on the network based on the expression of the MAPT gene for AD patients in the E-GEOD-84890 dataset.

The degree of divergence between the MAPT probes (Table 25) is significant, especially for probes 2 and 3. Probe 1 and 4, although having no common genes however, show significant similarities in their resulting protein products. Specifically, they appear to have the same degree of proteins related to nucleic acid binding, hydrolases, transferases and cytoskeletal proteins. Probes 2 and 3 on the other hand, are quite dissimilar to each other as well as probes 1 and 4. While hydrolases are a common feature between all datasets, the 2<sup>nd</sup> and 3<sup>rd</sup> probes show significantly less variety in their protein products, molecular functions and pathways. This results in the predicted genes being examined, but not considered as the most biologically relevant.

The most influential genes on the network include PKM2, a pyruvate kinase involved in glycolysis that catalyses the transfer of the phosphoryl group from phosphoenolpyruvate to ADP, generating ATP and pyruvate. PDG (Phosphogluconate Dehydrogenase) catalyses the oxidative decarboxylation of phosphogluconate and UCKL1, a uridine kinase, catalyze the phosphorylation of uridine to uridine monophosphate. It is clear that the similarities in function between these genes explains their presence in the drivers, even though they have no connection to the disease. UBE1 and UBA1 however, catalyze the first step in ubiquitin conjugation, which in turn marks cellular proteins for degradation through the ubiquitin-proteasome system. As seen in previous tests, this is linked with both ubiquitin conjugation as well as debris clearance. Other genes such as tubulin gamma, a key microtubule component, SEPT3, a member of the septin GTPase family and UBE2E2 which accepts the ubiquitin of the E1 complex have been discussed previously and their relevance to AD is being verified by the consistency of their presence.

Influenced Gene	Influenced Gene	Influenced Gene	Influenced Gene
Probe 1	Probe 2	Probe 3	Probe 4
LOC100132491	ZSWIM4	HS.505364	CALY
CPNE5	LOC442711	L3MBTL	SYN1
MAPT	LOC391142	ADCY3	NPTX1
C9ORF16	SSX5	LOC646569	RUNDC3A
RFNG	MAPK15	CNTNAP5	SEPT3
LOC100134734	LOC653635	BOLL	ERCC3
LOC100134530	MED22	HS.560896	OLFM1
PRKCSH	MIR1237	OSMR	SFRS14
BAIAP2	SGCA	VPS37A	ICAM5
KIAA0652	LOC652045	PCDHB9	ORC5L
RABGGTA	KLK10	S100PBP	NPTXR
PLXNB3	HS.544069	NEIL3	CHN1
RNF25	VARS	CLK2P	SYN1
DGKA	SDC3	C11ORF47	ROBO2
MEIS3	SIGLEC5	LOC649768	LOC652900
CCDC124	LOC648548	LOC649987	ELMO1
C16ORF67	GNAT1	LOC648581	DOC2A
MAZ	HS.149244	WDR5B	INA
IGSF8	RAMP2	C9ORF16	SLC12A5
DOC2A	PSENE1	TUBA1B	CPEB1
AP1B1	NCF4	INHBA	USP11
TAF6	ARPC2	DENND4A	KALRN
CPLX2	EPHA8	HS.518426	RTN3
RNF208	LOC644686	MAPT	PAK1
MADD	BMP1	LOC653270	LOC730744
PKD1	DPCR1	CCDC124	CDH8
LOC390298	GDPD4	ZNF322B	RND1
PPP2R5B	LOC641819	GPC1	RIMBP2
MACROD1	LYPD3	TTC39C	CCK
LOC645937	FLJ10357	ATP6V0A4	KIAA0513
GAMT	CD3EAP	PIK3C3	C1ORF128
APLP1	OR8G2	TANK	AK5
CAMTA2	TRIML1	LOC440786	TAGLN3
DEAF1	NHLH1	HS.555512	VSTM2B
ANKRD24	MIR1224	LOC727789	ATP6V1G2
C19ORF60	PCGF3	LGALS12	VSTM2B
LOC729495	C1QTNF6	ATP1A3	ADCY1
LOC100133673	EIF4EBP2	ANG	CYP2C8
LOC91316	TMEM102	IRF2BP1	ASNS
PCIF1	SULT2B1	LOC440804	CORO2B
TBC1D3C	RNF5	GPR141	CHRM1
MED16	HIVEP3	LOC388117	RTN1
ZNF574	ARSD	LOC440280	CNTNAP1
CDK5RAP3	SLC16A2	AKT1S1	PRKCE
MGRN1	LOC653073	CTNNA1	LOC286411
TSC2	LOC644086	LOC100128392	ITGB5
LOC729021	ACPT	NTN1	ARL3
OSBPL7	LOC643665	C9ORF95	RBP4
C21ORF56	C9ORF62	NT5M	CLTB
STRN4	RUNX2	LOC649864	LOC345630

**Table 26:** Driver analysis showing the top 50 most influenced genes according to their impact on the network based on the expression of the MAPT gene for AD patients in the E-GEOD-84890 dataset.

The most influenced genes targeted in the interactome (Table 26) show little commonalities between the four probes, but with cytoskeletal, nucleic acid binding proteins, transferases and hydrolases still making a significant percentage of the predicted protein products. It is worth noting that unlike the source genes, the first probe shows less variety in protein classes predicted, but that is mostly due to the large number of unknown or undiscovered predicted genes. As the algorithm is not weighted to better predict known genes such as other validation software such as Metacore or methods such as Nanostring, it relies entirely on the expression value of each gene probe present in the dataset for each patient.

Even accounting for these genes however, there are significant results in the most affected genes, such as MAPK15, a mitogen-activated protein kinase which has been shown to phosphorylate MBP (Myelin Basic Protein), a major component of the myelin sheath and present in the immune system. MAPK15 itself, is a key feature of the TNF-signalling pathway, other genes related to which are common features in previous results, as well as the NFKB pathway. Moreover, less studied genes such as NPTX1, a neuronal pentraxin suspected of mediating the uptake of synaptic material during synaptic remodelling, or NTN1, a netrin, part of a family of laminin secreted proteins, which controls guidance in the central nervous system by causing axons to attract or repel each other as well as regulating apoptosis, and other genes involved in the base mechanisms of the nervous system could provide insight in the progression of the disease if studied further.

## 5.5 Conclusion

While the expansion of the standard methodology has allowed us to obtain ever more complex and biologically relevant answers to the questions on the progress and development of AD, it is increasingly evident that the current avenues of biomarker discovery are limiting. By assuming possible disproportionate significance of certain genes such as the APOE4, APP and MAPT analysed in section 5.4, the analysis drifts further from the disease and closer to the interaction between these genes and their peripheral interactions. Moreover, by increasing the bias to this degree it is possible to infer direct interactions but not the lack thereof. If a key gene that needs to stay suppressed to prevent overexpression is dysregulated due to no longer being downregulated by the system, but in healthy individuals the expression is low, it is disproportionately hard to

showcase the fact by directly analysing the gene using machine learning or more traditional techniques. Additionally, it is becoming increasingly evident that AD is a highly complex disease caused by dysregulation in multiple systems, not simply a few genes or proteins and should be studied as such.



## Chapter 6: Conclusions

### 6.1 Novel Methodology

As indicated multiple times in this thesis, there is a fundamental flaw in most methodologies currently in use for biomarker discovery; the bias inherent in a focused question. This “error” is independent from the validity of the question, the nature of the dataset or the quality of the results obtained. It is entirely possible to reach a correct conclusion simply by enforcing good laboratory practice, rigorous planning and meticulous analysis of available avenues of thought even with that increased bias. However, as the complexity of the conditions and diseases that need to be understood and cured increase, and as we move towards the era of personalised medicine, broad generalisations as well as too focused approaches show increasing potential to hurt progress rather than aiding it. Diseases such as cancer now must consider the entire tumour micro- and macroenvironment, whereas diabetes research shows increasing focus in the nutritional systems. The issue with AD is the lack of understanding in which systems are the most influential in the disease as a whole and while the current hypotheses suggest that the immune system is most likely to provide the necessary answers, the truth resists simplicity. Moreover, even if the perfect disease biomarkers are discovered and the correct drugs synthesised for each condition, it is almost guaranteed that they will not work for every patient due to up- or downstream dysregulations in peripheral genes only marginally connected to the disease and disrupting entire systems as a result. The goal of this methodology was to use systems biology alongside machine learning to attain a greater understanding of not just the genes directly related to AD, but of the dysregulation in the brain’s systems and provide further, more focused avenues of research.

#### 6.1.2 Predicted biomarkers

Regarding predicted biomarkers, the most promising marker appears to be the tubulin superfamily. Genes encoding for multiple beta-tubulin classes are a persistent feature of multiple AD interactomes and significantly differentiated from healthy ones by being more central and more connected in their respected networks. Additionally, there were a few examples of alpha- and gamma-tubulins and all of these genes appeared in multiple brain regions, across multiple datasets and through a large variety of tests, which can be useful in explaining the genetic dysregulations that lead to the hyperphosphorylation of the tau protein and formation of neurofibrillary tangles. Moreover, they are consistently

involved in interaction with NF $\kappa$ B inhibitors and considering the range of processes NF $\kappa$ B is involved in as well as their criticality in the normal function of cells, this is a promising avenue for research. Furthermore, Rho GTPases and their pivotal role in the regulation of actin cytoskeleton and microtubule dynamics (Etienne-Manneville and Hall, 2002) have been shown multiple times to be dysregulated in AD, which provides further support for the tubulin hypothesis. Finally, while the structure and functions of ubiquitin have been studied in depth (Pickart and Eddins, 2004) and the role of the ubiquitin-proteasome system has been investigated in AD (Hong *et al*, 2014, Upadhyaya and Hegde, 2007) it has been mostly in the context of the amyloid cascade hypothesis and was targeted due to its pivotal role in protein degradation (Gong *et al*, 2016). However, the ubiquitin group and related genes are highly conserved and have been shown to be equally active in AD as well as cognitively normal individuals. Thus, the systems affected by and affecting these genes need to be understood. For instance, possibilities include a fully functioning protein degradation mechanism that is unable to cope with accelerated deposition of amyloid instead of being unable to clear it, the system being applied at a different region or with different priorities or upstream downregulations cause it to chronically underperform.

It is clear that the suggested predicted markers are not individual genes with clearly defined, singular roles but families and superfamilies regulating a large number of cellular processes and are involved in multiple pathways. Considering the sheer complexity of the brain and the variety of personal responses to AD, it would be unwise to focus on predicting a single biomarker or even a small panel of them. Although this may arise in the future, at our current level of understanding and technology it is paramount to focus on the dysregulation of the systems involved in AD by understanding their normal function and comparing them differentially. The machine learning approach to systems biology is ideal for this task as it is highly cost efficient, reproducible and fast, and these features also make it a likely candidate for use in the clinical field.

## 6.2 Quality of Results

The novelty of the methodology, resulting from the unique combination of the non-parametric hypothesis free approach combined with and ANN stepwise and network inference analysis, is unquestionable. However, it is crucial to note that the results must

be validated if this approach is to be expanded and find use in the clinical setting. So, how can we verify the quality and validity of the results? The biggest factor is biological significance, and the easiest way to verify it is through gene ontology. While the causes of AD are still largely unknown, there has been enough work done on the subject to provide us with a wide range of data on genes and their function as well as their involvement in pathways related to the disease. Moreover, there have been multiple predicted pairwise gene interactions in this study that support current finding as explained in chapter 5. However, biological significance is a rather nebulous term. The goal of this study was to identify possible biomarkers for AD and create a unified in silico methodology to accurately, quickly and efficiently reduce the variance present in complex datasets without increasing the bias. The final step of this analysis would be wet lab validation of the results, which is highly recommended and there are multiple established methods to achieve that, but is beyond the scope of this study.

Of course, the consistency of the results obtained is another indicator of quality. Due to the nature of the data, the probability that the results will diverge, if a standard distribution is assumed or the sampling isn't truly random, is quite high. The role of the algorithm is to perform these tasks reliably and repeat them until convergence is reached. If the quality of the dataset is sufficient, the ANN can recognise patterns within the data based on the questions provided. As seen in the results presented, there is a great degree of consistency between similar questions, especially ones regarding variance between AD and cognitively normal individuals. If the algorithm wasn't good, enough, the data of low quality and the results not significant, the patterns identified would change and the results would be inconsistent. It is crucial to note that the algorithm isn't predisposed towards specific results; experiments performed on different conditions including AD, multiple cancer types, diabetes, tuberculosis and others have been internally consistent and distinct from each other.

In conclusion, the results obtained are of high quality and consistency, and can be used to further advance AD research in silico by further expansion of the proposed methodology as well as in vitro to examine the effects of specific biomarkers. The results can then be fed back through the system for validation before moving to in vivo testing.

### 6.3 Hypothesis Free Approach Evaluation

So, what were the benefits of the non-parametric hypothesis free approach and was it worth it, or would a more focused approach using a null hypothesis have resulted in an increase in quality? In order to fully understand the implication of this question, it is paramount to split it into its components, the non-parametric and the hypothesis-free approach, evaluate each and consider whether together, they are more than the sum of their parts.

The non-parametric approach allows the user to avoid specifying a hypothesis related to any distributions within the data but can also be used when the user want to avoid specifying the structure of the model used. The first part is incredibly useful for the analysis of biological data, and especially when applied to patient data. Considering the genetic variance in any given population it would be crippling to assume a standard distribution of highly sensitive and dysregulated genes in the diseased members. Moreover, the healthy cohorts tend to be rather nebulous; it is borderline impossible to specify at the genetic level what a “healthy” or “normal” human is, as they might suffer from other conditions, have silent mutation that help or hinder, making it harder to generate a panel of highly differential genes against patients. The second part, avoiding a rigid model structure, further enhances the power of the technique as the approaches used can be optimised for each step, even if it means changing essential parameters and avoid over- or underfitting for the algorithm. Additionally, the quality of the results obtained this way are significantly higher, as it is possible to avoid issues resulting from rigid parameterisation, although great care should be taken when this is attempted to allow for cross comparison of the results. Non-parametric approached have been commonly used in a wide variety of fields, especially ones that rely on population studies such as economics and biology, and include numerous methods such as bootstraps, logrank tests and Kaplan-Meier survival analyses.

The second component of the approach, hypothesis free, refers not to any particular methodology used, but the study as a whole. This is trickier to justify as a null hypothesis has formed the basis for most biological studies with great success, and indeed we are not advocating the replacement of the null hypothesis in biological data analysis. The issue arises when a study starts with a very specific null hypothesis. As discussed and shown

multiple times in this thesis, when studying complex diseases such as AD, the need to understand and eventually manipulate the systems involved in the disease far exceeds the potential benefits of proving or disproving the influence of one to a small panel of genes. Considering the magnitude of this task, as the sheer number of genes involved in essential systems, as well the unknown number aberrant genes that drive dysregulation, the null hypothesis should arise after rigorous testing to reduce the number of possibilities after considering all possible options. As this is not a task that is practical for direct experimentation, machine learning can be used to make possible what in the past was considered preposterous and resulted in the bias inherent in the null hypothesis being the only possible way to study complex conditions.

In conclusion, the non-parametric nature of the approach allows us to optimise out tools and correct for shortfalls in the data as well as the algorithm and the hypothesis free approach drastically reduces the bias inherent in most studies, without reducing the quality of the results. After initial testing a null hypothesis can be reached, and the results analysed in a traditional manner. As a result, this approach can be used as a powerful preselection strategy that can only add to the quality, variance and validity of the results. Combined with the cost-effectiveness and speed of the ever-improving ANN algorithm used, it is already revolutionising how clinical data affects biological research.

#### 6.4 Implications for AD

Throughout this study, it was made evident that the nature of AD and its poor characterisation is proving to be a significant challenge in trying to cure it. As shown in chapters 4 and 5 there doesn't seem to be a single gene, or even a sufficiently small group of genes that can readily explain and characterise the disease. This is further supported by most literature and clinical trials as no approach has been successful. However, in this study, as the scope and granularity of the questions increased, certain patterns started becoming evident. It is clear that not single genes, but entire interconnected groups are responsible for the dysregulations leading to AD as discussed in section 6.1.2.

However, one of the key findings of this study and the one with the largest implications for the field of AD research has not been the patterns of dysregulated genes, but the patterns of dysregulations in entire pathways caused by said genes. These patterns could

potentially explain the variance in gene expression levels, the effect this variance has on all elements of critical pathways connected to the disease and traced back to understand the dysregulations that eventually lead to the development of AD. The final step of this process would be to computationally reconstruct all possible pathways related to AD step by step and analyse them using the methods described in this thesis. Moreover, further methods will need to be developed to capture a greater amount of information and interrogate the data to achieve the desired results. Some of these possibilities are discussed in section 6.5.

## 6.5 Future expansion

While the methods used in this study have proven to be robust, there is always room for improvement. In this section multiple potential avenues for further expansion will be presented and evaluated. Some of them are theoretical, but others have already been trailed, although the results were deemed to be beyond the scope of this study as they were proven to be very complex and extensive.

### 6.5.1 Single cell sequencing

Single cell sequencing is one of the most advanced methods available among modern NGS techniques and it allows the user to obtain full sequencing data for a specific cell. This has the potential to truly revolutionise AD research, as one of the major problems is the lack of information as to the alterations at the genetic level in different brain regions. Furthermore, it is now a widely accepted fact that AD start developing up to twenty years before the first symptoms starts becoming visible, and then it affects people at wildly different speed. By regular sampling of neuronal cells, it is possible to track changes in gene expression and RNA levels over time and compare them to similar results from cognitively normal controls, providing insight in the development of the disease as well as the causes.

Additionally, this information can be used in conjunction with the ANN methodology outlined in this study. The temporal information can be used as a predictor by itself providing the algorithm with the following question: for a patient with AD, based on the expression levels for each gene in the array (DNA or RNA), which genes were responsible for driving dysregulation over time? The ANN can then be used to identify these genes,

which can in turn be used to perform network inference as well as any of the tests outlined in this study. The expected results of such a study are quite likely to provide novel biomarkers and targets for therapy simply due to the accuracy afforded by the high quality of the data produced by single cell sequencing. The challenges presented by this approach are the substantial cost and availability of long term patient data, and while not insurmountable, are rather significant.

### 6.5.2 Top 10 approach

A major limitation of analysing gene-gene interactions obtained via network inference is the bias inherent in creating a force directed network of a subset of interactions which are themselves the result of a previous analysis and a subset of a complete array. The outcome of this approach is inevitably that the genes the user is attempting to study may not be supported as hubs by network statistics such as centrality. Moreover, as discussed in this thesis, the genes that are known to be dysregulated in a disease, and even be key factors to its progression or culmination, may not be the drivers for it. Instead, the regulators for said genes are quite likely to be the deciding factors. A way to increase the variance and decrease the bias of such analyses would be to exponentially expand the interactome in a non-parametric manner.

During the preselection process, performed using the Stepwise algorithm, it is possible to select the most differentially expressed genes and use them as predictors for a subsequent series of tests. By selecting the top 10 most differentially expressed genes in a given condition in a categorical question, or the following 10 most differential genes in a continuous question, it is possible to use their results in conjunction with a commonality analysis and network inference to create multiple overlapping interactomes and driver analyses. The genes common across these 10 questions are going to be either highly mechanistic or crucial to the variance presented by the question. It should also be possible to overlay the interactomes in order to achieve convergence and obtain the most biologically relevant set of genes that can then be used for prognosis and therapy.

Preliminary results in AD are promising and it can be safely stated that this is an approach worth exploring in greater detail. Furthermore, it is a highly modular approach with no predefined cut-off points, allowing the user to select the level of expansion desired. Care

should be taken with the fact that the approach is exponential in nature and increasing the level of complexity by two or more would require more advanced techniques to ensure quality and interpretability in the results.

### 6.5.3 Commonalities between predictors in a panel

In sections 5.3 and 5.4.3 current known markers for AD, APOE4 driver of familial AD, MAPT, the gene encoding for the tau protein and APP, the gene encoding for the amyloid precursor protein, were analysed and commonalities between previous tests were analysed to look for a connection between this study's results and published literature. Sadly, no significant overlap was discovered, but that was not entirely unexpected. APOE4 is a driver for a distinct variant of AD, and while the mechanisms for progressions are expected to be similar, which was proven, significant differentiation was expected. Moreover, tau and amyloid beta are found in large quantities in the majority of AD patients, but not all, and it is likely that they are the most outward symptom of the disease as well as contributing to it, in addition to the fact that there is a struggle to prove a connection between the two. Additionally, the presence of multiple valid but distinct probes has made detailed analysis an inaccurate and laborious task.

To improve this methodology, it is paramount to first identify and construct a truly representative panel of genes that can be used to explain the variance in certain question within AD. For instance, a panel to explain the abnormal phosphorylation of Tau, rather than high level of the protein, a panel related to A $\beta$  deposition and problems with clearance etc. Then the genes in the panel can be analysed using the commonality method and added or removed to further increase the power of the approach. The process is likely to take some time as each gene will need to be analysed both in silico and in vitro and the panel is liable to completely change as new information is added.

### 6.5.4 Epigenetics

In a disease such as AD the impact of lifestyle genes is likely to be significant but very little research has been done of the subject and what little there is, hasn't produced usable results. Since AD develops over a long period of time, and progresses at varying rates between individuals, changes in their genetic makeup are not likely to explain this variance by themselves. Further insight is urgently needed to understand the impact of



such genes and how they can be controlled by lifestyle choices instead of medication. The answer to that is to study the changes in epigenetic triggers in AD.

To that effect, DNA methylation datasets can be used to provide the information required. In recent years methylation datasets about AD have started being used, but their numbers are low and quality inconsistent. The expected results of such an analysis are unknown, and while the ANN has been tested on methylation data and has proven that it can analyse it, the field of AD epigenetics is still young and, although promising, further research is required.

### 6.5.5 Complete gene analysis

The final expansion possible using current technology and a way to remove any flaws in the methodology would be the analysis of all the steps involved in the DNA to protein process. The ideal dataset for this analysis would include patient data from at least 100 patients and 100 controls. Cells from each brain region would be analysed using a technique similar to the single cell sequencing expansion described earlier. Moreover, gene expression, RNA and protein levels would be recorded. That way it would be possible to know where and when a crucial change occurs. Changes between gene expression and RNA would uncover transcription errors, while comparisons between RNA and protein would do the same for translation. This would also allow for better targeted therapy as we wouldn't have to rely on specific treatments. For instance, if one source of dysregulation is not present at the gene level but at the protein level, the protein itself can be targeted, but if the source is genetic mutation, treating the symptoms, in this case the protein, would be inferior to fixing the source. Of course, cost is a limiting factor, but as the required technology becomes affordable, there is a strong possibility that this analysis would output illuminating, high quality results.

## 6.6 Cracking the Algorithm

The last concern that needs to be addressed is the most common criticism to this approach; the “black box” nature of the ANN algorithm. This criticism is based on the fact that while ANNs can be used to solve a multitude of problems, it is impossible to gain insight on those problems by studying the structure of the ANN. Furthermore, as complexity increases and more parameters are added to training, it becomes harder to understand why

the algorithm reached the conclusion it did, making it a “black box”. ANNs however, are no longer a novel and untested technology. They have been in use for years and have been providing increasingly high quality results in a wide variety of fields from data mining and pattern recognition, to disease diagnosis, finance to detect credit card fraud and even cybersecurity with some success.

It should be noted that arguments about the specifics and parameterisation in each ANN are still valid and the lack of universality in both the user base and the wildly varying quality and nature of the data analysed, results in a constant need for reevaluation of the standard acceptable parameters. For instance, in mathematics increasing the size of the hidden layer above two nodes is considered pointless as it adds time without increasing the quality of the results, but in biology going as high as five nodes allows the algorithm to cope with fuzzy and incomplete data.

Nevertheless, while understanding the process is not essential to interpret the results, it is possible to obtain the details on the algorithm’s function and decision process. By collecting information on the algorithm’s performance throughout testing, such as a breakdown of the bootstraps and the intermediate results of each loop and step, it is possible to verify the validity of the results. While understanding the decision process is still unlikely, ensuring that the algorithm reached a correct conclusion based on the information it was given is entirely possible.

While there are still challenges to be overcome, such as improving readability of the results and further reducing training time, ANNs are a very powerful tool. As technology progresses, so do the applications of ANNs along with their speed and efficiency. Their opaque nature can even be considered an advantage as it allows non-specialists to use the algorithm and draw meaningful conclusions for their respective fields, thus transitioning from an engineering resource to a scientific one. Furthermore, it is possible to use other software and algorithms to rationalise the results provided by the ANN as seen in this thesis with simple tables to rank genes and Cytoscape to visualise interactomes. Thus the “black box” criticism should not be a reason to avoid using ANNs, but rather a justification to develop techniques to obtain the maximum amount of information possible from such a powerful tool.

## References

1. ABDEL-FATAH TM, AGARWAL D, LIU DX, RUSSELL R, RUEDA OM, LIU K, XU B, MOSELEY PM, GREEN AR, POCKLEY AG, **2016**, SPAG5 as a prognostic biomarker and chemotherapy sensitivity predictor in breast cancer: a retrospective, integrated genomic, transcriptomic, and protein analysis, *The Lancet Oncology*, 17, 1004-1018
2. AGARWAL D, **2017**, Systems biology approaches for the identification of molecular characteristics of proliferation in breast cancer, Nottingham Trent University, PhD Thesis
3. AGATONOVIC-KUSTRIN S, BERESFORD R, **2000**, Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research, *Journal of Pharmaceutical and Biomedical Analysis*, 22, 717-727
4. ALBERT A, **2007**, Network Inference, Analysis, and Modeling in Systems Biology, *The Plant Cell*, 19(11): 3327–3338
5. ALYASS A, TURCOTTE M, MEYRE D, **2015**, From big data analysis to personalized medicine for all: challenges and opportunities, *BMC Medical Genomics*, 8:33
6. Alzheimer's Association Report, **2014**, Alzheimer's disease facts and figures Alzheimer's Association, <https://www.alz.co.uk/research/WorldAlzheimerReport2014.pdf>
7. ATHREYA AP, KALARI KR, CAIRNS J, GAGLIO AJ, WILLS QF, NIU N, WEINSHILBOUM R, IYER KR, WANG L, **2017**, Model-based unsupervised learning informs metformin-induced cell-migration inhibition through an AMPK-independent mechanism in breast cancer, *Oncotarget*, 8, 27199-27215
8. AVDIC U, CHUGH D, OSMAN H, CHAPMAN K, JACKSON J, EKDAHL CT, **2014**, Absence of interleukin-1 receptor 1 increases excitatory and inhibitory scaffolding protein expression and microglial activation in the adult mouse hippocampus, *Cellular and Molecular Immunology*, 12, 645-647
9. BA M, KONG M, LI X, NG KP, ROSA-NETO P, GAUTHIER S, **2016**, Is ApoE  $\epsilon$  4 a good biomarker for amyloid pathology in late onset Alzheimer's disease?, *Translational Neurodegeneration*, 5:20
10. BALLMAN VK, **2015**, Biomarker: Predictive or Prognostic?, *Journal of Clinical Oncology*, 33(33):3968-71
11. BALSIS S, CHOUDHURY TK, GERACI L, BENGE JF, PATRICK CJ, **2017**, Alzheimer's Disease Assessment: A Review and Illustrations Focusing on Item Response Theory Techniques, *Assessment*, 25(3):360-373
12. BARABASI AL, GULBAHCE N, LOSCALZO J, **2011**, Network medicine: a network-based approach to human disease, *Nature Reviews Genetics*, 12, 56-68.

