Detecting condition-specific protein clusters in expression data sets

Dissertation
zur Erlangung des Grades
des Doktors der Naturwissenschaften
der Naturwissenschaftlich-Technischen Fakultät
der Universität des Saarlandes
von
Sudharshini Thangamurugan

Saarbrücken, January 2025

Tag des Kolloquiums: 28th August, 2025
Dekan: Prof. Dr.-Ing. Dirk Bähre
Berichterstatter: Prof. Dr. Volkard Helms
Prof. Dr. Bruce Morgan

Vorsitz: Prof. Dr. Jochen Hub
Akad. Mitarbeiter: Priv.-Doz. Dr. Jessica Hoppstädter

Abstract

This thesis explores aspects of proteomics, transcriptomics and dental research using computational and experimental approaches to address scientific challenges. In transcriptomics, distinguishing diseased from healthy cell states is crucial for understanding disease mechanisms and identifying therapeutic targets. Differential protein abundances and rewired protein-protein interactions (PPIs) provide insights into disease progression. RNA-seq data analysis combined with PPIXpress, PPICompare, and ClusterONE identified as effective tools for detecting significant differential networks and protein clusters relevant to melanoma progression, enriched by Reactome pathways for targeted therapies and diagnostics.

In dental research, the limited availability of human enamel has led to the use of hydroxyapatite (HAP) as a surrogate to study biofilm formation. A systematic comparison of HAP and enamel examined protein composition, molecular functions, isoelectric points and molecular weight patterns. The results highlighted similarities between HAP and enamel, supporting the suitability of HAP for preventive dental research.

The study also investigated peroxiredoxins (Prxs), a family of antioxidant enzymes. Several studies have suggested that Prx isoforms form hetero-oligomers. In silico modelling using HADDOCK and AlphaFold analysed their structural dynamics, contributing to our understanding of protein biochemistry.

Zusammenfassung

In dieser Arbeit werden Aspekte der Proteomik, der Transkriptomik und der Dentalforschung mit Hilfe rechnerischer und experimenteller Ansätze untersucht, um wissenschaftliche Herausforderungen zu bewältigen. In der Transkriptomik ist die Unterscheidung zwischen kranken und gesunden Zellzuständen entscheidend für das Verständnis von Krankheitsmechanismen und die Identifizierung von therapeutischen Zielen. Unterschiedliche Proteinhäufigkeiten und neu verdrahtete Protein-Protein-Interaktionen geben Aufschluss über das Fortschreiten von Krankheiten. Die Analyse von RNA-seq-Daten in Kombination mit PPIXpress, PPICompare und ClusterONE hat sich als wirksames Instrument zur Erkennung signifikanter differentieller Netzwerke und Proteincluster erwiesen, die für das Fortschreiten des Melanoms relevant sind und durch Reactome-Pfade für gezielte Therapien und Diagnosen angereichert werden.

In der zahnmedizinischen Forschung hat die begrenzte Verfügbarkeit von menschlichem Zahnschmelz dazu geführt, dass Hydroxylapatit (HAP) als Surrogat für die Untersuchung der Biofilmbildung verwendet wird. In einem systematischen Vergleich von HAP und Zahnschmelz wurden Proteinzusammensetzung, molekulare Funktionen, isoelektrische Punkte und Molekulargewichtsmuster untersucht. Die Ergebnisse zeigten Ähnlichkeiten zwischen HAP und Zahnschmelz auf, was die Eignung von HAP für die präventive Zahnforschung unterstreicht.

Die Studie untersuchte auch Peroxiredoxine (Prxs), eine Familie von antioxidativen Enzymen. Mehrere Studien haben gezeigt, dass Prx-Isoformen Hetero-Oligomere bilden. Die In-silico-Modellierung mit HADDOCK und AlphaFold analysierte ihre strukturelle Dynamik und trug so zu unserem Verständnis der Proteinbiochemie bei. (Translated using DeepL.com)

Acknowledgements

This period of personal development, spanning the past three years, has been a significant one. It is therefore appropriate to take this opportunity to express gratitude to all those who have played a role in the journey undertaken in Saarbrücken. This period has been crucial in imparting knowledge and fostering self-reliance, thus shaping the person I am today.

Firstly, I would like to express my sincere gratitude to my supervisor, Professor Volkhard Helms, for the invaluable opportunity, unwavering support and mentorship he has provided throughout the duration of my doctoral studies. His profound guidance has been instrumental in enabling me to develop and execute my projects with confidence and independence. I would also like to extend my sincere appreciation to Professor Bruce Morgan for taking time out of his busy schedule to review of my thesis. Additionally, I would like to express my profound gratitude to Deutsche Forschungsgemeinschaft for financially supporting my thesis work through SFB 1027.

In addition, I would like to thank the staff involved in various research projects, namely Dr Johanna Dudek, Dr Simone Trautmann, Dr Lilly Lemke and Dr Bruce Morgan.

I would also like to thank my colleagues Trang, Hanah, Debarshee, Aram, Andreas and Markus, who have helped to create a pleasant atmosphere at the university. Finally, I would like to express my deep gratitude to Ms Kerstin Gronow-Pudelek, whose constant support and guidance in organisational matters has been invaluable.

I would like to thank my mother, Sumathi, for travelling across continents and being there for me during the final stages of my PhD. I would also like to thank Muthu, who has played a pivotal role in my life. His constant support and encouragement has been instrumental in motivating me to strive for self-improvement. Without Muthu's constant motivation, I would not have been able to dedicate myself to learning and working with such determination. Thank you Muthu!

Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Motivation

Proteins are essential macromolecules in all living organisms, serving as the primary functional and structural units within cells [1]. They perform a multitude of roles, including biocatalysis, cell communications, immune response, molecular mobility, and structural support. The complex three-dimensional structures of proteins, which are determined by their amino acid sequences, underpin their specific biological functions [2]. Understanding proteins and their mechanisms of action is, therefore, a fundamental aspect of biological research. In particular, protein-protein interactions (PPIs) form the backbone of cellular networks, coordinating processes such as signal transduction, immune responses and metabolic regulation [3]. Unravelling these interactions is crucial to understanding the molecular basis of health and disease.

Despite notable advances in experimental and computational methodologies, the landscape of protein interactions remains only partially elucidated. The development of high-throughput experimental technologies has markedly enhanced our capacity to identify and characterise protein-protein interactions. Techniques such as yeast two-hybrid screening, affinity purification coupled with mass spectrometry, and proximity labelling have enabled the systematic mapping of interaction networks. Concurrently, computational approaches have become indispensable for the processing, modelling and interpretation of the vast datasets generated by these experimental methods [4, 5, 6]. The application of bioinformatics tools, molecular docking simulations, and network analysis frameworks has proved invaluable in elucidating the structure, dynamics, and functional implications of protein interactions. Although high-throughput techniques have provided insights into the vast networks of protein interactions, the datasets generated are often incomplete

or non reproducible, limiting their utility for comprehensive biological interpretation [7]. Moreover, the transient and context-dependent nature of many PPIs represents a significant challenge for experimental characterisation [8]. Computational tools and algorithms have begun to address these limitations, offering scalable and integrative methods for the analysis of protein data. However, as biological systems become increasingly complex, it is imperative that existing approaches evolve to meet the demands of more accurate predictions, deeper structural insights and improved functional annotation of interactions.

Protein abundance data, derived from technologies such as mass spectrometry based proteomics, provides a snapshot of the concentration of proteins within a system under specific conditions [9, 10]. These quantitative measurements are of great value for the exploration of differential expression patterns in response to external stimuli, stress, or disease [11, 12]. Nevertheless, the interpretation of protein abundance data is a challenging process. The presence of variability in experimental conditions, technical noise, and potential biases in protein detectability can result in the masking of genuine biological signals [13, 14]. The application of advanced data analysis techniques, including normalization, statistical modelling, enrichment analysis and machine learning, is essential for the extraction of meaningful insights [13] and the integration of protein abundance data with other omics layers, such as transcriptomics and metabolomics.

In addition to understanding and analysing protein abundance data and protein interactions, it is equally critical to examine the three-dimensional conformations of proteins. The spatial structure of a protein is pivotal to understanding its function, as it dictates the protein's interaction capabilities, stability, and activity. Conventionally, the determination of protein structures has been undertaken through the utilisation of experimental techniques, including nuclear magnetic resonance (NMR) spectroscopy, X-ray diffraction, and cryo-electron microscopy (cryo-EM). Whilst these methodologies yield high-resolution structural insights, they are often laborious, costly and reliant on the specialised expertise of structural biologists. Moreover, the experimental determination of protein structures is constrained by factors such as the size of the protein, its flexibility, and the difficulty of crystallization, which renders it inaccessible for many targets. In silico modelling has emerged as a powerful alternative for predicting protein structures and interactions. The employment of computational methods enables researchers to generate structural models with efficiency, thereby reducing dependency on labour-intensive experimental techniques. Notable tools in this field include AlphaFold [15], HADDOCK [16], and MODELLER [17]. These tools not only accelerate the process of structure determination but also enable the modelling of complexes, interactions, and conformational changes that are often challenging to capture experimentally.

There is a great need for more robust, integrative and interpretable tools for protein data analysis. Continued research in this field will not only address critical knowledge gaps but also facilitate new avenues for innovation

in drug discovery, precision medicine, and synthetic biology. This motivation highlights the necessity of developing methodologies for protein interaction analysis, which will facilitate both fundamental scientific inquiry and applied research.

1.2 Overview

The present thesis is primarily concerned with the downstream processing of transcriptomic and proteomic data. The objective was to integrate transcriptomic data with protein interaction data in order to identify clusters that are responsible for the observed change in cell state. In this project, we employed a melanoma RNA-seq dataset to identify clusters of genes that work together to convert a healthy cell into a melanoma cell. In addition to this project, I also contributed to the analysis of proteomic data derived from saliva and biofilm pellicles collected from the oral cavity of subjects, which were subjected to nano liquid chromatography-mass spectrometry. The objective was to undertake a comparative analysis of the proteins deposited on pellicles of different surfaces and conditions, as well as those present in saliva. The study yielded multiple conclusions. These include the potential use of HAP as an alternative material for enamel research studies, the observation that proteins remaining in saliva and not adsorbed to form a pellicle exhibit contrasting molecular functions, and the identification of a completely different composition of proteins in the pellicles of patients with caries compared to those of patients who have undergone treatment or have no history of caries. Furthermore, I contributed to the prediction of protein complex structures through the utilisation of in silico modelling techniques, namely HADDOCK [16] and AlphaFold [15]. The peroxiredoxin proteins were the subject of study, which are known to exist as homo-dimeric structures that homo-oligomerise to form a decameric structure based on redox-dependent equilibrium. In recent years, evidence has emerged indicating that two monomers belonging to two closely related peroxiredoxin proteins have undergone hetero-dimerisation. The docking of heterodimers demonstrates the potential for heterodimer formation at the corresponding active sites.

1.2.1 First author publication

S. Thangamurugan, V. Helms, "Comparing workflows for combining transcriptomic with protein interaction data", Submitted to IEEE Transactions on Computational Biology and Bioinformatics.

Abstract:
Characterising and understanding the differences between diseased and healthy cell states is a growing focus in biomedical research. As first step one may identify differentially expressed and construct differential networks of the proteins encoded by these genes to reveal their roles in biological processes. Alternatively, differences in cell states can be identified by constructing condition-specific networks and estimating differential networks to detect

rewired protein-protein interactions (PPIs). These rewired interactions, which are newly exhibited or lost in diseased states, have the potential to provide crucial insights for targeted therapy. This study compared six software pipelines that infer rewired protein-protein interactions based on RNA-seq data from diseased and healthy cells. In the last step, each pipeline identified cohesive protein clusters in the differential network to predict protein components that may jointly exert certain molecular functions. The biological relevance of these clusters was assessed through Reactome pathway enrichment. The PPIXpress and PPICompare tools in combination with the DESeq2 package were able to identify highly meaningful, compact protein clusters when comparing RNA seq data of nevi and melanoma samples. Such pipelines appear to be a good basis for enhancing our understanding of disease progression and for finding solutions for targeted therapies

1.2.2 Coauthor publications

J. Dudek, T. Faidt, C. Fecher-Trost, S. Thangamurugan, P. Bayenat, S. Trautmann, A. Holtsch, F. Müller, V. Helms, K. Jacobs, M. Hannig, "Synthetic hydroxyapatite – a perfect substitute for dental enamel in biofilm formation studies", Submitted to Scientific Reports.

Abstract:
Dental enamel consists mainly of hydroxyapatite (HAP). The chemical composition of dental apatite differs between individuals and influences enamel properties. In contact with saliva, enamel is covered immediately by salivary molecules. This initial biofilm is mostly composed of proteins, and in addition to its protective properties mediates bacterial adhesion. The resulting bacterial biofilm is a highly complex ecosystem, which can provoke the development of diseases. Therefore, dental biofilms are the focus of preventive research. Chemically standardized surfaces are a good choice for dental biofilm studies. Synthetic HAP pellets meet the criteria for such well-defined enamel-like model surfaces. Therefore, we compared the in situ biofilm formation on HAP and enamel. No differences in formation kinetics, microstructure, and thickness of the initial biofilm on both materials were detected. Differences in the proteome composition depended mainly on the volunteer and not on the surface material. Formation kinetics and morphology of the bacterial biofilm as well the coverage with bacteria were also not distinguishable. However, bacterial viability on enamel was lower, which might be due to the presence of fluorine in enamel. Overall, synthetic HAP can be considered as a full substitute substrate for enamel. For viability studies, synthetic HAP may even be the preferred substrate.

My contribution: I carried out the qualitative and quantitative proteomic analysis and the statistical test of similarity. I also wrote the first drafts of the of the corresponding method sections and prepared the corresponding figures in the manuscript.

S. Trautmann, S. Thangamurugan, C. Fecher-Trost, J. Dudek, V. Flockerzi, V. Helms, and M. Hannig, "Snapshot of the seconds-pellicle – first insights in its ultrastructure, proteomic composition and changes over time", planned to submit to Journal of dental research.

Abstract: The dental pellicle is a continuously growing layer present at the interface between the oral surfaces and saliva. It possesses protective properties by shielding the dental surfaces against chemical and mechanical damages. The pellicle represents the basis for all oral biofilms and its formation starts immediately after oral hygiene through the adsorption of mostly salivary proteins to all exposed surfaces. The objective of the current study was 1) to visualize and elucidate the individual proteomic composition of the seconds pellicle 2) to analyse changes in its proteomic composition over time.

The in situ-pellicle was formed on polished bovine enamel or ceramics specimens. Transmission electron microscopic analyses were used to depict its ultrastructure at different formation times starting from 5 s up to 30 min. Its complex proteomic composition and changes over time were analysed by gel electrophoresis in combination with coomassie-staining and nano-tandem mass spectrometry, analysing individual 10-s, 3-min and 30-min pellicle samples.

This resulted in very first insights in the ultrastructure and proteome of the very initial seconds-pellicle, possessing an unexpected high number of up to 841 proteins on individual level. The analysis of several pellicle formation times enabled the qualitative and quantitative analysis of changes in the pellicle proteome over time as well as the first direct verification of protein desorption processes occurring during pellicle formation. Individual analyses enabled the identification of the core pellicle proteome at the different formation times. Comparisons between the salivary and pellicle proteomes at the different formation times revealed insights in enriched and depleted proteins within the pellicle and variations in the amount of substance of single pellicle proteins over time.

These informations represent the basis for future selective modifications of the pellicle layer in order to control of the hence originating biofilm and develop preventive strategies for the oral biofilm management. **My contribution:** I conducted qualitative and quantitative proteomics analysis, including differential analysis, molecular weight and isoelectric point distributions and enrichment analysis of molecular functions. Furthermore, I drafted the initial versions of the corresponding method sections and prepared the relevant figures for the manuscript.

T. Do, S. Thangamurugan, V. Helms, "PPIXpress and PPICompare Webservers infer condition-specific and differential PPI networks", In press at Bioinformatics Advances.

Abstract: We present PPIXpress and PPICompare as two webservers that enable analysis of protein-protein interaction networks (PPINs). Given a reference PPIN and user-uploaded expression data from one or multiple samples, PPIXpress constructs context-dependent PPINs based on major transcripts and

high-confidence domain interactions data. To derive a differential PPIN that distinguishes two groups of contextualized PPINs, PPICompare identifies statistically significant altered interactions between multiple context-dependent PPINs from PPIXpress. We present a case study where PPIXpress and PPICompare webservers were used in combination to construct the PPINs specific for melanocytic nevi and primary melanoma cells, and to detect the rewired protein interactions between these two sample types
Availability and Implementation: PPIXpress and PPICompare webservers are available at `https://service.bioinformatik.uni-saarland.de/ppi-webserver/indexPPIXpress.jsp` and `https://service.bioinformatik.uni-saarland.de/ppi-webserver/indexPPICompare.jsp`, respectively. Alternatively, the webservers and application updates can be found at `https://service.bioinformatik.uni-saarland.de/ppi-webserver/`
My contribution: I conducted the case study using the PPIXpress and PPICompare webserver and significant results were identified which validate the functioning of the tool. Furthermore, I drafted the initial versions of the corresponding case study sections and prepared the relevant figures for the manuscript.

A. Papazian, S. Agerbaek, S. Thangamurugan, M. Bengtson Løvendorf, B. Dyring-Andersen, V. Helms, "Rewiring of protein-protein interactions inferred from proteomics data of nevi and melanoma samples", planned to resubmit this manuscript to Journal of Proteome Research
Abstract: Analysis of protein-protein interactions (PPIs) may provide deeper insight into protein functions and reveal how components of cellular pathways interact with each other. There are multiple software tools that infer condition-specific PPI networks based on transcriptomic data. However, it is well known that the transcriptome does not always reflect the proteome. To address this, we present here a new computational approach that is able to characterize rewiring events in PPI networks based on proteomic data exclusively. We applied this approach by comparing a cohort of 14 nevi samples to their corresponding nevus-associated melanomas. Thereby, we were able to identify differentially expressed proteins, deregulated protein interactions, protein clusters and hub proteins that were consistent with previously published evidence on the progression of nevi to melanoma. Additionally, we compared the proteomics data against matched spatial transcriptomics data (GeoMX data for 8 out of 14 samples). This comparison highlighted the different outcomes of each approach and further emphasize the importance of considering multiomics analysis.
My contribution: I conducted statistical analysis to compare the proteomic data with respective genomic data with the objective of inferring any potential overlap of results.

Chapter 2
Background

This chapter introduces the fundamental concepts pertaining to proteins, protein-protein interaction networks, and public repositories for these networks, and reviews the basic properties of human PPINs and those of model organisms such as *Saccharomyces cerevisiae*. It also considers the identification of complexes within protein-protein interaction networks and the diverse computational tools and software employed for the computational analysis of complex biological data. The sections 2.2 and 2.5 were adapted and expanded from Chapter 5: Identification of Putative Protein Complexes in Protein-Protein Interaction Networks, written by myself and Prof. Volkhard Helms as part of the book Protein Interactions: The Molecular Basis of Interactomics edited by Helms and Kalinina (2022). [18].

2.1 Proteins

The central dogma of molecular biology postulates that deoxyribonucleic acid (DNA) contains characteristic, defining genes and instructions for protein synthesis. The dogma provides an explanation of the unidirectional flow of genetic information from DNA to RNA to protein [19]. A schematic representation of this process can be found in Fig. 2.1. Proteins constitute a group of macromolecules, composed of sequences of amino acids, which are responsible for a range of cellular processes. For example, they are involved in metabolic processes that provide structural support to cells, they act as enzymes that catalyse biochemical reactions within cells, and they are involved in the transmission of signals that enable cells to interact with their environment. It is therefore evident that the formation of proteins and their interactions with other macromolecules are of paramount importance in determining a cell's function and overall health. The function of each protein is contingent upon its distinctive amino acid sequence and three-dimensional conformation.

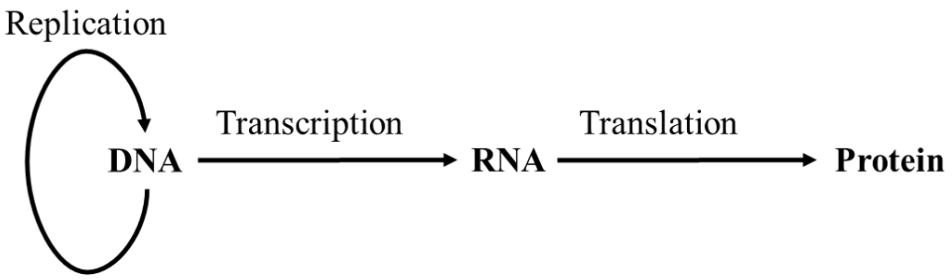


Figure 2.1: The central dogma of molecular biology, as pioneered by Francis Crick, postulates the transmission of genetic information within a biological system. This theory elucidates the process of unidirectional transfer of information from nucleic acid to nucleic acid or nucleic acid to proteins.