13. BARABASI AL, OLTVAI ZN, **2004**, Network biology: understanding the cell's functional organization, *Nature Reviews Genetics*, 5, 101-13.
14. BARRETT T, WILHITE ES, LEDOUX P, EVANGELISTA C, KIM FK, TOMASHEVSKY M, MARSHALL AK, PHILLIPPY HK, SHERMAN MP, HOLKO M, YEFANOV A, LEE H, ZHANG N, ROBERTSON LC, SEROVA N, DAVIS S, SOBOLEVA A, **2013**, NCBI GEO: archive for functional genomics data sets—update, *Nucleic Acids Research*, 41, 991-995
15. BATEMAN EJ, XIONG C, BENZINGER LST, FAGAN AM, GOATE A, FOX NC, MARCUS DS, CAIRNS NJ, XIE X, BLAZEY TM, HOLTZMAN DM, SANTACRUZ A, BUCKLES V, OLIVER A, MOULDER K, AISEN PS, GHETTI B, KLUNK WE, MCDADE E, MARTIN RN, **2012**, Clinical and Biomarker Changes in Dominantly Inherited Alzheimer's Disease, *The New England Journal of Medicine*, 367, 795-804
16. BERCHTOLD NC, COLEMAN PD, CRIBBS DH, ROGERS J, GILLEN DL, COTMAN CW, **2013**, Synaptic genes are extensively downregulated across multiple brain regions in normal human aging and Alzheimer's disease, *Neurobiology of Aging*, 34, 1653-1661
17. BERTOLACCINI L, SOLLI P, PARDOLESI A, PASINI A, **2017**, An overview of the use of artificial neural networks in lung cancer research. *Journal of Thoracic Disease*, 9, 924-931.
18. BISHOP CM, **1995**, Neural networks for pattern recognition, *Oxford University Press*
19. BISHOP CM, **2006**, Pattern Recognition and Machine Learning, *Springer*
20. BLAIR LJ, NORDHUES BA, HILL SE, SCAGLIONE KM, O'LEARY JC 3RD, FONTAINE SN, BREYDO L, ZHANG B, LI P, WANG L, COTMAN C, PAULSON HL, MUSCHOL M, UVERSKY VN, KLENGEL T, BINDER EB, KAYED R, GOLDE TE, BERCHTOLD N, DICKEY CA, **2013**, Accelerated neurodegeneration through chaperone-mediated oligomerization of tau, *The Journal of Clinical Investigation*, 123, 4158-4169
21. BLENNOW K, **2014**, CSF biomarkers for Alzheimer's disease: use in early diagnosis and evaluation of drug treatment, *Expert Review of Molecular Diagnostics*, 5, 661-672
22. BRAAK H, DEL TREDICI K, **2012**, Where, when, and in what form does sporadic Alzheimer disease begin?, *Current Opinions on Neurology*, 25, 708-714
23. BROWNE F, WANG H, ZHENG H, **2014**, An integrative network-driven pipeline for the prioritization of Alzheimer's disease genes, *2014 IEEE International Conference on Bioinformatics and Biomedicine*
24. BRUGGEMAN JF, WESTERHOFF VH, **2006**, The nature of systems biology, *Trends in Microbiology*, 15(1):45-50
25. CANU E, MCLAREN DG, FITZGERALD ME, BENDLIN BB, ZOCCATELLI G, ALESSANDRINI F, PIZZINI FB, RICCIARDI GK, BELTRAMELLO A, JOHNSON

- SC, FRISONI GB, **2011**, Mapping the structural brain changes in Alzheimer's disease: the independent contribution of two imaging modalities, *Journal of Alzheimer's Disease*, 26, 263-274
26. CARLSON NR, BIRKETT AM, **2017**, Physiology of Behaviour, *Pearson*
  27. CHATZIMICHAIL E, MATTHAIOS D, BOUROS D, KARAKITSOS P, ROMANIDIS K, KAKOLYRIS S, PAPASHINOPOULOS G, RIGAS A, **2014**, gamma-H2AX: A novel prognostic marker in a prognosis prediction model of patients with early operable non-small cell lung cancer. *International Journal of Genomics*, 2014, 160236
  28. CHEN PY, **2005**, Bioinformatics Technologies, *Springer*
  29. CINCO P, JING Y, WALDVOGEL HJ, CURTIS MA, ZHANG H, ABRAHAM WA, FAULL RLM, LIU P, **2015**, Arginine decarboxylase and agmatinase immunoreactivity in Alzheimer's superior frontal gyrus, *Alzheimer's and Dementia*, 11, P773
  30. CLARKE R, RESSOM HW, WANG A, XUAN J, LIU MC, GEHAN EA, WANG Y, **2008**, The properties of high-dimensional data spaces: implications for exploring gene and protein expression data, *Nature Reviews Cancer*, 8, 37-49
  31. CONESA A, MADRIGAL P, TARAZONA S, GOMEZ-CABRERO D, CERVERA A, MCPHERSON A, WOJCIECH SZCZEŚNIAK M, GAFFNEY DJ, ELO LL, ZHANG X, MORTAZAVI A, **2016**, A survey of best practices for RNA-seq data analysis, *Genome Biology*, 17:13
  32. COOK CE, TODD BERGMAN M, FINN RD, COCHRANE G, BIRNEY E, APWEILER R, **2015**, The European Bioinformatics Institute in 2016: Data growth and integration, *Nucleic Acids Research*, 44, D20-D26
  33. CRESPI GAN, HERMANS SJ, PARKER MW, MILES LA, **2015**, Molecular basis for mid-region amyloid- $\beta$  capture by leading Alzheimer's disease immunotherapies, *Scientific Reports*, 9649
  34. DAS P, MURPHY MP, YOUNKIN LH, YOUNKIN SG, GOLDE TE, **2001**, Reduced effectiveness of Abeta1-42 immunization in APP transgenic mice with significant amyloid deposition, *Neurobiology of Aging*, 22, 721-727
  35. DE CALIGNON A, POLYDORO M, SUÁREZ-CALVET M, WILLIAM C, ADAMOWICZ HD, KOPEIKINA JK, PITSTICK R, SAHARA N, ASHE HK, CARLSON AG, SPIRES-JONES LT, HYMAN TB, **2012**, Propagation of tau pathology in a model of early Alzheimer's Disease, *Neuron*, 73, 685-697
  36. DE JONG D, JANSEN RW, KREMER BP, VERBEEK MM, **2006**, Cerebrospinal fluid amyloid beta42/phosphorylated tau ratio discriminates between Alzheimer's disease and vascular dementia, *The Journals of Gerontology*, 61, 755-758
  37. DE STROOPER B, KARRAN E, **2016**, The Cellular Phase of Alzheimer's Disease, *Cell*, 164, 603-615

38. DOIG AJ, DEL CASTILLO-FRIAS MP, BERTHOUMIEU O, TARUS B, NASICALABOUZE J, STERPONE F, NGUYEN PH, HOOPER NM, FALLER P, DERREUMAUX P, **2017**, Why Is Research on Amyloid- $\beta$  Failing to Give New Drugs for Alzheimer's Disease?, *ACS Chemical Neuroscience*, 8, 1435-1437
39. DOIG AJ, DERREUMAUX P, **2015**, Inhibition of protein aggregation and amyloid formation by small molecules, *Current Opinion in Structural Biology*, 30, 50-56
40. DUBOIS B, FELDMAN HH, JACOVA C, HAMPEL H, MOLINUEVO JL, BLENNOW K, DEKOSKY ST, GAUTHIER S, SELKOE D, BATEMAN R, CAPPA S, CRUTCH S, ENGELBORGH S, FRISONI GB, FOX NC, GALASKO D, HABERT MO, JICHA GA, NORDBERG A, PASQUIER F, RABINOVICI G, ROBERT P, ROWE C, SALLOWAY S, SARAZIN M, EPELBAUM S, DE SOUZA LC, VELLAS B, VISSER PJ, SCHNEIDER L, STERN Y, SCHELTENS P, CUMMINGS JL, **2014**, Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria, *Lancet Neurology*, 13, 614-629
41. DUNCKLEY T, BEACH TG, RAMSEY KE, GROVER A, MASTROENI D, WALKER DG, LAFLEUR BJ, COON KD, BROWN KM, CASELLI R, KUKULL W, HIGDON R, MCKEEL D, MORRIS JC, HULETTE C, SCHMECHEL D, REIMAN EM, ROGERS J, STEPHAN DA, **2006**, Gene expression correlates of neurofibrillary tangles in Alzheimer's disease, *Neurobiology of Aging*, 27, 1359-1371
42. ELSHEIKH S.E., GREEN A.R., RAKHA E.A., POWE D.G., AHMED R.A., COLLINS H.M., SORIA D., GARIBALDI J.M., PAISH C.E., AMMAR A.A., GRAINGE M.J., BALL G.R., ABDELGHANY M.K., MARTINEZ-POMARES L., HEERY D.M. AND ELLIS I.O., **2009**, Global histone modifications in breast cancer correlate with tumor phenotypes, prognostic factors, and patient outcome, *Cancer Research*, 69 (9), pp. 3802-3809.
43. EMILSSON L, SAETRE P, JAZIN E, **2006**, Alzheimer's disease: mRNA expression profiles of multiple patients show alterations of genes involved with calcium signalling, *Neurobiology of Disease*, 21, 618-625
44. ENGELBORGH S, DE VREESE K, VAN DE CASTEELE T, VANDERSTICHELE H, VAN EVERBROECK B, CRAS P, MARTIN JJ, VANMECHELEN E, DE DEYN PP, **2008**, Diagnostic performance of a CSF-biomarker panel in autopsy-confirmed dementia, *Neurobiology of Aging*, 29, 1143-1159
45. FAGGIOLI F, VIJG J, MONTAGNA C, **2011**, Chromosomal aneuploidy in the aging brain, *Mechanisms of Ageing and Development*, 132, 429-436
46. FEHLBAUM-BEURDELEY P, JARRIGE-LE PRADO AC, PALLARES D, CARRIÈRE J, GUIHAL C, SOUCAILLE C, ROUET F, DROUIN D, SOL O, JORDAN H, WU D, LEI L, EINSTEIN R, SCHWEIGHOFFER F, BRACCO L, **2010**, Toward an

- Alzheimer's disease diagnosis via high-resolution blood gene expression, *Alzheimer's & Dementia*, 6, 25-38
47. FRAGKOULI A, TSILIBARY EC, TZINIA AK, **2014**, Neuroprotective role of MMP-9 overexpression in the brain of Alzheimer's 5xFAD mice, *Neurobiology of Disease*, 70, 179-189
  48. FRISONI BG, BLENNOW K, **2013**, Biomarkers for Alzheimer's: the sequel of an original model, *The Lancet Neurology*, 12, 126-128
  49. GALTON CJ, PATTERSON K, GRAHAM K, LAMBON-RALPH MA, WILLIAMS G, ANTOUN N, SAHAKIAN BJ, HODGES JR, **2001**, Differing patterns of temporal atrophy in Alzheimer's disease and semantic dementia, *Neurology*, 57, 216-225
  50. GEEKIYANAGE H, JICHA GA, NELSON PT, CHAN C, **2012**, Blood serum miRNA: non-invasive biomarkers for Alzheimer's disease, *Experimental Neurology*, 235, 491-496
  51. GEURTS P, **2010**, Bias vs Variance Decomposition for Regression and Classification, *Data Mining and Knowledge Discovery Handbook*, pp 749-763, *Springer*
  52. GIULIANI A, FILIPPI S, BERTOLASO M, **2014**, Why network approach can promote a new way of thinking in biology, *Frontiers in Genetics*, 83
  53. GÓMEZ-ISLA T, HOLLISTER R, WEST H, MUI S, GROWDON HJ, PETERSEN CR, PARISI EJ, HYMAN TB, **1997**, Neuronal loss correlates with but exceeds neurofibrillary tangles in Alzheimer's disease, *Annals of Neurology*, 41, 17-24
  54. GONG B, RADULOVIC M, FIGUEIREDO-PEREIRA EM, CARDOZO C, **2016**, The Ubiquitin-Proteasome System: Potential Therapeutic Targets for Alzheimer's Disease and Spinal Cord Injury, *Frontiers in Molecular Neuroscience*
  55. GONG C, IQBAL K, **2009**, Hyperphosphorylation of Microtubule-Associated Protein Tau: A Promising Therapeutic Target for Alzheimer Disease, *Current Medical Chemistry*, 15, 2321-2328
  56. GRAY KR, **2012**, Machine learning for image-based classification of Alzheimer's disease, Imperial College London, PhD Thesis
  57. GROSS AL, JONES RN, HABTEMARIAM DA, FONG TG, TOMMET D, QUACH L, SCHMITT E, YAP L, INOUYE SK, **2012**, Delirium and Long-term Cognitive Trajectory Among Persons with Dementia, *Archives of Internal Medicine*, 172, 1324-1331.
  58. GU Y, SCHUPF N, COSENTINO SA, LUCHSINGER JA, SCARMEAS N, **2012**, Nutrient intake and plasma  $\beta$ -amyloid, *Neurology*, 78, 1832-1840
  59. HAHR JY, **2015**, Physiology of the Alzheimer's disease, *Medical Hypotheses*, 85, 944-946
  60. HARR B, SCHLÖTTERER C, **2006**, Comparison of algorithms for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons, *Nucleic Acids Research*, 34(2): e8.

61. HASTIE T, TIBSHIRANI R, FRIEDMAN J, **2009**, *The Elements of Statistical Learning*, Springer
62. HEBB DO, **1949**, *The organization of behaviour*, Wiley
63. HOKAMA M, OKA S, LEON J, NINOMIYA T, HONDA H, SASAKI K, IWAKI T, OHARA T, SASAKI T, LAFERLA FM, KIYOHARA Y, NAKABEPPU Y, **2014**, Altered expression of diabetes-related genes in Alzheimer's disease brains: the Hisayama study, *Cerebral Cortex*, 24, 2476-2488
64. HOLZINGER A, DEHMER M, JURISICA I, **2014**, Knowledge Discovery and interactive Data Mining in Bioinformatics - State-of-the-Art, future challenges and research directions, *BMC Bioinformatics*, 15 Suppl 6:I1
65. HONG L, HUANG HC, JIANG ZF, **2014**, Relationship between amyloid-beta and the ubiquitin-proteasome system in Alzheimer's disease, *Neurology Research*, 36, 276-282
66. HORVATH S, DONG J, **2008**, Geometric Interpretation of Gene Coexpression Network Analysis, *PLoS One*
67. HUCKA M, FINNEY A, SAURO HM, BOLOURI H, DOYLE JC, KITANO H, ARKIN AP, BORNSTEIN BJ, BRAY D, CORNISH-BOWDEN A, CUELLAR AA, DRONOV S, GILLES ED, GINKEL M, GOR V, GORYANIN II, HEDLEY WJ, HODGMAN TC, HOFMEYR JH, HUNTER PJ, JUTY NS, KASBERGER JL, KREMLING A, KUMMER U, LE NOVÈRE N, LOEW LM, LUCIO D, MENDES P, MINCH E, MJOLSNESS ED, NAKAYAMA Y, NELSON MR, NIELSEN PF, SAKURADA T, SCHAFF JC, SHAPIRO BE, SHIMIZU TS, SPENCE HD, STELLING J, TAKAHASHI K, TOMITA M, WAGNER J, WANG J, **2003**, The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models, *Bioinformatics*, 19, 524-531
68. HUMPEL C, **2011**, Identifying and validating biomarkers for Alzheimer's disease, *Trends in Biotechnology*, 29, 26-32
69. HUYNH AR, MOHAN C, **2017**, Alzheimer's Disease: Biomarkers in the Genome, Blood, and Cerebrospinal Fluid, *Frontiers in Neurology*, 8:102
70. IOUROV IY, VORSANOVA SG, LIEHR T, YUROV YB, **2009**, Aneuploidy in the normal, Alzheimer's disease and ataxia-telangiectasia brain: differential expression and pathological meaning, *Neurobiology of Disease*, 34, 212-220
71. IRIZARRY AR, BOLSTAD MB, COLLIN F, COPE ML, HOBBS B, SPEED PT, **2003**, Summaries of Affymetrix GeneChip probe level data, *Nucleic Acids Research*, 4, 31(4):e15.
72. JACK RC, KNOPMAN SD, JAGUST JW, PETERSEN CR, WEINER WM, AISEN SP, SHAW ML, VEMURI P, WISTE JH, WEIGAND DS, LESNICK GT, PANKRATZ SV, DONOHUE CM, TROJANOWSKI QTJ, **2013**, Tracking pathophysiological processes



- in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers, *The Lancet Neurology*, 12, 207-216
73. JACK RC, KNOPMAN SD, JAGUST JW, SHAW ML, AISEN SP, WEINER WM, PETERSEN CR, TROJANOWSKI QTJ, **2010**, Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade, *The Lancet Neurology*, 9, 119-128
  74. JIANG X, JAO J, NEAPOLITAN R, **2015**, Learning Predictive Interactions Using Information Gain and Bayesian Network Scoring, *PLoS One*, 10(12): e0143247
  75. JOTHEESWARAN AT, WILLIAMS JD, PRINCE MJ, **2010** The predictive validity of the 10/66 dementia diagnosis in Chennai, India: a 3-year follow-up study of cases identified at baseline, *Alzheimer Disease and Associated Disorders*, 24, 296-302
  76. KAFETZOPOULOU LE, BOOCOCK DJ, DHONDALAY GKR, POWE DG, BALL GR, **2013**, Biomarker Identification in Breast Cancer: Beta-Adrenergic Receptor Signaling and Pathways to Therapeutic Response, *Computational and Structural Biotechnology Journal*, 6: e201303003
  77. KAPAKI E, PARASKEVAS GP, ZALONIS I, ZOURNAS C, **2003**, CSF tau protein and beta-amyloid (1-42) in Alzheimer's disease diagnosis: discrimination from normal ageing and other dementias in the Greek population, *European Journal of Neurology*, 10, 119-128
  78. KEMPF SJ, METAXAS A, IBÁÑEZ-VEA M, DARVESH S, FINSSEN B, LARSEN MR, **2016**, An integrated proteomics approach shows synaptic plasticity changes in an APP/PS1 Alzheimer's mouse model, *Oncotarget*, 7, 33627-33648
  79. KERR JD, HALLER GD, VAN DE VELDE JHC, **2016**, Baumann M, Oxford Textbook of Oncology
  80. KHAN AU, LIU L, PROVENZANO AF, BERMAN ED, PROFACI PC, SLOAN R, MAYEUX R, DUFF EK, SMALL AS, **2014**, Molecular drivers and cortical spread of lateral entorhinal cortex dysfunction in preclinical Alzheimer's disease, *Nature Neuroscience*, 17, 304-311
  81. KHATRI P, SIROTA M, BUTTE AJ, **2012**, Ten years of pathway analysis: current approaches and outstanding challenges, *PLoS One Computational Biology*, 8(2): e1002375
  82. KHOL P, NOBLE D, 2009, Systems biology and the virtual physiological human, *Molecular Systems Biology*, 5, 292
  83. KHOL P, CRAMPIN EJ, QUINN TA, Noble D, 2010, Systems Biology: An Approach, *Clinical Pharmacology & Therapeutics*, 88, 25-33
  84. KIM D, TSAI L, **2009**, Bridging Physiology and Pathology in AD, *Cell*, 137, 997-1000
  85. KOLESNIKOV N, HASTINGS E, KEAYS M, MELNICHUK O, TANG YA, WILLIAMS E, DYLAG M, KURBATOVA N, BRANDIZI M, BURDETT T, MEGY K,

- PILICHEVA E, RUSTICI G, TIKHONOV A, PARKINSON H, PETRYSZAK R, SARKANS U, BRAZMA A, **2015**, ArrayExpress update—simplifying data submissions, *Nucleic Acids Research*, 43, 1113-1116
86. LAMBERT JC, IBRAHIM-VERBAAS CA, HAROLD D, NAJ AC, SIMS R, BELLENGUEZ C, DESTAFANO AL, BIS JC, BEECHAM GW, GRENIER-BOLEY B, RUSSO G, THORTON-WELLS TA, JONES N, SMITH AV, CHOURAKI V, THOMAS C, IKRAM MA, ZELENIKA D, VARDARAJAN BN, KAMATANI Y, LIN CF, GERRISH A, SCHMIDT H, KUNKLE B, DUNSTAN ML, RUIZ A, BIHOREAU MT, CHOI SH, REITZ C, PASQUIER F, CRUCHAGA C, CRAIG D, AMIN N, BERR C, LOPEZ OL, DE JAGER PL, DERAMECOURT V, JOHNSTON JA, EVANS D, LOVESTONE S, LETENNEUR L, MORÓN FJ, RUBINSZTEIN DC, EIRIKSDOTTIR G, SLEEGERS K, GOATE AM, FIÉVET N, HUENTELMAN MW, GILL M, BROWN K, KAMBOH MI, KELLER L, BARBERGER-GATEAU P, MCGUINNESS B, LARSON EB, GREEN R, MYERS AJ, DUFOUIL C, TODD S, WALLON D, LOVE S, ROGAEVA E, GALLACHER J, ST GEORGE-HYSLOP P, CLARIMON J, LLEO A, BAYER A, TSUANG DW, YU L, TSOLAKI M, BOSSÙ P, SPALLETTA G, PROITSI P, COLLINGE J, SORBI S, SANCHEZ-GARCIA F, FOX NC, HARDY J, DENIZ NARANJO MC, BOSCO P, CLARKE R, BRAYNE C, GALIMBERTI D, MANCUSO M, MATTHEWS F; EUROPEAN ALZHEIMER'S DISEASE INITIATIVE (EADI); GENETIC AND ENVIRONMENTAL RISK IN ALZHEIMER'S DISEASE; ALZHEIMER'S DISEASE GENETIC CONSORTIUM; COHORTS FOR HEART AND AGING RESEARCH IN GENOMIC EPIDEMIOLOGY, MOEBUS S, MECOCCHI P, DEL ZOMPO M, MAIER W, HAMPEL H, PILOTTO A, BULLIDO M, PANZA F, CAFFARRA P, NACMIAS B, GILBERT JR, MAYHAUS M, LANNEFELT L, HAKONARSON H, PICHLER S, CARRASQUILLO MM, INGELSSON M, BEEKLY D, ALVAREZ V, ZOU F, VALLADARES O, YOUNKIN SG, COTO E, HAMILTON-NELSON KL, GU W, RAZQUIN C, PASTOR P, MATEO I, OWEN MJ, FABER KM, JONSSON PV, COMBARROS O, O'DONOVAN MC, CANTWELL LB, SOININEN H, BLACKER D, MEAD S, MOSLEY TH JR, BENNETT DA, HARRIS TB, FRATIGLIONI L, HOLMES C, DE BRUIJN RF, PASSMORE P, MONTINE TJ, BETTENS K, ROTTER JI, BRICE A, MORGAN K, FOROUD TM, KUKULL WA, HANNEQUIN D, POWELL JF, NALLS MA, RITCHIE K, LUNETTA KL, KAUWE JS, BOERWINKLE E, RIEMENSCHNEIDER M, BOADA M, HILTUENEN M, MARTIN ER, SCHMIDT R, RUJESCU D, WANG LS, DARTIGUES JF, MAYEUX R, TZOURIO C, HOFMAN A, NÖTHEN MM, GRAFF C, PSATY BM, JONES L, HAINES JL, HOLMANS PA, LATHROP M, PERICAK-VANCE MA, LAUNER LJ, FARRER LA, VAN DUJIN CM, VAN BROECKHOVEN C, MOSKVINA V,

- SESHADRI S, WILLIAMS J, SCHELLENBERG GD, AMOUYEL P., **2013**, Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease, *Nature Genetics*, 45, 1452-1458
87. LANCASHIRE LJ, LEMETRE C, BALL GR, **2009**, An introduction to artificial neural networks in bioinformatics--application to complex microarray and mass spectrometry datasets in cancer studies, *Briefings in Bioinformatics*, 10, 315-29.
88. LANCASHIRE LJ, POWE DG, REIS-FILHO JS, RAKHA E, LEMETRE C, WEIGELT B, ABDEL-FATAH TM, GREEN AR, MUKTA R, BLAMEY R, PAISH EC, REES RC, ELLIS IO, BALL GR, **2010**, A validated gene expression profile for detecting clinical outcome in breast cancer using artificial neural networks, *Breast Cancer Research Treatment*, 120, 83-93.
89. LANCASHIRE LJ, REES RC, BALL GR, **2008**, Identification of gene transcript signatures predictive for estrogen receptor and lymph node status using a stepwise forward selection artificial neural network modelling approach, *Artificial Intelligence in Medicine*, 43, 99-111
90. LANCASHIRE, **2006**, Artificial neural Network Predictive Modelling in Bioinformatics, PhD Thesis, NTU
91. LAURÉN J, GIMBEL AD, NYGAARD BH, GILBERT WJ, STRITTMATTER MS, **2009**, Cellular Prion Protein Mediates Impairment of Synaptic Plasticity by Amyloid- $\beta$  Oligomers, *Nature*, 457, 1128-1132
92. LEMETRE C, **2010**, Artificial neural network techniques to investigate potential interactions between biomarkers, Nottingham Trent University
93. LEMETRE C, LANCASHIRE LJ, REES RC, BALL GR, **2009**, Artificial Neural Network Based Algorithm for Biomolecular Interactions Modeling, *International Work-Conference on Artificial Neural Networks*, 877-885
94. LI Z, WANG W, ZHOU F, GAO X, PENG G, XU H, CHEN Y, **2005**, Interaction of stathmin-like 2 protein with the APP intracellular domain, *Tsinghua Science and Technology*, 10(4)
95. LIANG WS, DUNCKLEY T, BEACH GT, GROVER A, MASTROENI D, WALKER GD, CASELLI JR, KUKULL AW, MCKEEL D, MORRIS JC, HULETTE C, SCHMECHEL D, ALEXANDER EG, REIMAN ME, ROGERS J, STEPHAN AD, **2007**, Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain, *Physiological Genomics*, 28, 311-322
96. LIANG WS, DUNCKLEY T, BEACH TG, GROVER A, MASTROENI D, RAMSEY K, CASELLI RJ, KUKULL WA, MCKEEL D, MORRIS JC, HULETTE CM, SCHMECHEL D, REIMAN EM, ROGERS J, STEPHAN DA, **2010**, Neuronal gene

- expression in non-demented individuals with intermediate Alzheimer's Disease neuropathology, *Neurobiology of Aging*, 31, 549-566
97. LIANG WS, REIMAN EM, VALLA J, DUNCKLEY T, BEACH TG, GROVER A, NIEDZIELKO TL, SCHNEIDER LE, MASTROENI D, CASELLI R, KUKULL W, MORRIS JC, HULETTE CM, SCHMECHEL D, ROGERS J, STEPHAN DA, **2008**, Alzheimer's disease is associated with reduced expression of energy metabolism genes in posterior cingulate neurons, *Proceedings of the National Academy of Sciences of the United States of America*, 105, 4441-4446
  98. LIBBRECHT MW, NOBLE WS, **2015**, Machine learning applications in genetics and genomics, *Nature Reviews Genetics*, 16, 321-332
  99. LIU CC, LIU CC, KANEKIYO T, XU H, BU G, **2013**, Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy, *Nature Reviews Neurology*, 9, 106-118
  100. LIU JS, LAWRENCE CE, **1999**, Bayesian inference on biopolymer models, *Bioinformatics*, 15, 38-52
  101. LORD J, CRUCHAGA C, **2014**, The epigenetic landscape of Alzheimer's disease, *Nature Neuroscience*, 17, 1138-1140
  102. LOWERY AJ, MILLER N, DEVANEY A, MCNEILL RE, DAVOREN PA, LEMETRE C, BENES V, SCHMIDT S, BLAKE J, BALL G, KERIN MJ, **2009**, MicroRNA signatures predict oestrogen receptor, progesterone receptor and HER2/neu receptor status in breast cancer, *Breast Cancer Research*, 11(3):R27
  103. MAO H, BROWN EH, SILVER LD, **2017**, Mouse models of Casc3 reveal developmental functions distinct from other components of the exon junction complex, *RNA*, 23(1), 23-31
  104. MAES OC, XU S, YU B, CHERTKOW HM, WANG E, SCHIPPER HM, **2007**, Transcriptional profiling of Alzheimer blood mononuclear cells by microarray, *Neurobiology of Aging*, 28, 1795-1809
  105. MAJUMDER D, MUKHERJEEB A, **2011**, A passage through systems biology to systems medicine: adoption of middle-out rational approaches towards the understanding of therapeutic outcomes in cancer, *Analyst*, 136(4):663-78
  106. MANNEVILLE ES, Hall A, **2002**, Rho GTPases in cell biology, *Nature*, 420, 629-635
  107. MATTSON MP, **2008**, Glutamate and neurotrophic factors in neuronal plasticity and disease, *Annals of New York Academy of Science*, 144, 97-112
  108. MATTSSON N, CARRILLO CM, DEAN AR, DEVOUS DM, NIKOLCHEVA T, PESINI P, SALTER H, POTTER ZW, SPERLING SR, BATEMAN JR, BAIN JL, LIUM E, **2015**, Revolutionizing Alzheimer's disease and clinical trials through biomarkers, *Alzheimer's and Dementia*, 1, 421-419

- 109.MCCALL J, 2005, Genetic algorithms for modelling and optimisation, *Journal of Computational and Applied Mathematics*, 184, 205-222
- 110.MCDERMOTT JE, WANG J, MITCHELL H, WEBB-ROBERTSON BJ, HAFEN R, RAMEY J, RODLAND KD, **2013**, Challenges in Biomarker Discovery: Combining Expert Insights with Statistical Analysis of Complex Omics Data, *Expert Opinion on Medical Diagnostics*, 7, 37-51
- 111.MEAD S, POULTER M, UPHILL J, BECK J, WHITFIELD J, WEBB EFT, CAMPBELL T, ADAMSON G, DERIZIOTIS P, TABRIZI JS, HUMMERICH H, VERZILLI C, ALPERS PM, WHITTAKER CJ, COLLINGE J, **2009**, Genetic risk factors for variant Creutzfeldt–Jakob disease: a genome-wide association study, *The Lancet Neurology*, 8, 57-66
- 112.MIETELSKA-POROWSKA A, WASIK U, GORAS M, FILIPEK A, NIEWIADOMSKA G, **2014**, Tau protein modifications and interactions: their role in function and dysfunction, *International Journal of Molecular Sciences*, 15, 4671-4713
- 113.MILLER JA, WOLTJER RL, GOODENBOUR JM, HORVATH S, GESCHWIND DH, **2013**, Genes and pathways underlying regional and cell type changes in Alzheimer's disease, *Genome Medicine*, 5(5):48
- 114.MIOTTO R, WANG F, WANG S, JIANG X, DUDLEY TJ, **2017**, Deep learning for healthcare: review, opportunities and challenges, *Briefings in Bioinformatics*, bbx044
- 115.MORRIS JK, UY RAZ, VIDONI ED, WILKINS HM, ARCHER AE, THYFAULT JB, MILES JM, BURNS JM, **2017**, Effect of APOE  $\epsilon$ 4 Genotype on Metabolic Biomarkers in Aging and Alzheimer's Disease, *Journal of Alzheimer's Disease*, 58:4, 1129-1135
- 116.MUFSON JE, IKONOMOVIC DM, COUNTS ES, PEREZ ES, MALEK-AHMADI M, SCHEFF WS, GINSBERG DS, **2016**, Molecular and cellular pathophysiology of preclinical Alzheimer's disease, *Behavioural Brain Research*, 311, 54-69
- 117.MULDER C, VERWEY NA, VAN DER FLIER WM, BOUWMAN FH, KOK A, VAN ELK EJ, SCHELTENS P, BLANKENSTEIN MA, **2010**, Amyloid-beta(1-42), total tau, and phosphorylated tau as cerebrospinal fluid biomarkers for the diagnosis of Alzheimer disease, *Clinical Chemistry*, 56, 248-253
- 118.MURPHY K, **2012**, Machine Learning: A Probabilistic Perspective, *MIT Press*
- 119.NAGELE E, HAN M, DEMARSHALL C, BELINKA B, NAGELE R, **2011**, Diagnosis of Alzheimer's disease based on disease-specific autoantibody profiles in human sera, *PLoS One*, 6
- 120.NIKOLAEV A, MCLAUGHLIN T, O'LEARY D, TESSIER-LAVIGNE M, **2009**, N-APP binds DR6 to cause axon pruning and neuron death via distinct caspases, *Nature*, 457, 981-989