The analysis of proteins offers a fundamental understanding of human physiology and the progression and diagnosis of diseases. Given the diversity of proteins, which are distinguished by their specific combination of amino acids, their physicochemical properties exhibit a range of characteristics. Consequently, a variety of analytical, separation and identification techniques are employed in order to exploit the different properties of proteins. The most commonly employed technique for protein identification and quantification is liquid chromatography coupled with mass spectrometry called LC-MS. This combination of techniques is most effective when employed for targeted and untargeted protein analysis. Liquid chromatography is a technique employed for the separation of a mixture of diverse substances, including molecules and proteins, in a mobile liquid phase that traverses a column. During this process, the components interact with the stationary phase to varying degrees based on their distinct physicochemical properties, resulting in their elution from the column in a specific order. For example, the interaction between the molecules and the stationary phase may vary depending on their molecular size. Larger molecules may elute from the column at a faster rate than smaller molecules, potentially due to their ability to traverse the pores of the column more effectively. Similarly, other physicochemical properties are employed to separate the proteins based on charge, hydrophobicity, affinity, and so forth. This method is renowned for its high efficiency, accuracy and selectivity [20].

As an additional benefit, reducing the diameter of the chromatographic column allows for more precise analysis and quantification. This is called the nano LC-MS technique. A reduction in diameter results in a decrease in the quantity of sample and solvents required, thereby enabling the analysis of smaller volumes and less concentrated samples. Moreover, reducing the diameter of the column enables more precise regulation of the flow rates of the liquid mobile phase, thereby ensuring a linear speed of the molecules [21]. The eluted molecules are then converted to ions for quantitation in mass spectrometry. These ions are subsequently separated according to their mass-to-charge ratio, and their ratio and abundances are measured qualitatively and quantitatively.

2.2 Protein-Protein Interactions

In biological systems, the majority of cellular and molecular mechanisms are dependent on protein activity. Infrequently, a single protein serves as the regulator or executor of an entire mechanism. Alternatively, proteins frequently bind to other biomolecules, frequently other proteins, in order to execute cellular functions. Protein-protein interactions (PPIs) are defined as highly specific physical contacts between two or more proteins. These interactions are formed due to the conformational and physico-chemical properties of the involved proteins.

The majority of biological cells are composed of water, with the remaining dry weight consisting of proteins, which account for 40–55% of the total [22]. It can be thus surmised that freely diffusing cytosolic proteins frequently collide with other cellular proteins and may occasionally remain bound to each other for a short time as a non-specific assembly. It can be reasonably assumed that only a small proportion of these interactions will involve two or more proteins that are naturally destined to bind with each other. Specific PPIs can be classified according to their lifetime, with transient and stable interactions representing the two main categories. Transient (specific) interactions between proteins are of short duration and serve to perform functions such as signal transduction or to instigate further changes (for example, the sodium-potassium pump). Stable protein interactions are characterised by a long lifetime and frequently serve the purpose of forming macromolecular machines, such as haemoglobin or RNA polymerase.

For a single protein, all its physical interactions with other proteins can be represented as a mathematical graph, where the vertices represent the proteins and the undirected edges connecting the vertices represent the physical interactions between the proteins. A protein-centred network offers insight into the protein complexes in which the protein of interest may be involved and their associated biological functions. For instance, the enzyme aspartate semialdehyde dehydrogenase from Arabidopsis thaliana has been observed to be involved in three distinct protein complexes that are active in either an oxidation-reduction process, in methionine biosynthesis, or in lysine biosynthesis [23]. In contrast, protein-protein networks (PPINs) are global PPI graphs or networks that provide an overview of all PPIs existing in an organism. These comprehensive networks are catalogued by several established databases, including the Biological General Repository for Interaction Datasets (BioGRID) [24], Mentha [25], the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) [26], the Molecular Interaction Database (MINT) [27], the Protein Interaction Database (IntAct) [28], and others. Fig. 2.2 illustrates the connectivity of a small toy protein-protein interaction network (PPIN).

In graphs, the degree of a vertex is defined as the number of edges connected to it. In PPINs, this value thus represents the number of interactions involving the protein represented by the vertex. One method for examining the general connectivity and topology of a PPIN is to compute its degree

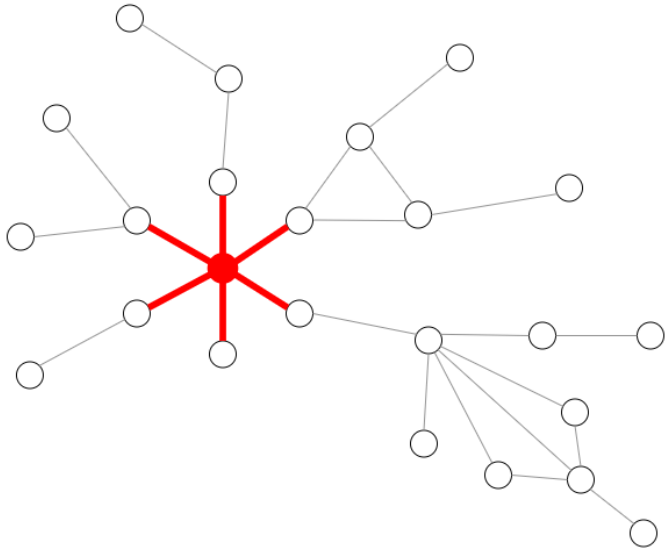


Figure 2.2: Schematic representation of a protein-protein interaction network. The circles are called the vertices of the network and represent all copies of an individual protein type. The lines connecting the vertices are called edges and represent physical contact between two proteins. The degree of a vertex measures the number of edges connected to it. The vertex highlighted in red has 6 edges connected to it or 6 binding partners, and hence its degree is 6. Note that this representation does not carry information on whether multiple interactions of one protein can occur simultaneously, potentially leading to the formation of a larger protein complex, or not.

distribution, which describes the frequency of each vertex degree occurring in the given network. The visual representation of degree distributions is frequently achieved through the use of plots, wherein the vertex degrees are displayed on the x-axis and their respective frequencies are represented on the y-axis. The analysis of multiple PPINs from diverse species revealed that these networks exhibit a "scale-free" topology, irrespective of the species in question [29]. In scale-free networks, the degree distribution follows a power law with a negative exponent λ, whereby the probability of vertex degree k is given by $P(k) = k^{-\gamma}$. Consequently, the highly connected proteins, referred to as hubs, are observed to occur at a significantly higher frequency than would be expected in an exponentially decaying scenario, where the probability is given by $P(k) = e^{-\gamma k}$. This scale-free nature implies that the average length of the shortest pathway between any two vertices increases at a much slower rate as a function of network size than would be expected under an exponentially decaying model.

2.3 Integration of various PPI resources in public data repositories

The field of protein interaction databases is characterised by two distinct categories: primary databases and meta databases. Primary databases collate the findings of numerous experimental interaction assays. Notable examples include the Biomolecular Interaction Network Database (BIND) [30], the IntAct molecular interaction database, the Molecular INTeraction Database (MINT), the Database of Interacting Proteins (DIP) [31], and the Biological General Repository for Interaction Datasets. In contrast, metadatabases typically integrate data from multiple primary databases. To illustrate, the Integrated Interactions Database (IID) [32] collates data from BIND, BioGRID, DIP, MINT, IntAct, and a few other sources, whereas the Agile Protein Interactomes DataServer [33] contains interactions from BioGRID, DIP, IntAct, MINT, and the Human Protein Reference Database (HPRD) [34]. In the case of model organisms, the meta-database Mentha integrates evidence-based interactions from BioGRID, DIP, IntAct, and MINT.

The International Molecular Exchange (IMEx) consortium, an international collaboration of major contributors of PPI data, has established guidelines to maintain a consistent set of uniquely defined molecular identifiers and interactions[35]. A number of prominent databases, including IntAct, MINT, DIP and BIND, are active members of the IMEx consortium. The STRING database is of particular interest as it provides interactions from both the IMEx consortium and BioGRID.

Database	Type	PPI Source	Species	Website
BIND	Primary	Evidence	Multiple	http://binddb.org
BioGRID	Primary	Evidence	Multiple	thebiogrid.org
DIP	Primary	Evidence	Multiple	dip.doe-mbi.ucla.edu
IntAct	Primary and meta	Evidence	Multiple	ebi.ac.uk/intact
MINT	Primary	Evidence	Multiple	mint.bio.uniroma2.it
APID	Meta: BioGRID, DIP, HPRD, IntAct, and MINT	Evidence	Multiple	apid.dep.usal.es
IID	Meta: IntAct, MINT, BioGRID, BIND, DIP, and others	Evidence and predicted	Multiple	iid20.ophid.utoronto.ca
mentha	Meta: BioGRID, DIP, IntAct, MINT and others	Evidence	Multiple	mentha.uniroma2.it
STRING	Meta: IMEx consortium and BioGRID	Evidence and predicted	Multiple	string-db.org

Table 2.1: Overview of selected primary and meta protein-protein interaction databases.

2.4 Protein-protein interaction networks of model organisms

As previously stated, the majority of biological processes within an organism are facilitated by protein interactions. An understanding of the interactome of an organism facilitates the identification of the proteins and genes associated with a specific process or disease. This, in turn, facilitates more accurate and comprehensive identification and mechanistic comprehension of disease-related pathways and potential regulatory mechanisms. Consequently, a significant aspect of computational systems biology is the examination and comparison of PPINs in one or more organisms, with the objective of elucidating the mechanisms and regulations of biological systems.

2.4.1 PPIN of *Saccharomyces cerevisiae*

In order to characterise the protein-protein interaction network of the eukaryotic model organism *Saccharomyces cerevisiae*, Gavin et al. [36] employed Tandem-Affinity Purification (TAP) and mass spectrometry. Uetz et al. [37] and Fields et al. [38] used the yeast two-hybrid method. The results obtained by these methods yielded protein-protein interaction networks (PPINs) comprising 16,000–40,000 interactions, involving the majority of the 6,000 yeast proteins. As previously stated, the network displays a power-law connectivity distribution, whereby a small number of proteins exhibit high connectivity and form hubs, while the majority of proteins interact with only a limited number of other proteins. In the initial experiments, the coverage of PPIs was relatively limited, at approximately 10%. This prompted concerns about the true scale-free nature of these networks, and whether they appeared to be scale-free due to other factors [39]. Nevertheless, the subsequent expansion of the coverage demonstrated that they do, in fact, possess a scale-free topology [40].

A crucial question is which vertices are the most significant in a PPIN. One method of defining importance is to ascertain whether a gene product is essential for the cell. If an "essential" gene is removed from the genome, the result is the death of the cell. In contrast, cells in which non-essential genes have been knocked out remain viable. In their experimental studies, Winzeler et al. [41] and Giaever et al. [42] demonstrated that approximately 1120 (19%) of the protein-coding genes in *S. cerevisiae* are "essential." A Gene Ontology analysis of these genes revealed that approximately 74% are involved in metabolic processes, while at least 14% are involved in cell cycle regulation. These appear to be the two essential functions for cell survival [43]. Fig. 2.3 illustrates the interconnectivity of yeast proteins based on the data in the latest version of the Mentha database `http://www.mentha.uniroma2.it/`. The proteins are coloured according to their essentiality (red) or non-essentiality (green).

Interestingly, when information about protein connectivity was combined with information about gene essentiality, it was found that highly

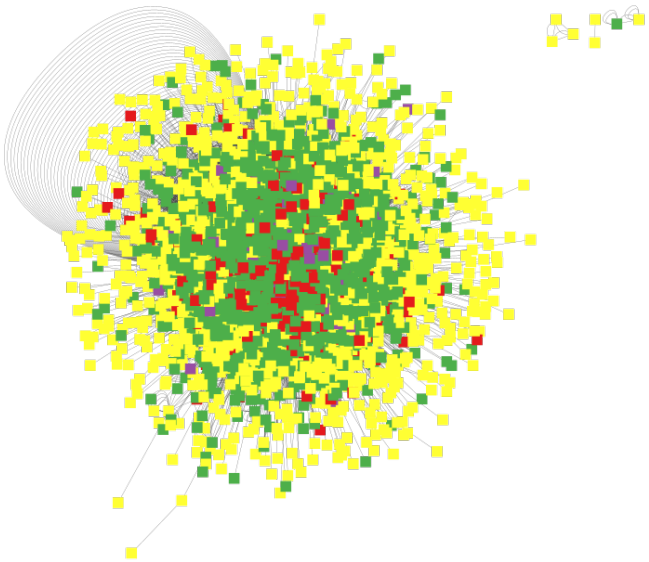


Figure 2.3: Complete interactome of *S. cerevisiae* derived from the mentha database and constructed using Cytoscape [44]. The interactome contains 6,342 genes and 233,322 interactions. Red node vertices represent essential genes (948 essential genes identified), green node vertices represent non-essential genes (3,583 non-essential genes identified), purple node vertices represent conditional genes (270 conditional genes identified) and yellow node vertices represent unknown essentiality (1,541 genes have unknown essentiality).

connected "hub proteins" are much more likely to be encoded by essential genes (about 60%) than low-degree proteins (about 15%) [29]. This makes intuitive sense. Knocking out a highly connected protein is likely to cause a major disruption in cellular processes. This behaviour is illustrated in Fig. 2.4, which plots the proportions of essential vs. non-essential genes as a function of the connectivity of the proteins they encode in the yeast PPIN.

Following the initial studies based on yeast two-hybrid screens mentioned above, various high-throughput methods have been used to determine PPIs in *S. cerevisiae*, such as HT mass spectrometric protein complex identification (HMS-PCI) [45] correlated mRNA expression, in silico predicted interactions, etc. In total, there is now confirmed evidence for approximately 80,000 interactions between *S. cerevisiae* proteins [46]. When the early data were pooled, about 2,400 of the 80,000 interactions were common to more than one high-throughput method [47]. This may be due to certain biases in the detection assays. Some methods, such as Y2H, were reported to have relatively high false positive rates (around 59%) or to be unable to detect certain types of interactions. For example, the Y2H method was found to detect comparatively fewer proteins that regulate translation [47]. Therefore, [48] constructed a 'filtered yeast interactome' (FYI) dataset by intersecting data from different methods. This interactome consists of 2,493 high-confidence

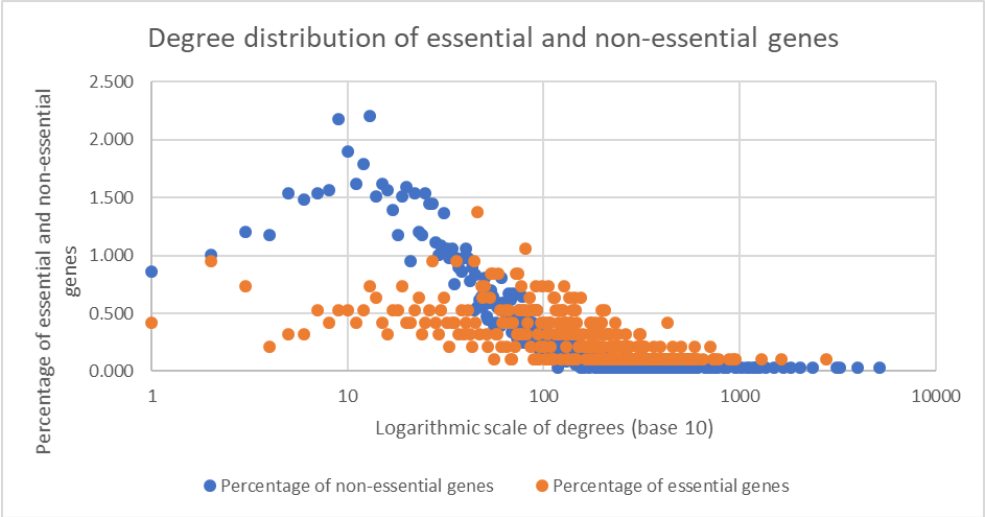


Figure 2.4: For the degree distributions for essential proteins and non-essential proteins (shown on the x-axis on a log scale), we colour-coded the respective fractions of essential (orange) and non-essential (blue) genes for different degrees. This analysis recovers the previously observed enrichment of essential genes/proteins among the high-degree vertices of the PPIN.

interactions (observed in common by at least two methods to rule out false positives), 1,379 proteins with an average of 3.6 degrees of interaction per protein, and a large connected component of 778 proteins. For each hub in the FYI, an average Pearson correlation coefficient (avPCC) was calculated, correlating the hub and its binding partners under different conditions. The hubs with a degree greater than 5 showed a bimodal probability distribution for a few conditions. The hubs with a degree greater than 5 showed a bimodal probability distribution for a few conditions. The hubs with a degree of 5 or less showed a normal distribution centred at 0.1. It was understood that the bimodal distribution suggests two kinds of hub types, static hubs and dynamic hubs, based on their expression profiles, see Fig. 2.4. In the 91 identified static hubs, the binding partners interact at the same time and are involved in the main functional part of the complex. In the 108 identified dynamic hubs, the binding partners interact with each other at different times or in different locations and rather tend to connect separate modules of the PPIN. The hubs with degree 5 or less showed a normal distribution centred at 0.1. It was assumed that the bimodal distribution suggested two types of hubs, static hubs and dynamic hubs, based on their expression profiles, see Fig. 2.5. In the 91 static hubs identified, the binding partners interact simultaneously and are involved in the main functional part of the complex. In the 108 dynamic hubs identified, the binding partners interact at different times or locations and tend to connect separate modules of the PPIN.

To the best of our knowledge, the distribution of essential proteins in either dynamic or static hubs has yet to be analysed. Batada et al. [49] proposed that the current data set for the protein-protein interaction network (PPIN)

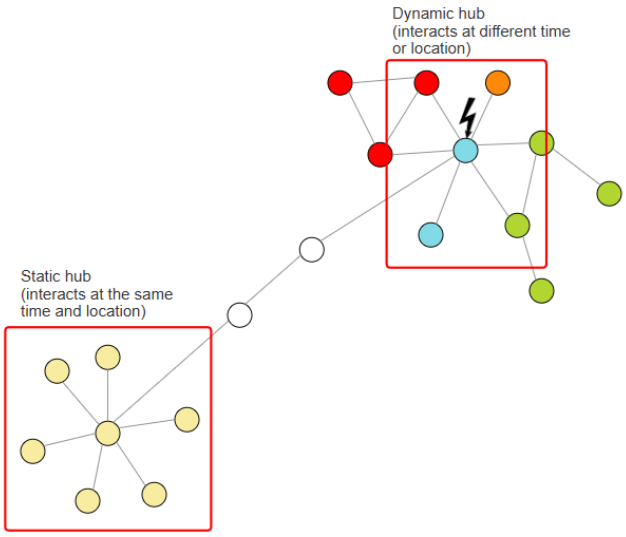


Figure 2.5: Schematic representation of static and dynamic hubs of the network. Proteins of the static hub interact with each other at the same time and location. Proteins of dynamic hubs interact with each other at different times or locations.

of *S. cerevisiae* is insufficient to draw definitive conclusions or differentiate between hubs. A number of the 5 conditions from the compendium utilised only 10 data points to differentiate the hubs, which may be insufficient for accurate differentiation. Agarwal et al. [50] reported that calculating avPCC for hubs in all conditions of the compendium, rather than using only five conditions, yielded 59 dynamic hubs with the same degree of threshold as 5. This demonstrates that the differentiation of hubs is primarily based on the expression profile and can vary with different experimental conditions. It is therefore questionable whether avPCC is an appropriate parameter for differentiating hubs. Furthermore, based on their functions, the hubs exhibited a spectrum of structural roles, which makes it challenging to differentiate them as static and dynamic hubs.

2.4.2 PPIN of Human

According to data from the GTEx consortium, 20,532 potential protein-coding human genes have been annotated [51]. It is a significant challenge to comprehensively map the interaction network among human proteins. In the initial phase of the project, Stelzl et al. [52] constructed a partial human protein interaction network based on yeast two-hybrid screening of 4,456 bait and 5,632 prey proteins. This yielded 3,186 novel interactions between 1,705 proteins. In a more recent study, Agarwal et al. [53] collated data from the Menche et al. [54] and Chatr-Aryamontri et al. [55] studies, as well as fifteen additional databases. This resulted in the construction of a comprehensive network comprising 342,353 interactions between 21,557 proteins.

In the study Shin and colleagues [56] examined the process of identifying drug targets within the human protein-protein interaction (PPI) network. The development of any pharmaceutical agent commences with the identification of a drug target, that is to say, a receptor protein that possesses a druggable binding pocket. As previously stated, PPIs are instrumental in regulating biological pathways, including the onset and progression of disease. It has been proposed that consideration of the PPI network of humans is advantageous for the identification of novel drug targets [57]. In recent years, approximately 40 PPIs have been identified as potential drug targets for drug development from the human interactome, which comprises 130,000-650,000 PPIs [58]. New computational structure-based approaches have been presented for the determination of inhibitors of PPIs, which are termed Small Molecule Protein-Protein Interaction Inhibitors (SMPPIs). To illustrate, the small ubiquitin-like modifier (SUMO) protein forms a covalent interaction with proteins that possess a SUMO interaction motif (SIM) through a process known as sumoylation. This process regulates a number of general cellular processes, including cell proliferation, chromosome winding, DNA replication and DNA repair. Additionally, it is involved in the development of neurodegenerative diseases and cancer. In light of the electrostatic similarity with the native binding partner protein, as identified by the software Elekit [59], an inhibitor was discovered that binds to the SUMO protein with low micromolar activity and interferes with the SUMO-SIM interaction [60]. Fig. 2.6 provides a schematic illustration of the rationale behind the design of mimicking SMPPIs.

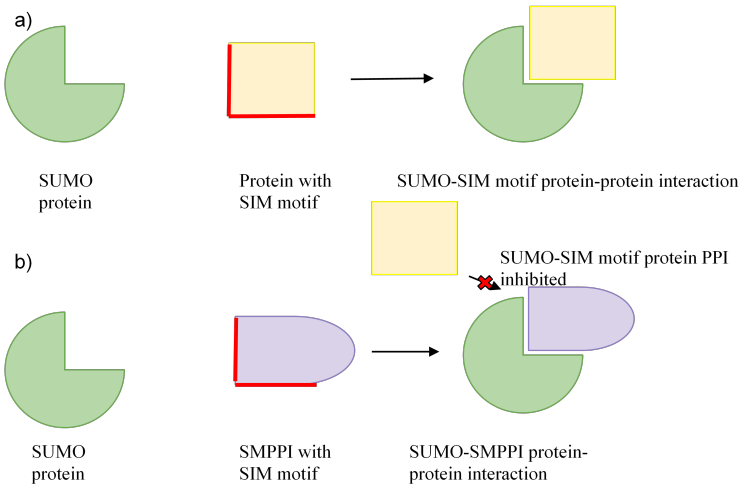


Figure 2.6: Overview mechanism of SUMO protein interactions. a) Proteins with SIM motifs (highlighted in red) interact with SUMO proteins by forming covalent bonds by the process called sumoylation. b) The SUMO proteins are targeted by SMPPIs (with SIM motifs highlighted in red) by binding to them and hence inhibiting proteins with SIM motifs to bind.

2.5 Complex identification in PPINs

It was initially recognised that protein complexes frequently function as macromolecular machines, playing a pivotal role in numerous cellular processes. To illustrate, RNA polymerase is a protein complex comprising 10 individual protein units and serves as the principal enzyme in gene transcription and the synthesis of a copy of mRNA from a DNA template. In order to gain a deeper comprehension of cellular mechanisms, a multitude of mathematical algorithms have been devised with the objective of identifying potential protein complexes based on interactomics data. The intuitive notion was that putative protein complexes could be identified from a protein-protein interaction network (PPIN) by detecting dense regions containing a multitude of connections or regions with considerable weights in a weighted PPIN [61]. The following section presents an overview of the various methods and algorithms employed for the detection of protein complexes within a protein-protein interaction network (PPIN).

2.5.1 Clustering with Overlapping Neighborhood Expansion (ClusterONE)

In the study, Nepusz and colleagues introduced the ClusterONE graph clustering algorithm, which takes a weighted protein-protein interaction network (PPIN) as input and constructs overlapping protein complexes [62]. The algorithm identifies clusters by discerning densely connected regions within a network and categorising them as non-overlapping complexes. However, in the case of a PPIN, proteins often possess multiple functions and may, therefore, be assigned to more than one complex on a situational basis. ClusterONE addresses the combinatorial nature of overlapping complexes, thereby accounting for the possibility that a single protein may participate in multiple complexes.

The ClusterONE algorithm employs a three-step process to detect overlapping complexes. (1) Proteins are grouped based on high cohesiveness by a greedy algorithm, which is run repeatedly from different starting proteins with the objective of identifying multiple and overlapping complexes. (2) The extent of overlap between complexes is quantified for each pair of groups, and those groups with an overlap score exceeding a preset threshold are merged. (3) Ultimately, any complexes formed by fewer than three proteins or with a density below the preset threshold are discarded.

Definitions

In this algorithm, the input PPIN is represented as a graph G, comprising three sets: V, which contains the vertices representing proteins; E, which contains the edges representing protein interactions; and W, which contains the weights associated with edges.

$$G = (V, E, W)$$

The overall density d_G of a graph G is commonly defined as the fraction of the number of edges $|E|$ over the maximum number of edges:

$$d_G = \frac{|E|}{|E|_{max}}$$

For a group of selected proteins V, one can distinguish between two types of edges: internal edges, which represent interactions between members of V, and outgoing edges, which represent interactions between members of V and proteins in the rest of the protein-protein interaction network (Fig. 2.7). The cohesiveness of the selected proteins relative to the remainder of the network can be evaluated by comparing the aggregated weight of the internal edges $(w^{(in)}(V))$ to that of the outgoing edges $(w^{(out)}(V))$.

$$f(V) = \frac{w^{(in)}(V)}{w^{(in)}(V) + w^{(out)}(V) + p|V|}$$

As a result of incomplete knowledge regarding protein interactions, a supplementary term, $p|V|$, has been introduced to account for the p unknown outgoing interactions of each member of V. The cohesiveness is a measure of the density of physical interactions among a group of proteins relative to the average density observed in their environment. Two scenarios may be posited with regard to high cohesiveness: (1) The group of proteins V forms dense and reliable edges among themselves (high $w^{(in)}(V)$), or (2) the group of proteins is more or less separated from the rest of the network (low $w^{(out)}(V)$). Protein groups with cohesiveness values exceeding one-third can be regarded as promising candidates for putative complexes, given that above this threshold, the internal weights begin to exceed the external weights.

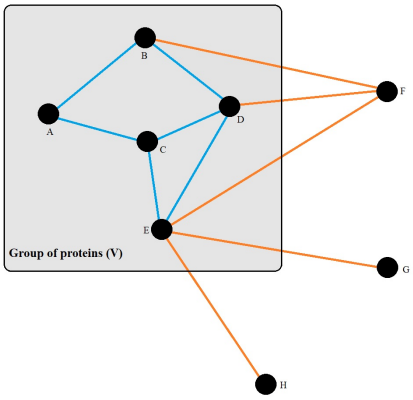


Figure 2.7: Schematic representation of a group of proteins selected within a PPIN. The blue lines represent the internal edges within the group. The orange lines represent the edges that connect vertices inside the group to the rest of the network. For example, with all edge weights set to 1 and $p = 0$, this group would have $w^{in}(V) = 6$ and $w^{out}(V) = 5$, resulting in a cohesiveness of $f(V) = 6/11$.

Algorithm

The initial stage of the process is the formation of the groups. The ClusterONE algorithm employs a greedy approach to the assembly of protein groups that are highly cohesive. The initial composition of group V is constituted by the protein with the highest degree in the PPIN that has not yet been visited. In each iteration, all proteins engaged in outgoing interactions are evaluated. An external protein, designated as v, is incorporated into the group if this action enhances the cohesiveness of the group, as indicated by the function $f(V)$, that is, when $f(V+v) > f(V)$. Conversely, an internal protein, identified as v, is removed from the group if its exclusion improves the cohesiveness of the group, that is, when $f(V-v) > f(V)$. Once no further enhancements to $f(V)$ can be achieved, the current group is deemed to be a local optimum, and the algorithm commences the assembly of the subsequent group, continuing until all proteins have been examined. The aforementioned process is exemplified in Fig. 2.8.

The second step is the assembly of candidate complexes. As ClusterONE permits proteins to be included in more than one group, this phase of the process examines the extent of the overlap between the locally optimal cohesive groups that were identified in the previous phase. The overlap score, denoted by $\omega(A, B)$, is a measure of the similarity between two groups, A and B. It is defined as the number of proteins that are common to both groups, $|A \cap B|$, divided by the total number of proteins in each group, A and B. The overlap score $\omega(A, B)$ of two groups A and B is calculated as follows:

$$\omega(A, B) = \frac{|A \cap B|^2}{|A||B|}$$

If the overlap score is greater than 0.8, then all pairs of cohesive groups are labelled as connected. Furthermore, if two groups are directly or indirectly connected, then they are merged to form candidate complexes. Finally, if a cohesive group does not overlap with and is not connected to other groups, then it is classified as a candidate complex without merging.

The third step is the filtering of candidates. The final step entails the assessment of the size and density (d_C) of each candidate complex (C) in accordance with the specified threshold (δ), with d_C as defined in section 2.5.1. Only those candidate complexes comprising a minimum of four proteins and a density greater than the specified threshold are retained for further consideration; all others are discarded.

For the same protein-protein interaction network (PPIN) from yeast (see Fig. 2.2), the ClusterONE plugin of Cytoscape identified 842 clusters as putative protein complexes when using the default parameters. These parameters included a minimum cluster size of 3, a minimum density set to auto-tuned (0.3 for weighted graphs and 0.5 for unweighted graphs), and a vertex-node penalty set to 2. Fig. 2.9 shows three examples of these putative protein complexes.

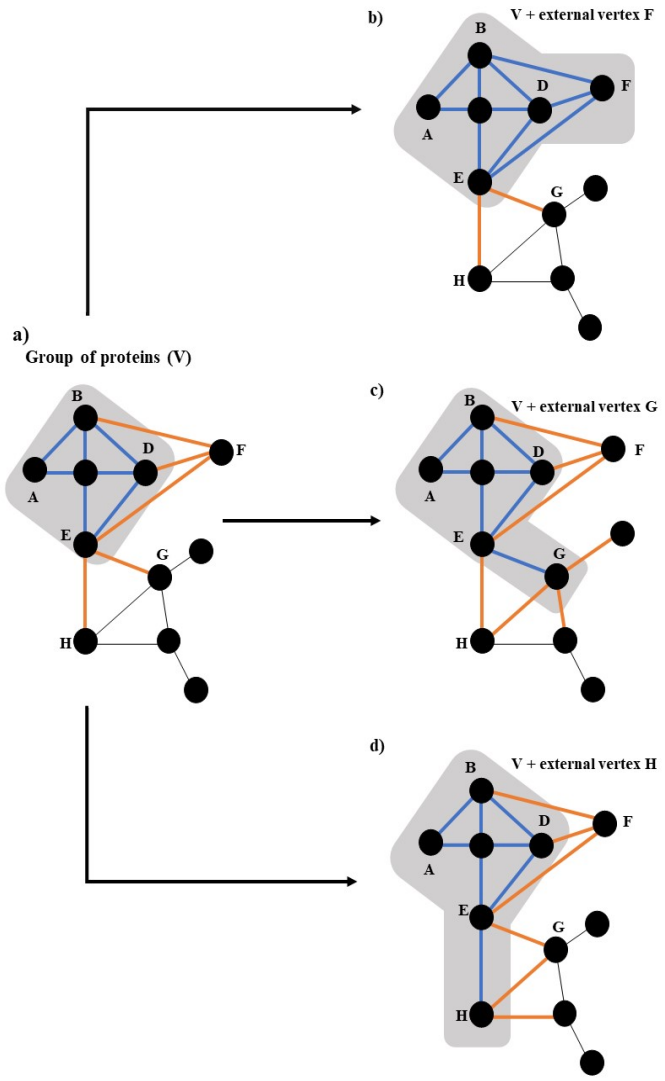


Figure 2.8: Workflow of the ClusterONE algorithm showing the cohesive growth of protein group V. (a) In this example, the group V consists of the five vertices A, B, C, D, and E. Assuming that all edge weights are set to 1 and the penalty $p = 0$, the cohesiveness of this group is $f(V) = 7/12$ with $w^{in}(V) = 7$ and $w^{out} = 5$. The group starts to grow by adding external vertices to or removing internal vertices from V based on the resulting changes in cohesiveness. The greedy algorithm adds an external vertex v only if $f(V + v) > f(V)$. Panels b), c) and d) show different options for expanding V. (b) Adding the external vertex F would increase $f(V)$ to $f(V + F) = 10/12$. In contrast, (c) adding vertices G would lower $f(V)$ to $f(V + G) = 8/15$, and (d) adding H would similarly lower it to $f(V + H) = 8/14$. The greedy algorithm thus only adds vertex F to group V and the new group cohesiveness is $f(V) = 10/12$. In the next iteration, the expansion process terminates with $V = \{A, B, C, D, E, F\}$ as a locally optimal cohesive group, since adding G or H in addition to F would not increase $f(V)$ any further, with $f(V+G) = 11/15$ and $f(V+H) = 11/14$ both $< 10/12$. The algorithm then restarts the expansion process by selecting the yet unvisited protein with the highest degree.

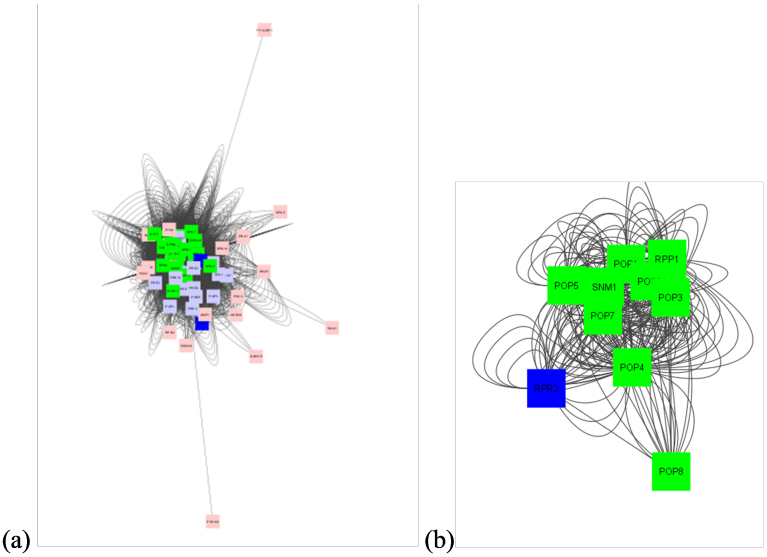


Figure 2.9: Protein complex prediction based on the *S. cerevisiae* PPIN according to data from Mentha using the ClusterONE algorithm. (a) Cluster with quality of 0.461, a p-value of 0.00022, 57 nodes vertices and 4346 interactions. Green vertices belong to the 19/22s regulator complex (22 out of 22 in CYC2008 detected by ClusterONE)., Purple vertices belong to the 20S proteasome complex (14 detected out of 14), and blue vertices belong to the Png1p/Rad23p complex (2 detected out of 2,). As for MCODE, the cluster identified by ClusterONE contains further proteins that are not known to be part of this complex. b) Protein complex identified by ClusterONE with a quality of 0.613, a p-value of 0.00016, 10 vertice nodes and 245 interactions. The green vertices belong to the ribonuclease MRP complex, the blue vertex belongs to the nucleolar ribonuclease P complex. CYC2008 lists 10 proteins for the ribonuclease MRP complex. Out of them, ClusterONE identified 9.

2.6 Docking of proteins structures

Protein docking is a computational method that employs a combination of algorithms and mathematical models to predict the protein complex from the individual binding partners that comprise it, thereby determining the nature of the interaction and binding between them. This allows the nature of the interaction and binding between them to be determined. In order to achieve this, the algorithms traditionally focus on the steric complementarity at the binding interfaces of the defined binding partners [63]. In the case of protein-protein complexes, the conformation of the proteins in the complex state differs from that observed in the unbound state. During the process of protein-protein complex docking, these discrepancies in protein conformation can be disregarded. This is accomplished through the utilisation of rigid body docking, which accounts for rotations and Cartesian coordinates. An alternative approach is to employ flexible docking, which accounts for a greater

number of coordinates, including internal coordinates. As the field of docking algorithms has advanced, researchers have sought to account for a wider range of properties beyond rigid and flexible conformations. For instance, algorithms such as High Ambiguity Driven biomolecular Docking (HADDOCK) [16], RosettaDock [64], and pyDockDNA [65] utilise rigid body docking in conjunction with an energy-based approach. Global RAnge Molecular Matching (GRAMM-X) [66] and ClusPro [67] employ a rigid body strategy in combination with Fourier fast Fourier transform (FFT). Shape-based incorporated with Fourier fast Fourier transform (FFT) docking techniques include Hex [68], PatchDock and SymmDock [69]. This section provides a comprehensive explanation of the functionality of HADDOCK and AlphaFold used in section 5.

2.6.1 HADDOCK

High Ambiguity Driven biomolecular Docking (HADDOCK) is an integrative modelling platform that utilises a multifaceted approach, integrating a range of data types, including biochemical and biophysical interaction data, experimental and bioinformatic information, binding interfaces, and orientation of binding information regarding the interactors. This approach enables comprehensive and accurate modelling of complex structures [70]. The data are derived from a variety of sources, including nuclear magnetic resonance (NMR) titrations, cross-linking and other chemical modification data, mutagenesis, hydrogen-deuterium (H/D) exchange, and data from available literature [71]. The information derived from NMR titrations provides insight into the binding sites or allosteric sites of proteins, as indicated by shifts in signals resulting from the binding of N^{15} to the proteins in question. In cross-linking data, the artificial linkers connecting two proteins at corresponding lysine residues, whether intramolecularly or extramolecularly, provide information about the maximum distance between the proteins' structures and the length of the linkers. This is identified through the use of mass spectroscopy. The mutagenesis data facilitate the identification of the amino acids that are crucial for binding. In a known protein structure, when the amino acids are mutated, if the mutations affect the protein interaction, it can be inferred that the mutated amino acid is responsible for binding. Conversely, if the amino acid is not involved in the interaction, it can be concluded that the amino acid is not involved in binding. In the H/D exchange data, when a protein is dissolved in heavy water, the exchangeable protons tend to undergo a chemical transformation, whereby they change into deuterons. Protected regions that do not undergo deuteronisation are deemed to represent binding sites for a given protein.

A review of the literature allows HADDOCK to consider additional information regarding protein structure, including the identification of active residues, residues involved in binding, and passive residues, which may also be involved in binding. These residues are of great importance to HADDOCK, as they are used to calculate an essential parameter called Ambiguous Interaction Restraints (AIR). This parameter provides information regarding the

minimum distance that an active or passive residue of a protein must be from the active or passive residues of another protein with which it is to interact. In other words, it defines the distance between the interface of binding between two proteins. Considering two proteins A and B, this effective distance parameter is then estimated by the formula below:

$$d_{iAB}^{eff} = (\sum_{m_{iA}=1}^{N_{Aatom}} \sum_{k=1}^{N_{resB}} \sum_{m_{kB}=1}^{N_{Batom}} \frac{1}{d_{m_{iA}n_{kB}}^{6}})^{-1/6}$$

where N_{Aatom} refers to the total number of atoms of the residue in protein A, N_{resB} refers to the total number of residues defined at the binding interface of molecule B and N_{Batom} refers to the total number of atoms of the residue in protein B.

HADDOCK performs docking in three steps, utilising information from the AIR parameter and other resources. In the initial phase (it0), the rigid bodies undergo energy minimisation. The two protein structures are positioned at a distance of 150A° and undergo random rotation in four cycles to identify the orientation that results in the lowest intermolecular energy function. The conformations with the lowest energy functions are then docked, which typically yields 1,000 conformations. From this initial set, 200 conformations are selected for the subsequent phase. In the subsequent phase (it1), semi-rigid refinement in torsion angles is employed to gradually optimise the rigid bodies of the two proteins. In the initial stage, the side chains at the interface are permitted to be flexible and undergo movement in order to identify an optimised structure. Subsequently, the side chains and the backbone structure of the two proteins are permitted to undergo movement in order to facilitate the refinement of the optimisation process. In the final stage of the process (itw), the structure is refined further by solvation with water molecules. The system is subjected to heating, and positional restraints are permitted for all atoms, with the exception of those belonging to flexible chains and the backbone at the interface of the structure. Subsequently, the system is cooled, with the position restraints being applied solely to the non-interface backbone of the structure. The resulting conformations are then grouped or clustered together based on the lowest root-mean-square deviation (RMSD) values observed in the backbone.

Finally, the resulting clusters of conformations are ranked at each stage based on van der Waals (E_{vdW}), electrostatics (E_{Elec}), restraint violation (E_{AIR}) and desolvation (E_{Desolv}) given by the following scoring functions:

$$it0 : 0.01E_{vdW} + 1E_{elec} + 0.01E_{AIR} - 0.01E_{BSA} + 1E_{Desolv} + 0.1E_{Sym}$$

$$it1 : 1E_{vdW} + 1E_{elec} + 0.01E_{AIR} - 0.01E_{BSA} + 1E_{Desolv} + 0.1E_{Sym}$$

$$itw : 1E_{vdW} + 0.2E_{elec} + 0.1E_{AIR} + 1E_{Desolv} + 0.01E_{Sym}$$

The results are clustered based on RMSD with a cut off of 7.5A°. At each stage, HADDOCK allows the user to set the maximum number of clusters, temperature, maximum time steps and symmetry restraints [71]. HADDOCK can handle maximum of 20 molecules.

2.6.2 AlphaFold

Two interrelated issues have constituted a significant challenge for researchers in the field of protein studies: firstly, the determination of the procedure by which proteins fold, and secondly, the determination of the final folded structure of any given protein. The concept of Anfinsen's dogma [72] was subsequently proposed, which postulates that the structure of a protein can be estimated based on the amino acid sequence that constitutes it. This dogma was a significant advancement in the field, as it enabled researchers to attempt prediction of the structure of proteins based on the reliable sequence data that was becoming increasingly available. However, due to the multitude of potential conformations that can arise from a given sequence, in 1960 Cyrus Levinthal [73] argued that it was not a straightforward task to predict the structure of proteins from sequence data alone. Consequently, advanced methodologies, such as machine learning and artificial intelligence (AI), are employed to predict structures and complexes based on the available information.