- 121.OKAZAKI T, WANG H, MASLIAH E, CAO M, JOHNSON SA, SUNDSMO M, SAITOH T, MORI N, **1995**, SCG10, a neuron-specific growth-associated protein in Alzheimer's disease, *Neurobiology of Aging*, 16, 883-894
- 122.PICKART MC, EDDINS JM, **2004**, Ubiquitin: structures, functions, mechanisms, *BBA – Molecular Cell Research*, 1695, 55-72
- 123.PICCOLO SR, SUN Y, CAMPBELL JD, LENBURG ME, BILDA AH, JOHNSON WE, **2012**, A single-sample microarray normalization method to facilitate personalized-medicine workflows, *Genomics*, 100(6), 337-344
- 124.POCHWAT B, NOWAK G1, SZEWCZYK B, **2015**, Relationship between Zinc (Zn (2+)) and Glutamate Receptors in the Processes Underlying Neurodegeneration, *Neural Plasticity*, 2015:591563
- 125.POLYDORO M, DZHALA IV, POOLER MA, NICHOLLS BS, MCKINNEY AP, SANCHEZ L, PITSTICK R, CARLSON AG, STALEY JK, SPIRES-JONES LT, HYMA BT, **2014**, Soluble pathological tau in the entorhinal cortex leads to presynaptic deficits in an early Alzheimer's disease model, *Acta Neuropathologica*, 127, 257-270
- 126.POULIAKIS A, KARAKITSOU E, MARGARI N, BOUNTRIS P, HARITOU M, PANAYIOTIDES J, KOUTSOURIS D, KARAKITSOS P, **2016**, Artificial Neural Networks as Decision Support Tools in Cytopathology: Past, Present, and Future, *Biomedical Engineering and Computational Biology*, 7:1-18
- 127.PRECHELT L, **1998**, Automatic early stopping using cross validation: quantifying the criteria, *Neural Networks*, 11, 761-767
- 128.PRINCZ A, TAVERNARAKIS N, **2017**, The role of SUMOylation in ageing and senescent decline, *Mechanisms of Ageing and Development*, 162, 85-90
- 129.QAZI TJ, QUAN Z, MIR A, QING H, **2017**, Epigenetics in Alzheimer's Disease: Perspective of DNA Methylation, *Molecular Neurobiology*, 55(2):1026-1044
- 130.RASKIN J, CUMMINGS J, HARDY J, SCHUH K, DEAN R, **2015**, Neurobiology of Alzheimer's Disease: Integrated Molecular, Physiological, Anatomical, Biomarker, and Cognitive Dimensions, *Current Alzheimer Research*, 12, 712-722
- 131.RIM JH, KIM SH, HWANG IS, KWON SS, KIM J, KIM HW, CHO MJ, KO A, YOUN SE, KIM J, LEE YM, CHUNG HJ, LEE JS, KIM HD, CHOI JR, LEE ST, KANG HC, **2018**, Efficient strategy for the molecular diagnosis of intractable early-onset epilepsy using targeted gene sequencing, *BMC Medical Genomics*, 11(1):6.
- 132.RITTER A, CUMMINGS J, **2015**, Fluid Biomarkers in Clinical Trials of Alzheimer's Disease Therapeutics, *Frontiers in Neurology*, 6:186
- 133.ROADKNIGHT CM, PALMER-BROWN D, MILLS GE, **2005**, Correlated activity pruning (CAPing), Computational Intelligence Theory and Applications. Fuzzy Days 1997. *Lecture Notes in Computer Science*, 1226

134. ROCCHI A, PELLEGRINI S, SICILIANO G, MURRI L, **2003**, Causative and susceptibility genes for Alzheimer's disease: a review, *Brain Research Bulletin*, 61, 1-24
135. ROJAS R, **1996**, Neural Networks, *Springer*
136. ROSÉN C, HANSSON O, BLENNOW K, ZETTERBERG H, **2013**, Fluid biomarkers in Alzheimer's disease - current concepts, *Molecular Neurodegeneration*, 8:20
137. ROSENBLATT F, **1958**, The perceptron: a probabilistic model for information storage and organization in the brain, *Psychological Review*, 65, 386
138. SALLOWAY S, SPERLING R, FOX NC, BLENNOW K, KLUNK W, RASKIND M, SABBAGH M, HONIG LS, PORSTEINSSON AP, FERRIS S, REICHERT M, KETTER N, NEJADNIK B, GUENZLER V, MILOSLAVSKY M, WANG D, LU Y, LULL J, TUDOR IC, LIU E, GRUNDMAN M, YUEN E, BLACK R, BRASHEAR HR, **2014**, Two phase 3 trials of bapineuzumab in mild-to-moderate Alzheimer's disease, *New England Journal of Medicine*, 370, 322-333
139. SALTER-TOWNSHEND M, WHITE A, GOLLINI I, MURPHY TB, **2012**, Review of statistical network analysis: models, algorithms, and software, *Statistical Analysis and Data Mining*, 5(4)
140. SAVONENKO AV, MELNIKOVA T, LI T, PRICE DL, WONG PC, **2015**, Neurobiology of Brain Disorders, *Academic Press*, 321-338
141. SCHENA M, SHALON D, DAVIS RW, BROWN PO, **1995**, Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science*, 270, 467-470
142. SCHNEIDER LS, MANGIALASCHE F, ANDREASEN N, FELDMAN H, GIACOBINI E, JONES R, MANTUA V, MECOCCI P, PANI L, WINBLAD B, KIVIPELTO M, **2014**, Clinical trials and late-stage drug development for Alzheimer's disease: an appraisal from 1984 to 2014, *Journal of Internal Medicine*, 275, 251-283
143. SCHNEIDER SL, MANGIALASCHE F, ANDREASEN N, FELDMAN H, GIACOBINI E, JONES R, MANTUA V, MECOCCI P, PANI L, WINBLAD B, KIVIPELTO M, **2014**, Clinical trials and late-stage drug development for Alzheimer's disease: an appraisal from 1984 to 2014, *Journal of Internal Medicine*, 275, 251-283
144. SELKOE DJ, HARDY J, **2016**, The amyloid hypothesis of Alzheimer's disease at 25 years, *EMBO Molecular Medicine*, 8, 595-608
145. SHAHZAD K, LOOR JJ, **2012**, Application of Top-Down and Bottom-up Systems Approaches in Ruminant Physiology and Metabolism, *Current Genomics*, 13, 379-394
146. SHANKAR GM, LI S, MEHTA TH, GARCIA-MUNOZ A, SHEPARDSON NE, SMITH I, BRETT FM, FARRELL MA, ROWAN MJ, LEMERE CA, REGAN CM, WALSH DM, SABATINI BL, SELKOE DJ, **2008**, Amyloid-beta protein dimers isolated directly from Alzheimer's brains impair synaptic plasticity and memory, *Nature Medicine*, 14, 837-842





147. SHANKAR GM, LI S, MEHTA TH, GARCIA-MUNOZ A, SHEPARDSON NE, SMITH I, BRETT FM, FARRELL MA, ROWAN MJ, LEMERE CA, REGAN CM, WALSH DM, SABATINI BL, SELKOE DJ, **2008**, Amyloid-beta protein dimers isolated directly from Alzheimer's brains impair synaptic plasticity and memory, *Nature Medicine*, 14, 837-842
148. SHARMA N, SINGH NA, **2016**, Exploring Biomarkers for Alzheimer's Disease, *Journal of Clinical and Diagnostic Research*, 10(7): KE01–KE06
149. SILVA ART, GRINBERG LT, FARFEL JM, DINIZ BS, LIMA LA, SILVA PJS, FERRETTI REL, ROCHA RM, FILHO WJ, CARRARO DM, BRENTANI H, **2012**, Transcriptional Alterations Related to Neuropathology and Clinical Manifestation of Alzheimer's Disease, *PLoS One*, 7(11): e48751
150. SIMPSON JE, INCE PG, SHAW PJ, HEATH PR, RAMAN R, GARWOOD CJ, GELSTHORPE C, BAXTER L, FORSTER G, MATTHEWS FE, BRAYNE C, WHARTON SB, **2011**, Microarray analysis of the astrocyte transcriptome in the aging brain: relationship to Alzheimer's pathology and APOE genotype, *Neurobiology of Aging*, 32, 1795-1807
151. SNIKERS S, STRINGER S, WATANABE K, JANSEN PR, COLEMAN JRI, KRAPOHL E, TASKESSEN E, HAMMERSCHLAG AR, OKBAY A, ZABANEH D, AMIN N, BREEN G, CESARINI D, CHABRIS CF, IACONO WG, IKRAM MA, JOHANNESSON M, KOELLINGER P, LEE JJ, MAGNUSSON PKE, MCGUE M, MILLER MB, OLLIER WER, PAYTON A, PENDLETON N, PLOMIN R, RIETVELD CA, TIEMEIER H, VAN DUIJN CM, POSTHUMA D1, **2017**, Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence, *Nature Genetics*, 49, 1107-1112
152. SOMMER C, GERLICH DW, 2013, Machine learning in cell biology - teaching computers to recognize phenotypes, *Journal of Cell Science*, 126, 5529-5539
153. SOOD S, GALLAGHER JI, LUNNON K, RULLMAN E, KEOHANE A, CROSSLAND H, PHILLIPS EB, CEDERHOLM T, JENSEN T, VAN LOON JCL, LANNFELT L, KRAUS EW, ATHERTON KP, HOWARD R, GUSTAFSSON T, HODGES A, TIMMONS AJ, **2015**, A novel multi-tissue RNA diagnostic of healthy ageing relates to cognitive health status, *Genome Biology*, 16:185
154. SOROKINA SY, KUPTZOV VN, URBAN YN, FOKIN AV, POJARKOV SV, IVANKOV MY, MELNIKOV AI, KULIKOV AM, **2013**, Databases as instruments for analysis of large-scale data sets of interactions between molecular biological objects, *Biology Bulletin*, 40, 223-242
155. STRIMBU K, TAVEL AJ, **2011**, What are Biomarkers?, *Current Opinion in HIV and AIDS*, 5, 463–466



156. SWAN AL, STEKEL DJ, HODGMAN C, ALLAWAY D, ALQAHTANI MH, MOBASHERI A, BACARDIT J, **2015**, A machine learning heuristic to identify biologically relevant and minimal biomarker panels from omics data, *BMC Genomics*, 16 Suppl 1:S2
157. UPADHYA CS, HEGDE NA, **2007**, Role of the ubiquitin proteasome system in Alzheimer's disease, *BMC Biochemistry*
158. VAFADAR-ISFAHANI B, BALL G, COVENEY C, LEMETRE C, BOOCOCK D, MINTHON L, HANSSON O, MILES AK, JANCIAUSKIENE SM, WARDEN D, SMITH AD, WILCOCK G, KALSHEKER N, REES R, MATHAROO-BALL B, MORGAN K, **2012**, Identification of SPARC-like 1 protein as part of a biomarker panel for Alzheimer's disease in cerebrospinal fluid, *Journal of Alzheimer's Disease*, 28, 625-636
159. VEUGELEN S, SAITO T, SAIDO CT, CHÁVEZ-GUTIÉRREZ L, DE STROOPER B, **2016**, Familial Alzheimer's Disease Mutations in Presenilin Generate Amyloidogenic A $\beta$  Peptide Seeds, *Neuron*, 90, 410-416
160. VOS SJ, VAN ROSSUM IA, VERHEY F, KNOL DL, SOININEN H, WAHLUND LO, HAMPEL H, TSOLAKI M, MINTHON L, FRISONI GB, FROELICH L, NOBILI F, VAN DER FLIER W, BLENNOW K, WOLZ R, SCHELTENS P, VISSER PJ, **2013**, Prediction of Alzheimer disease in subjects with amnesic and nonamnesic MCI, *Neurology*, 80, 1124-1132
161. VURAL S, WANG X, GUDA C, **2016**, Classification of breast cancer patients using somatic mutation profiles and machine learning approaches, *BMC Systems Biology*, 10, 62
162. WALKER DC, SOUTHGATE J, **2009**, The virtual cell—a candidate co-ordinator for 'middle-out' modelling of biological systems, *Briefings in bioinformatics*, 4(10), 450-461
163. WANG HS, PAN Z, SHI W, BROWN BS, WYMORE RS, COHEN IS, DIXON JE, MCKINNON D, **1998**, KCNQ2 and KCNQ3 potassium channel subunits: molecular correlates of the M-channel, *Science*, 282, 1890-1893
164. WANG Z, GERSTEIN M, SNYDER M, **2009**, RNA-Seq: a revolutionary tool for transcriptomics, *Nature Reviews: Genetics*, 10, 57-63
165. PRINCE M, JACKSON J, **2009**, World Health Organization, World Alzheimer Report. *Alzheimer's Disease International*, London  
<https://www.alz.co.uk/research/files/WorldAlzheimerReport.pdf>
166. XU Q, LIANG Y, **2001**, Monte Carlo cross validation, *Chemometrics and Intelligent Laboratory Systems*, 56, 1-11
167. YANG Y, SONG W, **2013**, Molecular links between Alzheimer's disease and diabetes mellitus, *Neuroscience*, 250, 140-150

168. YU J, SMITH VA, WANG PP, HARTEMINK AJ, JARVIS ED, **2004**, Advances to Bayesian network inference for generating causal networks from observational biological data, *Bioinformatics*, 20, 3594-3603
169. ZANG H, ZHANG S, HAPESHI K, **2010**, A review of nature-inspired algorithms, *Journal of Bionic Engineering*, 7, 232-237
170. ZENG Z, JIANG X, NEAPOLITAN R, 2019, Discovering causal interactions using Bayesian network scoring and information gain, *BMC Bioinformatics*, 17, 221
171. ZHANG B, GAITERI C, BODEA LG, WANG Z, MCELWEE J, PODTELEZHNIKOV AA, ZHANG C, XIE T, TRAN L, DOBRIN R, FLUDER E, CLURMAN B, MELQUIST S, NARAYANAN M, SUVER C, SHAH H, MAHAJAN M, GILLIS T, MYSORE J, MACDONALD ME, LAMB JR, BENNETT DA, MOLONY C, STONE DJ, GUDNASON V, MYERS AJ, SCHADT EE, NEUMANN H, ZHU J, EMILSSON V, **2013**, Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease, *Cell*, 153, 707-720
172. ZHANG Y, DONG Z, PHILLIPS P, WANG S, JI G, YANG J, YUAN TF, **2015**, Detection of subjects and brain regions related to Alzheimer's disease using 3D MRI scans based on eigenbrain and machine learning, *Frontiers in Computational Neuroscience*, 9:66.
173. ZIEGLER A, KOCH A, KROCKENBERGER K, GROßHENNIG A, **2012**, Personalized medicine using DNA biomarkers: a review, *Human genetics*, 131, 1627-1638
174. ZOU D, MA L, YU J, ZHANG Z, **2015**, Biological Databases for Human Research, Genomics, *Proteomics & Bioinformatics*, 13, 55-63
175. ZOU D, MA L, YU J, ZHANG Z, **2015**, Biological databases for human research, *Genomics, proteomics & bioinformatics*, 13, 55-63

## Discovery and application of immune biomarkers for hematological malignancies

Dimitrios Zafeiris, Jayakumar Vadakekolathu , Sarah Wagner, Alan Graham Pockley , Graham Roy Ball  and Sergio Rutella 

John van Geest Cancer Research Centre, College of Science and Technology, Nottingham Trent University, Nottingham, United Kingdom

### ABSTRACT

**Introduction:** Hematological malignancies originate and progress in primary and secondary lymphoid organs, where they establish a uniquely immune-suppressive tumour microenvironment. Although high-throughput transcriptomic and proteomic approaches are being employed to interrogate immune surveillance and escape mechanisms in patients with solid tumours, and to identify actionable targets for immunotherapy, our knowledge of the immunological landscape of hematological malignancies, as well as our understanding of the molecular circuits that underpin the establishment of immune tolerance, is not comprehensive.

**Areas covered:** This article will discuss how multiplexed immunohistochemistry, flow cytometry/mass cytometry, proteomic and genomic techniques can be used to dynamically capture the complexity of tumour-immune interactions. Moreover, the analysis of multi-dimensional, clinically annotated data sets obtained from public repositories such as Array Express, TCGA and GEO is crucial to identify immune biomarkers, to inform the rational design of immune therapies and to predict clinical benefit in individual patients. We will also highlight how artificial neural network models and alternative methodologies integrating other algorithms can support the identification of key molecular drivers of immune dysfunction.

**Expert commentary:** High-dimensional technologies have the potential to enhance our understanding of immune-cancer interactions and will support clinical decision making and the prediction of therapeutic benefit from immune-based interventions.

### ARTICLE HISTORY

Received 7 July 2017  
Accepted 15 September 2017

### KEYWORDS



Hematological malignancies; leukemia; lymphoma; multiple myeloma; immunotherapy; biomarker; gene expression profiling; prognosis

## 1 Introduction

Tumors are organized tissues that are infiltrated with immune cell populations of both the lymphoid and myeloid lineage [1] and possess both tumor-promoting and tumor-inhibiting properties. Compelling evidence indicates that preexisting immunological features contribute to the ability of patients with solid tumors to respond to immunotherapy with immunomodulatory agents such as checkpoint inhibitors [2]. The Immune Biomarkers Task Force of the Society for Immunotherapy of Cancer (SITC) recently published recommendations on the discovery of immune-related biomarkers, in which it highlighted the complexity of the tumor microenvironment (TME) and discussed novel tools to analyze the diversity of immune genes, proteins, cells, and pathways [3]. A broader understanding of baseline immunity, both in the periphery and in the TME, and of immune escape mechanisms is likely to expedite the identification of biomarkers that are predictive of clinical outcome and elucidate why cancer patients might fail to respond to immunotherapy [4,5]. Powerful technologies such as genome-wide association studies, multiplexed immunohistochemistry, high-dimensional blood profiling of immune cells by flow cytometry and mass cytometry are increasingly being integrated in this nascent, but rapidly evolving field. The aim of these approaches is to

assess immune competence and the likelihood of patients with solid tumors to respond to immunotherapy. In general, tumor infiltration by leukocyte subsets such as CD8<sup>+</sup> T cells and CD45RO<sup>+</sup> memory T cells with specific gene signatures and increased B-cell receptor (BCR) diversity is associated with an improved overall survival (OS), as has been demonstrated by mRNA sequencing data from The Cancer Genome Atlas (TCGA) in 11 solid tumor types encompassing breast, lung, melanoma, and lung adenocarcinoma and representing 3485 patients [6]. In contrast, macrophage signatures predicted poorer survival in most tumor types. The presence of T-cell infiltration contributes to a higher 'immunoscore' in patients with colorectal cancer (CRC), which correlates with improved patient prognosis [7].

Whereas the role of antitumor immunity in shaping clinical responses to therapy has been thoroughly investigated in melanoma and CRC, our understanding of the role played by individual immune cell types in the control of hematological malignancies remains limited. In principle, hematological malignancies are amenable to immune-mediated therapeutic effects, as suggested by the curative potential of allogeneic hematopoietic stem cell transplantation. Although immune checkpoint blockade has only been pursued recently in patients with Hodgkin and non-Hodgkin lymphoma [8,9], the field is expected to advance exponentially, as has already

**CONTACT** Sergio Rutella  [sergio.rutella@ntu.ac.uk](mailto:sergio.rutella@ntu.ac.uk)  John van Geest Cancer Research Centre, Nottingham Trent University, Clifton Campus, NG11 8NS, Nottingham, UK

© 2017 Informa UK Limited, trading as Taylor & Francis Group

occurred in solid tumor oncology. This will entail a paradigm shift in our current treatment modalities. An imperative for the correct design of clinical trials would be to dissect the determinants of response and resistance to checkpoint blockade and to decipher the architecture and composition of the TME, as well as the functional orientation of peripheral blood immune cells in patients with leukemia, lymphoma, and multiple myeloma (MM). Challenges to identifying biomarkers have recently been reviewed [10]. Despite the reciprocal relationship between tumors and the patient's immune system, it is presently unknown whether measurements in blood may correlate with findings from tumor sites, including lymph nodes and bone marrow (BM) [3,11]. In this respect, peripheral blood markers reflecting immune function at baseline ('peripheral immunoscore') have successfully predicted progression-free survival (PFS) in patients receiving vaccines for metastatic breast cancer and prostate cancer [12].

This review will focus on current strategies to interrogate the immunological TME in patients with hematological malignancies, with the objective to subvert cancer-induced immune suppression and identify targets for treatment.

## 2 Structure and function of the TME

Neoplastic cells activate gene expression programs in the TME that are supportive of tumor growth and inherently immune suppressive [4]. The TME is increasingly viewed as an attractive candidate for the discovery of predictive and prognostic immune biomarkers [11,13]. For instance, intra-tumoral levels of IL-15 strongly correlate with immune cell proliferation and disease recurrence in patients with CRC [14]. An 'immunome' compendium of mRNA transcripts specific for innate and adaptive immune cell populations has characterized the immune composition of the TME in CRC [15]. The patterns of gene expression were remarkably different in patients with significantly prolonged disease-free survival and in those with unfavorable outcome. The former showed an overrepresentation of T-cell-related genes, including  $\gamma\delta$  T cells and cytotoxic T cells, macrophages, and mast cells. Follicular helper T cells (T<sub>fh</sub>) and B cells also exerted a favorable effect on patient outcome. In contrast, patients with poor outcomes showed an overrepresentation of genes specific for eosinophils, Th2 cells, Th17 cells, Treg cells, and NK cells. Interestingly, the *in situ* immune reaction evolved with tumor progression from stages T1 to T4, with most of the T-cell markers decreasing with tumor stage.

Programmed Death Ligand (PDL)-L1 is expressed by cells in the TME, engages PD1 on T cells and triggers inhibitory signaling which prevents T-cell effector function and cytotoxicity [16]. PD-L1 expression in response to cytokine stimuli, most importantly IFN- $\gamma$ , has been termed 'adaptive immune resistance' [17]. Colocalization of inflammatory responses with CD8 and PD-L1 expression has been correlated with improved clinical outcome in patients with metastatic, but not localized, melanoma, implying that 'inflamed' tumors expressing PD-L1 might be more amenable to respond to immunotherapy [17]. A pragmatic classification of solid tumors based on their PD-L1 status and presence or absence of tumor-infiltrating lymphocytes (TILs) has been recently proposed [18]. Type I (PD-L1-expressing with TILs)

and type II TMEs (PD-L1 negative with no preexisting TILs) account for approximately 80% of human melanomas, with type I tumors having the best prognosis [17]. Other tumor types may exhibit a type III TME, in which constitutive PD-L1 expression is driven by oncogenic events rather than adaptive immune resistance, as shown in gliomas with loss of PTEN function [19] and in T-cell lymphomas [20]. Finally, although type IV tumors contain TILs, these show no expression of PD-L1, thereby suggesting a potential role for other immune suppressive circuits in driving immune dysfunction [18].

Intriguingly, three immune profiles have been revealed by clinical studies indicating that patients with 'inflamed' melanomas were more likely to respond to immunotherapy with checkpoint blocking agents [21,22]. The immune-inflamed phenotype is characterized by the presence of both CD4<sup>+</sup> and CD8<sup>+</sup> T cells, often accompanied by myeloid and monocytic cells, and by staining for PD-L1 on TILs and, in some cases, on tumor cells. The immune-excluded phenotype is characterized by tumors in which immune cells are retained in the stroma and fail to migrate and penetrate the tumor itself, and is unlikely to respond to immunotherapy. The third profile, the immune-desert phenotype, is characterized by a paucity of T cells, which is indicative of the absence of pre-existing antitumor immune responses, and by the presence of myeloid-derived suppressor cells (MDSCs), M2 macrophages and regulatory T (Treg) cells, which mediate immune suppression or tolerance. The importance of preexisting, clonally restricted CD8 T-cell responses and of physical proximity between PD1<sup>+</sup> and PD-L1<sup>+</sup> cells in the TME for tumor regression after immunotherapy with PD1 blocking agents has again been demonstrated in patients with metastatic melanoma [23].

In hematological malignancies, the BM represents not only the site of disease initiation and progression, but also a distinctive immunologic microenvironment that contains most developing and mature immune cell types, including long-lived CD4<sup>+</sup> and CD8<sup>+</sup> T cells [24]. A recent study identified landmark populations of BM-resident immune cells in mice [25]. Similar cells were grouped into clusters according to their expression of the measured proteins. The scaffold maps allowed the unsupervised visualization of the immune composition and complexity of murine BMs. In comparison, maps for secondary lymphoid organs exhibited an immune landscape dominated by mature T and B lymphocytes, as well as by myeloid cell clusters mapping closely to the macrophage and dendritic cell (DC) zones. The integration of human mass cytometry data from four healthy donors into the reference map revealed a similar overlay pattern between the two species [25].

In light of their origin from primary and secondary lymphoid tissues, hematological malignancies might be characterized by distinctive mechanisms of immune evasion compared with solid tumors [26]. In principle, hematological malignancies are poorly immunogenic and highly immune suppressive. For instance, acute leukemias disseminate rapidly and constrain protective antitumor immune responses through a plethora of immune subversive mechanisms, including the downregulation of MHC class I and class II expression, the consumption of essential amino acids through arginase-2

(ARG2) [27] and indoleamine 2,3-dioxygenase-1 (IDO1) [28], the induction of DC dysfunction, the expansion of Treg cells [29], and the upregulation of PD-L1 and other negative checkpoint molecules, such as cytotoxic T-lymphocyte-associated antigen-4 (CTLA-4) and lymphocyte activation gene 3 (LAG-3). PD-L1 expression might represent a general strategy of immune evasion among aggressive B-cell lymphomas [30]. The analysis of formalin-fixed, paraffin-embedded (FFPE) tissue biopsies from 237 primary lymphomas has detected PD-L1 protein expression in most nodular sclerosis and mixed cellularity classical Hodgkin's lymphomas (HL), primary mediastinal large B-cell lymphomas, Epstein-Barr Virus (EBV)-positive and EBV-negative posttransplantation lymphoproliferative disorders and EBV-associated diffuse large B-cell lymphomas (DLBCL). This group of neoplasms should then be considered for PD-1/PD-L1-directed therapies, as further discussed below.

Insights into the molecular mechanisms sustaining PD-L1 expression in lymphoma tissues have recently been provided [31]. Conditioned media from T-cell and B-cell lymphoma cell lines were shown to induce PD-L1/PD-L2 expression on macrophages in a signal transducer and activator of transcription (STAT)-3-dependent manner. *In vitro* studies pointed to a potential role of lymphoma-derived IL-27B in PD-L1/PD-L2 overexpression, suggesting that an IL-27/STAT-3 axis might be a target for immunotherapy in patients with NHL.

## 2 Immune gene signatures

Innate and adaptive immune responses within the TME can be assessed by gene expression profiling [32]. Immune gene signatures, especially those induced by IFN- $\gamma$ , are likely to be powerful biomarkers of response to checkpoint blockade. A considerable body of scientific evidence suggests that tumors responsive to immunotherapies display an inflammatory status which is associated with the concomitant counter-activation of immune suppressive circuits, thereby reflecting immune escape mechanisms. The implication of these observations is that preexisting immune responses are a prerequisite for the efficacy of immune checkpoint blockade. For instance, a 10-gene IFN- $\gamma$  score, including genes encoding *IDO1*, *LAG3*, *PRF1*, *GZM*, and other immune-related genes, showed a significant correlation with best overall response (OR) and PFS in patients with advanced melanoma, as well as a nonsignificant association with OS [33].

Importantly, immune-related gene signatures, and not tumor-related gene expression patterns, have been identified as being the main parameters associated with dissemination of CRC to distant metastases [34]. Specifically, patients without synchronous metastasis had a significantly increased expression of Th1-related genes, immune cytotoxicity-related genes and

MHC class II-related genes compared with patients having metastasis at the time of diagnosis. This study highlights the concept that immune phenotypes, as measured on the basis of multiple parameters, might be a crucial determinant for preventing the metastatic dissemination of tumors to distant sites. Although immune and genomic landscapes in pretreatment tumor biopsies correlate with response in patients with melanoma and other solid cancers, robust biomarkers that do not overlap between responders and nonresponders have not yet been

identified. An interesting study in 53 patients with metastatic melanoma initially treated with CTLA-4 blockade followed by programmed death-1 (PD-1) blockade at the time of progression analyzed immune gene signatures in longitudinal biopsies collected at multiple time points during therapy, using a 12-marker immunohistochemistry panel and targeted gene expression profiling on a nanoString platform [35]. Adaptive immune gene signatures in tumor samples obtained early during treatment, including the upregulation of cytolytic markers, *HLA* molecules, *IFN- $\gamma$*  pathway effector genes, and chemokines, were highly predictive of response to immune checkpoint blockade. Importantly, unique gene expression profiles observed in the TME of patients receiving monotherapy with anti-CTLA-4 or anti-PD-1 antibodies provided insights into the mechanisms of response to distinct forms of immune checkpoint blockade, as well as a compelling rationale for the design of combination immunotherapies.

The genomic landscape of tumors has been linked with tumor immunity, with neo-antigens that are predicted by tumor genome meta-analyses being implicated in driving T-cell responses and somatic mutations associated with immunological infiltrates being identified [36,37]. A recent analysis of TCGA data sets has allowed the identification of correlates of immune cytolytic activity in thousands of TCGA solid tumors [37]. On the basis of transcript levels of two tightly co-expressed cytolytic effector molecules, granzyme A and perforin, differences in cytolytic activities across tumor types were identified, with the highest levels being detected in kidney clear cell carcinoma and cervical cancers. Interestingly, cytolytic activities and expression of IFN-stimulated chemokines (*CXCL9*, *CXCL10*, and *CXCL11*) were associated with the counter-regulatory increase of immune suppressive molecules, including *IDO1*, *IDO2*, *PDL2*, and the *C1Q* complex, and with a modest, but significant, pan-cancer survival benefit [37].

Finally, immune gene co-expression patterns have been used to identify a subset of high-confidence marker genes in 9986 solid tumor samples from TCGA [38]. Immune cell scores derived from gene measurements were compared with flow cytometry and IHC data. Cell type scores calculated from a list of 60 marker genes measuring 14 immune cell populations were concordant with flow cytometry and IHC readings, and allowed comparisons of immune cell abundance across different tumor types. Further analyses in an immunotherapy data set (derived from patients receiving anti-CTLA-4 antibodies) showed that cell type gene signatures separated responders from nonresponders. Importantly, immune cell scores represent a convenient technique for extracting critical information on the immune contexture of a given tumor in those patients from whom sufficient material for flow cytometry studies is not available [38].

## 3 Immune biomarkers in hematological malignancies

The discovery and validation of immune biomarkers is an area of intense investigation. This section of the article provides examples of individual immune suppressive molecules that could be targeted to improve treatment outcome in patients with leukemia, lymphoma, and MM. We will highlight how online tools could expand our predictive capabilities [39] and support the identification of TME immune gene signatures

and key molecular drivers implicated in the progression of hematological malignancies, and allow the *in-silico* validation of experimental findings across multiple data sets (Table 1 and Figure 1) [41,43,45].