This section provides an overview of AlphaFold. AlphaFold is an artificial intelligence system that employs generative machine learning models and artificial neural networks to predict protein structures and complexes based on sequence information. The system employs a variety of data sources, including sequence and structure data from the Protein Data Bank (PDB), alignment data from Multiple Sequence Alignments (MSAs), and information about similar mutations in MSAs that determine the spatial proximity of folded proteins. These data are integrated into a neural network, which is then trained using the evolutionary relationships and sequence alignments. In the latest iteration, AlphaFold 3, the system is capable of processing information pertaining to other biomolecules, including DNA, RNA, and ligands.

The neural network is a collection of multiple layers of linked simulated nodes, the links between which can be strengthened or weakened. Upon the input of a FASTA file or a Macromolecular Crystallographic Information File (mmCIF), the tool proceeds to extract the pertinent metadata, such as sequence information (for FASTA files) and sequence, atom coordinates, and resolution information (in the case of mmCIF files) from these files. Utilising this information, a number of genetic databases are searched for MSAs. The MSA information is then used to search for structural templates from the PDB. The sequence of the templates is then compared with the input sequences. In the event that the two sequences do not exhibit a match, they are subsequently aligned using Kalign [74]. The identified structural templates are then incorporated into the model for training purposes. It should be noted that not all of the identified templates are included in the training process, as those that exhibit complete or partial identity with the input sequence are excluded. Templates with short sequences, comprising fewer than 10 amino acids or less than 10% of the input sequence, are excluded from further consideration. In the event that the input mmCIF files indicate a resolution in excess of 9°A, they are excluded from further consideration. In the event that the sequences exhibit a sequence identity of less than 40% when clustered with PDB clusters,

they are subsequently excluded.

The training is repeated several times over two sets of data: the self-distillation set and the known structures from the PDB. The self-distillation dataset is created by computing MSA of every cluster in Unbiclust30 against the same database. The sequences that were found in another sequence's MSA, the sequences with more than 1,024 amino acids, and the sequences with less than 200 amino acids were all filtered out. During the training of the model, each iteration entails the preprocessing of the multiple sequence alignment (MSA) and the cropping of residues. In the MSA preprocessing step, sequences that are closely related and are likely to be in close proximity are removed, as they tend to result in the deletion of branches of the phylogenetic tree. Subsequently, in the MSA clustering phase, AlphaFold 3 replaces the Evoformer module that was previously utilised in AlphaFold 2. The computational complexity of this module is proportional to $N_{seq}^2 \times N_{res}$, where N_{seq} is the number of sequences and N_{res} is the number of residues. This necessitates a reduction in the number of sequences. Therefore, a random fixed-size subset of the sequences was selected for input into the module. This resulted in the loss of some sequences, which in turn affected the accuracy of the prediction. Nevertheless, in the AlphaFold 3 iteration, the random subset remains selected but is substituted with the nearest sequence. This process is referred to as clustering. This ensures that all sequences have an equal opportunity to influence the prediction.

In the residue cropping step, two modes of cropping are employed: unclamped loss mode and clamped loss mode. In 90% of the training data, the backbone of the structures is clamped by setting $e_{max} = 10A$. Consequently, in the clamped loss mode, the residue crop positions are randomly selected from the set $Uniform[1, n + 1]$, where n represents the length of the sequence minus the crop size. The initial sequence crop size is set to 256, and then, through a process of fine-tuning, it is increased to 384. In the unclamped loss mode, the crop commences at $Uniform[1, n - x + 1]$, where x represents the value drawn from $Uniform[0, n]$.

The AlphaFold model is trained using a number of features, including the number of amino acids in the sequence, MSA cluster center sequences, binary values representing deletions to the left of every position in the MSA cluster, and so forth from the input information of amino acid sequences, MSA and structural templates. The final output comprises five predicted structures, accompanied by information on atom coordinates, confidence scores and distograms. Five confidence metrics are evaluated: predicted local distance difference test (pLDDT), predicted aligned error (PAE) score, predicted transmembrane (pTM) score, inner-pair pTM score, and per-chain pTM and per-chain pair ipTM. The pLDDT score is an indicator of the degree of confidence with which the local position of a specific atom in the structure can be predicted. The PAE score pertains to the confidence with which the chains, domains, or other biomolecules within the structure are packed. The pTM score, in turn, represents the overall accuracy of the predicted structure. The ipTM score indicates the degree of accuracy in the prediction of the interface.

The per-chain pTM and per-chain pair ipTM metrics pertain to the confidence with which chains or pairs of chains can be predicted. AlphaFold is capable of processing a maximum of 5,000 tokens as input. Each token represents either an amino acid residue present in proteins, a nucleotide base present in DNA or RNA structures, an atom present in a ligand or post-translational modification (except for ligands), or an atom in conjunction with the residue to which it is attached in the case of glycans, or an ion.

2.7 Computational tools and analysis

This section introduces several key concepts in the field of data analysis, including statistical hypothesis testing, differential expression analysis and enrichment analysis. These concepts are widely employed in the computational processing of biological data.

2.7.1 Statistical hypothesis testing

In essence, hypothesis testing is an inference method that is employed to ascertain whether sample data adheres to a presumed formulated hypothesis in comparison to the population data. In general terms, the sample data represents a subset of the population [75]. It follows that the hypothesis represents the pivotal element of this methodology, constituting a quantitative declaration that describes the comparison of the data to the population. This presumed formulated statement is referred to as the null hypothesis. The null hypothesis typically reflects the prevailing assumption or commonly accepted belief considered true unless evidence suggests otherwise.

The process of hypothesis testing is typically conducted in seven steps [76]. In the first step of the process, a statistical measure or quantity that maps the data to a numeric value is selected as the foundation for formulating a null hypothesis. This measure is called test statistic. Examples of test statistics that can be employed for hypothesis testing include the mean, median and the variance. In the second step, the null hypothesis (H_0) and alternative hypothesis (H_A) are formulated. The alternative hypothesis is a formulated statement that attempts to prove the alternative to the null hypothesis. For example, if the test statistic is the population mean (μ), the hypotheses can be structured as follows:

$$Null\ hypothesis\ (H_0) : \mu = \mu_0$$

$$Alternative\ hypothesis\ (H_A) : \mu \neq \mu_0$$

where μ_0 is a randomly defined number. The null hypothesis (H_0) and the alternative hypothesis (H_A) are inherently mutually exclusive, meaning that if one is true, the other must be false [76]. The objective, therefore, is to determine whether the value of μ is equal to μ_0. The intention is to establish whether there is compelling evidence to dismiss the H_0 [77]. This is because the process of invalidating the H_0 is more concrete than that of finding numerous pieces of evidence to validate it. Consequently, an attempt to invalidate the H_0 may

result in one of four possible outcomes. 1) where the H_0 is invalidated and the H_A is accepted, 2) where there is no evidence to invalidate the H_0, and thus it is not rejected and could be accepted. 3) An error occurs, where the H_0 is rejected, but should be accepted. This is known as a Type I error, 4) Another error, where the H_0 is accepted, but should be rejected. This is known as a Type II error [78, 79]. The four potential outcomes are presented in Table 2.2.

		Predicted condition	
		H_0 accepted	H_0 rejected
Actual Condition	H_0 accepted	H_0 accepted	Type I error
	H_0 rejected	Type II error	H_0 rejected

Table 2.2: Summary of the possible outcomes from hypothesis testing

The specific formulation of the H_0 and H_A hypotheses, as determined by the scenario under consideration, can be expressed as either one-sided or two-sided. In the above example, the hypotheses were formulated in a two-sided manner, as is typical in such cases. An example of a hypothesis formulated in a one-sided manner is as follows [80]:

Right-tailed test:

$$Null\ hypothesis\ (H_0) : \mu = \mu_0$$

$$Alternative\ hypothesis\ (H_A) : \mu > \mu_0$$

Left-tailed test:

$$Null\ hypothesis\ (H_0) : \mu = \mu_0$$

$$Alternative\ hypothesis\ (H_A) : \mu < \mu_0$$

In the third step, the distribution of the sample is determined. The probability distribution of the sample data, as determined by the chosen test statistic, is in accordance with the stated hypotheses. Accordingly, in order to calculate the sample distribution in accordance with the H_0, the probability distribution of the sample data derived from a population is estimated in a manner that would substantiate the veracity of H_0 with respect to the selected test statistic. Similarly, for the H_A, the probability distribution of the sample data derived from a population is estimated in a manner that would corroborate the veracity of H_A with respect to the aforementioned test statistic. Consequently, two distinct sampling distributions are prepared for the two hypotheses.

In the fourth step, a significance level value is determined. As demonstrated in Table 2.2, the potential for error is inherent in any process. It is therefore essential to establish a level that accurately reflects the risk of probability of such errors occurring. A Type I error occurs when the null hypothesis (H_0) is incorrectly rejected, and the probability of this error is denoted by α. Conversely, a Type II error takes place when an invalid null hypothesis (H_0) is mistakenly accepted, with the probability of this error represented by β [81]. In general, the value of α is regarded as the maximum acceptable level of error,

and it is referred to as the significance level [82]. For example, a significance level (α) of 0.05 implies a 5% chance of incorrectly rejecting the null hypothesis (H_0) when it is actually true. The most commonly used significance levels are 0.05 and 0.01 [83]. The α and β values are calculated by [81]:

$$\alpha \;=\; P(Type\ I\ error) \;=\; P(Rejecting\ H_0 \mid H_0\ is\ true)$$

$$\beta \;=\; P(Type\ II\ error) \;=\; P(Accepting\ H_0 \mid H_0\ is\ false)$$

In the fifth step, a critical value is determined. It is defined as the value that is compared with the test statistic and acts as a threshold that determines if the H_0 must be rejected. If the test statistic value exceed the critical value, the H_0 is rejected. This is because it is unlikely for the test statistic value to exceed the critical value if the H_0 is true. And, if the test statistic value falls within the the critical value, the H_0 there is no evidence to reject H_0 [84]. This value is mostly derived from the significance level and sampling distribution of sample data. The graphical representation of the critical value determining the rejection of H_0 on the sampling distribution curve of samples in the one-tailed and two-tailed cases are represented in the Fig. 2.10

In the sixth step, the numerical values of the test statistics are evaluated and the p-values are estimated. In the event that the test statistic was set as the mean, the mean of the sample data is calculated using the following formula:
Consider a sample $X \;=\; \{x_1, x_2, ..., x_n\}$, then

$$Mean\ (\mu) \;= \frac{1}{n}\sum_{n}^{i=1} x_i$$

where, n is the total number sample data, x_i is the sample data values. In the case of a selected test statistic of the median, the median value is calculated using the formula:
Consider a sample $X \;=\; \{x_1, x_2, ..., x_n\}$, then

$$Median(X) \;= \begin{cases} \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}, & if\ \text{n}\ is\ even \\ x_{\frac{n+1}{2}}, & if\ \text{n}\ is\ odd \end{cases}$$

where, n is the total number sample data, x is the sample data value. In the case of a selected test statistic of the variance, the variance value is calculated using the formula:

$$Variance\ (S^2) \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2$$

where, n is the total number sample data, x_i is the sample data values, and μ is the mean of sample data X.

The sample distribution and the numerical value of the test statistic calculated are utilised for the determination of p-values. In other words, a comparison is made between the numerical test statistic value and the theoretical assumptions regarding the sampling distribution in accordance with the hypotheses. A p-value provides an estimation of the probability of

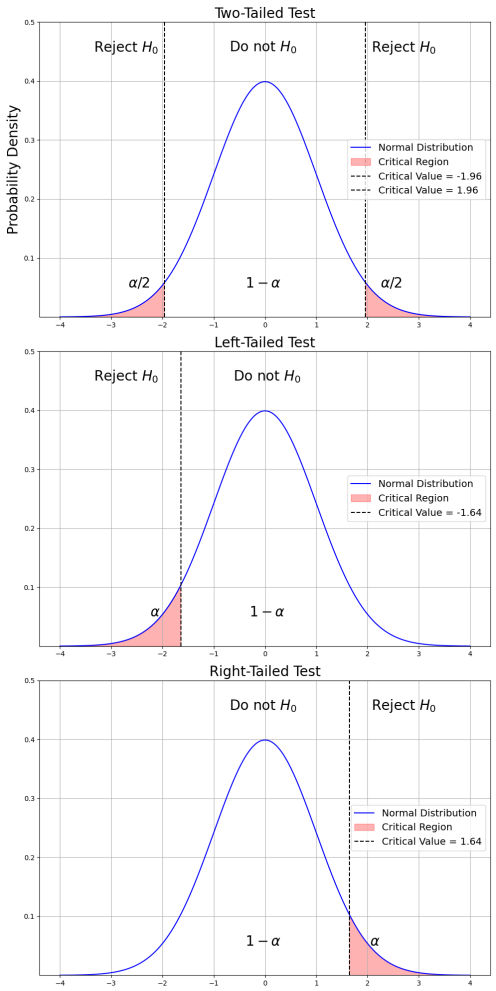


Figure 2.10: Representation of critical regions on probability distribution of sample in two-tailed (left), left-tailed (middle) and right-tailed (right) cases.

observing extreme values that are greater the numerical test statistic value, given that the H_0 is true [77].

In the final step, a decision is made if the H_0 should be rejected based on the obtained p-value. If the p-value obtained in less than the defined significance level, we then reject the H_0. If the p-value is greater than the defined significance level, we do not have enough evidence to reject H_0 and hence we accept it.

2.7.2 Common statistical tests

The principal objective of statistical tests in the biological context is to quantify the degree of similarity between data samples. In other words, these tests are employed to evaluate whether the two sample distributions exhibit a similarity by test statistics. A plethora of statistical tests are available for the significant conclusion of sample similarity. It is important to note, however, that these

tests are susceptible to certain conditions and that the accuracy of the results is contingent upon the selection of the appropriate test in accordance with the specific conditions of the data sample in question. A number of tests are based on the assumption that the data set under consideration follows a specific, predefined distribution, most commonly a normal distribution. There also exists some test that check if the sample data is normally distributed (DATAtab Team (2023), DATAtab e.U. Graz, Austria, https://datatab.de) like Shapiro-Wilk test [85], Kolmogrorov-Smirnov test [86], and Anderson-Darling test [87]. These tests are classified as parametric tests. Examples of parametric tests are Student's t-test [88], Welch t-test [89], and Analysis of variance (ANOVA) [90]. Conversely, some tests are characterised by greater flexibility in their conditions, and are not predicated on presumptions regarding the sample distribution patterns. Tests of this nature are designated as non-parametric. A few examples of non parametric tests are Mann Whitney U test [91], and Kruska-Wallis test [92]. Further conditions are based on the degree of correspondence between the data samples. In the event that the data samples being compared are matched data collected from the same source, they are referred to as paired data [93]. In such instances, a specific test must be utilised which examines the difference between the matched pairs, as opposed to testing the similarities between the unmatched data points [94].A number of tests employ paired data, including the paired sample t-test [88] and the Wilcoxon signed-rank test [95]. This section provides a detailed explanation of a few commonly used statistical tests.

Shapiro-Wilk test

In order to ascertain whether a parametric or non-parametric statistical test is required for the comparison of sample data, it is first essential to determine whether the data in question are normally distributed. In instances where the data are found to be of a normal distribution, parametric statistical tests are employed; conversely, non-parametric statistical tests are utilised in instances where this is not the case. Consequently, the Shapiro-Wilk test can be utilised to assess the normality of data distribution. It should be noted, however, that the suitability of this test is limited to small to moderate-sized datasets, comprising approximately 50 data points. The null hypothesis H_0 and the alternative hypothesis H_A for this test are stated below:

H_0: The sample is normally distributed
H_A: The sample is not normally distributed.

In order to establish whether a given sample is normally distributed, the Shapiro-Wilk test employs a three-step process. In the initial step, the data points of the sample are arranged in ascending order. Let us consider a sample data denoted by X comprising n number of data points, that is, $X = \{x_1, x_2, x_3, ..., x_n\}$. The sample data, arranged in ascending order, is denoted by Y, where $Y = \{y_1, y_2, y_3, ..., y_n\}$. In this case, $y_1 < y_2 < y_3 < ... < y_n$. The second step involves estimating the W statistic using the following for-

mula:

$$W = \frac{(\sum_{i=1}^{n} a_i y_i)^2}{\sum_{i=1}^{n}(y_i - \mu_y)^2}$$

where y_i represents the sorted sample data points, μ_y represents the mean of the sorted sample data, and a_i represents the tabulated coefficients that were reported in the study conducted by Shapiro and Wilk (1965) [85]. Once the W statistic has been estimated, the final step is to compare the W value to the corresponding critical value at the specified significance level (typically 0.05). This can be found in the Shapiro-Wilk p-value table, which is referenced in the same paper [85]. Should the W statistic value be less than the corresponding critical value at the 0.05 level, the null hypothesis H_0 is rejected and the alternative hypothesis H_A is accepted.The data set exhibits a notable departure from the expected normal distribution. Should the W statistic value exceed the corresponding critical value at 0.05, no evidence will be found to reject the H_0, and thus the sample will be deemed to exhibit a normal distribution.

Mann Whitney U test

The Mann-Whitney U test, which belongs to the category of non-parametric statistics, is a method employed to evaluate the significance of the differences between two independent samples of data. In order to ascertain the likelihood of a difference, the rank sums of the two samples are considered as the test statistic and compared on the basis of central tendency. In order to utilise the Mann-Whitney U test, several conditions must be met. As previously stated, the dataset in question must not adhere to a normal distribution, as this is a non-parametric test. The sample data must be continuous. The samples being compared must be independent and must not be related to each other; they should not be matched samples from the same source. The sample must have at least five data points and can be used with large sample data. The null hypothesis, H_0, and the alternative hypothesis, H_A, for this test are stated below:

H_0: There is no difference between the datasets
H_A: There is a difference between the datasets

The test is conducted in five steps to assess the similarity of the two sample datasets. In the initial step, the rank sums of both sample datasets are estimated through the assignment of ranks to both combined sample data set based on all the data points. Let us consider two sample datasets, denoted by X and Y, respectively. Each dataset comprises n_1 and n_2 data points, respectively. Thus, X is given by the set of data points $X = \{x_1, x_2, x_3, ..., x_{n_1}\}$, and similarly, Y is defined by the set of data points $Y = \{y_1, y_2, y_3, ..., y_{n_2}\}$. The combined data points of X and Y are assigned ranks, resulting in the set $R = \{r_1, r_2, r_3, ..., r_{n1+n2}\}$. Subsequently, the assigned ranks are summed within each group, thereby yielding the rank sum values for each group. Let

us define the sets of assigned ranks as follows: $R_1, \in R$ denotes the ranks allocated to the data points of X, and $R_2, \in R$ denotes the ranks allocated to the data points of Y. Let us consider the rank sums of R_1 and R_2, respectively, as T_1 and T_2.

$$T_1 = \sum_{i=1}^{|R_1|} r_i$$

$$T_2 = \sum_{i=1}^{|R_2|} r_i$$

In the second step, the U value is estimated from the rank sums, $U1$ and $U2$ values for each sample dataset in accordance with the following formula:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - T_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - T_2$$

$$U = min(U_1, U_2)$$

If the sample size of both the datasets are equal, the $U_1 = U_2 = U$.

In the third step, the mean μ_U and standard error σ_U values of U are calculated using the following formula:

$$\mu_U = \frac{n_1 n_2}{2}$$

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

In the fourth step, the z-score (z) is calculated by:

$$z = \frac{U - \mu_U}{\sigma_U}$$

In the final step of the process, the p-value is derived from the z-score by employing the z-distribution table. Consequently, it is possible to draw a conclusion regarding the similarity of the datasets on the basis of the p-value and the pre-defined significance level. In the event that the p-value is less than the pre-defined significance level, it can be deduced that the H_0 is rejected and the H_A is accepted, thereby concluding that the datasets are different. In the event that the p-value exceeds the significance level, no evidence exists to reject the H_0. Consequently, the H_0 is accepted, and the datasets are deemed to be similar.

Wilcoxon signed-rank test

The Wilcoxon signed-rank test is a type of non-parametric statistical test. It is used to judge whether the mean values of two dependent datasets are significantly different from one another, based on their central tendency. As this is a non-parametric test, the datasets need not adhere to a normal distribution. The data must be in the form of paired data, whereby the datasets are derived from the same source or group and exhibit a one-to-one correspondence. It is, however, important to note that the data points are recorded independently. In general, the distribution of the datasets should exhibit a similar symmetrical pattern. The number of samples in the datasets must be a minimum of 20 [96], and may be applied to large datasets. The H_0, and the H_A, for this test are stated below:

H_0: No difference between the datasets
H_A: There is a difference between the datasets

The test is conducted in five steps and ultimately determines whether the mean values of the datasets are similar. In the initial phase of the process, the discrepancies between the values of the datasets are determined. The absolute difference values are employed for the purpose of ranking. For the purposes of this discussion, we will consider two sample datasets, which we will denote by X and Y, respectively. The two datasets consist of the same number of data points, denoted by n. Thus, the set of data points comprising X is given by $X = \{x_1, x_2, x_3, ..., x_n\}$, and similarly, the set of data points comprising Y is given by $Y = \{y_1, y_2, y_3, ..., y_n\}$. Let us define the set of difference values between the datasets X and Y as D. Thus, $D = \{d_1, d_2, d_3, ..., d_n\}$ where $d_1 = x_1 - y_1$, $d_2 = x_2 - y_2$, and so on, up to $d_n = x_n - y_n$. The absolute values of D are ranked, resulting in the set $R = \{r_1, r_2, r_3, ..., r_n\}$.