A pan-cancer resource (PREdiction of Clinical Outcomes from Genomic profiles, PRECOG; <http://precog.stanford.edu>) has recently been developed to identify commonalities in prognostic genes from approximately 18,000 human tumors from 166 publicly available cancer data sets with survival outcomes across 39 cancer types, including different types of hematological malignancies [40]. The statistical associations between genes and clinical outcomes were assessed by z-scores, which are directly related to *p* values and represent the number of standard deviations from the mean of a normal distribution. Survival-associated z-scores for individual studies were combined to yield meta-z-scores for the prognostic significance of each gene in each cancer type. One of the two clusters identified was associated with inferior clinical outcomes and was functionally linked to cell proliferation [40]. However, proliferation genes were not adversely prognostic in AML. The other large tumor cluster was associated with favorable survival and was enriched in immunological processes and immune-response genes. A new machine-learning tool, known as CIBERSORT [45], was subsequently applied to PRECOG data to comprehensively map compositional differences in tumor-infiltrating leukocytes in relation to patient outcome. Expression profiles for 22 distinct leukocyte subsets were used as input.

CIBERSORT revealed remarkable differences in relative leukocyte composition between hematopoietic and solid tumors. As shown in Figure 2, CIBERSORT inferred high frequencies of plasma cells in MM specimens and the predominance of B-cell signatures in B-cell malignancies, thereby underpinning its utility for identifying the cell of origin in diverse tumor types [40]. Pooling cancer types allowed the identification of global leukocyte prognostic patterns. Higher frequencies of estimated T cells, especially intra-tumor  $\gamma\delta$  T cells, correlated with superior survival. In contrast, infiltration with polymorphonuclear cell fractions was the most significant adverse prognostic factor. Finally, signatures of polarized M2 macrophages predicted worse clinical outcome than pro-inflammatory M1 macrophages.

**Acute myeloid leukemia.** Immune responses are defective in patients with AML due to the presence of powerful immune suppressive circuits that are activated by soluble factors and immune checkpoint molecules, including PD-L1, TIM-3, and IDO1 [28,46]. Serum kynurenine and tryptophan levels at diagnosis, a measure of systemic IDO1 activity, correlate with patient outcome [47]. Testing of checkpoint blockade is currently being pursued in patients with AML ([www.clinicaltrials.gov](http://www.clinicaltrials.gov), NCT02892318; NCT02508870; NCT02532231; NCT02771197; NCT03065400; and NCT03066648). Although the mutational burden and immunogenicity of AML are inherently low, immunotherapies boosting T-cell functions might be effective, especially in the setting of minimal residual disease, and particularly when combined with checkpoint inhibition or other strategies to overcome leukemia-induced immune dysfunction. Importantly, genetic mutations such as *t(8;21)* and *inv(16)* directly affect the expression of CD200 (a suppressor of macrophage and NK cell function) and CD48 (the ligand for the activating NK receptor CD244), respectively.

**Chronic lymphocytic leukemia.** Chronic lymphocytic leukemia (CLL) is characterized by profound immune defects that are already present in the early stages of the disease and these lead to a heightened vulnerability to severe infections. The frequency of PD-L1-expressing monocytic MDSCs might be significantly increased in untreated CLL patients compared with healthy controls [48]. MDSCs from patients with CLL have been shown to modulate T-cell function *in vitro* and to induce Treg cell differentiation, partly through their expression of IDO1. Plasmacytoid DCs, which play an undisputed role in antiviral immunity as well as antileukemia responses, are reduced in number and function in patients with CLL as a result of decreased expression of FMS-like tyrosine kinase 3 receptor (Flt3) and Toll-like receptor 9 (TLR9) [49]. These represent molecular targets for restoring immune competency. Functional screening assays have identified multiple inhibitory ligands in CLL which impair actin synapse formation in T cells, including CD200, CD270, CD274, and CD276 [50]. Importantly, lenalidomide, an immune-modulatory drug, can downregulate tumor cell inhibitory molecule expression, thus preventing the induction of T-cell defects. Blockade of the PD1 pathway with pembrolizumab has been successfully pursued in patients with CLL and Richter transformation into DLBCL [51]. Objective responses were documented in four out of nine patients with Richter transformation and in 0 out of 16 patients with relapsed CLL. Analyses of pretreatment tumor specimens showed increased expression of PD-L1 and a trend toward increased expression of PD1 in the TME of patients with confirmed clinical responses. All responding patients with Richter transformation had received prior therapy with ibrutinib, a Bruton's tyrosine kinase inhibitor (TKI).

**Chronic myeloid leukemia.** Targeted treatment with TKIs has revolutionized the fate of patients with chronic myeloid leukemia (CML). Intriguingly, TKIs exert a variety of off-target immunological effects (comprehensively reviewed in Ref. [52]), suggesting that novel combinations of molecularly targeted agents and immunotherapies may further improve clinical success rates for CML. Mass cytometry has enabled the identification of prognostic immune biomarkers in longitudinally collected samples from patients with CML receiving TKIs [53]. An increase of circulating CD8<sup>+</sup> cytotoxic T cells occurred after 7 days of TKI therapy and, importantly, changes in single-cell transduction events, including downregulation of phosphorylated CREB S133 and upregulation of phosphorylated STAT3, reflected molecular response at 3 and 6 months.

**Hodgkin's lymphoma (HL).** Classical HL is characterized by a paucity of malignant Hodgkin and Reed-Sternberg cells in lymphoid tissues, accompanied by a massive infiltrate of reactive cells, including leukocytes and stromal cell types. Modulators of innate and adaptive immune responses such as galectin-1 (Gal-1), a member of a highly conserved family of carbohydrate-binding proteins, are overexpressed by Reed-Sternberg cells, thereby leading to depletion of Th1, Th17, and cytotoxic T cells, with an expansion of Treg cells in the TME [54]. Gal-1 levels are elevated in patient serum in association with clinical parameters such as Ann Arbor stage, areas of nodal involvement and International Prognostic Score. In classical HL, tumor-associated macrophage and monocyte signatures in diagnostic FFPE lymph node specimens have been

Table 1. Online resources for the meta-analysis of the prognostic value of immune genes in patients with hematological malignancies.

Database	Link	Diseases	Column 1	Source	Sample numbers	Array	Survival analysis	URL
Bloodspot	<a href="http://servers.binf.ku.dk/bloodspot/">http://servers.binf.ku.dk/bloodspot/</a>	Adult AML	Gene expression	TCGA	183	RNA sequencing	Kaplan-Meier	Network, C. G. A. R. Ref. [44].
Prognoscan	<a href="http://www.prognoscan.org/">http://www.prognoscan.org/</a>	Normal karyotype AML	Gene expression	GSE12417	163 (A and B), 79 Plus	HGU133AB, HGU133Plus2	Cox survival analysis	Ref. [41]
		Relapsed and refractory AML	Gene expression	GSE5122	58	HGU133A		<a href="http://clincancerres.aacrjournals.org/content/13/7/2254.long">http://clincancerres.aacrjournals.org/content/13/7/2254.long</a>
		AML: Response to farnesyltransferase inhibitor (FTI) treatment	Gene expression	GSE8970	34	HGU133A		<a href="http://www.bloodjournal.org/content/111/5/2589.long">http://www.bloodjournal.org/content/111/5/2589.long</a>
		Burkitt lymphoma and DLBCL	Gene expression	GSE4475	158	HGU133A		<a href="http://www.nejm.org/doi/full/10.1056/NEJMoa055351">http://www.nejm.org/doi/full/10.1056/NEJMoa055351</a>
		DLBCL, response to treatment	Gene expression	E-TABM-346	53	HGU133A		<a href="http://europepmc.org/abstract/MED/18615101">http://europepmc.org/abstract/MED/18615101</a>
		FL with/without translocation t(14;18)	Gene expression	GSE16131	184	HGU133AB		<a href="http://www.bloodjournal.org/content/114/4/826.long?ssochecked=true">http://www.bloodjournal.org/content/114/4/826.long?ssochecked=true</a>
		Multiple myeloma	Gene expression	GSE2658	559	HGU133Plus2		<a href="https://www.nature.com/leu/journal/v20/n7/full/2404253a.html">https://www.nature.com/leu/journal/v20/n7/full/2404253a.html</a>
PRECOC	<a href="https://precog.stanford.edu/">https://precog.stanford.edu/</a>	Normal karyotype AML	Gene expression	GSE12417	163 (A and B), 79 Plus	HGU133AB, HGU133Plus2	Kaplan Meier	<a href="https://www.nature.com/nm/journal/v21/n8/full/nm.3909.html">https://www.nature.com/nm/journal/v21/n8/full/nm.3909.html</a>
		AML/RAEB	Gene expression	GSE1427	198	HGU133AB		<a href="http://cgp.iaijournals.org/content/3/3-4/169.abstract">http://cgp.iaijournals.org/content/3/3-4/169.abstract</a>
		FLT3 gene-expression signature in normal karyotype AML	Gene expression	Bullinger, AML, GSE8043	137	HGU133AB		<a href="http://www.bloodjournal.org/content/111/9/4490.long">http://www.bloodjournal.org/content/111/9/4490.long</a>
		Adult AML	Gene expression	ca00119	170	HGU95Av2		<a href="http://www.bloodjournal.org/content/108/2/685.long?ssochecked=true">http://www.bloodjournal.org/content/108/2/685.long?ssochecked=true</a>
		Adult AML	Gene expression	GSE10358	304	HGU133Plus2		<a href="http://www.bloodjournal.org/cgi/pmidlookup?view=long&amp;pmid=18270328">http://www.bloodjournal.org/cgi/pmidlookup?view=long&amp;pmid=18270328</a>
		Adult <i>de novo</i> AML	Gene expression	TCGA	183	HGU133Plus2		<a href="http://www.nejm.org/doi/full/10.1056/NEJMoa1301689#article">http://www.nejm.org/doi/full/10.1056/NEJMoa1301689#article</a>
		Gene expression profiling of CEBPA double and single mutant and CEBPA wild-type AML	Gene expression	GSE14468	526	HGU133Plus2		<a href="http://www.bloodjournal.org/content/113/13/3088.long">http://www.bloodjournal.org/content/113/13/3088.long</a>
		Adult ALL	Gene expression	GSE5314	54	GPL3999		<a href="http://ascopubs.org/doi/abs/10.1200/JCO.2006.09.3534?url_ver=Z39.88-2003&amp;rft_id=ori:rid:crossref.org&amp;rft_dat=cr_pub%3dpubmed">http://ascopubs.org/doi/abs/10.1200/JCO.2006.09.3534?url_ver=Z39.88-2003&amp;rft_id=ori:rid:crossref.org&amp;rft_dat=cr_pub%3dpubmed</a>
		Children with high-risk B-cell precursor acute lymphoblastic leukemia (BCP-ALL)	Gene expression	GSE11877	207	HGU133Plus2		<a href="http://www.bloodjournal.org/content/116/23/4874.long?ssochecked=true">http://www.bloodjournal.org/content/116/23/4874.long?ssochecked=true</a>
		Burkitt lymphoma	Gene expression	GSE4475	221	HGU133A		<a href="http://www.nejm.org/doi/full/10.1056/NEJMoa055351">http://www.nejm.org/doi/full/10.1056/NEJMoa055351</a>
		Burkitt lymphoma	Gene expression	GSE4732	303	GPL3706		<a href="http://www.nejm.org/doi/full/10.1056/NEJMoa055759">http://www.nejm.org/doi/full/10.1056/NEJMoa055759</a>



Table 1. (Continued).

Database	Link	Diseases	Column 1	Source	Sample numbers	Array	Survival analysis	URL
		DLBCL	Gene expression	GSE21846	29	GPL1708		<a href="https://www.ncbi.nlm.nih.gov/pubmed/21633089">https://www.ncbi.nlm.nih.gov/pubmed/21633089</a>
		DLBCL	Gene expression	E-TABM-346	53	HGU133A		<a href="http://europepmc.org/abstract/MED/18615101">http://europepmc.org/abstract/MED/18615101</a>
		Burkitt lymphoma	Gene expression	GSE4475	158	HGU133A		<a href="http://www.nejm.org/doi/full/10.1056/NEJMoa055351">http://www.nejm.org/doi/full/10.1056/NEJMoa055351</a>
		DLBCL	Gene expression	Shipp_DLBCl	58	Hu6800_EntrezCDF		<a href="https://www.nature.com/nmj/journal/v8/n1/full/nm0102-68.html">https://www.nature.com/nmj/journal/v8/n1/full/nm0102-68.html</a>
		DLBCL	Gene expression	Monti_DLBCl	176	HGU133AB		<a href="http://www.bloodjournal.org/con tent/105/5/1851.long?ssochecked=true">http://www.bloodjournal.org/con tent/105/5/1851.long?ssochecked=true</a>
		DLBCL treated with chemotherapy plus Rituximab	Gene expression	GSE10846	420	HGU133Plus2		<a href="http://www.nejm.org/doi/full/10.1056/NEJMoa0802885">http://www.nejm.org/doi/full/10.1056/NEJMoa0802885</a>
		FL	Gene expression	Glas_FL	106 samples (80 patients)	Glas_FL		<a href="http://www.bloodjournal.org/con tent/105/1/301.long">http://www.bloodjournal.org/con tent/105/1/301.long</a>
		FL with/without t(14;18)	Gene expression	GSE16131	184	HGU133AB		<a href="http://www.bloodjournal.org/con tent/114/4/826.long?ssochecked=true">http://www.bloodjournal.org/con tent/114/4/826.long?ssochecked=true</a>
		Mantle cell lymphoma	Gene expression	GSE10793	71	GPL3278		<a href="https://bmccancer.biomedcentral.com/articles/10.1186/1471-2407-8-106">https://bmccancer.biomedcentral.com/articles/10.1186/1471-2407-8-106</a>
		Mantle cell lymphoma	Gene expression	Rosenwald_MCL	101	Rosenwald_MCL		<a href="http://www.sciencedirect.com/science/article/pii/S153561080300028X?via%3Dihub">http://www.sciencedirect.com/science/article/pii/S153561080300028X?via%3Dihub</a>
		Multiple myeloma	Gene expression	GSE6477	162	HGU133A		<a href="http://cancerres.aacrjournals.org/con tent/67/7/2982.long">http://cancerres.aacrjournals.org/con tent/67/7/2982.long</a>
		Multiple myeloma	Gene expression	GSE9782	528	HGU133AB		<a href="http://www.bloodjournal.org/con tent/109/8/3177.long">http://www.bloodjournal.org/con tent/109/8/3177.long</a>
		Multiple myeloma	Gene expression	GSE24080	559	HGU133Plus2_EntrezCDF		<a href="https://breast-cancer-research.bio medcentral.com/articles/10.1186/bcr2468">https://breast-cancer-research.bio medcentral.com/articles/10.1186/bcr2468</a>
	PROGeneV2 <a href="http://watson.compbio.iupui.edu/chtray/proggene/data base/index.php">http://watson.compbio.iupui.edu/chtray/proggene/data base/index.php</a>	DLBCL treated with chemotherapy plus Rituximab	Gene expression	GSE10846	420	HGU133Plus2	Kaplan Meier	<a href="http://www.nejm.org/doi/full/10.1056/NEJMoa0802885">http://www.nejm.org/doi/full/10.1056/NEJMoa0802885</a>
		FL with/without t(14;18)	Gene expression	GSE16131	184	HGU133AB		<a href="http://www.bloodjournal.org/con tent/114/4/826.long?ssochecked=true">http://www.bloodjournal.org/con tent/114/4/826.long?ssochecked=true</a>
		Normal karyotype AML	Gene expression	GSE12417	163 (A and B), 79 Plus	HGU133AB, HGU133Plus2		Ref. [41]
		CLL	Gene expression	GSE22762	44 (A and B), 107 Plus	HGU133AB, HGU133Plus2		<a href="https://www.nature.com/leu/journal/v25/n10/full/leu2011125a.html">https://www.nature.com/leu/journal/v25/n10/full/leu2011125a.html</a>



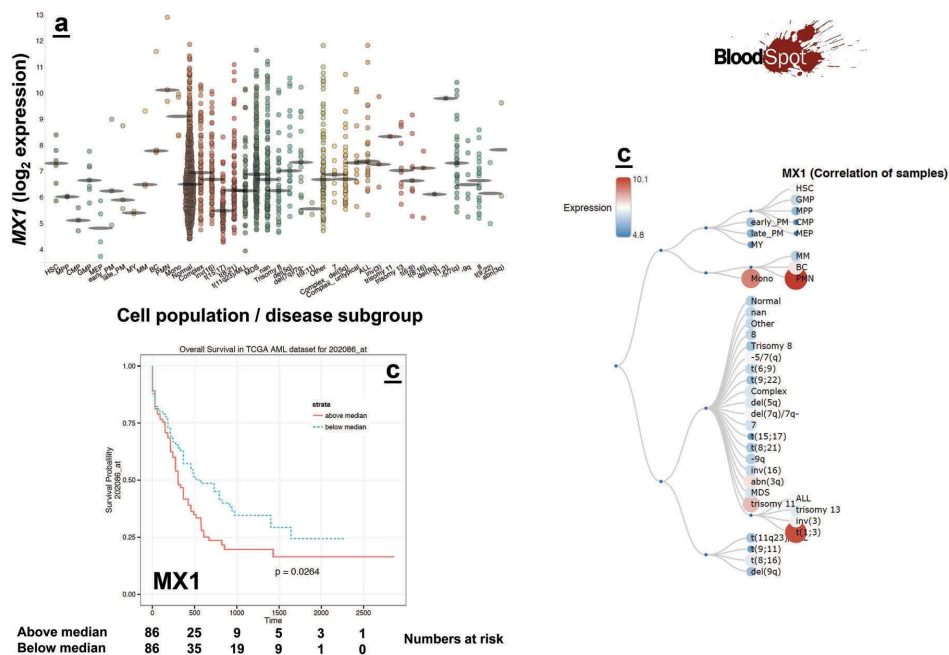


Figure 1. Identification of potential immune biomarkers using publicly-available on-line tools.

Blood-Spot (<http://servers.binf.ku.dk/bloodspot/>) provides plots of gene expression in normal and malignant hematopoietic cells at different maturation stages based on curated microarray data [42]. We selected MX1 (Myxovirus [Influenza] Resistance 1), an interferon (IFN)-inducible gene, as an example of use of Blood-Spot to interrogate human AML data sets. Panel A: mRNA expression levels are depicted across a broad range of normal hematopoietic differentiation stages (first 11 columns on the left; data derived from Gene Expression Omnibus Series GSE42519) and in patients with different cytogenetic subgroups of AML (data derived from Gene Expression Omnibus Series GSE13159, GSE15434, GSE1804, GSE14468, and from The Cancer Genome Atlas [TCGA]). HSC = hematopoietic stem cell; MPP = multi-potential progenitor; CMP = common myeloid progenitor; GMP = granulocyte-monocyte progenitor; MEP = megakaryocyte-erythroid progenitor; PM = promyelocyte; BC = band cell; MM = metamyelocyte; MY = myelocyte; MDS = myelodysplastic syndrome; NA = not available.

Panel B shows an interactive hierarchical tree summarising the relationship between the samples displayed. Expression level are visualized by size and colour of the nodes, as intuitively indicated by the colour legend. The full name of cell type abbreviations can be obtained by moving the mouse over the individual nodes. Moreover, nodes can be clicked to collapse a branch of the tree.

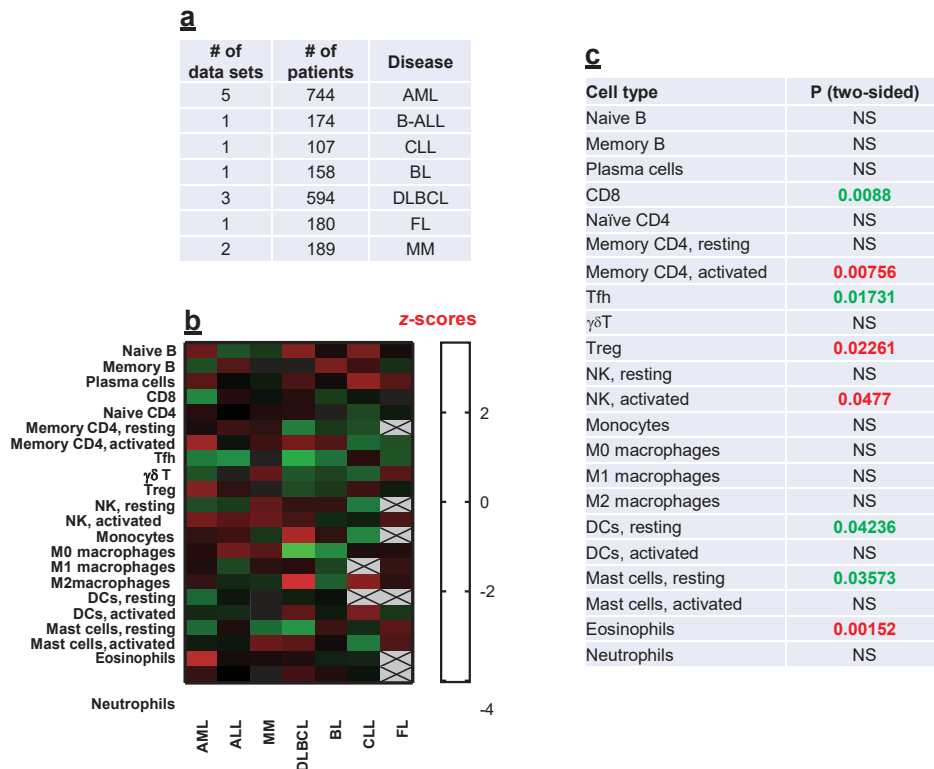
Panel C shows a survival plot (Kaplan Meier analysis) based on a high-quality AML dataset from TCGA. MX1 expression levels were dichotomized (above or below median). Other built-in tools allow the removal of cell populations from the graphs, the export of plots as a PDF file and the comparison of paired populations in the default expression plots using the Student's *t* test.

associated with high risk of primary treatment failure and with decreased PFS and OS [55,56]. Among the 27 individual genes with a discriminative power for outcome prediction exceeding that of the best clinical variable (patient age), matrix metallo-peptidase-1 was overexpressed in patients with treatment failure.

**Non-Hodgkin's lymphoma (NHL).** NHLs are typically associated with chronic inflammatory and autoimmune conditions, with severe immune dysregulation being an established risk factor and a hallmark of the disease. For instance, high pretreatment plasma levels of CXCL13, IL-6, and IL-10 predict worse PFS and OS in patients with AIDS-related NHL (AIDS-NHL) receiving intensive multi-agent chemotherapy and immunotherapy with rituximab [57]. Longitudinal monitoring of cytokine levels 1–5 years preceding NHL diagnosis has identified cytokines and other molecules associated with chronic immune activation, such as IL-6, IL-10, and TNF- $\alpha$ , as predictors of the development of systemic AIDS-NHL [58,59]. Similarly, circulating levels of B-cell attracting chemokine 1, soluble TNF receptor 2, and soluble vascular endothelial growth factor 2 have been

correlated with the risk of NHL in advance of diagnosis [60]. Similarly, genetic variants of TLR9 which lead to increased transcriptional activity in mononuclear cells might increase NHL susceptibility [61]. Finally, three independent population-based case-control studies have revealed a correlation between NHL risk and single-nucleotide polymorphisms within 12 innate immunity genes, including IL-1 receptor antagonist and IgG Fc receptor 2A [62].

**Follicular lymphoma (FL).** FL is the second most common type of NHL, accounting for approximately 20% of all cases. The malignant B cells in FL are of germinal-center origin. FL is clinically heterogeneous, with some patients experiencing an indolent clinical course and others having rapidly progressive disease. A multivariate model of survival was constructed using whole-genome microarray data from lymph node tissues from 191 patients with untreated FL [63]. This study identified two distinct immune response gene signatures, immune-response 1 and immune-response 2, which reflected the biological characteristics of the nonmalignant immune cells within the biopsy specimens and were molecular predictors of the length of survival in patients with FL. The immune-



**Figure 2.** Immune PRECOG; correlation between immune gene expression levels and survival in haematological malignancies (<https://precog.stanford.edu/about.php>). Details about available data sets, patient numbers and disease type are provided in panel A. AML = acute myeloid leukaemia; B-ALL; B-cell precursor acute lymphoblastic leukaemia; CLL = chronic lymphocytic leukaemia; BL = Burkitt lymphoma; DLBCL = diffuse large B-cell lymphoma; FL = follicular lymphoma; MM = multiple myeloma.

Panel B summarises the correlation between publicly available immune gene expression levels and overall survival (z-scores). Grey boxes in the heat map denote missing values. The 22 immune cell populations shown here were identified by Newman and co-workers based on the expression of 'signature genes' [40,45]. Tfh = follicular helper T cells; Treg = regulatory T cells; DCs = dendritic cells; NK = natural killer.

Panel C shows two-sided  $p$  values that were calculated from z-scores. Green denotes correlation with better clinical outcome and red indicates correlation with worse clinical outcome. The abundance of immune cell populations was inferred from transcriptomic data sets using a recently developed analytical tool (CIBERSORT; Cell type Identification By Estimating Relative Subsets Of known RNA Transcripts). DCs = dendritic cells; NS = not significant.

response 1 signature included genes associated with T cells and genes which were highly expressed in macrophages. Genes in the immune-response 2 signature were preferentially expressed in macrophages and DCs. Importantly, the gene expression-based model predicted patient survival independently of clinical variables such as the International Prognostic Index (IPI) and the presence or absence of B symptoms [63].

Other immune cell types, such as tumor-associated mast cells and tumor-associated macrophages, have prognostic importance in FL. Mast cell infiltration was detected using immunohistochemistry and was shown to negatively affect PFS in patients with FL receiving a combination of immunotherapy (rituximab) and chemotherapy (CHOP) [64]. The prognostic impact of mast cell infiltration was again independent of the FL IPI. The mechanisms by which mast cells reduce the efficacy of antibody-based therapies in FL remain to be determined and might include the negative regulation of macrophage activity and antibody-dependent cellular cytotoxicity through the expression of Fcγ receptors which can engage rituximab [64].

*Diffuse large B-cell lymphoma (DLBCL).* DLBCL is the most common subtype of NHL, representing more than 30% of all adult NHL cases diagnosed in Western countries, and is characterized by an aggressive clinical course. In spite of improved response and survival rates after the addition of rituximab to the therapeutic armamentarium, up to 40% of patients with DLBCL experience relapse and have a poor prognosis.

Gene expression profiling and next-generation sequencing have been instrumental to the identification of molecular subtypes of DLBCL, which are not obviously related to histological subtypes of DLBCL and are associated with a remarkable divergence in clinical behavior. Patients with activated B-cell-like (ABC) gene signatures have a shorter survival compared with patients with the other two molecular subtypes, that is, germinal center B-cell (GCB) and primary mediastinal B-cell lymphoma signatures [65]. Non-GCB type DLBCLs are enriched with PD-L1-expressing tumors and might benefit from targeted immunotherapies [66].

DLBCLs have a heterogeneous immune infiltrate, which includes macrophages, DCs, NK cells, T-cell subsets, and B

cells. Interestingly, pretreatment gene expression of *CD68* as well as immunohistochemically defined *CD68*<sup>+</sup> macrophages might correlate with better outcome in patients with DLBCL receiving chemo-immunotherapy, independently of IPI scores or molecular subgroups [67]. In contrast, macrophage infiltration was negatively correlated with OS in patients treated without rituximab, leading to the hypothesis that rituximab administration might switch macrophage profile toward a tumor-promoting phenotype.

Tissue microarray immunohistochemistry with automated scoring of FoxP3, CD68, and micro-vessel (CD34) density (MVD) has been shown to stratify patients with DLBCL into risk groups and to predict prognosis [68]. Patients in the high-risk group had significantly worse EFS and PFS, suggesting that TME components should be considered as an important tool to predict patient survival. The NanoString digital hybridization approach for RNA quantification has been employed to detect immune effector and checkpoint genes in FFPE biopsies from patients with DLBCL [69]. The product of the immune effectors (*CD4* × *CD8*) in a ratio with the product of checkpoints (*PD-L1* × *M2* macrophages) was used to identify low-immune and high-immune groupings of patients with significant differences in 4-year survival. Patients with a GCB or an ABC molecular subtype of DLBCL and a high immune ratio had a significantly extended survival compared with GCB and ABC patients with a low immune ratio, suggesting that the balance of anti-tumoral immunity, that is, the ratio of immune effector cells to negative checkpoint molecules, might have an important prognostic value in DLBCL.

**Primary mediastinal large B-cell lymphoma (PMLBCL).** PMLBCL, a distinct and uncommon subtype of DLBCL, is more frequent in young females and originates in the mediastinum, presenting with features of local invasion [70]. Aberrations consisting of structural genomic rearrangements, missense, nonsense, and frame-shift mutations involving the major histocompatibility complex (MHC) class II trans-activator *CIITA* have been detected in approximately 50% of patients with PMLBCL [71]. Genomic lesions in *CIITA* resulted in decreased protein expression and reduction of MHC class II surface expression, favoring the establishment of an immune-privileged microenvironment in PMLBCL.

PMLBCL has a unique transcriptomic signature which is close to classical HL and is characterized by constitutive expression of *PD-L1* and *PD-L2*. Amplification and/or translocations involving chromosome 9p24.1, a region that includes *PDCD1LG2*-encoding *PD-L2*, are a common event in PMLBCL but not in DLBCL [72]. This observation entails that PMLBCLs might be susceptible to *PD1* blockade. A recent clinical trial run as part of the KEYNOTE-013 multicenter phase 1b study has shown decreases in target lesion in approximately 80% of patients evaluable by imaging [73]. Overall, median survival was not reached for treated patients. Drug-related adverse effects were observed in 60% of the patients and were manageable. Other immune suppressive circuits in patients with PMLBCL include the downregulation of *HLA-DR* expression and the decrease of cytotoxic *CD8*<sup>+</sup>*TIA1*<sup>+</sup> T cells, features which correlate with shorter PFS [74].

**Multiple myeloma (MM).** Patients with MM suffer from severe and complex defects of humoral and cellular immunity,

including an increased production of immune suppressive cytokines [75] and an expansion of immune regulatory cell types [76]. *IL-17*, *IL-21*, *IL-22*, *IL-23*, and *Th17* cells are increased in patients with MM compared with healthy donors [77]. In particular, *IL-17* might promote MM growth, colony formation, and development in a murine xenograft model.

*PD1* and its ligands are broadly expressed in the TME of MM, in which they may mediate immune evasion mechanisms [78]. Similarly, *PD-L1* expression, as well as *IDO1* function, is increased in patients with MM compared with healthy controls [79,80]. Of interest, *PD-1/ID-1* blockade may abrogate bone marrow stromal cell (BMSC)-induced MM growth, an effect which is further potentiated by lenalidomide and correlates with the induction of intracellular expression of *IFN-γ* and granzyme B in effector cells. BMSCs from patients with MM also inhibit the lysis of MM cells in a cell contact-dependent fashion by inducing the expression or surviving, a caspase-3 inhibitor, and downregulating *CD95* expression [81].