In the second step, the W statistic value is determined from positive and negative ranks of differences, and T^+ and T^- in accordance with the following formula:

$$T^+ = \sum_{i=1}^{|R|} r_i \quad where\, r_i \,>\, 0$$

$$T^- = \sum_{i=1}^{|R|} r_i \quad where\, r_i \,<\, 0$$

$$W = min(T^+, T^-)$$

In the third step, the expected value μ_W and standard error σ_W values of W are calculated using the following formula:

$$\mu_W = \frac{n(n+1)}{4}$$

$$\sigma_W = \sqrt{\frac{n(n+1)(2n+1) - \sum \frac{t_i^3 - t_i}{2}}{24}}$$

where t refers to the number of data points sharing a rank i.

In the fourth step, the z-score (z) is determined by:

$$z = \frac{W - \mu_W}{\sigma_W}$$

In the final step, the p-value is calculated from the z-score using the z-distribution table, in a manner analogous to that employed in the Mann Whitney U test. Consequently, a conclusion regarding the similarity of the datasets can be drawn based on the p-value and the pre-defined significance level. Should the p-value be less than the significance level, the H_0 is rejected and the H_A is accepted, thereby concluding that the datasets are different. In the event that the p-value exceeds the significance level, no evidence exists to reject the H_0. Consequently, the H_0 is accepted, and the datasets are considered to be similar.

2.7.3 Differential expression analysis

In recent times, high-throughput sequencing of omics data has provided researchers with a wealth of valuable information, including quantitative count data of the number of reads of transcript, viable binding regions on DNA from RNA-seq and chip-seq assays [97]. The generation of count data from an organism in different conditions gives rise to an important question: is the abundance of the transcript similar in both conditions? If not, do these variances play a role in phenotypic, physiological, or disease progression? As the reads of a transcript are typically mapped to a target gene, and the number of reads represents the active expression levels of the corresponding gene, the difference in reads in different conditions can be used to identify changes in the abundance of gene expression levels of specific genes [98, 99]. This analysis, which determines genes with significant differential abundance of expression levels is referred to as differential expression analysis. In this analysis, normalised count data is considered for statistical testing, with the aim of determining whether there is a significant difference in the gene abundances in the different conditions [100]. This analysis is conducted with the aim of gaining insights into potential targeted therapeutics and diagnostics, as well as for identifying disease development and progression. A number of tools are available for conducting this statistical test, with the aim of identifying genes that exhibit significant differential abundances in different conditions. Examples of such tools include DESeq2 (Bioconductor R package)[101], TREAT (Bioconductor R limma package) [102], SAMseq (samr R package) [103], and so forth. This section will discuss the functioning of a few of these tools in detail.

DEseq2

The Deseq2 tool employs an algorithmic approach to effectively fit a model to the data, thereby estimating the expression levels of differentially expressed

genes. The tool essentially fits the samples as a negative binomial distribution [104], with the corresponding mean and variance being fitted by regression. The workflow of this algorithm can be described in five steps.

As previously stated, the algorithm examines the count matrix and develops a model in which the read counts are assumed to follow a negative binomial distribution [105]. Let us consider a count matrix with n number of genes and m number of samples, designated as K, in which the rows represent genes and the columns represent samples. Therefore, the term K_{ij} denotes the read counts, non negative numbers, mapped to the gene i of sample j. It is assumed that the reads of sample j, mapped to gene i, are modelled by a negative binomial (NB) distribution given by :

$$K_{ij} \sim NB(\mu_{ij}, \sigma_{ij}^2)$$

where μ_{ij} and σ_{ij}^2 are the two parameters of the negative binomial distribution, mean and variance respectively. However, in the present context of the count matrix, the two parameters remain unknown. Consequently, the two parameters are estimated from the count matrix (K_{ij}).

In the first step, the mean parameter, or the expectation value, of the counts mapped to gene i from sample j is estimated using the following formula:

$$\mu_{ij} = q_{i,\rho(j)} s_j$$

where $q_{i,\rho(j)}$ is the condition-dependent value of gene i of sample j under a specific condition ρ and s_j is the size factor representing the sampling depth and coverage of library j. It is typical for s_j to be a constant, which allows it to be used for all the genes in a sample. The rationale behind utilising the size factor s_j is to account for the disparate sequencing depths observed across different samples. These discrepancies could be resolved by normalising with the total reads value; however, it was observed that the genes with higher expression values diminished the impact of genes with lower expression values. This illustrates that normalising with total reads is not a viable approach and thus, size factors were introduced. These size factor for a sample s_j represent the median of ratios of observed counts and geometric mean of the counts across all samples [106], as determined by the following formula:

$$s_j = median \frac{K_{ij}}{(\sum_{x=i}^{m} k_{ix})^{\frac{1}{m}}}$$

The value of $q_{i,\rho}$ is calculated by taking the mean of the counts from the samples that exhibit the condition ρ, as detailed in the formula provided below:

$$q_{i\rho} = \frac{1}{m_\rho} \sum_{j:\rho(j)=\rho} \frac{K_{ij}}{s_j}$$

where m_ρ number of samples exhibiting the condition ρ. $q_{i,\rho}$ is a value that is proportional to the fragments of DNA from gene i.

In the second step, the variance parameter of negative binomial distribution is calculated. It estimates the spread of reads in a sample. The variance

is estimated using a combination of a specific measure of dispersion (α) and the mean μ_{ij} given by:

$$Var\ K_{ij} = \mu_{ij} + \alpha_i \mu_{ij}^2$$

Dispersion is used to describe the variability that can be estimated with a high degree of accuracy if each condition had several replicate samples. However, when the number of sample replicates for a given condition is insufficient, high dispersion variability for each gene can compromise the accuracy of the results. Therefore, it is assumed that genes with similar expression levels have similar dispersions. Consequently, in the third step, the initial estimation of dispersion is calculated using maximum likelihood, and a curve is fitted to represent the estimated dispersion value based on the read counts of the replicate samples in a condition. However, this does not provide insight into the extent of deviation of each gene from the fitted curve. In the fourth step, the empirical Bayes method is employed to reduce the discrepancy between the initial estimation of the dispersion and the curve, thereby improving the fit. The shrinkage method is a crucial technique in reducing the number of false positives in differential expression analysis. It also reduces the variability of genes with low expression levels.

In the fifth step, a statistical hypothesis test for differential gene expression is conducted. The null hypothesis H_0 of this test posits that the expression levels of a gene in one condition are identical to those in another condition.

$$H_0 : q_i A = q_i B$$

where $q_i A$ refers to the expression level parameter of gene i in condition A and $q_i B$ refers to the expression level parameter of the same gene in condition B. The test statistic described below is employed to test the hypothesis [97].

$$K_i A = \sum_{j:\rho(j)=A} K_{ij}$$

$$K_i B = \sum_{j:\rho(j)=B} K_{ij}$$

$$K_i S = K_i A + K_i B$$

where s represents the overall sum of samples in both conditions.

The sample data from both conditions were fitted to a generalised linear model (GLM) in order to estimate the effect sizes of the conditions. This is represented for each gene as follows:

$$log_2(q_{ij}) = \sum_s x_{js} \beta_{is}$$

where x_{js} is a design matrix element representing the presence of the sample in either condition A or B, and β_{is} are the coefficients from the GLM fitting, which represent the overall strength of the fit and the log fold change values. It has been observed that genes with low expression levels tend to exhibit a high degree of variance in the estimated log fold change values. Accordingly,

the algorithm utilises a shrinkage approach to the log fold change estimates, employing an empirical Bayes method. This serves to reduce the high degree of variability.

Subsequently, the Wald hypothesis test is conducted following the fitting of the GLM for each gene. This is achieved by calculating the z-statistic, which is obtained by dividing the shrunken estimates of log fold change values by their standard error. The p-values are derived by comparing the z-statistic with the values in the standard normal distribution tables. Given that multiple comparisons are made, there is a tendency for the false discovery rate of the genes to be high. Therefore, the issue of multiple testing is addressed by filtering the genes based on their average expression levels across all samples. Consequently, the filtered p-values assist in determining whether the null hypothesis (H_0) should be rejected and thus concluded to be significantly differentially expressed, or accepted on the grounds that there is no evidence to reject H_0.

T-tests relative to a threshold (TREAT)

T-tests relative to a threshold represent another algorithm that is used to estimate whether a gene is significantly differentially expressed in different conditions. In contrast to the other algorithms, which merely indicate whether a gene is or is not differentially expressed without providing insight into the significance of the observed difference, this algorithm enables the user to define a threshold, thereby facilitating the determination of whether the observed difference is biologically meaningful.In the previous method, we described how False Discovery Rate (FDR) filters genes and adjusts p-values. However, the threshold defined for a log fold change to be considered significant does not account for variability, which results in a lack of reliability in terms of reproducibility. The TREAT algorithm represents an extension of the empirical Bayes statistic initially proposed by Smyth [107] and incorporates the fold change threshold, taking into account the FDR correction. The TREAT workflow can be defined in four steps.

In the first step, the hypotheses underlying the statistical tests are established. In consideration of the log fold change values (β_i) estimated for a given gene, designated as i, between two distinct conditions, denoted as A and B, the null hypothesis, represented as H_0, and the alternative hypothesis, represented as H_A, can be formulated as follows:

$$H_0 : |\beta_i| \leq \tau$$

$$H_A : |\beta_i| > \tau$$

where τ is a value specified by the user and used to determine the significance of differential gene expression. In other words, the log fold change values that exceed the specified threshold are deemed to be biologically meaningful. The null hypothesis (H_0) is employed to assess whether a log fold change value falls within the specified interval of values, defined as $[-\tau, \tau]$.

In the second step, the read counts for each gene are fitted into a linear model. Let us consider a gene, designated as g, present in m number of samples. The count data of gene g from all m samples can be represented as a vector, $y_g = \{y_{g1}, y_{g2}, ..., y_{gm}\}$, where y_{gj} is the count of gene g from sample j. The linear model used to fit the sample data can be represented as follows:

$$y_g = X\alpha_g$$

where X is a design matrix element that represents the presence of a sample in a specific condition or group, and α_g represents the unknown coefficients that are to be determined.

In the third step, the variance of the fit model is shrunk to increase accuracy. In the linear model fitted, the estimated variance of the gene may be noisy, particularly in cases where the gene exhibits low expression values, which can result in a high degree of disparity. Consequently, the algorithm incorporates a hierarchical model that postulates the requisite form of the prior distribution of the variance. This is achieved by adjusting the gene variances to a posterior variance, which serves to shrink the observed variances to a prior estimate. This is achieved by combining the read count of a gene with the average variance across all genes. The degree of shrinkage is contingent upon the discrepancy between the degrees of freedom of the observed and prior distributions. Subsequently, the moderated t-statistic is defined in accordance with the posterior variance, as illustrated below:

$$t_g = \frac{\beta_g}{posterior\ var\ SE(\beta_g)}$$

where β_g is the log fold change value of gene g between conditions A and B, and $posterior\ var\ SE(\beta_g)$ is the standard error of the posterior variance.

In the final step, the p-values are calculated from the observed t-statistic, denoted as t_g. Given that the null hypothesis H_0 defines the potential for observing β_g within a specified range, the p-value seeks to estimate the probability of rejecting H_0 under the most extreme conditions. Consequently, the following probabilities are subjected to analysis:

$$p = P(T > t_{obs} + \delta) + P(T < t_{obs} - \delta)$$

where δ represents the minimum value of the observed log fold change and the defined threshold, T represents the theoretical t-distribution value for the corresponding degree of freedom. The p-value thus determines whether the null hypothesis (H_0) must be rejected in favour of the alternative hypothesis (H_A), thereby concluding that the gene g is significantly differentially expressed between conditions A and B. Alternatively, the null hypothesis (H_0) may be accepted in the absence of evidence to the contrary.

2.7.4 Biomolecule annotations

The advancement of high-throughput technology has led to an exponential increase in the amount of sequencing data obtained for various biomolecules,

including DNA, RNA, and proteins [108, 109, 110]. From these sequencing data, biologists postulate the unification of biology, whereby it is proposed that a significant proportion of genes and proteins are conserved in the majority of living cells, and that the functional information can be shared between diverse organisms [111]. However, these sequence data sets are typically comprised of a lengthy list of genes, which can be tedious to interpret in terms of their biological relevance through manual literature mapping. It was therefore crucial to create a repository of all scientific findings regarding biomolecules and to tag them with terms that define their roles, biological tendencies and locations in any organism.

A number of databases have been developed with the specific purpose of annotating genes and proteins, thereby facilitating the interpretation and analysis of lists of data. One such database is the Gene Ontology Consortium [111], which originated from three other databases: FlyBase [112], Mouse Genome Informatics [113, 114] and Saccharomyces Genome Database [115] and have been extended and now encompass important databases for other organisms, including plants, animals and microbes [116]. As the volume of information about biomolecules continues to expand and evolve, it is vital that the database is able to accommodate this growth in a way that avoids any overlap or contradiction between the annotation terms. Accordingly, the database employs a structured approach, comprising three non-overlapping categories for annotations: cellular component, molecular functions, and biological processes. The term "biological process" is used to denote the specific biological function to which a particular gene or gene product contributes, thereby highlighting its role within the cellular context. The term "molecular function" is used to describe the role of a gene product in biochemical activity. The term "cellular component" can be defined as the anatomical location in which the activity of a gene product is observed. The annotation terms are also structured in a directed acyclic graph (DAG) representation, where each term is connected to at least one parent term and zero, one or multiple child terms [117]. The more general terms are considered to be parent terms, with each subsequent level of child terms becoming increasingly specific. The various terms are associated by five predefined relation types, which are based on the biological roles of genes and gene products. The following relations are used: "is a", "is part of", "regulates", "positively regulates" and "negatively regulates" [117, 118]. The "is a" relation is applied when a term is identified as a subclass (or child) of another term (the parent term). The term "part of" is applied when a term constitutes a division or member of another parent term. The term "regulates" is used to indicate that a term modulates or controls the biological processes of its parent term. The terms "positively regulates" and "negatively regulates" are used to denote the activation and deactivation of the parent term, respectively.

An illustrative example of the structured representation of the Gene Ontology (GO) term "peptidase inhibitor activity" is presented in Fig. 2.11. The term "peptidase inhibitor activity" is classified within the Molecular Function category of Gene Ontology (GO) annotations. The hierarchical level of

this term indicates that the genes or gene products annotated to it are also annotated to broader parent terms, such as "enzyme inhibitor activity" and "peptidase regulator activity." The black arrow represents the "is a" relation to the parent terms. It is evident that the term refers to a negative control or deactivation type of activity. Consequently, it is also related to parent terms such as "peptidase activity" with a "negatively regulates" relation, which is represented by the red arrow. The remaining terms that indicate a modulation type of activity in relation to the parent terms are associated with the term "regulates," as illustrated by the yellow arrow. he most recent consortium statistic, updated in November 2024, indicates that there are a total of 40,635 GO terms. Of these, 26,467 are classified under the Biological Process category, 10,146 under the Molecular Function category, and 4,022 under the Cellular Component category [119].

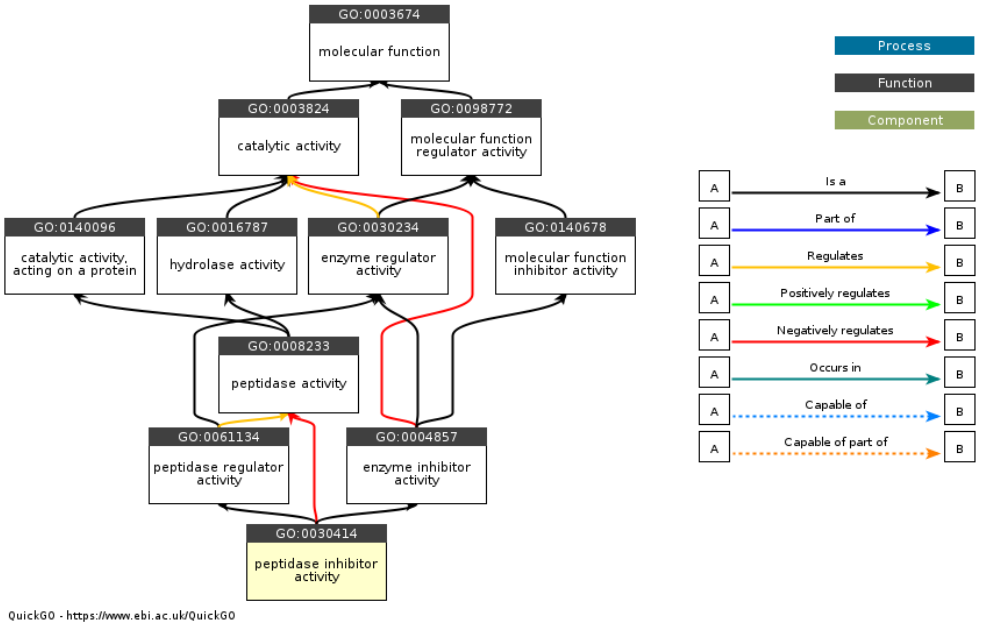


Figure 2.11: The figure represents the directed acyclic graph (DAG) of the molecular function category of the Gene Ontology (GO) terms, specifically the 'Peptidase inhibitor activity' branch. The black arrows indicate the 'is a' relation, the yellow arrow represents a 'regulates' relation, and the red arrow represents a 'negatively regulates' relation with the corresponding parent terms. This figure was adapted from the QuickGO web service [117].

In addition to the ontology terms based on molecular function, biological processes and cellular components, the biomolecules are also annotated in accordance with the biological pathways in which they participate. This would facilitate the analysis and decoding of bottlenecks in systems biology, the identification of targeting therapy and the interpretation of their roles in disease-related pathways.

One of the most widely used open-source and peer-reviewed databases for the annotation of pathways of genes and gene products is the Reactome

pathway database [120]. The database catalogues a range of biochemical processes and interactions between biomolecules, including those involved in transportation, replication, and metabolism. The functional relations and curated biologically processes are combined and presented in an interactive map, facilitating visualisation and analysis. The database permits researchers to upload a list of genes or gene products, and the tool generates a reaction map as a graphical representation of the subset of the uploaded list that is involved in any pathway. The most recent statistics, updated in September 2024, indicate that the database contains pathway information for 15 species. A total of 2,742 pathways in the Homo sapiens database are associated with 11,289 proteins and 15,492 reactions. For illustrative purposes, Fig. 2.12 shows an example of the oncogenic MAPK signalling from the Reactome database.

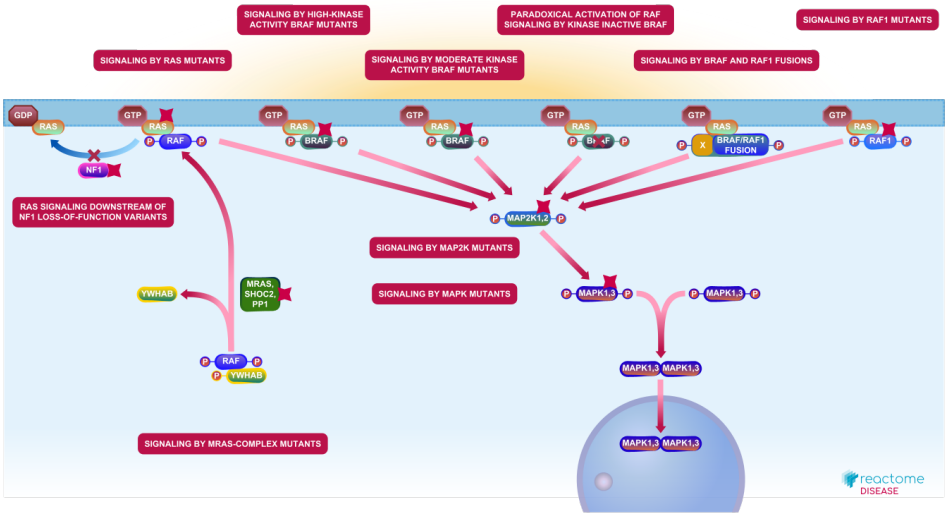


Figure 2.12: Oncogenic MAPK signalling pathways extracted from the web service reactome pathway viewer [121]

2.7.5 Overrepresentation statistical analysis

In addition to the annotation of the extensive list of genes or gene products derived from high-throughput sequencing, the overrepresentation statistical analysis is a prevalent and indispensable analytical tool that provides insights into whether the groups of genes or proteins that collaborate in a biological process or pathway are statistically overrepresented [122, 123]. This is of particular importance in many subsequent analyses, such as differential gene expression [124]. From the list of genes that are either upregulated or downregulated, the enrichment analysis can reveal whether there is a significant subset of genes that function together in a pathway in one condition, but which is essentially insignificant in the other condition. Consequently, this analysis can provide a substantial amount of information regarding a set of genes that may be responsible for disease progression or the development and

differentiation between conditions. There are many online tools that perform enrichment analysis, such as Gene Ontology Enrichment Analysis, which is linked to PANTHER [125], enrichR [126], gProfiler [127], etc. In this section you will find a detailed description of how this test works.