A thorough characterization of T cell, DC, and NK cell phenotypes has demonstrated a decreased expression of T-cell activation markers, Th1 cells, and proliferation markers in patients with high-risk 'smoldering' MM compared with healthy controls [82]. The fact that treatment with the immune-modulating drug lenalidomide translated into an increase of functionally active T cells, even when combined with low-dose dexamethasone, suggests that immune modulatory drugs might delay the progression of smoldering MM to overt MM.

Finally, MM can avoid immune surveillance via the transfer of membrane proteins in a process known as trogocytosis [83]. For instance, *CD86* and *HLA-G* from malignant plasma cells can be acquired by T cells residing in the BM compartment. *HLA-G*-expressing T cells exhibited a regulatory potency similar to that of natural Treg cells. Interestingly, the association of *CD86* or *HLA-G* expression with a poor prognosis suggests the induction of *in vivo* immune suppression.

## 5 Future immunotherapy approaches for hematological malignancies

T-cell engineering with synthetic chimeric antigen receptors (CAR) is revolutionizing current treatment paradigms for patients with B-cell malignancies. Durable clinical responses up to 24 months were induced by *CD19*-directed CAR T cells in 90% of children and adults with relapsed or refractory B-cell acute lymphoblastic leukemia (ALL) [84]. Remissions caused by *CD19*-specific CAR T cells were correlated with high serum levels of *IL-15* in patients with lymphoma [85]. *CD30*-specific CAR T cells have been safely and successfully administered to patients with HL [86]. Clinical responses to CAR T cells could be improved by targeting tumor-induced immune suppression with pembrolizumab [87] or by antagonizing *IDO1* activity with lymphodepleting drugs such as fludarabine and cyclophosphamide [88]. Innovative approaches are currently being developed to target T-cell malignancies with *CD7*-specific CAR T cells [89] and to eradicate antigen-loss relapses of myeloid malignancies with dual *CD19-CD123*-redirected CAR T cells [90]. Anti-myeloma activity of CAR T cells specific for B-cell maturation antigen has recently been shown in one patient with chemotherapy-resistant disease [91]. Intriguingly, clinical responses have been

achieved using CD19-specific CAR T cells in one patient with MM despite the absence of CD19 expression on malignant plasma cells [92]. Finally, a phase I clinical trial in 16 patients with relapsed or refractory B-cell malignancies (MM, NHL, and CLL) has shown complete clinical responses after the infusion of CAR T cells specific for malignancy-associated K light chains [93].

Bi-specific antibody constructs are also being implemented in patients with advanced acute leukemia and with NHLs. Treatment with blinatumomab, a CD3-CD19 bi-specific T-cell engager antibody, has resulted in significantly longer median OS than chemotherapy (7.7 months vs. 4.0 months) in a randomized clinical trial in adults with relapsed or refractory ALL [94]. Blinatumomab induces the expansion of both naïve and memory CD4<sup>+</sup> and CD8<sup>+</sup> T cells in patients and might skew T-cell receptor *repertoires* [95]. Immune biomarkers which predict clinical responses to blinatumomab have not been identified yet. Interestingly, PD-L1 expression levels may be higher in children with ALL refractory to blinatumomab [96].

Evidence from clinical trials in patients with solid tumors suggests that combination strategies that synergize with immune checkpoint blockade might be more effective than single-agent immunotherapy, as reviewed elsewhere [97]. It is anticipated that the rational development of personalized combination immunotherapy approaches for patients with hematological malignancies will be informed by the discovery and validation of immune biomarkers.

## 6 Multiplexed tissue biomarker imaging

The direct assessment of immune phenotypes and their spatial relationship by multiplexed techniques provides essential information which is highly complementary to gene expression profiling and may allow the discovery of composite predictive biomarkers [32].

Multiplexed immunofluorescence allows the detection of up to 30 proteins in regions of interest within the TME. Multiple fluorophores can be applied on a single tissue section and are interrogated using a multispectral microscope [11,98]. This technology enables a comprehensive characterization of the topography and spatial relationship between tumor cells and microenvironmental cell types, including immune cells. Of relevance, the density of CD8<sup>+</sup> T-cell infiltrates in the invasive margins of melanoma lesions has been associated with expression of the PD1/PD-L1 immune inhibitory axis and with clinical responses to anti-PD-1 immunotherapy [23]. Quantitative image analysis could also be valuable in dissecting the spatial distribution of DCs at different maturation stages within the tumor-draining lymph nodes, thus providing insights into actionable circuits of immune dysfunction [99].

NanoString Technologies (Seattle, USA) has recently developed a multiplexed immune profiling approach to measure the expression of up to 800 targets at protein and RNA level on a single FFPE tissue slide [100]. This Digital Spatial Profiling platform allows the analysis of tumor geography and the delivery of digital counts of biomarker expression with single-cell resolution. It is expected that multiplexed technologies can be applied to the investigation of immune cell distribution in tissue biopsies from patients with hematological malignancies.

However, the extensive data that are generated with the use of the above technologies will need to be integrated and 'converted' into useful information using novel bioinformatics approaches.

## 7 Machine learning

Advances in bioinformatics have led to a vast amount of data being generated at an accelerated pace. Next-generation RNA and DNA sequencing methods is providing access to incredibly detailed information on entire genomes and allowing us to interrogate more potential biomarkers with an increased level of accuracy. This massive volume of data creates a problem of complexity which makes it impossible to use traditional methodologies.

Machine learning is an interdisciplinary field of bioinformatics which employs a data-driven class of algorithms to find solutions to a given problem by studying, for example, gene expression patterns across many cases/patients. Although widely and successfully used in biology and biomarker discovery studies, the use of these approaches in hematological malignancy studies has, to date, been extremely limited.

Many approaches have been developed, each of which will be explained in terms of their utility here. These approaches can be broadly characterized in two distinct groups; supervised and unsupervised machine learning.

### 7.1 Supervised learning

Supervised learning approaches are widely applied and use source features to predict a target class [101]. The supervised approach allows the algorithm to train itself by detecting patterns in large data sets that are predictive of the target class, for example, how does *IFNG* behave in acute myeloid leukemia compared to acute lymphoblastic leukemia? We can make use of previous studies and adjust the algorithm parameters so that it accounts for this information. One major advantage is that such approaches are tolerant of the highly complex, nonlinear and noisy data that are often found in biological systems.

#### 7.1.1 Artificial neural networks

Artificial Neural Networks (ANNs) are statistical models emulating the function of a network of human neurons for the purposes of encapsulating information in order to analyze large, complex data sets. The learning process is based on the mathematical interconnections between the processing elements that constitute the network architecture [102]. This allows them to classify cases based on data by assigning a numerical weight value to each input and adjusting them as they sample the data, effectively learning the optimal solution. The main advantages of ANNs include their high fault and failure tolerance, scalability, and consistent generalization ability, all of which allow them to effectively predict or classify new, fuzzy, and unlearned data [102,103]. Additionally, they have been recently used to create panels of biomarkers that, when used in conjunction with each other, predict breast cancer [104].

The original ANN architecture, as proposed by Rosenblatt in 1958, was based on the concept of a single artificial processing neuron with an activation threshold, adjustable weights and bias. However, this could only be used for the classification of linearly separable patterns, as it only learns when an error occurs during testing. This is rarely the case with complex problems such as cancer, as patients do not typically fall into a standard distribution and variance in the data is often significant. Typically, ANNs make use of a multilayer perceptron which is made up of multiple perceptrons arranged in layers of three or more, consisting of input, hidden, and output layers. These consider the predictor variables, perform feature detection through an activation function, and output the results of the algorithm respectively.

ANNs have been successfully used to predict and classify data in different contexts, such as early detection [105], prediction of long-term survival [106] and biomarker discovery in breast cancer [104,107], classification of CRC tissues [108], and discrimination between benign and malignant endothelial lesions [109]. One of the major disadvantages of ANNs is their liability to overfit when the parameters have not been optimized. Moreover, they often receive criticism for their 'black box' approach which allows for little to no interpretation of the results and process.

#### 6.1.1 Support vector machines

Support vector machines (SVMs) are supervised classification and regression algorithms that are primarily designed to solve binary problems. They are focused on finding a hyperplane which separates two classes [110] and have been successfully used in pattern recognition and classification. The popularity of SVMs is a result of the availability of a large variety of kernels (functions that separate data) which can be broadly split into linear, polynomial, sigmoid, and radial basis function categories. The greatest advantage of SVMs when compared to similar machine learning methods is that selecting the correct kernel function enables the analysis of nonlinear data and overcomes the curse of dimensionality. However, the introduction of more features increases the complexity, and therefore the computing power required. Notwithstanding the practical issues, SVMs have been used for analyzing high density data, such as RNA, miRNA, and proteomics, and they remain one of the most popular classification methods, especially for cancer prediction and prognosis [111–114].

As indicated above, disadvantages of SVMs include the computational processing power and the time, although much like ANNs, these problems are quickly being addressed. A more crucial issue facing the application of SVMs is choosing the appropriate parameters and kernel that will allow for sufficient generalization because of the high algorithmic complexity which is required for 'real' data. As a result, the use of SVMs is less supported in settings which require interpretation and decision-making [110].

#### 6.1.2 Decision trees and random forests

Tree-based methods involve stratifying a data set into multiple categories (similar to hierarchical clustering) that can then be used to predict possible outcomes based on the values of the input variables. These methods can be used for both

classification and regression problems. Decision tree classification algorithms pose a series of questions based on the features of the data set and train to split those features into separate categories, thereby resulting in a dendrogram.

Although the advantages of these methods are that they are computationally efficient, have good predictive values, and their results are easy to interpret, their predictive accuracy tends to be lower than their counterparts. To mitigate this issue, methods such as random forests, bagging, and boosting are used to construct multiple trees in parallel. These can then be combined to provide a significant boost to their prediction accuracy at the cost of some of their interpretability.

#### 6.1.3 Bayesian networks

Bayes theory states that the conditional probability of A given B is the conditional probability of B given A scaled by the relative probability of A compared to B. Using Bayesian networks, the association between a set of variables or nodes can be determined through joint conditional probability distributions [115].

Although such approaches have been used for multiple biological applications such as inferring cellular networks, modeling protein signaling pathways, data integration, genetic data analysis, and classification [116–118], they are limited by the fact that they need larger than average data sets to obtain sufficient prior probabilities to produce an accurate outcome. This in turn makes them extremely computationally expensive. Moreover, they tend to perform poorly on high-dimensional data and their output tends to be complex and as such, can be hard to interpret for nonspecialists. Finally, it should be noted that Bayesian networks are not truly Bayesian in nature. They simply adhere to the basic rules of Bayesian statistics on probabilistic inference. It would be more accurate to say that Bayesian networks are directed graphical models with Bayesian elements.

### 7.2 Unsupervised learning

Unsupervised machine learning approaches are used when the desirable or predefined output is not available. The goal of unsupervised learning problems is to discover the structure of the data and define groups of similar examples, commonly called clustering. Clustering is one of the main unsupervised approaches and it functions by assigning data points to natural categorical classes or groups, based on similarity or difference of patterns without prior training [119].

Unsupervised learning approaches are best used when the subject is a very large data set with few known variables. This allows the user to find natural patterns in the data and discover novel groups that have not been previously established and using which training can be undertaken. They have been most commonly used to distinguish patterns in microarray data by clustering genes based on their expression levels [120–122].

#### 7.2.1 Hierarchical clustering

Hierarchical clustering, the most common unsupervised learning technique, has been widely used for the analysis of microarray data. It is based on measuring distances between data points and defining the first instance of each point as a single



cluster, followed by merging the clusters according to distance, with smaller distances between clusters indicating greater similarity. The process continues in an iterative manner until all samples have been used to produce a phylogenetic tree-like structure of the clusters (dendrogram), with individual samples at the bottom, and a cluster containing every element in the data set at the top [119]. Some of the most popular methods to determine cluster hierarchy include Single-linkage, Complete-linkage, Average-linkage, and Centroid distance.

The major limitation of the hierarchical clustering approach is that as the clusters grow, they might not be representative of the objects within, and it is hard to rectify mistakes that occur early in the clustering process.

### 7.2.1 K-means clustering

Much like hierarchical clustering, K-means clustering is a partition algorithm which works by arbitrarily grouping objects into a predetermined number of clusters in an iterative manner. The centroid-average expression of each cluster is assigned randomly, based on the Euclidean distance between each object and the closest cluster average. The algorithm then recalculates the average centroid expression, based on the mean of all objects assigned to it, and repeats the process until convergence is reached, where the average expression of each cluster does not change significantly [119]. Unlike hierarchical clustering, this method has the advantage of being able to deal with large data sets and as a result has been applied to more complex problems. However, the major drawback of this method is that repeating the test can produce significantly different results, as the final assignment of clusters is dependent on the initial random assignment of objects [123].

### 7.2.2 Principle component analysis

Reduction in dimensionality is often necessary for a visual inspection of high-dimensional data, as the number of variables being investigated often exceed the number of samples. This leads to data points being scarcely distributed in a high-dimensional feature space [124]. The aim of principle component analysis is to map the original data into its principle components by linearly transforming the data to reduce dimensionality. These principle components are orthogonally arranged, mutually uncorrelated linear combinations of the original variables, and are often ranked by the amount of variance they can explain in the data. The highest ranked components contain most of the relevant information, whereas low ranked principle components can be removed if they are not required. This approach is often used as a visualization tool and preprocessing step for classification and clustering [119].

## 7.3 Novel approaches

Two bioinformatics approaches developed recently have managed to provide novel solutions to common problems related to big data analysis.

### 7.3.1 CIBERSORT

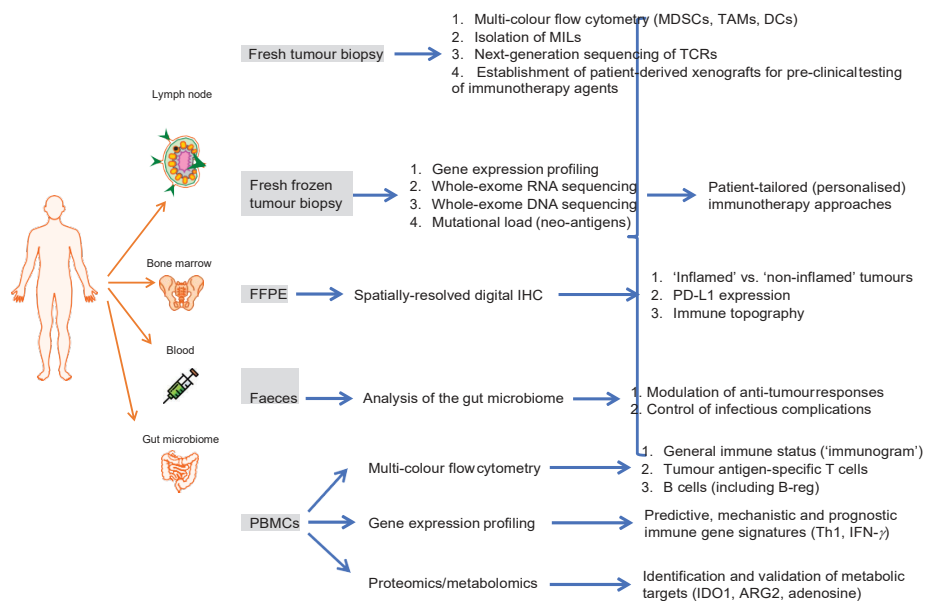
CIBERSORT is a platform for characterizing the cell composition of tissues based on their gene expression profiles [45]. Traditionally, immunohistochemistry and flow cytometry have been used to answer such questions and, although highly successful, they are limited by their reliance on known markers as well as the fact that these techniques are harmful to cells, likely altering the results. CIBERSORT manages to achieve similar results to these techniques using the RNA mixtures of the desired tissue. It is an SVM regression algorithm which allows the user to differentiate cell types in large data sets. CIBERSORT has been proven to have superior performance and be substantially more accurate over traditional machine learning methods when the samples studied were unknown, noisy, or closely related. However, limitations include its reliance on a reference database, the fidelity and size of which are considerable factors in the algorithm's ability to classify the cell samples, the lack of a  $p$  values for detection limits and a systematic over- and underestimation of certain cell types. Much like all major machine learning approaches, these problems are being mitigated as more computing power becomes available and the size and fidelity of databases increases.

### 7.3.2 Hive plots

One of the key challenges in the field of bioinformatics is the issue of visualization. Although the approaches discussed previously have expanded the field of biomarker discovery by allowing researchers to consider new possibilities, their use in diagnostics is limited by the fact that the results often require expert specialists to interpret. If these approaches are to achieve widespread use by clinicians for prognosis, it is paramount to have a clear and easily understandable output. Developed by Krzywinski et al. [125], hive plots offer an alternative network visualization method to traditional maps. These maps, usually produced by software such as Cytoscape, Gephi, Netminer and more recently, programming languages such as R, have a tendency to include an overwhelming amount of information, leading to networks that need to be analyzed with sorting algorithms to be readable and hard to interpret. Moreover, complexity increases exponentially as more information is included. Hive plots offer a rational visualization technique which groups nodes based on specific properties determined by the user. The properties can be inherent network statistics, or information such as features of clinical data.

## 8 Expert commentary

A patient's immunological profile should be considered a highly dynamic framework, which is affected by variations in tumor genetics, epigenetics and micro-RNA expression, age, microbiome composition, pharmacological agents, and environmental factors including infections and exposure to sunlight [21]. There is an emerging need to identify immune biomarkers of cancer response to immunotherapies [39]. High-dimensional technologies will also enhance our understanding of TME-cancer interactions and will support the prediction of therapeutic



**Figure 3.** Approaches to immune biomarker discovery in patients with haematological malignancies. Blood, bone marrow and lymph node samples should be interrogated using genomics and proteomics approaches, immunohistochemistry and flow cytometry to collect comprehensive and personalised profiles on neo-antigen expression, topography and functional orientation of immune cells, tumour specificity of T cells and prognostic immune gene signatures. Lymphoid tissue-resident T cells hold promise as immune effector cells for immunotherapy clinical trials, analogous to the tumour-infiltrating T cells from patients with melanoma [126], in light of recent evidence that *ex vivo*-expanded marrow-infiltrating lymphocytes (MILs) can be safely administered to patients with high-risk myeloma early after autologous CD34-selected haematopoietic stem cell transplantation [127]. Patients on immunotherapy clinical trials should be sampled sequentially in order to discover and validate mechanistic immune gene signatures associated with response to treatment and/or failure to respond. The gut microbiome could be manipulated to optimise immunotherapeutic responses to checkpoint blockade, as reviewed elsewhere [128].

FFPE = formalin-fixed paraffin-embedded; PBMCs = peripheral blood mononuclear cells; MDSCs = myeloid-derived suppressor cells; TAMs = tumour-associated macrophages; DCs = dendritic cells; MILs = marrow-infiltrating lymphocytes; IHC = immunohistochemistry; B-reg = regulatory B cells; TCRs = T-cell receptors; Th1 = T-helper type 1; IDO1 = indoleamine 2,3-dioxygenase-1; ARG2 = arginase 2.

benefit from immune-based interventions (Figure 3). Immune assays for biomarker discovery, as well as sample collection and handling, must be harmonized and standardized for investigators to be able to compare and share results [3].

Although the role of immune gene signatures in stratifying patients with haematological malignancies and in supporting clinical decision-making remains to be investigated, efforts are being devoted to the discovery of prognostic signatures (to predict outcome independent of therapy), predictive signatures (to assist in treatment selection according to therapeutic effectiveness), and mechanistic immune signatures in patients with solid tumours [129,130]. Prognostic signatures help predict outcome independent of therapy, whereas predictive biomarkers and signatures (before treatment) might assist in treatment selection according to therapeutic effectiveness. Mechanistic signatures should capture the maximal intensity of immune responses which occur in tumor lesions that are about to regress after immunotherapy administration [129]. Importantly, comprehensive analyses have indicated that prognostic, predictive, and mechanistic immune signatures across different immunotherapeutic strategies might overlap qualitatively and converge into a common pathway [129]. It is becoming evident that solid tumours which are responsive to treatment generally have an inflammatory status, indicative of preexisting immune responses, as well as expression of cytolytic markers with concomitant counter-activation of immune suppressive and immune escape circuits, which should be

targeted with rational combinatorial approaches (for instance, PD-L1 blockade coupled with small-molecule IDO1 inhibitors [131]).

Because of inherent limitations of gene expression profiles, other approaches, such as flow cytometry, quantitative immunohistochemistry, and next-generation sequencing for T-cell antigen receptors or similar technologies (multi-N-plex quantitative PCR, spectratyping, and immune phenotyping) are recommended to thoroughly characterize the immunological landscape of the TME and to establish predictive models [23], as recently reviewed by the Immune Biomarkers Task Force of the Society for Immunotherapy of Cancer [11]. Conceivably, the analysis of multidimensional data sets will be instrumental to mapping the immunological landscape of haematological malignancies, to revealing potential immune biomarkers and informing the rational design of immune therapies. A combination of personalized transcriptomic and proteomic measurements will likely be required to develop accurate immune gene signatures in individual patients (Figure 4). The collection of comprehensive immunological profiles or 'cancer-immune set points' will inform personalized clinical trials and support the prediction of anticancer responses to immunotherapy [21].

## 7 Five-year view

Immune profiling of patients with haematological malignancies is expected to underpin the discovery and validation of new

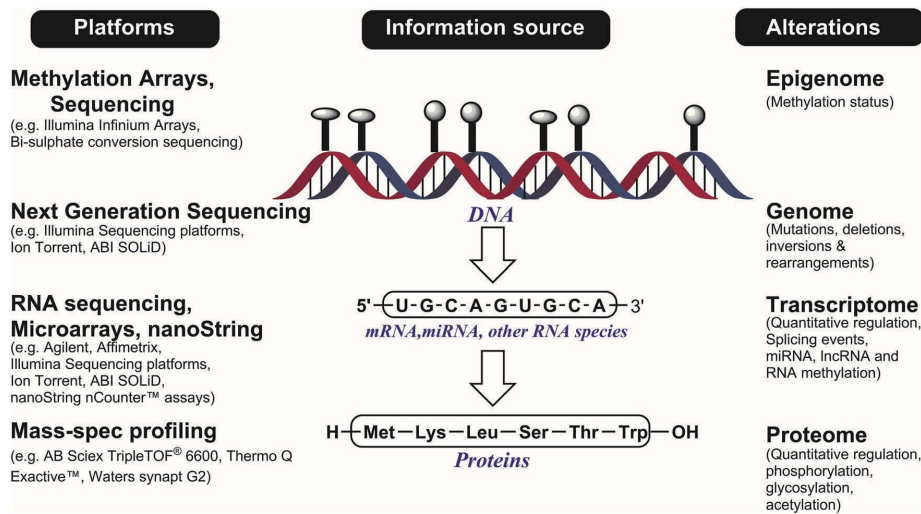


Figure 4. Technological platforms underpinning the discovery and validation of immune gene signatures in patients with cancer.

biomarkers, and to foster the clinical implementation of a more refined and personalized approach to immune-based interventions. Immune parameters could be used to build dynamic frameworks and to support treatment allocation to cancer patients, such as the recently proposed ‘cancer immunogram’ [2], the aim of which would be to visualize the state of cancer-immune interactions in individual patients with cancer and to discuss treatment options in a personalized manner. The information required to build a cancer immunogram should include tumor foreignness, patients’ immunological status, evidence for tumor infiltration with T cells, expression of checkpoints and other molecules inhibiting T-cell function, and tumor cell sensitivity to immune effectors, including the inactivation of antigen processing machinery components [2]. The above parameters should be collected from the blood and/or tumor tissues using transcriptomic approaches, high-resolution immune phenotyping, spatially resolved immunohistochemistry, and standard immunological assays [11].

Strategies that combine different methods of capturing the immunological status of the TME may particularly support the development of composite predictive biomarkers for immune checkpoint inhibition in the Hematology clinic, an area that is expected to flourish during the next few years [32]. For example, gene expression profiling approaches, such as nanoString Technologies’ digital platform [100], coupled with multiplexed immunohistochemistry techniques, will allow investigators to quantify mRNA species and multiple proteins expressed in cell populations within morphologically defined regions of interest in the TME, thus providing crucial information about the topography and spatial localization of immune cells at different tumor stages or after treatment with immunotherapies.

Finally, new bioinformatics approaches are being developed to unravel the complexity and multidimensionality of data sets obtained through transcriptomic, sequencing, and proteomic techniques, to identify responders and

nonresponders and to stratify and select patients based on immune gene signatures in the TME [132]. In the foreseeable future, immune biomarkers might guide the development and personalization of combination immunotherapy approaches [10]. As machine learning is becoming an integral part of biomarker discovery, it presents its own set of challenges with the first one being the constant need for higher computational power. As the size of the available data sets and the complexity of the platform technologies (e.g. the move to 1million SNP probes on a chip, or the advent of RNA deep-Seq. studies) increases, computational requirements will increase exponentially. While current advances in GPU-accelerated parallel computing, solid-state drives and the availability of highly parallel cloud computing solutions have allowed for a significant increase in processing power, it is proving insufficient to handle some of the more complex questions. There is also a trend occurring where the processing power increases so the analyses that are conducted become deeper and more detailed.

The quality and size of the data sets is a key factor in ensuring high quality results. Not only have the standards for size been raised, with data sets like METABRIC and databases like TCGA, TARGET, ADNI, and others providing access to data from thousands of cases, but the quality desired in such data is going to keep increasing as well. This is compounded by the fact that as more data becomes publically available it can be used to validate tests results with ever-increasing accuracy. If comparative analysis is conducted across multiple cancers of different tissue origin (so called pan cancer studies) or between the ever-increasing number of molecular subtypes of given cancers a greater need for processing will be required.

Finally, further research is required in the more recent areas of machine learning, primary among them being network inference studies and the so called deep learning and deep mining strategies. Understanding how questions of interest



interact and affect each other, such as how genes regulate each other in a given disease, and use machine learning to model more possibilities than could be reasonably studied manually [77] will further increase the potential venues of research.

### Key issues

- Identification of predictive/prognostic immune biomarkers in the blood and TME of patients with hematological malignancies
- Development of prognostic and mechanistic immune gene signatures in patients with hematological malignancies receiving immunotherapies, including checkpoint blockade
- Handling and analysis of multi-dimensional data sets using artificial neural network models
- Prospective validation and incorporation of immunological parameters into personalised routine clinical practice (patient stratification, treatment allocation)

### Funding

The Authors' work is supported through research grants from the Roger Counter Foundation (Dorset, UK), the Qatar National Research Fund (NPRP8-2297-3-494) and the John and Lucille van Geest Foundation.

### Declaration of interest

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

### ORCID

Jayakumar Vadakekolathu  <http://orcid.org/0000-0002-2671-4285>  
 Alan Graham Pockley  <http://orcid.org/0000-0001-9593-6431>  
 Graham Roy Ball  <http://orcid.org/0000-0001-5828-7129>  
 Sergio Rutella  <http://orcid.org/0000-0003-1970-7375>

### References

Papers of special note have been highlighted as either of interest (·) or of considerable interest (•) to readers.

- Palucka AK, Coussens LM. The basis of oncoimmunology. *Cell*. 2016;164:1233–1247.
- Blank CU, Haanen JB, Ribas A, et al. The “cancer immunogram”. *Science*. 2016;352:658–660.
  - This paper discussed an innovative conceptual framework to inform the delivery of personalized immunotherapies to patients with cancer.
- Gnjatic S, Bronte V, Brunet LR, et al. Identifying baseline immune-related biomarkers to predict clinical outcome of immunotherapy. *J Immunother Cancer*. 2017;5:44.
- Anderson KG, Stromnes IM, Greenberg PD. Obstacles posed by the tumor microenvironment to T cell activity: A case for synergistic therapies. *Cancer Cell*. 2017;31:311–325.
- Chen DS, Mellman I. Oncology meets immunology: the cancer-immunity cycle. *Immunity*. 2013;39:1–10.
- Iglesia MD, Parker JS, Hoadley KA, et al. Genomic analysis of immune cell infiltrates across 11 tumor types. *J Natl Cancer Inst*. 2016;108(11).
- Galon J, Costes A, Sanchez-Cabo F, et al. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science*. 2006;313:1960–1964.
- Armand P, Shipp MA, Ribrag V, et al. Programmed death-1 blockade with pembrolizumab in patients with classical Hodgkin lymphoma after brentuximab vedotin failure. *J Clin Oncol*. 2016;34:3733–3739.
- Ansell SM, Lesokhin AM, Borrello I, et al. PD-1 blockade with nivolumab in relapsed or refractory Hodgkin's lymphoma. *N Engl J Med*. 2015;372:311–319.
- Melero I, Berman DM, Aznar MA, et al. Evolving synergistic combinations of targeted immunotherapies to combat cancer. *Nat Rev Cancer*. 2015;15:457–472.
- Stronck DF, Butterfield LH, Cannarile MA, et al. Systematic evaluation of immune regulation and modulation. *J Immunother Cancer*. 2017;5:21.
- Farsaci B, Donahue RN, Grenga I, et al. Analyses of pretherapy peripheral immunoscore and response to vaccine therapy. *Cancer Immunol Res*. 2016;4:755–765.
- Wargo JA, Reddy SM, Reuben A, et al. Monitoring immune responses in the tumor microenvironment. *Curr Opin Immunol*. 2016;41:23–31.
- Mlecnik B, Bindea G, Angell HK, et al. Functional network pipeline reveals genetic determinants associated with in situ lymphocyte proliferation and survival of cancer patients. *Sci Transl Med*. 2014;6:228ra237.
- Bindea G, Mlecnik B, Tosolini M, et al. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity*. 2013;39:782–795.
- Pardoll DM. The blockade of immune checkpoints in cancer immunotherapy. *Nat Rev Cancer*. 2012;12:252–264.
- Taube JM, Anders RA, Young GD, et al. Colocalization of inflammatory response with B7-H1 expression in human melanocytic lesions supports an adaptive resistance mechanism of immune escape. *Sci Transl Med*. 2012;4:127ra137.
  - This study establishes an important link between inflammatory responses and tumor immune escape mechanisms in patients with melanoma.
- Teng MW, Ngiew SF, Ribas A, et al. Classifying cancers based on T-cell infiltration and PD-L1. *Cancer Res*. 2015;75:2139–2145.
- Parsa AT, Waldron JS, Panner A, et al. Loss of tumor suppressor PTEN function increases B7-H1 expression and immunoresistance in glioma. *Nat Med*. 2007;13:84–88.
- Marzec M, Zhang Q, Goradia A, et al. Oncogenic kinase NPM/ALK induces through STAT3 expression of immunosuppressive protein CD274 (PD-L1, B7-H1). *Proc Natl Acad Sci U S A*. 2008;105:20852–20857.
- Chen DS, Mellman I. Elements of cancer immunity and the cancer-immune set point. *Nature*. 2017;541:321–330.
  - This review article thoroughly summarizes the translational potential of the ‘hot’ versus ‘cold’ tumor paradigm.
- Herbst RS, Soria JC, Kowanetz M, et al. Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. *Nature*. 2014;515:563–567.
- Tumeh PC, Harview CL, Yearley JH, et al. PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature*. 2014;515:568–571.
- Mercier FE, Ragu C, Scadden DT. The bone marrow at the crossroads of blood and immunity. *Nat Rev Immunol*. 2011;12:49–60.
- Spitzer MH, Gherardini PF, Fragiadakis GK, et al. An interactive reference framework for modeling a dynamic immune system. *Science*. 2015;349:1259425.
- Curran EK, Godfrey J, Kline J. Mechanisms of immune tolerance in leukemia and lymphoma. *Trends Immunol*. 2017;38:513–525.
- Mussai F, De Santo C, Abu-Dayyeh I, et al. Acute myeloid leukemia creates an arginase-dependent immunosuppressive microenvironment. *Blood*. 2013;122:749–758.
- Folgiero V, Goffredo BM, Filippini P, et al. Indoleamine 2,3-dioxygenase 1 (IDO1) activity in leukemia blasts correlates with poor outcome in childhood acute myeloid leukemia. *Oncotarget*. 2014;5:2052–2064.

1. Curti A, Pandolfi S, Valzasina B, et al. Modulation of tryptophan catabolism by human leukemic cells results in the conversion of CD25- into CD25+ T regulatory cells. *Blood*. 2007;109:2871–2877.
2. Chen BJ, Chapuy B, Ouyang J, et al. PD-L1 expression is characteristic of a subset of aggressive B-cell lymphomas and virus-associated malignancies. *Clin Cancer Res*. 2013;19:3462–3473.
3. Horlad H, Ma C, Yano H, et al. An IL-27/Stat3 axis induces expression of programmed cell death 1 ligands (PD-L1/2) on infiltrating macrophages in lymphoma. *Cancer Sci*. 2016;1696–1704:107.
4. Gibney GT, Weiner LM, Atkins MB. Predictive biomarkers for checkpoint inhibitor-based immunotherapy. *Lancet Oncol*. 2016;17: e542–e551.
5. Ribas A, Robert C, Hodi FS, et al. Association of response to programmed death receptor 1 (PD-1) blockade with pembrolizumab (MK-3475) with an interferon-inflammatory immune gene signature. *J Clin Oncol*. 2015;33:3001–3001.
6. Mlecnik B, Bindea G, Kirilovsky A, et al. The tumor microenvironment and Immunoscore are critical determinants of dissemination to distant metastasis. *Sci Transl Med*. 2016;8:327ra326.
7. Chen PL, Roh W, Reuben A, et al. Analysis of immune signatures in longitudinal tumor samples yields insight into biomarkers of response and mechanisms of resistance to immune checkpoint blockade. *Cancer Discov*. 2016;6:827–837.
8. Brown SD, Warren RL, Gibb EA, et al. Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. *Genome Res*. 2014;24:743–750.
9. Rooney MS, Shukla SA, Wu CJ, et al. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell*. 2015;160:48–61.
10. Danaheer P, Warren S, Dennis L, et al. Gene expression markers of tumor infiltrating leukocytes. *J Immunother Cancer*. 2017;5:18.
  - This study comprehensively evaluates the prognostic importance of immune biomarkers across a broad range of tumor types.
11. Church SE, Galon J. Tumor microenvironment and immunotherapy: the whole picture is better than a glimpse. *Immunity*. 2015;43:631–633.
12. Gentles AJ, Newman AM, Liu CL, et al. The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat Med*. 2015;21:938–945.
13. Mizuno H, Kitada K, Nakai K, et al. PrognoScan: a new database for meta-analysis of the prognostic value of genes. *BMC Med Genomics*. 2009;2:18.
14. Bagger FO, Sasivarevic D, Sohi SH, et al. BloodSpot: a database of gene expression profiles and transcriptional programs for healthy and malignant haematopoiesis. *Nucleic Acids Res*. 2016;44:D917–924.
15. Goswami CP, Nakshatri H. PROGeneV2: enhancements on the existing database. *BMC Cancer*. 2014;14:970.
16. Ley TJ, Miller C, Ding L, et al. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med*. 2013;368:2059–2074.
17. Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods*. 2015;12:453–457.
18. Austin R, Smyth MJ, Lane SW. Harnessing the immune system in acute myeloid leukaemia. *Crit Rev Oncol Hematol*. 2016;103:62–77.
19. Hara T, Matsumoto T, Shibata Y, et al. Prognostic value of the combination of serum L-kynurenine level and indoleamine 2,3-dioxygenase mRNA expression in acute myeloid leukemia. *Leuk Lymphoma*. 2016;57:2208–2211.
20. Jitschin R, Braun M, Buttner M, et al. CLL-cells induce IDOhi CD14+HLA-DRlo myeloid-derived suppressor cells that inhibit T-cell responses and promote TRegs. *Blood*. 2014;124:750–760.
21. Saulep-Easton D, Vincent FB, Le Page M, et al. Cytokine-driven loss of plasmacytoid dendritic cell function in chronic lymphocytic leukemia. *Leukemia*. 2014;28:2005–2015.
22. Ramsay AG, Clear AJ, Fatah R, et al. Multiple inhibitory ligands induce impaired T-cell immunologic synapse function in chronic lymphocytic leukemia that can be blocked with lenalidomide: establishing a reversible immune evasion mechanism in human cancer. *Blood*. 2012;120:1412–1421.
23. Ding W, LaPlant BR, Call TG, et al. Pembrolizumab in patients with CLL and Richter transformation or with relapsed CLL. *Blood*. 2017;129:3419–3427.
24. Zitvogel L, Rusakiewicz S, Routy B, et al. Immunological off-target effects of imatinib. *Nat Rev Clin Oncol*. 2016;13:431–446.
25. Gullaksen SE, Skavland J, Gavasso S, et al. Single cell immune profiling by mass cytometry of newly diagnosed chronic phase chronic myeloid leukaemia treated with nilotinib. *Haematologica*. 2017;102:1361–1367.
26. Ouyang J, Plutschow A, von Strandmann P, et al. M.A. Galectin-1 serum levels reflect tumor burden and adverse clinical features in classical Hodgkin lymphoma. *Blood*. 2013;121:3431–3433.
27. Steidl C, Lee T, Shah SP, et al. Tumor-associated macrophages and survival in classic Hodgkin's lymphoma. *N Engl J Med*. 2010;362:875–885.
28. Tan KL, Scott DW, Hong F, et al. Tumor-associated macrophages predict inferior outcomes in classic Hodgkin lymphoma: a correlative study from the E2496 Intergroup trial. *Blood*. 2012;120:3280–3287.
29. Epeldegui M, Lee JY, Martinez AC, et al. Predictive value of cytokines and immune activation biomarkers in AIDS-related non-Hodgkin lymphoma treated with rituximab plus infusional EPOCH (AMC-034 trial). *Clin Cancer Res*. 2016;22:328–336.
30. Vendrame E, Hussain SK, Breen EC, et al. Serum levels of cytokines and biomarkers for inflammation and immune activation, and HIV-associated non-Hodgkin B-cell lymphoma risk. *Cancer Epidemiol Biomarkers Prev*. 2014;23:343–349.
31. Breen EC, Hussain SK, Magpantay L, et al. B-cell stimulatory cytokines and markers of immune activation are elevated several years prior to the diagnosis of systemic AIDS-associated non-Hodgkin B-cell lymphoma. *Cancer Epidemiol Biomarkers Prev*. 2011;20:1303–1314.
32. Purdue MP, Hofmann JN, Kemp TJ, et al. A prospective study of 67 serum immune and inflammation markers and risk of non-Hodgkin lymphoma. *Blood*. 2013;122:951–957.
33. Carvalho A, Cunha C, Almeida AJ, et al. The rs5743836 polymorphism in TLR9 confers a population-based increased risk of non-Hodgkin lymphoma. *Genes Immun*. 2012;13:197–201.
34. Hosgood HD 3rd, Purdue MP, Wang SS, et al. A pooled analysis of three studies evaluating genetic variation in innate immunity genes and non-Hodgkin lymphoma risk. *Br J Haematol*. 2011;152:721–726.
35. Dave SS, Wright G, Tan B, et al. Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. *N Engl J Med*. 2004;351:2159–2169.
  - This study establishes a correlation between immune gene signatures and survival which is independent of conventional prognosticators in patients with follicular lymphoma.
36. Taskinen M, Karjalainen-Lindsberg ML, Leppa S. Prognostic influence of tumor-infiltrating mast cells in patients with follicular lymphoma treated with rituximab and CHOP. *Blood*. 2008;111:4664–4667.
37. Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*. 2000;403:503–511.
38. Xing W, Dresser K, Zhang R, et al. PD-L1 expression in EBV-negative diffuse large B-cell lymphoma: clinicopathologic features and prognostic implications. *Oncotarget*. 2016;7:59976–59986.
39. Riihijarvi S, Fiskvik I, Taskinen M, et al. Prognostic influence of macrophages in patients with diffuse large B-cell lymphoma: a correlative study from a Nordic phase II trial. *Haematologica*. 2015;100:238–245.
40. Gomez-Gelvez JC, Salama ME, Perkins SL, et al. Prognostic impact of tumor microenvironment in diffuse large B-cell lymphoma

- uniformly treated with R-CHOP chemotherapy. *Am J Clin Pathol.* 2016;145:514–523.
1. Keane C, Vari F, Hertzberg M, et al. Ratios of T-cell immune effectors and checkpoint molecules as prognostic biomarkers in diffuse large B-cell lymphoma: a population-based study. *Lancet Haematol.* 2015;2:e445–455.
  2. Martelli M, Ferreri A, Di Rocco A, et al. Primary mediastinal large B-cell lymphoma. *Crit Rev Oncol Hematol.* 2017;113:318–327.
  3. Mottok A, Woolcock B, Chan FC, et al. Genomic alterations in CIITA are frequent in primary mediastinal large B cell lymphoma and are associated with diminished MHC class II expression. *Cell Rep.* 2015;13:1418–1431.
  4. Shi M, Roemer MG, Chapuy B, et al. Expression of programmed cell death 1 ligand 2 (PD-L2) is a distinguishing feature of primary mediastinal (thymic) large B-cell lymphoma and associated with PDCD1LG2 copy gain. *Am J Surg Pathol.* 2014;38:1715–1723.
  5. Zinzani PL, Ribrag V, Moskowitz CH, et al. Safety and tolerability of pembrolizumab in patients with relapsed/refractory primary mediastinal large B-cell lymphoma. *Blood.* 2017;130:267–270.
  6. Steidl C, Gascoyne RD. The molecular pathogenesis of primary mediastinal large B-cell lymphoma. *Blood.* 2011;118:2659–2669.
  7. Wang H, Wang L, Chi PD, et al. High level of interleukin-10 in serum predicts poor prognosis in multiple myeloma. *Br J Cancer.* 2016;114:463–468.
  8. Gorgun GT, Whitehill G, Anderson JL, et al. Tumor-promoting immune-suppressive myeloid-derived suppressor cells in the multiple myeloma microenvironment in humans. *Blood.* 2013;121:2975–2987.
  9. Prabhala RH, Pelluru D, Fulciniti M, et al. Elevated IL-17 produced by TH17 cells promotes myeloma cell growth and inhibits immune function in multiple myeloma. *Blood.* 2010;115:5385–5392.
  10. Atanackovic D, Luetkens T, Kroger N. Coinhibitory molecule PD-1 as a potential target for the immunotherapy of multiple myeloma. *Leukemia.* 2014;28:993–1000.
  11. Gorgun G, Samur MK, Covens KB, et al. Lenalidomide enhances immune checkpoint blockade-induced immune response in multiple myeloma. *Clin Cancer Res.* 2015;21:4607–4618.
  12. Bonanno G, Mariotti A, Procoli A, et al. Indoleamine 2,3-dioxygenase 1 (IDO1) activity correlates with immune system abnormalities in multiple myeloma. *J Transl Med.* 2012;10:247.
  13. De Haart SJ, Van De Donk NW, Minnema MC, et al. Accessory cells of the microenvironment protect multiple myeloma from T-cell cytotoxicity through cell adhesion-mediated immune resistance. *Clin Cancer Res.* 2013;19:5591–5601.
  14. Paiva B, Mateos MV, Sanchez-Abarca LI, et al. Immune status of high-risk smoldering multiple myeloma patients and its therapeutic modulation under LenDex: a longitudinal analysis. *Blood.* 2016;127:1151–1162.
  15. Brown R, Kabani K, Favaloro J, et al. CD86+ or HLA-G+ can be transferred via trogocytosis from myeloma cells to T cells and are associated with poor prognosis. *Blood.* 2012;120:2055–2063.
  16. Maude SL, Frey N, Shaw PA, et al. Chimeric antigen receptor T cells for sustained remissions in leukemia. *N Engl J Med.* 2014;371:1507–1517.
  - This study documents a significant clinical benefit for CAR-transduced T cells in patients with relapsed and refractory leukemia.
  17. Kochenderfer JN, Somerville RPT, Lu T, et al. Lymphoma remissions caused by anti-CD19 chimeric antigen receptor T cells are associated with high serum interleukin-15 levels. *J Clin Oncol.* 2017;35:1803–1813.
  18. Ramos CA, Ballard B, Zhang H, et al. Clinical and immunological responses after CD30-specific chimeric antigen receptor-redirection lymphocytes. *J Clin Invest.* 2017;127:3462–3471.
  19. Chong EA, Melenhorst JJ, Lacey SF, et al. PD-1 blockade modulates chimeric antigen receptor (CAR)-modified T cells: refueling the CAR. *Blood.* 2017;129:1039–1041.
  20. Ninomiya S, Narala N, Huye L, et al. Tumor indoleamine 2,3-dioxygenase (IDO) inhibits CD19-CAR T cells and is downregulated by lymphodepleting drugs. *Blood.* 2015;125:3905–3916.
  21. Gomes-Silva D, Srinivasan M, Sharma S, et al. CD7-edited T cells expressing a CD7-specific CAR for the therapy of T-cell malignancies. *Blood.* 2017;130:285–296.
  22. Ruella M, Barrett DM, Kenderian SS, et al. Dual CD19 and CD123 targeting prevents antigen-loss relapses after CD19-directed immunotherapies. *J Clin Invest.* 2016;126:3814–3826.
  23. Ali SA, Shi V, Maric I, et al. T cells expressing an anti-B-cell maturation antigen chimeric antigen receptor cause remissions of multiple myeloma. *Blood.* 2016;128:1688–1700.
  24. Garfall AL, Maus MV, Hwang WT, et al. Chimeric antigen receptor T cells against CD19 for multiple myeloma. *N Engl J Med.* 2015;373:1040–1047.
  25. Ramos CA, Savoldo B, Torrano V, et al. Clinical responses with T lymphocytes targeting malignancy-associated kappa light chains. *J Clin Invest.* 2016;126:2588–2596.
  26. Kantarjian H, Stein A, Gokbuget N, et al. Blinatumomab versus chemotherapy for advanced acute lymphoblastic leukemia. *N Engl J Med.* 2017;376:836–847.
  - This randomized clinical trial shows the survival benefit of immunotherapy over conventional chemotherapy in patients with acute lymphoblastic leukemia.
  27. Bertaina A, Filippini P, Bertaina V, et al. Immune cell phenotype and function after treatment with blinatumomab for childhood relapsed B-cell precursor acute lymphoblastic leukemia (BCP-ALL). *Blood.* 2013;122:2668–2668.
  28. Feucht J, Kayser S, Gorodezki D, et al. T-cell responses against CD19 + pediatric acute lymphoblastic leukemia mediated by bispecific T-cell engager (BiTE) are regulated contrarily by PD-L1 and CD80/CD86 on leukemic blasts. *Oncotarget.* 2016;7:76902–76919.
  29. Mahoney KM, Rennert PD, Freeman GJ. Combination cancer immunotherapy and new immunomodulatory targets. *Nat Rev Drug Discov.* 2015;14:561–584.
  30. Stack EC, Foukas PG, Lee PP. Multiplexed tissue biomarker imaging. *J Immunother Cancer.* 2016;4:9.
  31. Chang AY, Bhattacharya N, Mu J, et al. Spatial organization of dendritic cells within tumor draining lymph nodes impacts clinical outcome in breast cancer patients. *J Transl Med.* 2013;11:242.
  32. Cesano A. nCounter(R) PanCancer immune profiling panel (NanoString Technologies, Inc., Seattle, WA). *J Immunother Cancer.* 2015;3:42.
  33. Miotto R, Wang F, Wang S, et al. Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform.* 2017. doi: 10.1093/bib/bbx044.
  34. Chatzimichail E, Matthaios D, Bouros D, et al. gamma-H2AX: A novel prognostic marker in a prognosis prediction model of patients with early operable non-small cell lung cancer. *Int J Genomics.* 2014;2014:160236.
  35. Bertolaccini L, Solli P, Pardolesi A, et al. An overview of the use of artificial neural networks in lung cancer research. *J Thorac Dis.* 2017;9:924–931.
  36. Abdel-Fatah TM, Agarwal D, Liu DX, et al. SPAG5 as a prognostic biomarker and chemotherapy sensitivity predictor in breast cancer: a retrospective, integrated genomic, transcriptomic, and protein analysis. *Lancet Oncol.* 2016;17:1004–1018.
  37. Mehdy MM, Ng PY, Shair EF, et al. Artificial neural networks in image processing for early detection of breast cancer. *Comput Math Methods Med.* 2017;2017:2610628.
  38. Huang SH, Loh JK, Tsai JT, et al. Predictive model for 5-year mortality after breast cancer surgery in Taiwan residents. *Chin J Cancer.* 2017;36:23.
  39. Abdel-Fatah TM, McArdle SE, Agarwal D, et al. HAGE in triple-negative breast cancer is a novel prognostic, predictive, and actionable biomarker: a transcriptomic and protein expression analysis. *Clin Cancer Res.* 2016;22:905–914.
  40. Haj-Hassan H, Chaddad A, Harkouss Y, et al. Classifications of multi-spectral colorectal cancer tissues using convolution neural network. *J Pathol Inform.* 2017;8:1.

1. Makris GM, Pouliakis A, Siristatidis C, et al. Image analysis and multi-layer perceptron artificial neural networks for the discrimination between benign and malignant endometrial lesions. *Diagn Cytopathol.* **2017**;45:202–211.
2. van Belle V, Van Calster B, van Huffel S, et al. Explaining support vector machines: a color based nomogram. *PLoS One.* **2016**;11: e0164568.
3. Powell RT, Olar A, Narang S, et al. Identification of histological correlates of overall survival in lower grade gliomas using a bag-of-words paradigm: A preliminary analysis based on hematoxylin & eosin stained slides from the lower grade glioma cohort of the Cancer Genome Atlas. *J Pathol Inform.* **2017**;8:9.
4. Araujo T, Aresta G, Castro E, et al. Classification of breast cancer histology images using Convolutional Neural Networks. *PLoS One.* **2017**;12:e0177544.
5. Huang MW, Chen CW, Lin WC, et al. SVM ensembles in breast cancer prediction. *PLoS One.* **2017**;12:e0161501.
6. Cheerla N, Gevaert O. MicroRNA based pan-cancer diagnosis and treatment recommendation. *BMC Bioinformatics.* **2017**;18:32.
7. Zeng Z, Jiang X, Neapolitan R. Discovering causal interactions using Bayesian network scoring and information gain. *BMC Bioinformatics.* **2016**;17:221.
8. Zhu X, Ko YJ, Berry S, et al. A Bayesian network meta-analysis on second-line systemic therapy in advanced gastric cancer. *Gastric Cancer.* **2017**;20:646–654.
9. Field SL, Dasgupta T, Cummings M, et al. Bayesian modeling suggests that IL-12 (p40), IL-13 and MCP-1 drive murine cytokine networks in vivo. *BMC Syst Biol.* **2015**;9:76.
10. Luo Y, El Naqa I, McShan DL, et al. Unraveling biophysical interactions of radiation pneumonitis in non-small-cell lung cancer via Bayesian network analysis. *Radiother Oncol.* **2017**;123:85–92.
11. Sommer C, Gerlich DW. Machine learning in cell biology - teaching computers to recognize phenotypes. *J Cell Sci.* **2013**;126:5529–5539.
12. Stadler N, Dondelinger F, Hill SM, et al. Molecular heterogeneity at the network level: high-dimensional testing, clustering and a TCGA case study. *Bioinformatics.* **2017**;33:2890–2896.
13. Athreya AP, Kalari KR, Cairns J, et al. Model-based unsupervised learning informs metformin-induced cell-migration inhibition through an AMPK-independent mechanism in breast cancer. *Oncotarget.* **2017**;8:27199–27215.
14. Vural S, Wang X, Guda C. Classification of breast cancer patients using somatic mutation profiles and machine learning approaches. *BMC Syst Biol.* **2016**;10(Suppl 3):62.
15. Rodriguez A, Laio A. Machine learning. Clustering by fast search and find of density peaks. *Science.* **2014**;344:1492–1496.
16. Xanthopoulos P, Pardalos PM, Trafalis TB. Principal component analysis. Petros X, Panos MP, Theodore BT (Eds.) In: *Robust data mining.* New York, NY: Springer New York; **2013.** p. 21–26.
17. Krzywinski M, Biro I, Jones SJ, et al. Hive plots—rational approach to visualizing networks. *Brief Bioinform.* **2012**;13:627–644.
18. Goff SL, Dudley ME, Citrin DE, et al. Randomized, prospective evaluation comparing intensity of lymphodepletion before adoptive transfer of tumor-infiltrating lymphocytes for patients with metastatic melanoma. *J Clin Oncol.* **2016**;34:2389–2397.
19. Noonan KA, Huff CA, Davis J, et al. Adoptive transfer of activated marrow-infiltrating lymphocytes induces measurable antitumor immunity in the bone marrow in multiple myeloma. *Sci Transl Med.* **2015**;7:288ra278.
20. Roy S, Trinchieri G. Microbiota: a key orchestrator of cancer therapy. *Nat Rev Cancer.* **2017**;17:271–285.
21. Galon J, Angell HK, Bedognetti D, et al. The continuum of cancer immunosurveillance: prognostic, predictive, and mechanistic signatures. *Immunity.* **2013**;39:11–26.
22. Topalian SL, Taube JM, Anders RA, et al. Mechanism-driven biomarkers to guide immune checkpoint blockade in cancer therapy. *Nat Rev Cancer.* **2016**;16:275–287.
23. Beatty GL, O'Dwyer PJ, Clark J, et al. First-in-human phase I study of the oral inhibitor of indoleamine 2,3-dioxygenase-1 epacadostat (INC024360) in patients with advanced solid malignancies. *Clin Cancer Res.* **2017**;23:3269–3276.
- This clinical trial shows the safety of small-molecule drugs that target immune suppression in the tumor microenvironment.
24. Lesokhin AM, Callahan MK, Postow MA, et al. On being less tolerant: enhanced cancer immunosurveillance enabled by targeting checkpoints and agonists of T cell activation. *Sci Transl Med.* **2015**;7:280sr281.



# An Artificial Neural Network Integrated Pipeline for Biomarker Discovery Using Alzheimer's Disease as a Case Study

Dimitrios Zafeiris \*, Sergio Rutella, Graham Roy Ball

John van Geest Cancer Research Centre, College of Science and Technology, Nottingham Trent University, United Kingdom

## article info

### Article history:

Received 18 August 2017  
 Received in revised form 6 February 2018  
 Accepted 11 February 2018  
 Available online 21 February 2018

### Keywords:

Artificial neural network  
 Machine learning  
 Supervised learning  
 Network inference  
 Alzheimer's disease  
 Biomarker discovery

## abstract

The field of machine learning has allowed researchers to generate and analyse vast amounts of data using a wide variety of methodologies. Artificial Neural Networks (ANN) are some of the most commonly used statistical models and have been successful in biomarker discovery studies in multiple disease types. This review seeks to explore and evaluate an integrated ANN pipeline for biomarker discovery and validation in Alzheimer's disease, the most common form of dementia worldwide with no proven cause and no available cure. The proposed pipeline consists of analysing public data with a categorical and continuous stepwise algorithm and further examination through network inference to predict gene interactions. This methodology can reliably generate novel markers and further examine known ones and can be used to guide future research in Alzheimer's disease.

© 2018 Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

### 1.1. Machine Learning

One of the biggest challenges that has arisen as part of the recent advances in the field of bioinformatics, is the vast amount of data that is being generated at an ever-increasing pace [1–3]. Utilising techniques such as next generation RNA and DNA sequencing, researchers have been able to provide access to exceptionally precise information on entire genomes [4]. This massive volume of data has created a problem of complexity, making it impossible to interrogate the data with traditional methodologies and provide answers with the desired degree of accuracy.

Machine learning is an interdisciplinary field of bioinformatics that involves a data-driven class of algorithms that seek to find solutions to a given problem by studying patterns in datasets based on factors such as gene expression and clinical information across a multitude of cases. These approaches have been widely and successfully used in biology, particularly in biomarker discovery studies [5,6], due to the versatility and power afforded by them and has resulted in a wide variety of machine learning algorithms and methodologies. This review seeks to explore the potential of an Artificial Neural Network (ANN)

based pipeline to discover, analyse and validate novel biomarkers in diverse diseases. For this purpose, Alzheimer's disease (AD) will be used since the cause of the condition is poorly understood and there is no widely available cure or treatment.

### 1.2. Supervised Learning

Supervised learning approaches, the mechanisms of which are further discussed in chapter 3, are widely applied and use source features to predict a target class [7]. The supervised approach allows the algorithm to train itself by detecting patterns in large data sets that are predictive of the target class, such as highlighting the variance at the genetic level between AD and cognitively normal individuals. We can also make use of previous studies and adjust the algorithm parameters so that it accounts for this information, which allows the power of this approach to increase over time and produce more accurate and robust results. One major advantage of supervised learning is that such approaches are tolerant of the highly complex, nonlinear and noisy data that are often found in biological systems.

### 1.3. Artificial Neural Networks

ANNs are statistical models that emulate the function of a network of human neurons, for the purpose of encapsulating information in order to analyse large, complex datasets. The learning process is based on the mathematical interconnections between the processing elements that constitute the network architecture [8]. This allows them to classify cases based on data by assigning a numerical weight value to each input

*Abbreviations:* ANN, artificial neural network; AD, Alzheimer's disease; MLP, multi-layer perceptron; APP, amyloid precursor protein; A $\beta$ , beta amyloid; NFT, neurofibrillary tangles.

\* Corresponding author.

E-mail address: [dimitrios.kapsoulis@ntu.ac.uk](mailto:dimitrios.kapsoulis@ntu.ac.uk) (D. Zafeiris).



and adjust them as they sample the data, effectively learning the optimal solution. The main advantages of using ANNs include their high fault and failure tolerance, scalability and consistent generalisation ability, which allows them to predict or classify well for new, fuzzy and unlearned data [8,9]. This makes the ideal for biomarker studies which resulted in their use in generating panels of biomarkers that can be used as predictors in conjunction with each to aid prognosis in diseases such as breast cancer [10].

ANN architecture is based on the perceptron, coined by Rosenblatt in 1958, which is composed of a single artificial processing neuron with an activation threshold, adjustable weights and bias, but only usable for the classification of linearly separable patterns, as learning is achieved when an error occurs during testing. This is rarely the case with complex conditions such as AD, cancer or diabetes, as patients rarely fall in a standard distribution and the variance between them is potentially significant. Typically, ANNs make use of a Multi-Layer Perceptron (MLP) which is made up of multiple perceptrons arranged in layers of three or more, consisting of input, hidden and output layers, which consider predictor variables, perform feature detection through an activation function and output the results of the algorithm respectively.

Alternative ANN architectures include Recurrent Neural Networks, Radial Basis Function, Kohonen's self-organizing maps and Adaptive Resonance Theory but the focus of this review will be on the MLP.

ANNs have seen widespread success in predicting and classifying data in multiple cancer subtypes such as early detection [11], prediction of long term survival [12] and biomarker discovery in breast cancer [10,13], classifying colorectal cancer tissues [14] and discriminating between benign and malignant endothelial lesions [15]. Thus, we are confident that they will see similar success in AD.

The main ANN disadvantage is their liability to overfit when the parameters have not been optimised and often receive criticism for their "black box" approach that allows for little interpretation of the results and process.

### 1.1. Alzheimer's Disease

Alzheimer's disease is recognised as the most common form of dementia worldwide. This chronic neurodegenerative disease usually starts slowly, with the common early symptom being difficulty to remember short-term events and progressively getting worse, with severe degeneration of multiple brain regions including the hippocampus,

entorhinal cortex, neocortex, nucleus basalis, locus coeruleus and raphe nuclei (Fig. 1), leading to disruption in mental functions such as comprehension, judgement, language and calculation. Moreover, due to slow progression that characterises the disease as well as common misconceptions, it is common for patients and their families to assume that this degeneration is a normal part of ageing, thus delaying early prognosis. It is crucial to emphasise that AD is the abnormal degeneration of mental faculties and while age is indeed the biggest risk factor, it is far from the only one.

In addition to the enormous emotional cost the disease exerts on patients and their families, it has become a major public concern due to the high healthcare costs which, in combination with the overall rise in the elderly population has classified AD as a priority condition [16]. According to the World Health Organisation, in 2015 there were over 40 million people with dementia in the US, 15 million of which suffered from Alzheimer's disease. Healthcare costs have spiralled to over USD 900 billion, whereas in Europe the costs have risen to nearly 250 billion euros, a rise of almost 40% from 2008. Moreover, it is projected that by 2050, 22% of the world's population will be over the age of 60, and therefore at increased risk, with patients in third world countries accounting for 80% of the total.