The input list for the overrepresentation analysis is comprised of a list of genes or gene products from a population, which is uploaded into the tools. The list may comprise significant genes or genes that are either upregulated or downregulated as a result of an experiment. The ontology or pathway terms are then annotated for the uploaded list. Subsequently, the uploaded data is divided into groups corresponding to specific biological processes, molecular functions, cellular components, or pathway terms. In addition to the uploaded list, a reference list of genes or gene products is also provided. The specified reference list may be the population from which the subset of genes are typically derived. Subsequently, a binomial test is employed to investigate whether the uploaded genes or gene products belonging to the classified terms are statistically overrepresented or under-represented in comparison to the reference list. The null hypothesis (H_0) and the alternative hypothesis (H_A) of the binomial test employed are stated below:

H_0: The observed likelihood of a gene being annotated to a term from the upload list is the same as that in the reference list.
H_A: The observed likelihood of a gene being annotated to a term from the upload list is not the same as that in the reference list.

Let us consider M to be the total number of genes in the reference list and N to be the total number of genes in the input list. The observed likelihood, $p(C)$, of a gene being annotated to a term C is calculated by the following equation [128]:

$$p(C) = \frac{N(C)}{M}$$

where $N(C)$ refers to the number of genes from the input list that have been annotated to a term C.

The p-values are subsequently estimated using the probability of observing $i(C)$ or more genes (or a greater number, to determine the upper limit for rejecting H_0) being annotated to a given term C. The formula for estimation is provided below:

$$p\ value = \sum \binom{N}{i} p(C)^i (1 - p(C))^{N-i}$$

A p-value of less than 0.05 indicates that the observed results are not random and that the annotation of the input list of genes differs from that of the reference list. In the event that the number of observed genes is greater than the expected number of genes (i.e. the summation from $i(C)$ to N), it is deemed to be overrepresented. In the event that the number of observed genes is less than the expected number of genes (i.e. the summation ran from 0 to $i(C)$), it is considered to be under-represented. In addition to the estimation of

p-values using the binomial distribution, other tests are employed, including the hypergeometric test, Fisher's exact test, and the chi-squared test, among others [124, 129].

Chapter 3
Downstream analysis of transcriptomic data

This chapter introduces six workflows for identifying and analysing protein clusters from differential protein-protein interaction (PPI) networks in healthy and diseased conditions. The workflows were then compared in order to ascertain which one produced the most biologically meaningful results. The sections from 3.1 to 3.5 were adapted and expanded from the manuscript from Thangamurugan, S. and Helms, V. (2024), entitled "Comparing workflows for combining transcriptomic with protein interaction data", 2024. I was responsible for the design and implementation of the pipelines, the performance of computational analysis, and the preparation of the manuscript. Volkhard Helms provided assistance in the design of the study, the analysis of the data, and the editing of the manuscript.

3.1 Introduction

Analyzing protein-protein interaction networks (PPINs) is one key area of modern computational biology. PPINs often contain dense subnetworks or clusters. These are groups of proteins that interact with each other and concertedly play a role in processes such as gene expression, transport, signalling, apoptosis, and others [130, 131]. Protein clusters may manifest themselves as transient or permanent protein complexes. Alterations in components of these clusters, due to mutations or differential expression, can potentially lead to abnormal interactions and the activation of disease-related pathways [132]. Over the years, many algorithms have been developed to identify protein clusters in PPINs [133, 134, 135]. One noteworthy algorithm is called ClusterOne (Clustering with Overlapping Neighborhood Expansion) [62]. It employs a greedy-based search based on a cohesiveness score as a heuristic

to identify protein clusters. The algorithm allows for the overlap of proteins between clusters, accounting for the multifunctional nature of proteins. It has been shown that identification and analysis of protein complexes present opportunities to uncover functional and disease-related pathways, potentially benefiting the development of targeted therapies [130, 136, 137].

Transcriptional regulation of genes varies between cell types and cell states, leading to phenotypic, physiological, and even trajectory variations towards a diseased state [138, 139]. Investigating these variations between cell types and states provides a profound basis for understanding what distinguishes healthy from diseased states. In recent years, high-throughput technologies have frequently been applied to capture active genes, transcripts, proteins, metabolites and the entire interactome of biological samples. For example, one often examines the gene expression profiles of two conditions followed by differential expression analyses to identify the differentially expressed (DE) genes in diseased conditions [140]. This is done to get insights into potential therapeutic targets, to monitor disease progression, and to identify biomarkers for diagnosis. To enhance the functional interpretation of the set of DE genes, some studies have annotated interactions between the proteins encoded by these DE genes by interactome data taken for example from the popular STRING database [26]. STRING compiles compendia of functional and/or physical interactions in many organisms. The network then comprises nodes representing the DE genes and edges representing the protein interactions between the encoded proteins as recorded in the STRING database. By identifying clusters in this network, one aims to identify groups of proteins that are associated with the rewired biological functions and possibly with disease-related pathways.

The complete PPIN or interactome of a multi-cellular organism provides an overview of all protein interactions captured in any cells or samples of this organism. In fact, in each cell type or condition, only a subset of these interactions is active. When comparing two cell states such as healthy and diseased states, a number of protein interactions may be "rewired" (e.g. gained or lost) either due to differential expression of the respective genes or alternative splicing. Then, some binding partners may not be available for interaction in that condition. Consequently, examining the full interactome is not suitable for identifying the rewiring events that distinguish the conditions of cells. Instead, previous studies pruned the interactome to those genes expressed in a specific condition and cell state [141, 142]. For example, our group developed a tool, PPIXpress, that constructs condition-specific PPINs by pruning the interactome of an organism based on provided gene/transcript expression data [143, 144]. PPIXpress has been used to construct condition-specific networks for healthy and diseased states. To facilitate a comparison of the networks and estimate rewiring events between the two states, our group developed another tool, PPICompare [145]. PPICompare performs pairwise comparison by assigning the networks of one state, for example, the healthy state, as a reference and reports the statistically significant rewired interactions that are lost or gained in the networks of the other state, for example, the diseased

state. Finally, the tool generates a differential network that contains all the significantly rewired events between the two states.

In the past years, two distinct methodologies were established to identify protein clusters in differential PPI networks. The first strategy starts with the estimation of differentially expressed (DE) genes, which are subsequently used as a basis to construct a PPIN. The second entails the construction of condition-specific networks and subsequent analysis of the differentially expressed interactions. In this study, we evaluated the comparison of different workflows when applied to the same case-study dataset, with the objective of identifying the workflow that yields the most biologically meaningful results. To this end, six pipelines were tested in this study, comprising three pipelines following the first technique and three pipelines following the second technique. These pipelines were then subjected to cluster prediction on the networks, with computing enrichment of Reactome pathways in individual protein clusters being the final step. The main objective was to identify which pipeline is most effective at accurately elucidating the enriched pathways that underpin the transition from a healthy to a diseased state.

The accuracy and suitability of the pipelines were evaluated based on a number of criteria. Firstly, an analysis was conducted on the global properties of the networks. Secondly, the degree of overlap between the predicted clusters and reported protein complexes was estimated. Finally, we analysed which pipeline was most effective at determining pathways that are most closely related to its biological state.

3.2 Materials and method

3.2.1 Gene expression dataset

An RNA sequencing (RNA-seq) dataset comprising 80 samples [146] was retrieved from the GEO database (GSE112509). It includes 57 primary melanoma samples (M) and 23 benign melanocytic nevi samples (N) biopsied from patients. We kept the split of the original authors into two transcriptomic subtypes of melanocytic nevi (13 samples of N1 and 10 samples of N2) and melanoma (26 samples of M1 and 31 samples of M2).

3.2.2 Pipelines

Melanoma and nevi data were processed using six distinct software pipelines that each generate a differential analysis of PPI inferred from gene expression data, as illustrated in Fig 3.1. Pipelines 0, I and II follow a workflow that is similar to that of conventional pipelines. Initially, the gene expression data underwent pre-processing and normalization, after which a list of significantly differentially expressed genes between melanoma and nevi samples was generated. This was estimated either employing the Bioconductor R package DEseq2 [101] or a t-test relative to a threshold called TREAT [102]. Subsequently, we inferred interactions between the proteins encoded by the significantly differentially expressed genes using the databases STRING or

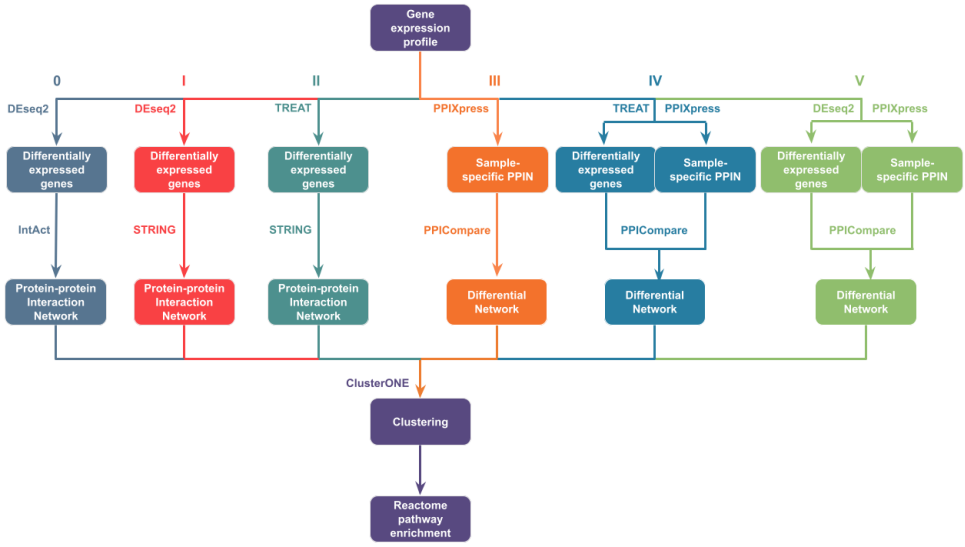


Figure 3.1: Summary of the six pipelines used to process the 80 samples of melanoma and nevi RNA-seq data.

IntAct [28]. Consequently, a network was constructed comprising the significantly differentially expressed genes that interact with each other on the proteome level. In contrast, pipelines IV, V and VI employ an inverse order of processing. Initially, condition-specific protein interaction networks were constructed using the tool PPIXpress, and then differentially abundant protein interactions were estimated using the tool PPICompare. In addition, pipelines IV and V refined the identification of differentially abundant protein interactions by additionally incorporating the criterion that at least one of the involved genes must be significantly differentially expressed.

In the second-last step of each pipeline, significant clusters were identified in the networks using the algorithm ClusterONE inside Cytoscape [44]. Finally, pathway enrichment analyses were conducted on these clusters in order to gain insight into the biological processes and pathways associated with each cluster.

In this study, we conducted enrichment analyses using the Gene Ontology tool, accessible via the following `http://geneontology.org/`. It is linked to the PANTHER classification system [125], which is a comprehensive, annotated library of gene families and protein-coding genes in a genome.

3.2.3 PPIXpress

PPIXpress is a platform-independent JAVA-based tool to infer condition and cell-specific PPINs based on respective gene-level or transcriptome-level expression data. The tool requires as inputs gene-level or transcript-level expression data of at least one sample and a reference complete PPIN of the respective species. The tool then infers a PPIN for each specific condition or transcriptomic sample via a mapping and a subsequent contextualization step.

Fig 3.2 provides an overview of the methodology.

In the mapping step, the tool evaluates the complete PPIN, which can be loaded manually by the user or automatically retrieved from the latest versions of the MENTHA [25] or IntAct databases. Each PPI is annotated with a corresponding domain-domain interaction (DDI). The longest isoform of each protein is considered, as it is typically the principal variant identified in experimental analyses and databases [147, 148]. Proteins are identified by querying UniProt [149], and this information is used to access the Ensembl [150] database for gene, protein, and transcript annotations. Pfam domain associations of each transcript are determined via Pfam domain annotation database [151] and InterProScan [152]. Subsequently, physical interactions between domains are retrieved from high-confidence data in the DOMINE [153] and IDDI [154] databases, complemented by info from 3did [155] and iPfam [156]. Consequently, PPIXpress generates a corresponding reference domain-domain interaction network (DDIN). If a specific PPI cannot be mapped to a known DDI, artificial domains are added to the interacting proteins to ensure a complete association between the protein-level and domain-level reference networks.

In the contextualization step, gene or transcript-level expression data is utilized. The user sets a lower threshold for the minimum abundance required for a gene or transcript to be considered. Here, we set the threshold to the smallest possible value of 1 read mapped to a gene. For gene-level expression data, genes expressed above or at this threshold are mapped to their longest coding transcripts via Ensembl, with these transcripts representing the associated proteins. The domain annotations and domain-domain interactions from the mapping step are used to infer a pruned, sample-specific domain-domain interaction network (DDIN) based on the expression data. Proteins not supported by domains are removed from the reference PPIN, generating a sample-specific PPIN. For transcript-level data, the tool performs the same mapping of proteins and node pruning process. Additionally, edge pruning is conducted by removing edges in the reference PPIN that lack support from the sample-specific DDIN, resulting in a sample-specific PPIN.

3.2.4 PPICompare

Each sample-specific PPIN generated by PPIXpress reflects the PPI network in one sample of a particular state. When aiming at identifying the biological processes that drive a cell towards differentiation or disease, one is interested in characterizing the rewiring of PPIs between two different states by comparing representative sets of samples for both states. To this end, the JAVA tool, PPICompare, enables downstream processing of the PPIXpress results. The tool requires two groups of sample-specific PPINs that can be generated either by PPIXpress or by other methods. PPICompare then detects the significantly rewired PPIs between the sample-specific PPINs of the two groups, estimates the causes for each rewired PPI, and identifies a small set of causes that can explain all rewired PPIs. This is done essentially in three steps: The first step is to examine the interactome differences between all pairs of samples between

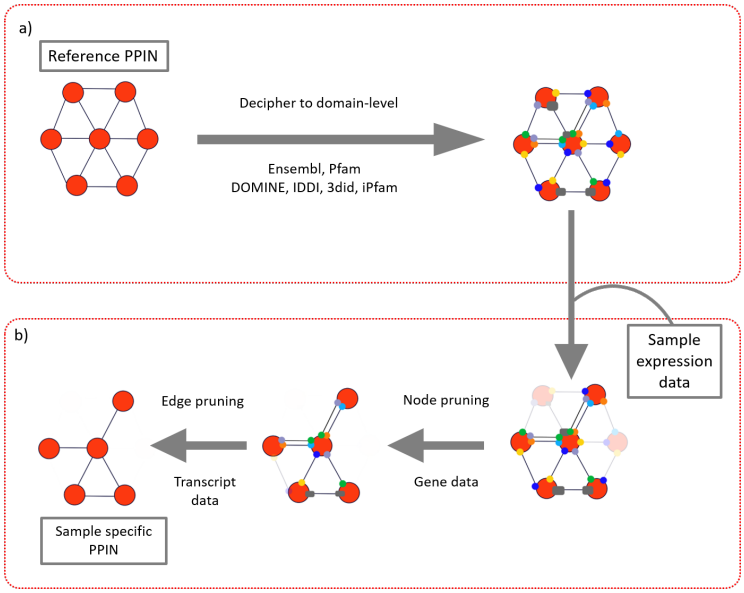


Figure 3.2: PPIXpress operates in two stages. In the initial mapping step, a), the protein nodes (red circles) of the reference PPIN (where black lines represent interactions between the proteins) are associated with respective protein domains (small coloured circles on the circumference of the red circles) yielding a corresponding domain-domain interaction network (DDIN). Artificial domains (grey rectangles) are introduced when a PPI cannot be mapped to an underlying DDI. b) represents the contextualisation step, where the sample expression profile is used to infer a corresponding sample-specific DDIN. The reference PPIN is then pruned by comparison to the sample-specific DDIN to create a subnetwork, the sample-specific PPIN.

groups. The second step is to assess the statistical significance and causes of each rewiring event. Finally, a small set of likely changes in the transcriptome is identified that explain all detected rewirings. These steps are illustrated in Fig. 3.3.

In the first step, one group is designated as the reference group, and each sample of the second group is compared to each sample of the reference group. The tool records PPIs lost or gained in the second group compared to the reference group, creating a differential network for each pairwise comparison. Edges in this network are annotated with +1 if a PPI is present in the second group but absent in the reference group, and -1 if a PPI is absent in the second group but present in the reference group. Summing these differential networks yields a global differential network, where edge annotations reflect the total number of changes for each PPI between the two groups. The proportion of rewiring events in each pairwise comparison is quantified using the Jaccard distance [157], referred to as the rewiring probability (P_{rew_i}). The overall rewiring probability between the two groups (P_{rew}) is the average of all individual pairwise probabilities:

$$P_{rew} = \frac{1}{N} \sum_{i}^{N} P_{rew\_i}$$

In the second step, PPICompare identifies statistically significant rewired interactions using a one-tailed binomial test, with p-values corrected for multiple testing via the Benjamini-Hochberg method at a user-defined FDR threshold. The tool outputs these significant rewired interactions as a differential network. It also reads the 'major_transcript' file from each sample, an output from PPIXpress, to determine transcriptomic changes underlying the differential interactions. This identifies whether rewiring is due to differential gene expression, alternative splicing, or both.

The final step involves identifying a small set of transcriptomic changes explaining all rewiring events. The significant rewiring events and their causes form a bipartite graph, with reasons for rewiring connected to corresponding events. Each reason is weighted based on the number of significant rewiring events it causes and the number of pairwise comparisons in which these events occur. A greedy algorithm [158] is then used to solve this set coverage problem, estimating a minimal subset of reasons that explain all rewiring events.

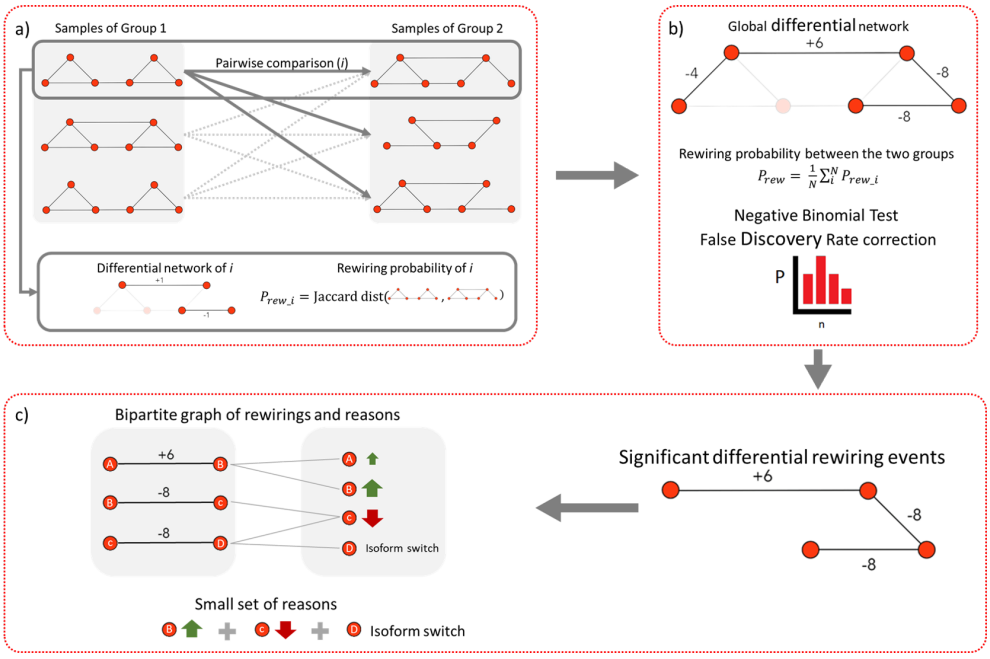


Figure 3.3: Workflow of PPICompare, where a) represents the examining of interactome differences between all pairs of samples between groups, b) represents assessing the significance and causes of each rewiring event and c) represents identifying a small set of likely changes that explain all the observed rewirings.

3.3 Results and discussion

3.3.1 Overall results

We executed all 6 pipelines for three different comparisons, either nevi subtype N1 samples against melanoma subtype M1 samples, nevi subtype N2 samples against melanoma subtype M2 samples, or all nevi samples against all melanoma samples (N vs M), respectively. An overview of the results from this analysis is presented in Table 3.1, namely the number of significant DE genes, the number of nodes and interactions in the differential networks and the number of clusters predicted in the differential network by ClusterONE. Pipelines II and IV, which incorporate the TREAT method, identified far fewer significant DE genes than the other pipelines. This resulted in smaller differential networks with fewer proteins and a smaller number of clusters identified. In pipelines III to V, which are based on the construction of PPINs using PPIXpress prior to DE analysis, more clusters were identified, but these were smaller in size in comparison to pipelines 0 to II. In the differential networks, the number of interactions determined using IntAct and STRING in pipelines 0 and I was considerably larger than the number of interactions constructed by PPIXpress and PPICompare. This may be one reason for the larger clusters found by pipelines 0 to II.