### 1.2. Theories and Treatments

Compounding the social and economic challenges presented by the disease is the fact that its root causes are unknown and there is no cure or effective treatment. While there is a small percentage of the population, 1–5% of all cases, that suffer from early onset AD, which is caused by mutations in the amyloid precursor protein gene (APP) and the two presenilin genes PSEN-1 and PSEN-2, the cause for the majority of late onset Alzheimer's cases is still unknown. In the last decade, clinically approved drugs for AD such as Cholinesterase inhibitors like Donepezil, Galantamine and Rivastigmine as well as *N*-methyl-D-aspartate antagonist Memantine [17] have not been able to make significant progress with the disorder.

Cholinesterase inhibitors, which target the cholinergic systems in the basal forebrain, were developed based on the theory that the loss of acetylcholine neurons during the early development of the disease inhibit the synthesis and degradation of acetylcholine, one of the major neurotransmitters in the brain. Therapy was targeted at patients with mild, moderate and severe AD but improvement of cognitive

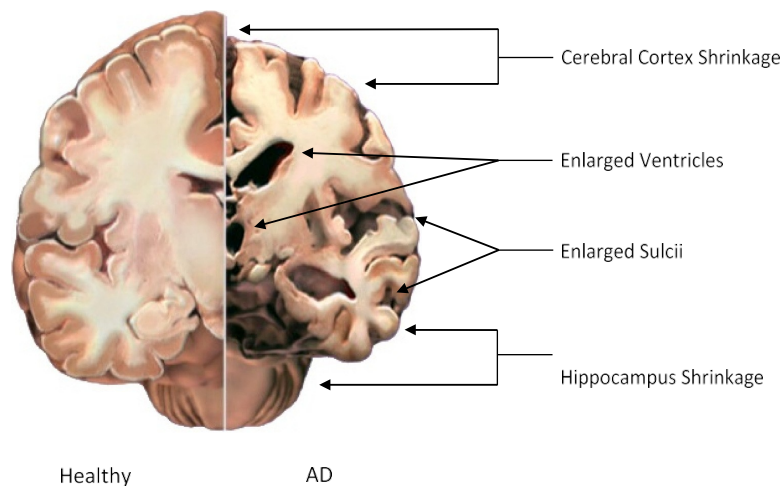


Fig. 1. Physiological differences between a healthy and AD brain section, demonstrating white matter shrinkage in the hippocampus and cerebral cortex. Source: [www.alz.org](http://www.alz.org).

functions was noticeably better in patients that started treatment early [18]. *N*-methyl-D-aspartate antagonist on the other hand, is an uncompetitive moderate affinity antagonist, targeted at moderate to severe AD cases, with the purpose of protecting neurons from excitotoxicity. Other forms of therapy have focused on combinations of these drugs and treatment of the behavioural and psychological symptoms of the disease.

More recently, therapeutic approaches have been based on the amyloid hypothesis, attempting to slow, stop and reverse the development of amyloid plaques by inhibiting production of beta amyloid, as well as the hyperphosphorylation and deposition of tau protein. Finally, further research has been focusing on the effects of oxidative damage and chronic inflammation in the brain to determine their effects in the development and progression of AD. It is evident by the variety of approaches as well as the failure of most forms of therapy to reverse or even significantly slow the disease progression, that a deeper understanding of the pathogenesis of AD is urgently needed to effectively combat it.

### 1.1. Physiology of Alzheimer's Disease

Historically, identification of AD could only be performed post mortem upon examination of the brain tissue. As a result, the physiological hallmarks of AD have been widely considered to be the presence of amyloid plaques, extracellular deposits of insoluble beta-amyloid ( $A\beta$ ) in the parenchyma of the brain as well as neurofibrillary tangles (NFT), intracellular deposits of hyper-phosphorylated tau protein which fill the neuron and take its shape, preventing it from functioning correctly (Fig. 2).

Amyloid plaques consist of a solid core of defective  $A\beta$  and are surrounded by degenerate axons and dendrites, activated microglia and astrocytes. This defective protein is a result of the cleaving of the amyloid precursor protein (APP) by secretases beta ( $\beta$ ) and gamma ( $\gamma$ ). The location APP is cleaved by  $\gamma$ -secretase determines whether  $A\beta$  will be the long or short form. The short form is the most common (~90%) but the long form is found as often as 40% in the brains of AD patients [19], and while small amounts can be cleared easily, the high rate of production leads to the system being unable to keep up. Moreover, soluble forms of the protein have been shown to be neurotoxic and synaptotoxic [20].

Neurofibrillary tangles are a result of the hyperphosphorylation of tau, a microtubule associated protein (MAP) whose role is to bind to tubulin and stabilise the structure of neurons to maintain their function. When hyperphosphorylated due to excessive amounts of phosphate

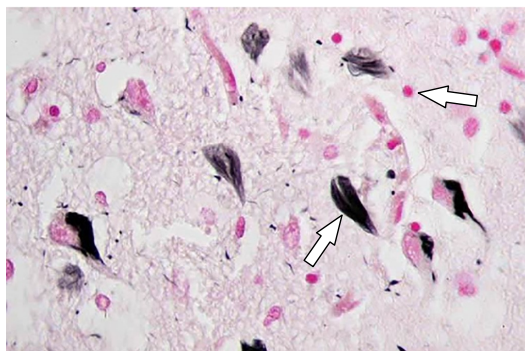


Fig. 2. Amyloid plaques (pink) and neurofibrillary tangles (black) in Alzheimer's disease brain tissue.  
Source: [www.alzheimers.org.uk](http://www.alzheimers.org.uk).

ions, it changes from its normal soluble form to oligomeric and fibrillized forms, does not bind to tubulin, inhibits microtubule structure and assembly and has been shown to have a neurotoxic effect [21].

### 1.2. The Amyloid Cascade Hypothesis

The leading theory for the cause of Alzheimer's disease is the amyloid cascade hypothesis, first proposed in 1992 and its influence on AD research cannot be understated. The hypothesis posits that mutations in the APP and presenilin genes PSEN1 and PSEN2 leads to the deposition of  $A\beta$  in the brain which subsequently leads to the formation of NFTs, cell death and dementia. Experiments in animal models have shown that chemically or damage induced lesions lead to an increase in APP levels and accelerate the development of AD [22,23]. Unfortunately, all approaches based on the amyloid cascade have failed at Phase III clinical trials - tramiprosate, tarenflurbil and semagacestat - and research has not been able to conclusively link the build-up of  $A\beta$  to the formation of NFTs (Fig. 3) [24].

While it has been made clear that the amyloid cascade hypothesis is not enough to sufficiently explain the development of AD or aid in its detection and consequently, is currently under heavy scrutiny, it is also not possible to accept the null hypothesis, as autosomal dominant mutations in the aforementioned APP, PSEN1 and PSEN2 genes along with the apolipoprotein E4 (APOE4) allele have been proven to be the key components in familial, or early onset, Alzheimer's disease. Instead, the amyloid cascade hypothesis has to be modified to account for the rate of  $A\beta$  deposition and clearance, the connection with the development of NFTs and the effect of inflammation in the development of AD. Karran et al. [25] have attempted to update the hypothesis for use in therapeutics by presenting four distinct scenarios describing the role of  $A\beta$  in AD. These scenarios are:

1.  $A\beta$  could trigger development of the disease and further accumulation has little to no effect
2. development starts once  $A\beta$  reaches a certain, as yet unknown, threshold
3.  $A\beta$  is a key driver of AD and its continued deposition accelerates the effect
4.  $A\beta$  is irrelevant and the presence of plaques and increased levels of  $A\beta$  are a side effect of a different cause.

It should be noted that a major limitation of this hypothesis is that it fails to account for AD patients with little to no AD pathology [26] and thus amyloid plaques as identified by PET scan. In recent years, mice studies have shown that  $A\beta$  deposition is a potential driver for tau hyperphosphorylation, fixing one the major limitations of the amyloid hypothesis. Crossing APP transgenic mice with tau knockout mice, resulted in offspring with significantly fewer behavioural deficits [27] while other studies have shown that soluble oligomers of  $A\beta$  can lead to alterations in tau, potentially cascading to AD [28] although the mechanisms are still unclear. Strooper and Karran [29] attempted to provide alternatives including proteostatic stress during the biochemical phase when  $A\beta$  aggregates at an abnormally fast pace, defections in the amyloid and tau clearance mechanisms and a decrease in synaptic plasticity. As Selkoe and Hardy [27] suggest, the amyloid hypothesis, for all its limitations, is essential for therapeutics due to the fact that the complexity of the disease increases drastically after initiation due to the rise in complexity of downstream pathogenic processes, the most likely point of the disease where treatment will be at its most successful.

### 1.3. Inflammation in Alzheimer's Disease

Recent research has also been focused on investigating the role of inflammation in AD in an attempt to explain the development of the disease. The inflammation hypothesis posits that deposition of  $A\beta$  causes chronic activation of the immune system and disrupts microglial

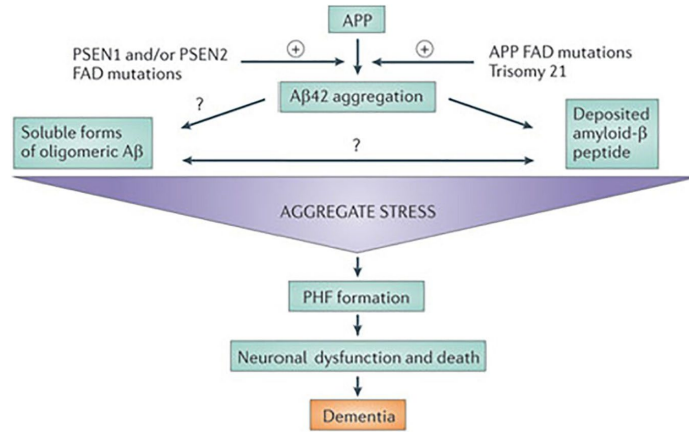


Fig. 3. Diagram of the amyloid cascade hypothesis showing the theorised links between the aggregation of Aβ to cell death and dementia. Source: Karran et al. [25].

clearance functions. Microglia are immune cells located in the parenchyma of the brain, making up 20% of the total glial population. Their functions include phagocytosis, induction of inflammation, and antigen presentation to lymphocytes [30]. However, their roles also include clearance of extracellular deposits of Aβ, and microglial receptors TLR2, TLR4, TLR6 and co-receptors CD36, CD14 and CD47 activated upon detection of the protein. These receptors can also sense pathogen-associated molecular patterns such as bacterial lipopolysaccharides and viral surface proteins and thus are instrumental for mediating the immune response. Certain bacteria have similar surface amyloids, such as curli fibers, which resemble Aβ aggregates and thus activate toll-like receptors (TLR) and CD36, which in turn triggers the formation of a TLR4-TLR6 heterodimer and results in signalling activation via the transcription factor NF-κB. This leads to a cytokine cascade which further attracts immune cells to the site of the perceived infection (Fig. 4).

Moreover, certain cytokines such as IL-1β, damage the synaptic plasticity by disrupting the formation of dendritic spines, with high cytokine expression being able to disrupt normal hippocampus function. This leads to the hypothesis that chronic activation of the immune systems leads to chronic inflammation and microglial cell death, resulting in increased proliferation and accelerated senescence.

## 1. Artificial Neural Networks and Systems Biology

### 1.1. Artificial Neural Networks

As explained previously, ANNs are a form of machine learning, statistical models emulating the function of a neuron, able to identify patterns and linearly separate them by assigning a numerical weight value to each input and adjust them as they sample the data, effectively learning the optimal solution. They can make use of parallel processing in order to predict solutions to complex and non-linear data (Fig. 5) [31].

The ANN used for this project is a Multi-Layer Perceptron (MLP) with a back-propagation (BP) algorithm. It is organised in several layers, each with a number of mathematical processing elements depending on the complexity of the problem and the BP algorithm is responsible for feeding the error back through the model, allowing it to adjust the training weights accordingly and stop early if no gains can be made.

### 1.2. Stepwise Analysis

The stepwise ANN approach developed by Lancashire [33] allows for the identification of a gene or set of genes with the best predictive performance to classify samples based on a certain question by data mining the complete transcriptome. The ANN model functions by modifying the network weights and subsequently adding variables in an iterative manner to find a model with the lowest predictive error. The architecture consists of a single hidden layer, feed forward MLP with a variable number of hidden nodes and a sigmoidal transfer function, using a back-propagation algorithm incorporating supervised learning for updating the network weights. A Monte Carlo Cross Validation (MCCV) strategy was applied to produce a more generalized model with an improved predictive ability for unseen or future cases. The MCCV randomly divides the samples into training, test and validation subsets in 60:20:20 proportion for 50 iterations to provide the most consistent models. The parameters selected for this series of tests are 1 step, 10 loops with a momentum of 0.5, learning rate of 0.1 and threshold of 0.01 [34]. These parameters have been thoroughly tested and successfully used in other studies [10]. The dataset used for this experiment is [dataset] E-GEOD-48350 [35].

The dataset is publicly available and has been accessed using ArrayExpress [36] as well as the Gene Expression Omnibus (GEO) [37]. It was selected based on the following parameters to ensure high quality results:

- Human samples only
- Patient size of N80
- Genes in array N40,000
- A minimum of four brain region samples
- Healthy controls between 33% and 66% of the dataset
- Recent Publication
- Raw data in the form of CEL files available.

The methodology flowchart is included in the Supplemental Fig. 1. The outcome of the stepwise analysis is a list of genes, ordered from the most to least likely to explain the variance in the population based on AD status.



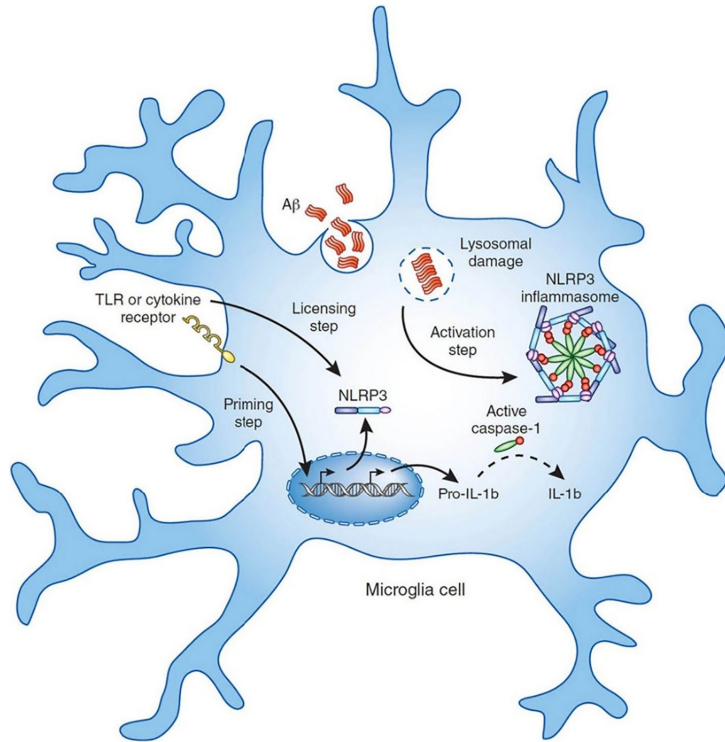


Fig. 4. Microglial cell diagram showing the formation of the NLRP3 inflammasome and cytokine cascade as a result of Aβ detection. Source: Heneka et al. [32].

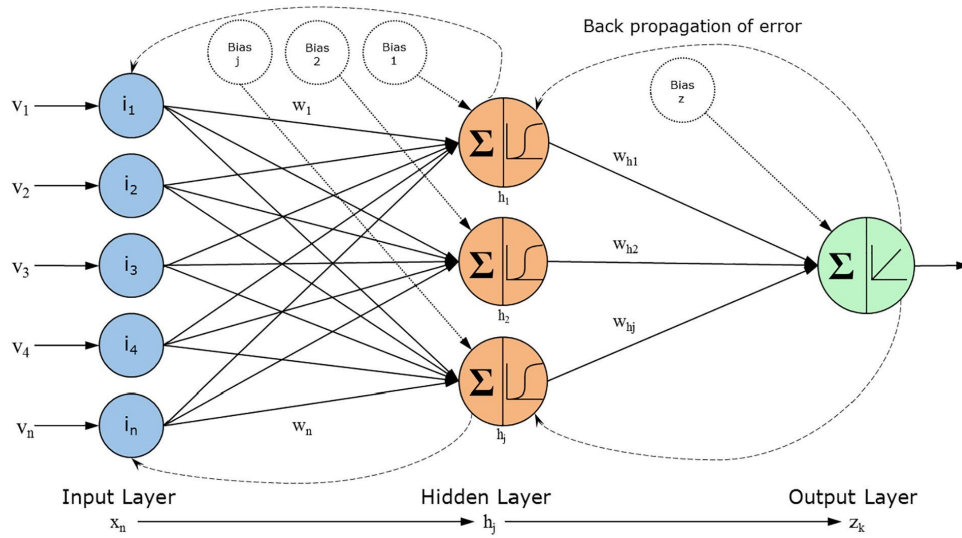


Fig. 5. Workflow diagram of the artificial neural network algorithm developed by Lancashire et al. [31] used for this project. The parameters for the hidden and output layer nodes are in their paper.

### 1.1. Categorical and Continuous

It is worth noting that two distinct versions of the algorithm were used – categorical and continuous. The categorical version seeks to interrogate the dataset using two predictors 0 and 1 for two distinct possibilities. This is based on known clinical information and a multitude of questions were considered. These questions include examining the differences between a healthy and an AD brain based on the overall gene expression as well as the differences between different regions in the brain, most notably the hippocampus. The continuous version of the algorithm allows us to consider every gene as its own independent predictor. This was used to examine the currently accepted biomarkers for AD [38,39] APP (amyloid beta precursor protein), MAPT (microtubule associated protein tau) and APOE (apolipoprotein E) and compare them to biomarkers discovered by the categorical algorithm.

### 1.2. Network Inference

The results obtained from the stepwise ANN approach were further analysed with an interaction algorithm developed by Lemetre et al. [34] to perform network inference. The interaction algorithm allows for the iterative quantification of the influence that multiple genes might have on the expression level of a single gene, until all the genes within the data have been quantified this way, using the same parameter values as those utilized for the ANN stepwise algorithm [34]. This allows for the determination of the central role of the most influential genes selected by the stepwise ANN within a system. The interaction algorithm predicts a single probe and assigns a weighted score which is directly proportional to the intensity of linkage between itself and the expression values of all other gene probes [35], while the intensity and directionality of the interaction between a source and target are determined based on the sum of the weights from an input to an output. The association between gene pairs can be bi- or unidirectional and be either stimulatory or inhibitory. This process was repeated until all gene probes were used as an output iteratively and a large matrix of interaction scores was generated by averaging values across 10 iterations. The results were visualised using Cytoscape. The methodology, proposed

by Tong et al. [40], is a novel ANN designed to infer directed gene-gene interactions in a pairwise manner, allowing the user to observe how changes in a given genes leads to changes in other genes and the network as a whole. The flowchart is included in the Supplemental Fig. 2.

### 1.3. Interaction Matrix

One of the greatest problems encountered during the previous approach when they are used to predict a single best marker is the fact that the selection process is stochastic; there is a random probability element and while the results can be statistically significant, it makes the process imprecise. To counter that effect and increase the power of this method, the top 500 genes selected by the stepwise process were split into 5 datasets of 100 genes each and combined into 16 sets of 200 genes each for network inference. This specific number was selected as the stepwise algorithm performance started to plateau after the first 400 genes indicating that the differentiation between the given conditions – AD and healthy – was decreasing. Once the network inference was completed, the data was consolidated and the top 1000 strongest interactions were selected and visualised with Cytoscape.

The reasoning behind developing this technique is that the normal single marker approach only focuses on a small subset (~0.1%) of the genes actively influencing a given condition. Moreover, by only selecting the 100 strongest interactions, it is guaranteed that in the resulting network, the biggest hubs, hence the most like drivers of the disease and targets for therapy, will be kept to a minimum and will be biased towards the most differentiated genes as seen in Fig. 6. It is important to note however, that for a highly focused system such as studying a specific subset of genes in a subset of a disease, such as proliferation markers in untreated breast cancer patients, the very nature of the data would result in a network where all the hubs are equally important. Thus, in such cases, identifying key markers and drivers using the strongest interactions is still the superior choice.

As seen in Fig. 6, upon separating the data to only include gene expression data exclusively from the hippocampus from AD patients only, selected as it is the area most strongly affected in AD, a rarely

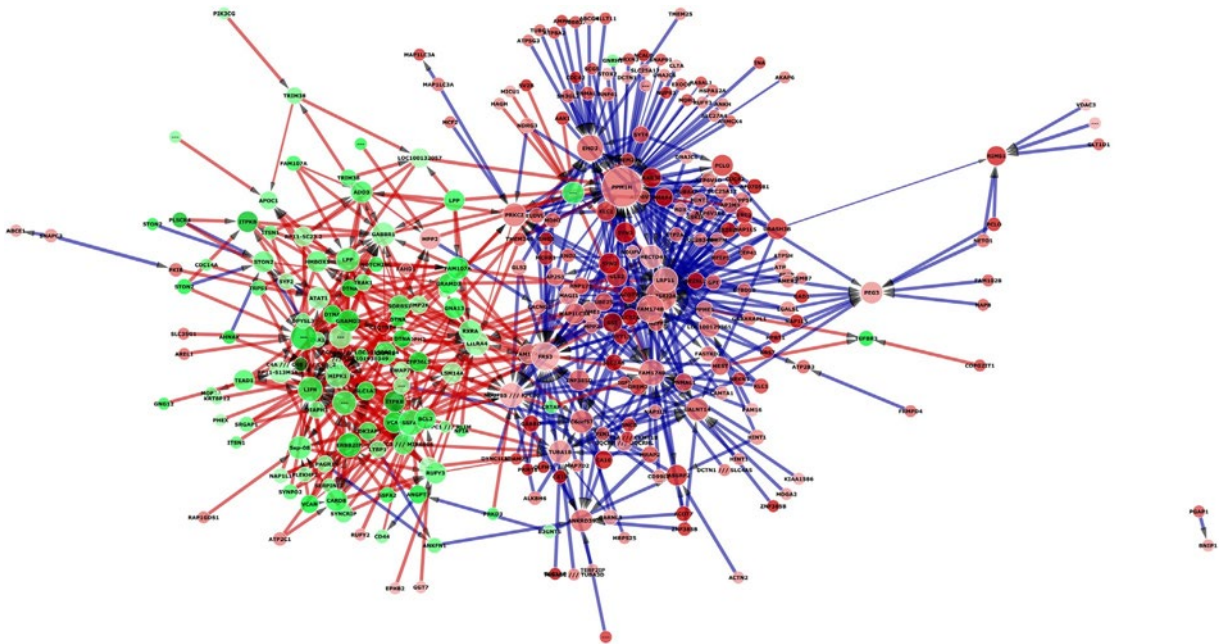


Fig. 6. Force directed interactome encompassing 500 gene probes and 1000 predicted interactions of the hippocampus in the E-GEOD-48350 AD cohort. Red edges indicate an inhibitory effect, whereas blue edges indicate promotion. Edge thickness is directly proportional to the strength of the interaction. Green nodes are upregulated genes while red ones are downregulated. The intensity of the colour is directly proportional to the degree of up- or downregulation.

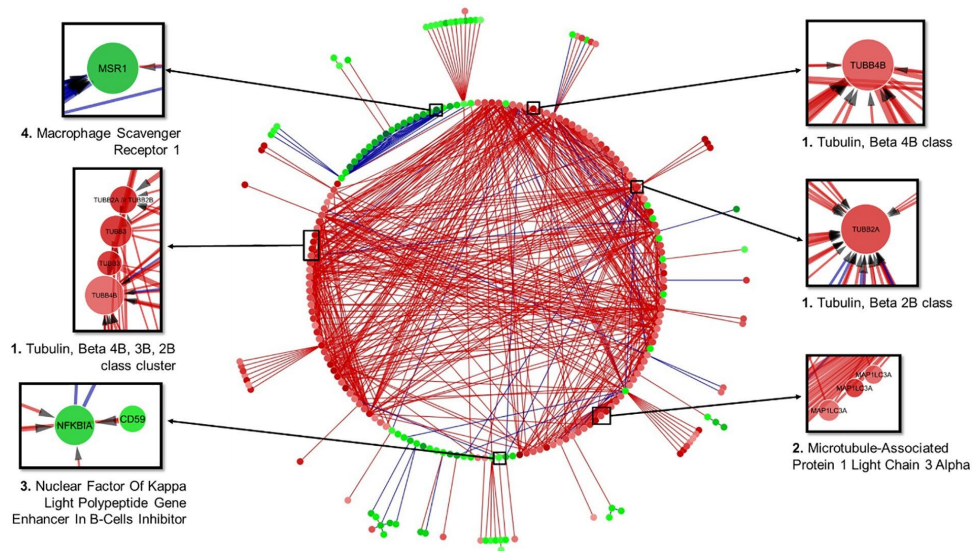


Fig. 7. Alternative circular layout interactome of the 1000 strongest interactions between 500 genes in AD independent of the brain region in the E-GEOD-48350 dataset. Based on the overall expression of all brain regions. Novel targets identified. Red edges indicate and inhibitory effect, whereas blue edges indicate promotion. Edge thickness is directly proportional to the strength of the interaction. Green nodes are upregulated genes while red ones are downregulated. The intensity of the colour is directly proportional to the degree of up- or downregulation.

seen duality presents itself. In most complex diseases such as cancer, the dysregulation that is represented in such interactomes is a direct result of the mechanisms of the disease. Successful cancers can hijack the body's immune response, avoid detection and proliferate uncontrollably. This in turn, leads to the body mounting a very strong response by attempting to upregulate anti-tumour factors and suppress proliferation factors among others in order to prevent the abnormal cells from disrupting the function of crucial organs [10]. Diabetes is similarly represented, as due to chronically high sugar levels the function of the organs affected get significantly damaged [41]. This leads to interactomes that are either mostly up- or down-regulated.

However, irrespective of the cause, non-familial AD is a direct result of the failure to regenerate damaged cells and clean away debris over a long period of time. Moreover, the isolated nature of the brain, the increased regulation of substances that can cross the blood brain barrier and most importantly the brain's plasticity, are crucial defence factors other organs lack. Plasticity is especially important as the brain can tolerate extensive damage before showing significant dysregulation, which is why AD is so hard to identify early [42]. As a result, the interactomes of affected regions show both up- and downregulation as it is possible to observe both suppression factors that could potentially be the direct cause of the disease and healing factors that are attempting to restore balance, as the mechanisms for it are still present and functional. In fact, dysregulation in the mechanisms involved in immune response and debris clearance could be used as predictors for early prognosis of AD as they are still functional, but increasingly ineffective.

This duality in the interactome however, reveals an interesting pattern within the data. Based on a fold change analysis of the original microarray data for AD in E-GEOD-48350, the genes that are over-expressed are downregulated overall. Conversely, underexpressed genes are predicted to be mostly downregulated. It is a fact that the hippocampus is the most dysregulated brain region in AD, so this is possible proof that the system is attempting to restore balance by suppressing the high expression of factors such as HIPK1 [43], a kinase which plays

an important role in senescence, ITPKB, a kinase that regulates inositol polyphosphates or BCL2, a protein phosphatase which is a crucial apoptosis factor. In short, the system is attempting to decrease the effect of genes involved in cell death.

The factors that are underexpressed on the other hand, appear to be upregulated and significantly more dysregulated, with an overall larger number and stronger individual interactions. The largest hub is PPM1H, another protein phosphatase which dephosphorylates CDKN1B, a CD kinase inhibitor involved in diseases such as Type IV Multiple Endocrine Neoplasia and familial Primary Hyperparathyroidism. Another such gene is FRS3, a fibroblast growth factor receptor substrate which is involved in regulation of RAS signalling.

While these genes and others like them seem to indicate that there is a significant effort to re-establish homeostasis, of further interest are the genes that do not fall inside these clearly defined categories. These genes include multiple tubulins such as TUBA1B and TUBB2A which are underexpressed but being simultaneously up- and downregulated, TGFBR3 which encodes for the transforming growth factor beta, type III receptor and plays a crucial role in cell adhesion and is associated with diseases such as familial cerebral saccular aneurysm. TGFBR3 itself activates transcription factors of the SMAD family, which in turn, regulates gene expression. ATP2C1 is an ATPase which catalyses the hydrolysis of ATP and is underexpressed while still attempting to down-regulate CARD8. CARD8 itself is caspase recruitment domain containing family of proteins and is involved in pathways negatively regulating the activation of NFkB, which as explained during the introduction, has a key role in the theory of neuroinflammation, and is quite likely an attempt to slow down or stop the chronic immune response leading to said neuroinflammation. Other irregularities include MAP1LC3A and MPP2 explained earlier and CD44, a cell-surface glycoprotein involved in cell-cell interactions, cell adhesion and migration and interacts with, among other things, matrix metalloproteinases (MMPs). MMPs, and MMP-9 in particular have long been suspected in playing a key role during AD and have been shown neuroprotective capabilities [44]. Finally, one of the most highly underexpressed and downregulated



genes is C1QTNF4, a complement-C1q tumour necrosis factor-related protein whose role is not clearly defined but has been suspected of acting like a pro-inflammatory cytokine, leading to the activation of NFKB and upregulate production of IL6.

Additionally, one of the major advantages of this method is that it generates a large and complex interactome that can be used to further examine a gene of interest as seen in Fig. 8.

In this example tubulin 2 beta (TUBB2A), a structural component of microtubules and a gene closely associated with tau, has consistently been in the top genes identified in AD across multiple tests. Due to the size of the previous interactome, there is enough complexity to be able to further analyse the way it interacts with other genes without having to use the algorithm again. If enough genes are identified as relevant to the question, then they can be used as predictors in the continuous ANN and then used for network inference. This also solves the major disadvantage of this methodology; it is computationally expensive and slow.

In Fig. 8 we can observe that TUBB2A is underexpressed but also downregulated by the clear majority of predicted interactions, including by other tubulin variants such as TUBB3 and TUBB4B as well as BRE which was discussed earlier. It is interesting however that both CASC3 and NFKBIA, both of which are overexpressed in this case, are attempting to upregulate TUBB2A, weakly in the case of NFKBIA but relatively strongly in the case of CASC3. CASC3 also appears to be very strongly downregulated by TUBB4B, MRPS25 a mitochondrial ribosomal subunit involved in mitochondrial translation and organelle maintenance and biosynthesis, and FARSB, a Phenylalanyl-TRNA Synthetase

Beta Subunit involved in tRNA aminoacylation and has been found to be associated with muscular dystrophy. Thus, it is possible to surmise that the dysregulated state of the TUBB2A gene in the network is directly correlated with mechanistic dysregulations in other genes that in turn affect genes responsible for regulation of TUBB2A itself. CASC3 and NFKBIA are failing to significantly upregulate TUBB2A back to normal levels due to dysregulation within themselves.