	Pipeline 0			Pipeline I			Pipeline II			Pipeline III			Pipeline IV			Pipeline V		
	N1 vs M1	N2 vs M2	N vs M	N1 vs M1	N2 vs M2	N vs M	N1 vs M1	N2 vs M2	N vs M	N1 vs M1	N2 vs M2	N vs M	N1 vs M1	N2 vs M2	N vs M	N1 vs M1	N2 vs M2	N vs M
No. of significant differentially expressed genes	10,103	3,963	12,318	10,103	3,963	12,318	1,919	29	1,534				1,919	29	1,534	10,103	3,963	12,318
No. of significant differentially expressed proteins	5,429	1,816	6,055	5,429	1,816	6,055	930	20	851				930	20	851	5,429	1,816	6,055
Differential network (nodes / interactions)	4,779 / 96,091	1,304 / 13,345	5,382 / 118,903	5,405/ 121,237	1,807/ 18,528	6,021/ 145,263	927/ 8,477	20/ 3	841/ 8,008	10,973/ 34,504	10,319/ 28,591	9,687/ 25,224	3,350/ 4,487	373/ 363	3,288/ 4,094	8,160/ 18,441	6,104/ 11,541	8,095/ 17,558
Network diameter	8	9	8	9	10	9	10	1	10	12	11	12	16	4	18	12	13	12
No. of significant clusters	8	24	11	17	20	20	19	0	32	86	96	107	157	3	152	149	161	155

Table 3.1: Overview of results obtained when processing transcriptomics data by the six pipelines shown in Fig 3.1

Then, we split up the differential networks generated by the six pipelines into two networks based on the direction of expression regulation and rewirings of protein interactions. For pipelines 0, I and II, the differential network was split based on the direction of protein regulation. One of the networks was formed by the significantly upregulated proteins and the other one by the significantly downregulated proteins. For pipelines III, IV and V, the differential networks were split based on rewiring events. All protein interactions that were expressed exclusively in the melanoma group were considered as one network, and all interactions that were expressed exclusively in the melanocytic nevi group were considered as another network. On all these new sub-networks, cluster prediction was performed using ClusterONE. In pipelines 0, I and II, these clusters represent clusters that are upregulated or downregulated in melanoma. In pipelines III, IV and V, represent clusters that are exclusively found or lost in melanoma. An overview of these results is provided in Tables 3.2 and 3.3.

	Pipeline 0			Pipeline I			Pipeline II		
	N1 vs M1	N2 vs M2	N vs M	N1 vs M1	N2 vs M2	N vs M	N1 vs M1	N2 vs M2	N vs M
No. of significant differentially expressed genes	10,103	3,963	12,318	10,103	3,963	12,318	1,919	29	1,534
No. of significantly downregulated genes in melanoma	3,774	1,029	4,596	3,774	1,029	4,596	711	11	824
No. of significantly downregulated proteins in melanoma	1,630	476	1,958	1,630	476	1,958	444	8	510
Network with downregulated proteins (nodes/interactions)	1,086/ 10,262	260/ 2,316	1,341/ 15,315	1,626/ 9,294	475/ 1,133	1,953/ 13,835	443/ 5,982	8/ 3	501/ 6,708
Network diameter	9	13	10	13	16	10	8	1	9
No. of significant clusters in the downregulated protein network.	21	5	18	37	23	52	12	0	18
No. of significantly upregulated genes in melanoma	6,329	2,934	7,722	6,329	2,934	7,722	1,208	18	710
No. of significantly upregulated proteins in melanoma	3,799	1,340	4,097	3,799	1,340	4,097	486	12	341
Network with upregulated proteins (nodes/interactions)	3,206/ 47,443	854/ 5,640	3,499/ 54,287	3,779/ 73,437	1,332/ 14,081	4,068/ 81,805	484/ 955	12/ 0	340/ 380
Network diameter	9	12	10	11	10	9	10	0	12
No. of significant clusters in the upregulated protein network.	12	35	18	17	12	19	25	0	20

Table 3.2: Prediction of clusters in upregulated and downregulated genes in melanoma from pipelines 0, I and II.

	Pipeline III			Pipeline IV			Pipeline V		
	N1 vs M1	N2 vs M2	N vs M	N1 vs M1	N2 vs M2	N vs M	N1 vs M1	N2 vs M2	N vs M
Differential network (nodes/interactions)	10,973/ 34,504	10,319/ 28,591	9,687/ 25,224	3,350/ 4,487	373/ 363	3,288/ 4,094	8,160/ 18,441	6,104/ 11,541	8,095/ 17,558
No. of interactions found only in melanoma (nodes/interactions)	7,687/ 16,369	8,993/ 20,322	7,577/ 15,263	2,068/ 2,416	70/ 60	2,645/ 3,143	5,973/ 10,373	5,132/ 7,872	6,401/ 11,139
Network diameter	13	13	12	17	4	15	13	14	14
No. of significant cluster from the melanoma only network	111	115	117	113	3	120	146	168	146
No. of interactions found only in nevi (nodes/interactions)	7,977/ 18,135	4,609/ 8,269	5,500/ 9,961	1,747/ 2,071	307/ 303	940/ 951	4,640/ 8,068	2,116/ 3,669	3,993/ 6,419
Network diameter	11	12	12	18	2	16	12	16	13
No. of significant cluster from the nevi only network	133	117	137	96	0	61	136	54	136

Table 3.3: Prediction of clusters by pipelines III, IV and V for networks that contain interactions exclusively detected in melanoma or exclusively in melanocytic nevi networks.

In both sub-network conditions, pipelines III to V generated networks with larger diameters than those obtained by pipelines 0 to II. This indicates that the networks constructed by pipelines III to V are less interconnected than those generated by pipelines 0 to II. This may also explain the tangible difference in the number of clusters identified. In both sub-network conditions, the number of clusters identified in networks from pipelines III to V is

considerably larger than the number of clusters identified in networks from pipelines 0 to II.

3.3.2 Overlap of clusters with known complexes

Next, we compared the clusters identified in the differential networks to reported protein complexes collected at the Complex Portal [159] by considering the Jaccard similarity score. In pipeline IV when N1 samples were compared against M1 samples, Cluster 48 contained all elements of the known chaperonin-containing T-complex and had the highest similarity score among all clusters of 0.666. Previous studies indeed indicated that this chaperonin-containing T-complex is highly expressed in melanoma tissues [160]. Furthermore, silencing of this complex was shown to influence multiple cancer-related pathways, including cyclin and cell cycle regulation signalling, PPAR-alpha/RXR-alpha signalling, RhoGDI signalling and PPAR signalling [161]. Table 3.4 lists the five clusters with the highest similarity scores. It was shown that neuronal nicotinic acetylcholine receptors and amiloride-sensitive sodium channel complexes, which exhibit a high Jaccard similarity score, are directly involved in the progression of melanoma [162]. Nicotine acetylcholine receptors are involved in melanoma metastasis through NOTCH1 signalling and are also responsible for PI3K/AKT and ERK signalling pathways [163]. Amiloride-sensitive sodium channel complexes are highly expressed in melanoma cells and promote cell migration and proliferation [164].

Predicted cluster number	Pipeline	Comparison	Size of predicted cluster	P-value of predicted cluster	Name of reported cluster	Size of reported cluster	No. of elements similar between predicted and reported clusters	Jaccard similar score
48	IV	N1 vs M1	12	0.00065	Chaperonin-containing T-complex	8	8	0.666
7	0	N1 vs M1	5	0.041	Neuronal nicotinic acetylcholine receptor complex, alpha3-alpha5-beta4	3	3	0.6
7	0	N1 vs M1	5	0.041	Neuronal nicotinic acetylcholine receptor complex, alpha3-alpha6-beta4	3	3	0.6
91	V	N1 vs M1	5	0.021	Cardiac Troponin complex	3	3	0.6
133	IV	N vs M	5	0.028	Amiloride-sensitive sodium channel complex, delta-alpha-beta-gamma	4	3	0.5

Table 3.4: Clusters predicted by ClusterONE on differential networks are compared with reported protein complexes. Number of overlapping proteins and the Jaccard similarity scores are reported. Shown are the top five clusters with the highest Jaccard similarity scores.

Similarly, we compared the clusters identified from the sub-networks of up or downregulated proteins with reported protein complexes. The five clusters with the highest Jaccard similarity scores are presented in Tables 3.5 and 3.6. It has been demonstrated that the E3 ubiquitin ligase complex plays an essential role in the progression of melanoma and is a promising drug

target, exhibiting reduced toxicity and enhanced selectivity [165, 166]. The E3 ligase has been demonstrated to influence the BRAF and MAPK pathways, melanoma metastasis and differentiation [167]. A somatic mutation in one of the NMDA receptors, GRIN2A, is prevalent in malignant melanoma cells and was shown to disrupt tumour-suppressing characteristics and increase cell migration [168]. The NuA4 histone acetyltransferase complex was shown to be modulated by Inhibitor of growth protein 3 (ING3) to induce apoptosis in melanoma cells following UV irradiation [169].

Predicted cluster number	Pipeline	Comparison	Size of predicted cluster	P-value of predicted cluster	Name of reported cluster	Size of reported cluster	No. of elements similar between predicted and reported clusters	Jaccard similar score
12	IV	N1 vs M1	14	0.0000047	Chaperonin-containing T-complex	8	8	0.57
58	III	N vs M	8	0.0017	TRAPP II complex, TRAPPC2 variant	10	6	0.5
126	III	N vs M	5	0.028	Amiloride-sensitive sodium channel complex, delta-alpha-beta-gamma	4	3	0.5
25	III	N2 vs M2	8	0.000103	TRAPP II complex, TRAPPC2 variant	10	6	0.5
56	V	N1 vs M1	6	0.0038	Amiloride-sensitive sodium channel complex, delta-alpha-beta-gamma	4	3	0.42857

Table 3.5: Clusters predicted by ClusterONE in the sub-networks of downregulated proteins (from pipelines 0, I and II) and sub-network of interactions lost in melanoma (from pipelines III, IV and V) were compared with reported protein complexes. Shown are the top five clusters with the highest Jaccard similarity scores.

Predicted cluster number	Pipeline	Comparison	Size of predicted cluster	P-value of predicted cluster	Name of reported cluster	Size of reported cluster	No. of elements similar between predicted and reported clusters	Jaccard similar score
37	V	N1 vs M1	13	0.000786	GID E3 ubiquitin ligase complex, RMND5A-RANBP10 variant	9	7	0.466
29	V	N1 vs M1	16	0.000055	NuA4 histone acetyltransferase complex	19	11	0.458
14	III	N1 vs M1	16	0.000056	CCR4-NOT mRNA deadenylase complex, CNOT6L-CNOT7 variant	10	8	0.444
66	III	N2 vs M2	27	0.018	NuA4 histone acetyltransferase complex	19	14	0.437
81	III	N vs M	7	0.018	NMDA receptor complex, GluN1-GluN2A-GluN2B	3	3	0.428

Table 3.6: Clusters predicted by ClusterONE in the sub-networks of upregulated proteins (from pipelines 0, I and II) and sub-network of interactions newly added in melanoma (from pipelines III, IV and V) were compared with reported protein complexes. Shown are the top five clusters with the highest Jaccard similarity scores.

3.3.3 Pathway enrichment of clusters

Finally, we analyzed the enrichment of Reactome pathways in all individual clusters predicted from any of the differential networks constructed by the six pipelines. A comprehensive list of all clusters and their top three significant enriched Reactome pathways is provided in the supplementary material (Tables A.1 to A.18). Recent studies have identified a number of key pathways that are thought to be involved in the progression and development of melanoma. For instance, the MAPK/ERK signalling pathway is upregulated in response to mutations in the BRAF and RAS genes, which result in cell proliferation, migration, and metastasis [170]. WNT signalling pathway is hyperactivated due to mutations in the encoding components or genes encoding ß-catenin [171]. The PI3K-AKT pathway is frequently activated in melanoma, due to mutations in the AKT1 or PIK3CA genes or changes in the copy numbers of pathway components [172]. Table 3.7 lists which one of these major signalling pathways that are associated with the initiation, development and migration of melanoma, were recovered and enriched in the predicted clusters. Cluster 8 associated with PI3K-AKT signaling was identified by pipeline 0 in the N vs M comparison. Furthermore, three clusters associated with WNT signalling were identified by pipeline I. All other clusters listed in Table 3.7 were identified by pipelines III to V.

Similarly, Reactome pathway enrichment analysis was performed on all clusters predicted for the up- or downregulated sub-networks constructed by the six pipelines. The pathways that are associated with melanoma and enriched in the predicted clusters are tabulated in Tables 3.8 and 3.9. Nevi are generally composed of hyper-proliferated melanocytes with activated BRAF or NRAS oncogenes that showed temporary proliferation but later underwent oncogene-induced senescence. However, in some cases, the tumour suppressor pathways are inactivated and proliferation-inducing pathways such as WNT signalling pathways are activated [173]. This can be observed in Table 3.8 where WNT signalling is significantly enriched and upregulated in the protein clusters detected in melanocytic nevi samples. Canonical WNT signalling regulates other oncogenic signalling pathways such as PI3K-AKT signalling and MAPK signalling due to ß-catenin stabilization [171]. In summary, pipelines III to V identified clusters enriched for these signalling pathways in all three comparisons (N vs M, N1 vs M and N2 vs M2). Pipeline I identified only WNT signalling, whereas pipelines 0 and II did not yield any hits.

Finally, we need to mention one caveat with regard to Pipeline III, which initially prompted our decision to carry out this research. When we initially analyzed the differential networks constructed by pipeline III, PPICompare reported statistically significant rewiring of the interactions of P35354 (Prostaglandin G/H synthase 2 (PTGS2)) with P04439 (HLA class I histocompatibility antigen, A alpha chain), Q5S007 (Leucine-rich repeat serine/threonine-protein kinase 2 (STK2)), O14939 (phospholipase D2) and Q9H0V9 (VIP36-like protein) when N1 was compared with N2, with N1 as the reference group. This was labelled by PPICompare to be due to the loss of P35354 in N2. Similarly, the same interactions were rewired when comparing

Pathway	Pipeline	Nevi vs Melanoma	Cluster No.	Size	Reactome Pathways	Adjusted P-value
TGF-ß signalling	III	N2 vs M2	33	13	Downregulation of TGF-beta receptor signalling (R-HSA-2173788)	6.320e-04
WNT signalling	I	N vs M	9	288	Signalling by WNT (R-HSA-195721) TCF dependent signalling in response to WNT (R-HSA-201681)	1.120e-14 4.510e-14
	I	N vs M	10	347	Signalling by WNT (R-HSA-195721) TCF dependent signalling in response to WNT (R-HSA-201681)	5.790e-11 1.610e-10
	III	N vs M	25	19	WNT ligand biogenesis and trafficking (R-HSA-3238698) Signalling by WNT (R-HSA-195721)	1.570e-20 3.170e-11
	IV	N vs M	123	9	Negative regulation of TCF-dependent signalling by WNT ligand antagonists (R-HSA-3772470)	4.580e-02
	V	N vs M	72	11	WNT ligand biogenesis and trafficking (R-HSA-3238698) Signalling by WNT (R-HSA-195721)	2.530e-09 2.190e-04
	I	N1 vs M1	15	262	Signalling by WNT (R-HSA-195721) TCF dependent signalling in response to WNT (R-HSA-201681)	1.630e-12 2.360e-10
	III	N1 vs M1	72	13	WNT ligand biogenesis and trafficking (R-HSA-3238698) Signalling by WNT (R-HAS-195721)	7.440e-17 2.120e-09
	IV	N1 vs M1	7	47	Repression of WNT target genes (R-HSA-4641265)	5.830e-03
	IV	N1 vs M1	117	7	Signalling by WNT (R-HSA-195721)	1.820e-03
	V	N1 vs M1	72	7	WNT ligand biogenesis and trafficking (R-HSA-3238698) Negative regulation of TCF-dependent signalling by WNT ligand antagonists (R-HSA-3772470)	1.610e-04 8.910e-03
MAPK signalling	III	N vs M	3	46	Oncogenic MAPK signalling (R-HSA-6802957)	4.32e-02
	IV	N1 vs M1	1	36	Signalling by MAP2K mutants (R-HSA-9652169)	2.290e-02
	III	N2 vs M2	37	18	p38MAPK events (R-HSA-171007)	4.270e-02
	III	N2 vs M2	67	16	p38MAPK events (R-HSA-171007)	3.770e-02
	V	N2 vs M2	73	5	IFNG signalling activates MAPKs (R-HSA-9732724)	3.400e-03
PI3K-AKT signalling	0	N vs M	8	18	Erythropoietin activates Phosphoinositide-3-kinase (PI3K) (R-HSA-9027276)	1.490e-04
	III	N vs M	74	24	MET activates PI3K/AKT signalling (R-HSA-8851907)	1.950e-02
	IV	N vs M	111	13	PIP3 activates AKT signalling (R-HSA-1257604)	1.230e-02
	IV	N vs M	145	5	Negative regulation of the PI3K/AKT network (R-HSA-199418) PI5P, PP2A and IER3 Regulate PI3K/AKT Signalling (R-HSA-6811558)	3.220e-06 4.790e-06
	V	N vs M	18	23	MET activates PI3K/AKT signalling (R-HSA-8851907)	8.490e-03
	IV	N1 vs M1	70	13	PIP3 activates AKT signalling (R-HSA-1257604)	1.230e-02
	V	N1 vs M1	19	25	MET activates PI3K/AKT signalling (R-HSA-8851907)	1.020e-02
	III	N2 vs M2	83	11	PI5P, PP2A and IER3 Regulate PI3K/AKT Signalling (R-HSA-6811558)	6.550e-03
	IV	N2 vs M2	60	13	AKT phosphorylates targets in the nucleus (R-HSA-198693)	1.420e-02
BRAF signalling	III	N vs M	3	46	Signalling by BRAF and RAF1 fusions (R-HSA-6802952) Signalling by moderate kinase activity BRAF mutants (R-HSA-6802946)	4.080e-02 4.780e-02
c-KIT-SCF signalling	IV	N1 vs M1	34	18	Signalling by SCF-KIT (R-HSA-1433557)	2.230e-03
RAS signalling	V	N vs M	50	10	CREB1 phosphorylation through NMDA receptor-mediated activation of RAS signalling (R-HSA-442742)	3.180e-02
	III	N2 vs M2	67	16	Signalling to RAS (R-HSA-167044)	4.940e-02
PTEN signalling	IV	N vs M	111	13	PTEN Regulation (R-HSA-6807070) Regulation of PTEN gene transcription (R-HSA-8943724)	3.300e-03 8.060e-03
	IV	N1 vs M1	70	13	PTEN Regulation (R-HSA-6807070) Regulation of PTEN gene transcription (R-HSA-8943724)	3.300e-03 8.060e-03

Table 3.7: Reactome pathways associated with the development of melanoma that are enriched in protein clusters identified in differential networks.

N2 with M2 groups, using N2 as the reference group. This was labelled by PPICompare to be due to the gain of P35354 in M2. This observation suggests that P35354 is either absent or expressed at low levels in the N2 samples.