### 1.1. Driver Analysis

One of the challenges faced when trying to elucidate a marker, driver or therapy target is the selection criteria used. It is crucial to point out that the data used in these experiments presents us with a "snapshot" of the condition investigated, a generalized picture of how each gene is affected by every other gene, while the biological system is in a state of imbalance. As a result, the biggest hubs of most interactomes tend to be either the genes most up- or down-regulated in the network at the time. This has two potential interpretations. The hub is the source of the imbalance and thus, the most likely driver of the disease and target for therapy, and the downregulation is a result of the system attempting to restore balance, or that the hub is the factor preventing the imbalance by working against the disease and is being upregulated in an effort to restore the system to its original state.

The purpose of the driver analysis is to provide a non-biased selection condition based on the sum of the weights each gene exerts on the network, quantifying the amount of influence on a target and the amount of influence of a target. As explained in Section 2.4 the

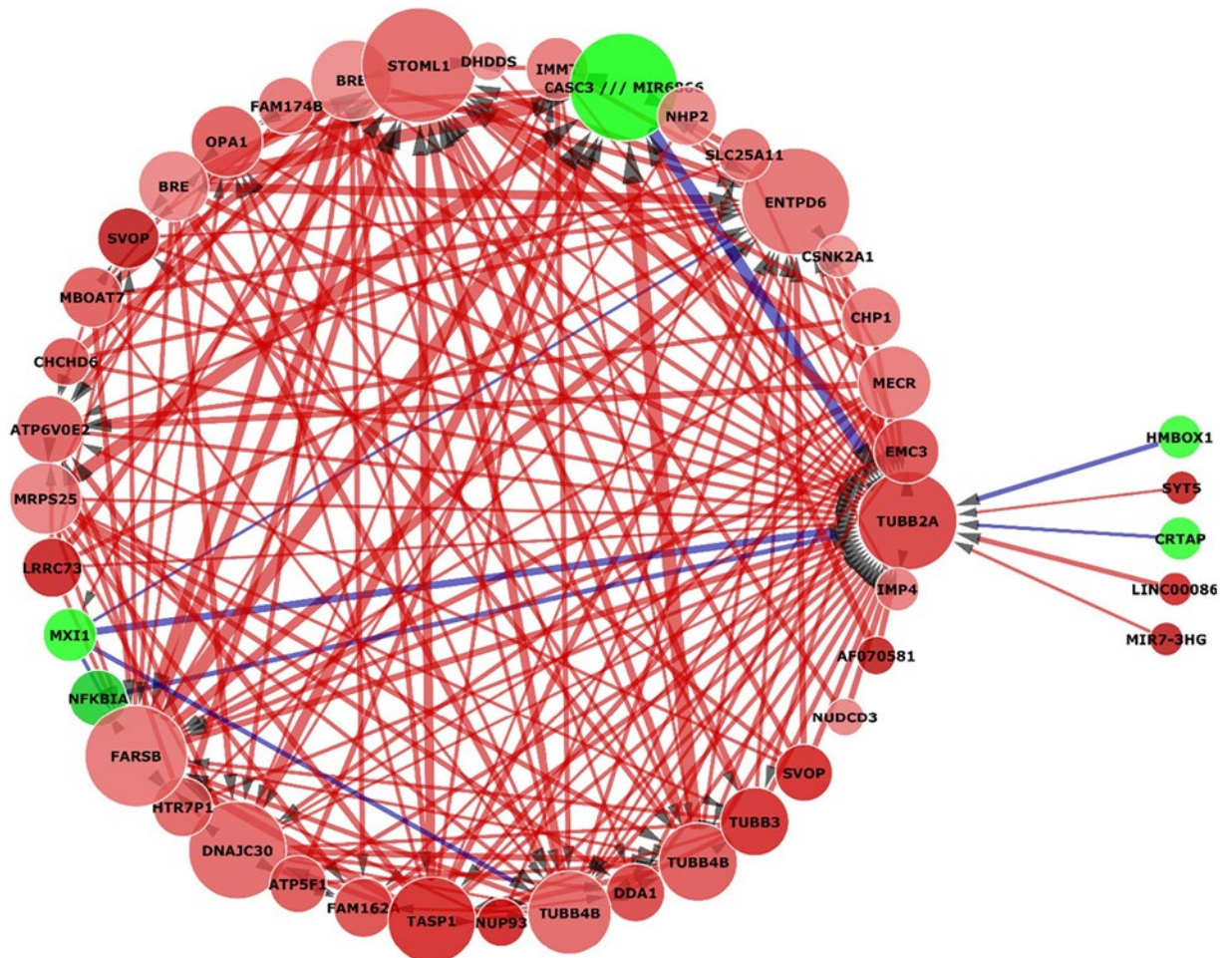


Fig. 8. Focused Tubulin interactome based on Fig. 7. Tubulin beta 2A interactions in AD. Of note is its positive regulation by an NFKB inhibitor.

interaction algorithm analyses the selected genes in a pairwise manner and assigns each of those pairs a value predicting how strongly their genes interact. Hence, by summing the weight that each source gene exerts on each target and vice versa, it becomes possible to rank them by which ones have the greatest overall effect on the network and which ones are the most affected.

The advantage of this method is the fact that it considers and gives equal importance to non-hubs as it only measures the total effect each gene has on the totality of the network. As such, it is possible to draw attention to genes with a multitude of weak interactions rather than only a few strong ones, which might otherwise not be visible. It is reasonable to assume that such genes may not be the greatest drivers of the disease, but crucial components of the system, and this method allows us to analyse those genes without

them being obscured by the hubs and most likely drivers, thus giving a wider and impartial view of the condition. Moreover, the driver analysis is not affected by the complexity of the question, being able to provide comparable results across multiple datasets, in both focused and general conditions.

The driver analysis was carried out on the 500 selected genes of the matrix interaction. The most influential source genes showed significant similarities and differences to the results of previous analyses on AD (Table 1). Genes identified in the interactome such as a ITPKB and CASC3 as well as trafficking proteins like TRAK1 and kinases like PRKD3 are expected. Of note is the disproportionate presence of BCL2 when compared to the interactome. However, the sources of interest include RHOBTB3, a member of the highly conserve family of Rho GTPases similar to RHOQ discovered during

**Table 1**  
Driver analysis showing the top 50 most influential and most influenced genes according to their unbiased impact on the network in the hippocampus in AD. The influence amount is the sum of all weights calculated by the interaction algorithm and is relative to the rest of the values. Probe IDs in red have not been annotated as of January 2017.

Amount of influence	Gene symbol	Amount influenced	Gene symbol
1213.081747	ITPKB	2381.32828	RP4-758J24.5
1155.143791	GNA13	1814.197348	PPM1H
1148.207036	RHOBTB3	1792.055655	C6orf57
1130.228993	VCAN	1754.53677	PRKCZ
1122.165399	PRKD3	1738.174507	FAM174B
1119.483983	ITPKB	1733.508537	FAM174B
1113.48562	TRAK1	1694.647661	LRP11
1108.416853	CASC3 /// MIR6866	1686.025743	CAPN2
1090.674524	SRGAP1	1643.902991	RASGRF2
1087.335279	LPP	1612.688333	FASTKD2
1028.750389	LIFR	1609.099447	1561158 at
1026.678888	GLIS3	1592.564443	RXRA
1025.359853	TEAD1	1560.277805	HIPK1
1018.653673	CARD8	1533.256201	SWAP70
1018.095788	ERBB2IP	1529.733365	GALNT14
1017.418527	RUFY3	1523.370265	LOC100129361
1012.441445	242611 at	1504.669747	PEG3
1010.030914	CRTAP	1473.99368	RP11-513M16.7
992.2756126	PABPC1 /// RLIM	1437.631403	HECTD4
982.4210022	SORBS1	1435.335801	SYF2
979.1729048	233323 at	1431.019129	1557286 at
973.5676705	SYNCRIP	1430.602853	TGFBR3
971.9687449	SEPT8	1419.648077	FAM107A
967.7392151	SSFA2	1390.736715	244457 at
967.402376	BCL2	1376.260249	BNIP1
966.0628739	DTNA	1366.651015	LTBP1
962.5225317	KLC1	1352.118161	B3GNT5
949.0374794	GRAMD3	1351.336497	CRTAP
935.7444619	FAM107A	1320.758301	RP11-5C23.2
933.9110942	SSFA2	1315.432274	ABCE1
930.2480972	HMBOX1	1314.897356	FAM174A
917.4727487	TRPS1	1312.652293	HMBOX1
913.4533421	PALLD	1310.768154	AP2S1
913.3942276	FAM107A	1302.110691	GPS1
909.7624557	BCL2L1	1289.970537	MOB1A
905.7671419	CDK2AP1	1282.458469	ALKBH6
904.8356429	VCAN	1270.746624	KRT8P12
904.3397815	CAPN2	1264.244601	MAG11
902.6081661	233323 at	1255.168137	ANKRD39
899.9210524	NOTCH2NL	1251.151278	DNAJC6
896.8863383	ZFP36L1	1240.782655	EHD3
893.9583626	ZNF385B	1231.946711	238466 at
888.5853093	ADD3	1227.9842	AREL1
880.9829708	WWTR1	1218.986806	ATAT1
876.8780354	PALLD	1211.954338	LILRA4
861.0770622	SYN2	1207.832696	LIFR
860.3346604	NFIA	1207.407361	TUBA1B
859.6319203	228297 at	1204.867524	GABBR2
851.0770584	DTNA	1195.223424	ITPKB
849.629429	AP2M1	1194.137548	PLEKHB2

earlier testing, as well as SRGAP1. SRGAP1 encodes for a GTPase activator and works in conjunction to CDC42, a GTPase of the same family, to negatively regulate neuronal cell migration. Moreover, when combined with receptor ROBO1, it can deactivate CDC42. Its presence so high on the source list as a downregulating factor, indicates that its function is being stronger than expected, resulting in slower cell migration and impediment of the regeneration process. CARD8, discussed earlier, has a strong, negative effect on the network, suppressing the expression of related genes.

Meanwhile, the most targeted genes on the network include PPM1H, a protein phosphatase, TGFBR3, multiple kinases, and an alpha-tubulin TUBA1B. More beta tubulins are included in the complete list. Also, although rarely seen, ATAT1, an alpha tubulin acetyltransferase, a neuronal cell component crucial to the microtubule growth appears to be negatively regulate. ATAT1 is involved in coenzyme binding and tubulin *N*-acetyltransferase activity and only acetylates older microtubules, being unable to act on unstable ones. Genes such as APGAT1 which fulfil similar purposes have been discovered in previous test, suggesting that slower/weaker acetylation of older microtubules could play a key role in the development of AD. Curiously, one of the upregulated factors is AREL1, apoptosis resistant  $\epsilon 3$  ubiquitin protein ligase 1, which inhibits apoptosis. It is possible that it is being upregulated in an attempt to keep the neurons alive and functioning to prevent further damage. Finally, the presence of ITPKB as both a significant source and target indicate that it is a crucial component of the system regardless of disease state. The results will be used for a functional analysis via the Bioconductor R package [45]. A second table regarding the driver analysis of the cohort of cognitively normal controls is available in the Supplemental Table 2 for comparison.

## 1. Conclusions and Future Developments

In conclusion, the results obtained by this series of experiments show promise for a greater understanding of the biology behind Alzheimer's disease, its progression and the mechanisms involved. By expanding to other brain regions and datasets and focusing the questions on the most relevant genes, it is possible to identify new markers and drivers of the disease that can be used alongside the current ones to improve prognosis and provide more targets for therapy.

It is worth noting that the results obtained and analysed with this pipeline have been generated without using a null hypothesis, in a non-parametric manner. The only question was the difference between AD and healthy brains and was expanded to include predictors as general as the presence of the disease down to the expression of individual genes. It is evident by the results that by reducing the bias introduced by datamining for very focused questions and increasing the variance, we are presented with multiple potential biomarkers as well as new discovery routes such as further evidence of the role on inflammation and microtubule stabilisation. The pipeline has thus managed to generate unbiased, varied and novel information that can be used to guide further, more targeted research as well as validation of these results experimentally.

Future development will focus on improving the speed and power of the algorithms and increase the interpretability of the results. Using general-purpose computing on graphics processing units, it is possible to reduce the time requirements by up to 75% at the cost of computational power, though recent advances in the field have made it significantly more likely and affordable. Further tests are being focused on the variance between different brain regions as well as the effect of individual genes on the system. Moreover, this series of tests is being repeated in RNA-seq and proteomic datasets in order to study the effect of AD pre and post translation, as well as other gene expression datasets to ensure consistency in the results.

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.csbj.2018.02.001>.

- [1] Lemetre C, Lancashire LJ, Rees RC, Ball GR. Artificial neural network based algorithm for biomolecular interactions modeling. *Bio-inspired systems: computational and ambient intelligence*. 2010. p. 877–85.
- [2] Blair JL, Nordhus BA, Hill SE, Scaglione KM, O'Leary JC, Fontaine SN, et al. Accelerated neurodegeneration through chaperone-mediated oligomerization of tau. *J Clin Invest* 2013;123(10):4158–69.
- [3] Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, et al. ArrayExpress update—simplifying data submissions. *Nucleic Acids Res* 2015;43:1113–6.
- [4] Barrett T, Wilhite ES, Ledoux P, Evangelista C, Kim FK, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* 2013;41:991–5.
- [5] Sharma N, Singh NA. Exploring biomarkers for Alzheimer's disease. *J Clin Diagn Res* 2016;10.
- [6] Blennow K. CSF biomarkers for Alzheimer's disease: use in early diagnosis and evaluation of drug treatment. *Expert Rev Mol Diagn* 2014;5:661–72.
- [7]

## References

- [1] Cook CE, Bergman MT, Finn RD, Cochrane G, Birney E, Apweiler R. The European bioinformatics institute in 2016: data growth and integration. *Nucleic Acids Res* 2015;44:D20–6.
- [2] Alyass A, Turcotte M, Meyre D. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Med Genomics* 2015;8.
- [3] Holzinger A, Dehmer M, Jurisica I. Knowledge discovery and interactive data mining in bioinformatics - state-of-the-art, future challenges and research directions. *BMC Bioinforma* 2014;15.
- [4] Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol* 2016;17.
- [5] Swan AL, Stekel DJ, Hodgman C, Allaway D, Alqahtani MH, Mobasher A, et al. A machine learning heuristic to identify biologically relevant and minimal biomarker panels from omics data. *BMC Genomics* 2015;16.
- [6] Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet* 2015;16:321–32.
- [7] Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform* 2017. <https://doi.org/10.1093/bib/bbx044>.
- [8] Chatzimichail E, Matthaios D, Bouros D, Karakitsos P, Romanidis K, Kakolyris S, et al. gamma-H2AX: a novel prognostic marker in a prognosis prediction model of patients with early operable non-small cell lung cancer. *Int J Genomics* 2014;160236.
- [9] Bertolaccini L, Solli P, Pardolesi A, Pasini A. An overview of the use of artificial neural networks in lung cancer research. *J Thorac Dis* 2017;9:924–31.
- [10] Abdel-Fatah TM, Agarwal D, Liu DX, Russell R, Rueda OM, Liu K, et al. SPAG5 as a prognostic biomarker and chemotherapy sensitivity predictor in breast cancer: a retrospective, integrated genomic, transcriptomic, and protein analysis. *Lancet Oncol* 2016;17:1004–18.
- [11] Mehdy MM, Ng PY, Shair EF, Saleh NIM, Gomes C. Artificial neural networks in image processing for early detection of breast cancer. *Comput Math Methods Med* 2017;2017:22610628.
- [12] Huang SH, Loh JK, Tsai JT, Houg MF, Shi HY. Predictive model for 5-year mortality after breast cancer surgery in Taiwan residents. *Chin J Cancer* 2017;36:23.
- [13] Abdel-Fatah TM, McArdle SE, Agarwal D, Moseley PM, Green AR, Ball GR, et al. HAGE in triple-negative breast cancer is a novel prognostic, predictive, and actionable biomarker: a transcriptomic and protein expression analysis. *Clin Cancer Res* 2016;22:905–14.
- [14] Haj-Hassan H, Chaddad A, Harkouss Y, Desrosiers C, Toews M, Tanougast C. Classifications of multispectral colorectal cancer tissues using convolution neural network. *J Pathol Inform* 2017;8:1.
- [15] Makris GM, Poulidakis A, Siristatidis C, Margari N, Terzakis E, Koureas N, et al. Image analysis and multi-layer perceptron artificial neural networks for the discrimination between benign and malignant endometrial lesions. *Diagn Cytopathol* 2017;45:202–11.
- [16] Duthley B. Alzheimer disease and other dementias; 2013.
- [17] Yiannopoulou GK, Papageorgiou GS. Current and future treatments for Alzheimer's disease. *Ther Adv Neurol Disord* 2013;6(1):19–33.
- [18] Farlow M, Anand R, Messina Jr J, Hartman R, Veach J. A 52-week study of the efficacy of rivastigmine in patients with mild to moderately severe Alzheimer's disease. *Eur Neurol* 2000;44(4):236–41.
- [19] Calignon A, Fox LM, Pitstick R, Carlson GA, Bacskai BJ, Spire-Jones TL, et al. Caspase activation precedes and leads to tangles. *Nature* 2010;464(7292):1201–4.
- [20] Mucke L, Selkoe DJ. Neurotoxicity of amyloid  $\beta$ -protein: synaptic and network dysfunction. *Cold Spring Harb Perspect Med* 2012;2(7).
- [21] Iqbal K, Grundke-Iqbal I. Opportunities and challenges in developing Alzheimer disease therapeutics. *Acta Neuropathol* 2011;122:543.
- [22] An R, Bienkowski MJ, Shuck ME, Miao H, Tory MC, Pauley AM, et al. Membrane-anchored aspartyl protease with Alzheimer's disease beta-secretase activity. *Nature* 1992;402(6761):533–7.
- [23] Wallace WC, Bragin V, Robakis NK, Sambamurti K, Vanderputten D, Merrill CR, et al. Increased biosynthesis of Alzheimer amyloid precursor protein in the cerebral cortex of rats with lesions of the nucleus basalis of Meynert. *Mol Brain Res* 1991;10(2):173–8.
- [24] Reitz C. Alzheimer's disease and the amyloid cascade hypothesis: a critical review. *Int J Alzheimers Dis* 2012. <https://doi.org/10.1155/2012/369808>.
- [25] Karran E, Mercken M, Strooper B. The amyloid cascade hypothesis for Alzheimer's disease: an appraisal for the development of therapeutics. *Nat Rev Drug Discov* 2011;10:698–721.
- [26] Salloway S, Sperling R, Fox NC, Blennow K, Klunk W, Raskind M, et al. Two phase 3 trials of bapineuzumab in mild-to-moderate Alzheimer's disease. *N Engl J Med* 2014;370:322–33.
- [27] Selkoe DJ, Hardy J. The amyloid hypothesis of Alzheimer's disease at 25 years. *EMBO Mol Med* 2016;8:595–608.
- [28] Shankar GM, Li S, Mehta TH, Garcia-Munoz A, Shepardson NE, Smith I, et al. Amyloid-beta protein dimers isolated directly from Alzheimer's brains impair synaptic plasticity and memory. *Nat Med* 2008;14:837–42.
- [29] De Strooper B, Karran E. The cellular phase of Alzheimer's disease. *Cell* 2016;164:603–15.
- [30] Aloisi F. Immune function of microglia. *Glia* 2001;36(2):165–79.
- [31] Lancashire LJ, Lemetre C, Ball GR. An introduction to artificial neural networks in bioinformatics—application to complex microarray and mass spectrometry datasets in cancer studies. *Brief Bioinform* 2009;10(3):315–29.
- [32] Heneka TM, Golenbock TD, Latz E. Innate immunity in Alzheimer's disease. *Nat Immunol* 2015;16:229–36.
- [33] Lancashire LJ, Rees RC, Ball GR. Identification of gene transcript signatures predictive for estrogen receptor and lymph node status using a stepwise forward selection artificial neural network modelling approach. *Artif Intell Med* 2008;43:99–111.
- [34] Tong DL, Boock DJ, Dhondalay GKR, Lemetre C, Ball GR. Artificial Neural Network Inference (ANNI): a study on gene-gene interaction for biomarkers in childhood sarcomas. *PLoS One* 2014;9:e102483.
- [35] Li J, Lee H, Wang Y, Tong AH, Yip KY, Tsui SK, et al. Interactome-transcriptome analysis discovers signatures complementary to GWAS loci of type 2 diabetes. *Sci Rep* 2016;6.
- [36] Kempf SJ, Metaxas A, Ibañez-Vea M, Darvesh S, Finsen B, Larsen MR. An integrated proteomics approach shows synaptic plasticity changes in an APP/PS1 Alzheimer's mouse model. *Oncotarget* 2016;7:33627–48.
- [37] Prinz A, Tavernarakis N. The role of SUMOylation in ageing and senescence decline. *Mech Ageing Dev* 2017;162:85–90.
- [38] Fragkouli A, Tsilibary EC, Tzinia AK. Neuroprotective role of MMP-9 overexpression in the brain of Alzheimer's 5xFAD mice. *Neurobiol Dis* 2014;70:179–89.
- [39] Huber W, Carey JV, Gentleman R, Anders S, Carlson M, Carvalho SB, et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* 2015;12:115–21.

# Appendix

## 1. Complete results tables

Due to the sheer size and amount of data present in the results tables, they have been submitted separately in portable storage. If required, they can be made available by contacting the project supervisor.

## 2. Dataset Breakdown

### E-GEOD-48350 (Blair *et al*, 2013)

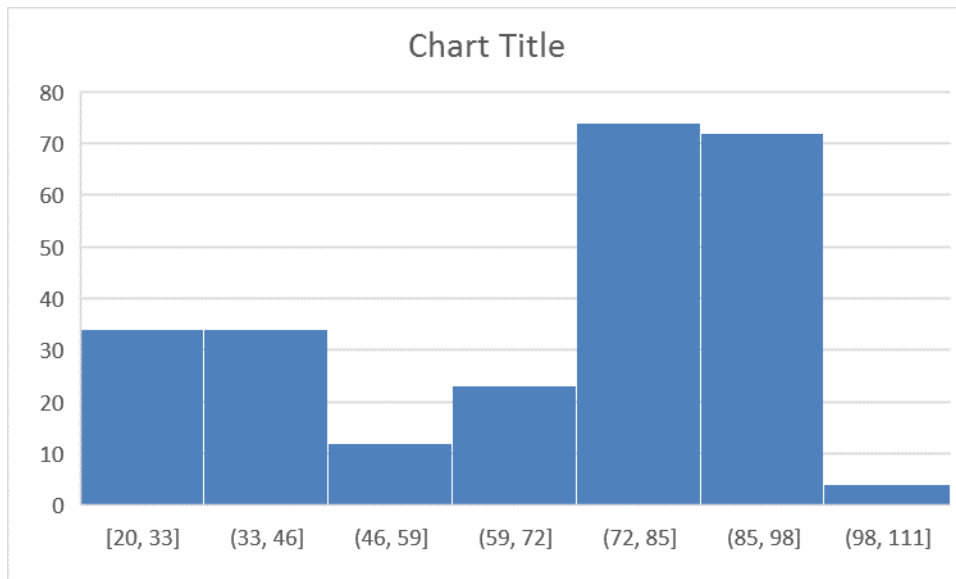
Array used: A-AFFY-44 - Affymetrix GeneChip Human Genome U133 Plus 2.0 [HG-U133\_Plus\_2]

Size: 253 samples, 54676 gene probes

Sample breakdown:

Entorhinal Cortex -	Healthy – 18 female, 21 male AD – 15 female, 7 male
Hippocampus -	Healthy – 20 female, 23 male AD – 10 female, 9 male
Postcentral Gyrus -	Healthy – 20 female, 23 male AD – 15 female, 10 male
Superior frontal gyrus -	Healthy – 24 female, 24 male AD – 14 female, 7 male

Age Breakdown:



E-GEOD-5281 (Liang *et al*, 2007)

Array used: A-AFFY-44 - Affymetrix GeneChip Human Genome U133 Plus 2.0 [HG-U133\_Plus\_2]

Size: 161 Samples, 54675 gene probes

Sample breakdown:

Superior frontal gyrus - Healthy – 10 female, 13 male

AD – 4 female, 7 male

Primary Visual Cortex - Healthy – 8 female, 11 male

AD – 3 female, 9 male

Posterior cingulate cortex - Healthy – 3 female, 6 male

AD – 4 female, 9 male

Middle temporal gyrus- Healthy – 6 female, 10 male

AD – 4 female, 8 male

Hippocampus - Healthy – 4 female, 6 male

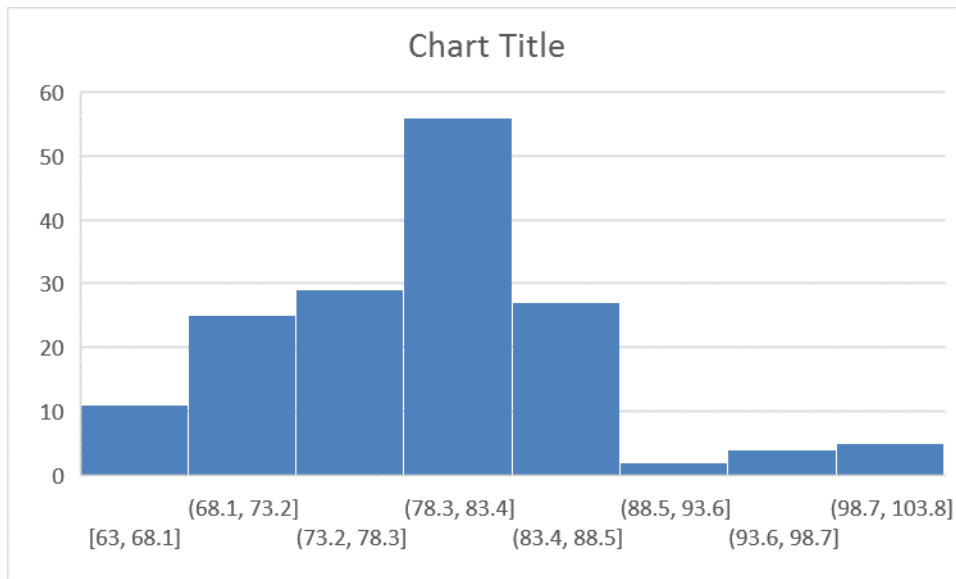
AD – 3 female, 10 male

Entorhinal Cortex - Healthy – 6 female, 4 male

AD – 3 female, 10 male

Age Breakdown





E-GEOD-9770

Array used: A-AFFY-44 - Affymetrix GeneChip Human Genome U133 Plus 2.0 [HG-U133\_Plus\_2]

Size: 35 samples, 54375 genes

All samples are MCI

breakdown

entorhinal cortex – 6

hippocampus – 6

middle temporal gyrus – 5

posterior cingulate cortex – 5

primary visual cortex – 5

superior frontal gyrus - 6

3. Commonality tables

<b>Commonalities AD-Healthy</b>	<b>Commonalities AD-Master</b>	<b>Commonalities Healthy-Master</b>
<b>Gene Symbol</b>	<b>Gene Symbol</b>	<b>Gene Symbol</b>
PHIP	PHIP	KLHL35
TUBB3	TUBB3	ITFG1

ENSA	ENSA	PHIP
ACP2	ACP2	TUBB3
MSANTD3-TMEFF1	MSANTD3-TMEFF1	ENSA
USP5	USP5	ACP2
MBOAT7	NHP2	SSSCA1
IMP4	MBOAT7	MSANTD3-TMEFF1
TUBB3	IMP4	DPYSL4
AGPAT1	TUBB3	USP5
DCTPP1	CHP1	SYT5
MECR	AGPAT1	SYT5
PCYOX1L	DCTPP1	BFSP1
TMEM59L	MECR	KCNF1
DLK2	TMEM59L	MBOAT7
HBQ1	DLK2	CSNK2A1
LOC100129361	HBQ1	IMP4
DNAJC30	LOC100129361	MAST3
MAP1LC3A	DNAJC30	TUBB3
CECR6	MAP1LC3A	SST
MRPS25	CECR6	AGPAT1
RPUSD3	MANBAL	DCTPP1
MAP1LC3A	MRPS25	DDA1
SVOP	RPUSD3	MECR
LRRC73	MAP1LC3A	TMEM59L
SPRN	SVOP	DLK2
AGPAT1	LRRC73	ADAMTS8
TRAPPC9	SPRN	HBQ1
	AGPAT1	LOC100129361
	TRAPPC9	DNAJC30
		MAP1LC3A
		CECR6
		MRPS25
		RPUSD3
		ZNRF1
		MANEAL
		C12orf73
		MAP1LC3A
		SVOP
		MAP1LC3A
		LRRC73

MSANTD1
SPRN
AGPAT1
TRAPPC9

<b>Commonalities AD-Healthy</b>	<b>Commonalities AD-Master</b>	<b>Commonalities Healthy-Master</b>
<b>Gene Symbol</b>	<b>Gene Symbol</b>	<b>Gene Symbol</b>
<b>CRTAP</b>	CRTAP	CRTAP
<b>NFKBIA</b>	NFKBIA	ITFG1
<b>SSSCA1</b>	KIF5B	RNF165
<b>TUBB2A</b>	ACP2	CDIPT
<b>CDKN2C</b>	CDKN2C	NUDCD3
<b>CASC3 /// MIR6866</b>	B4GALNT1	NFKBIA
<b>CAPN2</b>	CASC3 /// MIR6866	SSFA2
<b>PNISR</b>	TM7SF2	CDKN2C
<b>CD59</b>	RHOQ	ZNF443
<b>MPP2</b>	PNISR	CASC3 /// MIR6866
<b>TUBB3</b>	CD59	MAX
<b>CHP1</b>	MPP2	RAB27A
<b>SMAD5</b>	SLC8A2	PNISR
<b>RBM6</b>	RBM6	LPIN1
<b>GNRH1</b>	MS4A14	CD59
<b>CRIP1</b>	ITPKB	MPP2
<b>LARP7</b>	GNRH1	KIAA1551
	CRIP1	ERBB2IP
	LARP7	PRKD3
		MRPS25
		RBM6
		ZXDC
		GNRH1
		CRIP1
		ZRANB3
		RHOBTB3
		LARP7
		ATP5F1
		TRAPPC9