Pathway	Pipeline	Nevi vs Melanoma	Cluster No.	Size	Reactome Pathways	Adjusted P-value
TGF-ß signalling	III	N1 vs M1	26	20	Signalling by TGFB family members (R-HSA-9006936)	6.370e-03
WNT signalling	I	N vs M	40	50	Signalling by WNT (R-HSA-195721) TCF dependent signalling in response to WNT (R-HSA-201681)	8.740e-10 6.620e-09
	III	N vs M	26	13	WNT ligand biogenesis and trafficking (R-HSA-3238698) Signalling by WNT (R-HSA-195721)	6.440e-18 1.720e-09
	III	N vs M	136	5	Signalling by WNT in cancer (R-HSA-4791275)	3.400e-02
	V	N vs M	91	5	WNT ligand biogenesis and trafficking (R-HSA-3238698) Signalling by WNT (R-HSA-195721)	4.600e-05 2.470e-02
	III	N1 vs M1	29	12	WNT ligand biogenesis and trafficking (R-HSA-3238698) Signalling by WNT (R-HSA-195721)	2.480e-18 6.710e-10
	IV	N1 vs M1	4	39	Repression of WNT target genes (R-HSA-4641265)	1.620e-03
	IV	N1 vs M1	61	8	Signalling by WNT (R-HSA-195721)	3.600e-03
	IV	N1 vs M1	90	10	WNT ligand biogenesis and trafficking (R-HSA-3238698) Signalling by WNT (R-HSA-195721)	5.490e-04 1.060e-02
	V	N1 vs M1	88	6	WNT ligand biogenesis and trafficking (R-HSA-3238698) Negative regulation of TCF-dependent signalling by WNT ligand antagonists (R-HSA-3772470)	9.180e-05 6.370e-03
	I	N2 vs M2	21	6	Signalling by WNT (R-HSA-195721)	9.070e-06
	III	N2 vs M2	68	9	Negative regulation of TCF-dependent signalling by WNT ligand antagonists (R-HSA-3772470) Signalling by WNT in cancer (R-HSA-4791275) Signalling by LRP5 mutants (R-HSA-5339717)	6.750e-05 4.420e-04 2.190e-03
	III	N2 vs M2	115	5	Signalling by WNT in cancer (R-HSA-4791275)	3.400e-02
	V	N2 vs M2	53	6	Signalling by WNT in cancer (R-HSA-4791275)	3.390e-02
MAPK signalling	IV	N1 vs M1	2	40	Signalling by MAP2K mutants (R-HSA-9652169) Oncogenic MAPK signalling (R-HSA-6802957)	1.890e-02 2.460e-02
c-KIT-SCF signalling	V	N vs M	40	21	Signalling by KIT in disease (R-HSA-9669938)	1.190e-02
RAS signalling	V	N vs M	70	17	MET activates RAS signalling (R-HSA-8851805)	4.610e-02
	V	N vs M	77	11	CREB1 phosphorylation through NMDA receptor-mediated activation of RAS signalling (R-HSA-442742)	3.880e-02
	V	N vs M	131	10	CREB1 phosphorylation through NMDA receptor-mediated activation of RAS signalling (R-HSA-442742)	1.240e-04
	III	N1 vs M1	95	30	RAS signalling downstream of NF1 loss-of-function variants (R-HSA-6802953)	3.950e-02
	III	N1 vs M1	122	8	CREB1 phosphorylation through NMDA receptor-mediated activation of RAS signalling (R-HSA-442742)	2.380e-02
PTEN signalling	IV	N vs M	9	15	PTEN Regulation (R-HSA-6807070) Regulation of PTEN gene transcription (R-HSA-8943724)	6.230e-03 1.280e-02
	IV	N1 vs M1	8	17	PTEN Regulation (R-HSA-6807070) Regulation of PTEN gene transcription (R-HSA-8943724)	1.070e-02 1.900e-02

Table 3.8: When comparing transcriptomic samples from melanocytic nevi with melanoma samples, Reactome pathways associated with the development of melanoma were enriched in protein clusters identified in the sub-networks of downregulated proteins in melanoma (from pipelines 0, I and II) or in the sub-networks of interactions lost in melanoma (from pipelines III, IV and V).

However, the DisGeNET database documents that PTGS2 shows a significant association with melanocytic nevi [174]. Hence, we conducted a meticulous investigation of the expression level of this protein and questioned why it was flagged by PPICompare. We examined the gene expression profile of PTGS2 in all groups, with N2 samples exhibiting a considerably high expression level of PTGS2 (Fig: 3.4). This suggests that the identified "reason" for the rewiring of

Pathway	Pipeline	Nevi vs Melanoma	Cluster No.	Size	Reactome Pathways	Adjusted P-value
TGF-ß signalling	III	N vs M	48	14	Downregulation of TGF-beta receptor signalling (R-HSA-2173788)	5.78E-04
	V	N vs M	110	32	TGF-beta receptor signalling activates SMADs (R-HSA-2173789)	2.98E-02
	III	N1 vs M1	66	6	TGF-beta receptor signalling activates SMADs (R-HSA-2173789)	3.59E-02
	III	N2 vs M2	63	12	Downregulation of TGF-beta receptor signalling (R-HSA-2173788)	3.34E-04
WNT signalling	I	N1 vs M1	8	151	TCF dependent signalling in response to WNT (R-HSA-201681) Deactivation of the beta-catenin transactivating complex (R-HSA-3769402)	6.42E-05 1.15E-04
	IV	N1 vs M1	32	20	Repression of WNT target genes (R-HSA-4641265)	3.75E-04
MAPK signalling	III	N vs M	50	18	IFNG signalling activates MAPKs (R-HSA-9732724)	4.61E-02
	V	N vs M	95	10	IFNG signalling activates MAPKs (R-HSA-9732724)	1.53E-02
	V	N vs M	113	9	IFNG signalling activates MAPKs (R-HSA-9732724)	1.22E-02
	IV	N1 vs M1	8	33	Oncogenic MAPK signalling (R-HSA-6802957)	2.25E-02
	V	N2 vs M2	75	5	IFNG signalling activates MAPKs (R-HSA-9732724)	3.40E-03
PI3K-AKT signalling	III	N vs M	53	26	MET activates PI3K/AKT signalling (R-HSA-8851907)	1.67E-02
	IV	N vs M	91	9	Negative regulation of the PI3K/AKT network (R-HSA-199418) PI5P, PP2A and IER3 Regulate PI3K/AKT Signalling (R-HSA-6811558)	3.71E-07 5.10E-07
	V	N vs M	19	23	MET activates PI3K/AKT signalling (R-HSA-8851907)	8.49E-03
	III	N1 vs M1	29	30	MET activates PI3K/AKT signalling (R-HSA-8851907)	1.53E-02
	V	N1 vs M1	27	26	MET activates PI3K/AKT signalling (R-HSA-8851907)	1.67E-02
	V	N2 vs M2	57	13	AKT phosphorylates targets in the nucleus (R-HSA-198693)	1.42E-02
PTEN signalling	V	N2 vs M2	46	22	PTEN Regulation (R-HSA-6807070) Regulation of PTEN gene transcription (R-HSA-8943724)	3.22E-02 4.26E-02

Table 3.9: When comparing transcriptomic samples from melanocytic nevi with melanoma samples, Reactome pathways associated with the development of melanoma were enriched in protein clusters identified in the sub-networks of upegulated proteins in melanoma (from pipelines 0, I and II) or in the sub-networks of interactions only found in melanoma (from pipelines III, IV and V).

interactions involving P35354 due to the loss of the protein is not valid. In the N2 group, one of the samples exhibited zero gene expression levels for P35354. When PPICompare was applied to compare N1 samples with N2 samples, 130 pairwise comparisons were performed between the samples of the two groups (13 N1 melanocytic nevi samples, 10 N2 melanocytic nevi samples). Of these comparisons, 13 necessarily involved the loss of P35354, which was deemed to be significant in causing a rewiring event. Thus, in pipelines IV and V, we added an additional criterion for significant rewiring events by additionally requiring the differential expression of at least one of the binding partners.

3.4 Discussion

The results of our analysis demonstrate that the processing of melanoma and nevi transcriptomic samples yielded diverse outcomes depending on the software and pipelines used. Pipelines 0, I and II, where differential expression

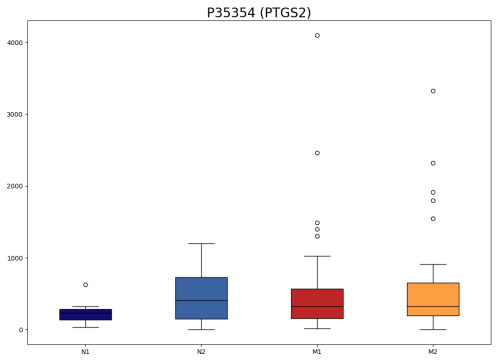


Figure 3.4: Box plot illustrating the associated gene expression levels of PTGS2 in the N1, N2, M1, and M2 groups.

analysis precedes network construction, yielded fewer clusters but they were fairly large in size. The networks were constructed using either the IntAct or the STRING databases. STRING comprises both predicted and experimental protein interaction data, whereas IntAct contains only experimental protein interaction data. Consequently, the networks constructed using STRING were more densely connected than IntAct networks. Conversely, pipelines III, IV and V, where PPI networks are constructed first and then subjected to differential analyses, yielded more clusters but these were smaller in size. In pipeline III, the networks were constructed based on the rewiring events between melanocytic nevi and melanoma samples. In contrast, in pipelines IV and V, these rewiring events were considered significant only if at least one of the binding partners was also significantly differentially expressed. Table 3.7 (and supplementary tables A.10, A.13, A.16) reveals that the clusters constructed by pipelines III, IV and V exhibit a greater degree of enrichment with specific Reactome pathways that are characteristic of melanoma initiation, development and metastasis. The majority of clusters predicted by pipelines 0, I and II were not enriched with pathways specific to melanoma progression. This may be attributed to the fact that certain supportive proteins that bind to other differentially expressed proteins may play a role in melanoma-related pathways, although they may not necessarily be differentially expressed.

We found that analysis of sub-networks provided a more detailed understanding of which clusters are upregulated, downregulated, exclusively found or lost in melanoma. The supplementary tables A.2, A.3, A.5, A.6 etc., also include enrichment analyses of the clusters, which provide insight into the Reactome pathways that lead melanocytic nevi cells to develop into melanoma cells. For instance, clusters predicted to be responsible for WNT signalling were found to be upregulated and exclusively present in nevi samples. This matches the findings of previous studies that demonstrated a crucial role of WNT signalling in increasing proliferation and undermining senescence signalling [171, 173]. This signalling pathway is responsible for the activation of one of the most crucial melanoma oncogenes, Microphthalmia-associated Transcription Factor (MITF), through ß-catenin.

In conclusion, we found it beneficial to separately analyze PPINs of upregulated or downregulated interactions. This approach led to the identification of protein clusters that were more meaningful and better interpretable than when analyzing the differential PPI network. The utilization of TREAT in the context of PPI rewiring is not recommended. The PPIXpress and PPI-Compare tools, in combination with DEseq2, were found to be an effective approach for the identification of compact protein clusters that exhibited good overlap with known protein complexes and meaningful enrichment of pathways.

3.5 Acknowledgement

This study was funded by Deutsche Forschungsgemeinschaft through CRC 1027 (project C3). The authors would like to thank Prof. Manfred Kunz, Dr Hans Binder and Dr Maria Schmidt for providing the RNA-seq primary melanoma and melanocytic nevi raw data.

Chapter 4
Downstream analysis of proteomic data

4.1 Introduction

In this chapter, the section 4.3.1, was adapted from the submitted manuscript from J. Dudek, T. Faidt, C. Fecher-Trost, S. Thangamurugan, P. Bayenat, S. Traut- mann, A. Holtsch, F. Müller, V. Helms, K. Jacobs, and M. Hannig (2024), entitled "Synthetic hydroxyapatite – a perfect substitute for dental enamel in biofilm formation studies". I was responsible for the bioinformatics analysis, which entailed qualitative analysis, statistical similarity tests, the writing of the corresponding methods, and the preparation of visualisations for the results of the manuscript. I was also responsible for describing the computational results in this chapter. Johanna Dudek was responsible for the design of the project. She also provided assistance in the planning of the data analysis. The section, 4.3.2, was adapted from the manuscript from S. Trautmann, S. Thangamurugan, C. Fecher-Trost, J. Dudek, V. Flockerzi, V. Helms, and M. Hannig, "Snapshot of the seconds-pellicle – first insights in its ultrastructure, proteomic composition and changes over time". I was responsible for the bioinformatics data analysis. Additionally, I drafted the initial versions of the corresponding method sections and prepared the relevant figures for the manuscript. I was also responsible for describing the computational results in this chapter. Simone Trautmann was responsible for the design of the project and also provided assistance in the planning of the data analysis. The section 4.3.3, describes unpublished research work carried out by myself in collaboration with Dr Lilia Lemke. I was responsible for the bioinformatic analysis, and wrote this subchapter. Dr. Lilia lemke assisted me with the design and planning for data analysis.

In the buccal cavity, shortly after teeth have been cleaned, a protein

film, known as the salivary pellicle, is actively formed on the surface of the oral cavity. The film serves to prevent the formation of tartar deposits and to protect the teeth from acid-producing microorganisms. This pellicle functions as a physical barrier, preventing the erosion and demineralisation of the teeth [175]. A number of specific proteins present in saliva undergo adsorption as a result of electrostatic calcium-phosphate ionic interaction, including glycoproteins, albumin, mucin, proline-rich proteins, cystatins, and so forth [176]. The adsorption of these proteins serves as the initial layer of the film, thereby initiating the formation of the pellicle. The formation of this layer occurs within a timeframe of 30 seconds to 3 minutes. Furthermore, additional interaction forces, including Van der Waals and hydrophobic interactions, exert a considerable influence on protein adsorption, thereby contributing to the formation of the initial layer of the pellicle [177]. Subsequently, further interactions occur between the salivary proteins and the adsorbed proteins, resulting in the adsorption of additional proteins and macromolecules, including lipids and carbohydrates, onto the film. This results in the formation of a pellicle matrix comprising a dense basal layer and a globular outer layer [178, 179]. This layer is formed within a timeframe of 30 to 120 minutes [180]. As time progresses, the continually expanding outer layer forms a complex ecosystem that provides numerous binding sites for microorganisms to adhere directly, thus facilitating the development of dental bacterial biofilms [181]. These biofilms can contain over 700 distinct bacterial species [182].In unfavourable conditions, these bacteria may become pathogenic, leading to dental diseases such as caries, gingivitis and periodontitis [183]. Consequently, an in-depth understanding of biofilm formation is crucial for the prevention of dental diseases.

The availability of human teeth for the purpose of understanding biofilms and conducting further research is severely restricted. Hence, dental studies are predominantly done on other alternative biomaterials. In recent decades, bovine enamel has been the preferred substitute for human enamel due to its accessibility and similarity to the latter's microstructure [184].It would therefore be optimal to utilise a well-defined, standardised material that closely resembles human enamel surfaces in order to facilitate the undertaking of dental preventive research. The composition of human enamel is 95% carbonated hydroxyapatite (HAP), a hard mineral that is tightly organised in a high-density structure [166]. The HAP pellets, which are prepared through pressureless sintering of HAP powder, have been investigated as a potential standardised enamel-like material due to their highly reproducible properties, which are conducive to the investigation of dental biofilms [185]. In recent years, numerous studies have been conducted to compare HAP with human enamel. In order to gain insight into the enamel microstructure and the anti-wear mechanism, a microtribology comparative study was conducted utilising the nano-scratch technique [186]. The fluoride uptake and the ability to withstand acid attacks were compared on enamel and HAP surfaces [187, 188]. Additionally, a comparative study on the adhesion properties of microorganisms, particularly *Staphylococcus aureus*, over the biofilms of HAP

and enamel was conducted [189]. Nevertheless, a comprehensive investigation of the entire process of pellicle formation, conducted in direct comparison with enamel under physiological conditions, has yet to be undertaken. In this study, we conducted a systematic comparative analysis of the pellicle formation on enamel and HAP, with the aim of determining whether HAP can be considered a suitable standardised substitute for enamel.

Dental caries is a multifactorial disease, the aetiology of which is the result of the complex interplay between dietary habits, host buccal hygiene, and the microbial flora present in the oral cavity. Dental caries is a prevalent disease affecting both children and adults. Adverse conditions may also result in chronic effects [190]. The disease is caused by a specific species of bacteria that colonise the enamel surface and overproduce acidic and proteolytic products, which lead to demineralisation and digestion of the surface organic matrix [191]. Nevertheless, a closer examination reveals that there is still much to be discovered about the various aspects of caries formation, including the composition, function and erosion properties of their proteinaceous structures. It is established that the adsorption of salivary proteins facilitates caries accumulation and lesion formation; [192] however, further study on the adhesion, surface interactions and protein components would undoubtedly contribute to a deeper understanding of the demineralisation processes and the development of more effective preventive measures.

4.2 Materials and method

4.2.1 Datasets

Proteins adsorbed on HAP vs enamel analysis

A total of five consenting healthy human volunteers (male and female) were recruited for the study. They were required to be free from gingivitis, caries, periodontal diseases, and any other potential dental diseases that could affect the composition of oral fluid. The oral pellicles formed in situ on HAP and enamel surfaces were collected from all the volunteers. The protocols for collecting the pellicles were approved by the ethics committee of the Medical Association of Saarland, Germany.

Enamel slabs were prepared from a slit-like space on the ventral surface of bovine incisors. The surfaces were polished and purified prior to oral exposure. Hydroxyapatite (HAP) pellets were prepared via pressureless sintering [185] of HAP powder compressed in a stainless steel mould. These pellets were also polished and purified in a similar manner to the enamel surfaces prior to oral exposure.

Two hours prior to the implantation of the enamel slab and HAP pellets, all volunteers were instructed to brush their teeth and rinse their oral cavities. In order to retrieve the data pertaining to the pellicle proteins adsorbed at 5 seconds and 3 minutes, the enamel slabs and the HAP pellets were placed, in a similar manner to previous studies [193], in the molar regions of the lower jaw. To obtain the pellicle protein data at 2, 24 and 48 hours, the enamel slabs and

HAP pellets were mounted on silicon splints and exposed intraorally. During the oral exposure period, the volunteers were refrained from using any oral hygiene or cleaning measures, including toothpaste or other chemical agents. Subsequently, the splints were removed and stored in a moist environment outside of the mouth during mealtimes.

The pellicles that had formed on the surfaces of the enamel slabs and HAP pellets were subjected to a series of chemical processes, including elution, precipitation, and electrophoresis on a NuPAGE Bis-Tris gel. The peptides from the gel were isolated and subjected to identification and quantitation using nano-mass spectrometry.

Proteins in saliva vs proteins adsorbed on pellicle analysis

A total of five consenting healthy non-smoking human subjects (male and female, aged between 32 and 47 years) were recruited for the study. As mentioned before, it was imperative that the subjects be free from any potential dental diseases, such as active caries, gingivitis, and periodontal disease, which were examined by experienced dentists to ensure that the composition of oral fluid was not affected. Furthermore, the subjects were selected to have not consumed any antibiotics, antimicrobial agents, or anti-inflammatory drugs, nor undergone radiotherapy, within the previous six months. Saliva samples and enamel pellicles formed on the surfaces of the volunteers' teeth were collected at different time points. In this case, the time points were 10 seconds, 3 minutes and 30 minutes. The initial pellicle is the subject of study in order to gain insight into the functions and processes of interactions between the enamel and the biomolecules or microorganisms present in oral fluids. The subsequent formation of pellicles at the 30-minute time point represents the focus of the study, with the objective of understanding the interactions and adsorption or desorption patterns of the oral fluid particles in relation to the initial biofilm of the pellicle. The methodology for the collection of pellicles was approved by the ethics committee of the Medical Association of Saarland, Germany.

The enamel slabs were prepared from bovine incisors, specifically from the labial surfaces. The surfaces were polished in a stepwise manner by means of wet grinding with abrasive paper in order to increase the grit size, purified, washed and rehydrated 12 hours prior to intraoral exposure. The enamel slabs were positioned, in a similar manner to previous studies [193], within the lower jaw, specifically in the premolar and molar teeth regions, for a period of 10 seconds, 3 minutes, and 30 minutes, respectively, to allow for pellicle formation.

In order to circumvent the potential circadian effects of salivary composition, the experiments commenced at 9:00 am. Two and a half hours prior to the commencement of the experiment, the subjects were instructed to refrain from consuming food and beverages, and to perform oral hygiene in accordance with standard practices. This was done to prevent any potential influence on the samples and to minimise subject variability. Prior to the collection and intraoral exposure of samples, the oral hygiene procedure was

followed using dental silk tooth brushing without the use of toothpaste for a period of 30 minutes.

Following the formation of the pellicle at varying time points, the slabs were individually rinsed with water and then air-dried. The proteins were eluted from the slabs by successive ultrasonication in urea/CHAPS buffer at 4°C, followed by incubation in Triton X-100 and subsequent ultrasonication in RIPA buffer at 4°C. The eluted proteins were precipitated, washed, air-dried and denatured for gel electrophoresis, after which they were subjected to identification and quantification. For saliva sample collection, 10ml of unstimulated saliva was collected on ice over 20 minutes. Subsequently, 90g of each sample were denatured for gel electrophoresis and subjected to identification and quantification.

Proteins in pellicle and saliva from active caries, treated caries and healthy conditions

A total of twenty-eight human subjects, comprising both male and female participants aged between four and six years old, were recruited for the study. Of the 28 subjects, 11 exhibited active caries, 9 had undergone treatment for caries, and 8 showed no history of caries. The subjects were evaluated by experienced dentists and the DMFT (Decayed, Missing due to caries, and Filled Teeth) classifcation of the permanent teeth was carried out to classifiy the characteristics of the subjects. The pellicles that formed on the teeth surfaces of the subjects were isolated and subjected to further investigation. The ceramic slabs, each measuring 8 cm^2, were prepared and mounted in situ on a holder. These slabs were then exposed to the oral cavity for a period of three minutes to allow for pellicle formation, after which they were removed. The methodology for the collection of the pellicles was approved by the ethics committee of the Medical Association of Saarland, Germany.

To eliminate the potential for circadian effects on salivary composition, the experiments commenced at 9:00 am, analogous to the previous studies [193]. Prior to the commencement of the protocol, the subjects were instructed to perform oral hygiene without the use of toothpaste. To eliminate the possibility of the paste exerting an influence on the samples. Furthermore, the subjects were instructed to refrain from consuming food for a period of two hours prior to the commencement of the protocol. This would serve to reduce the degree of subject-specific variability in the data.

Protein abundances

A nano-MS/MS analysis can readily identify the proteins and provide a range of quantitative data that illustrate the abundance of proteins. These include scores, hit ranks and the number of peptides per protein. Nevertheless, the mass spectrometry method is not without its limitations, particularly in regard to ion suppression effects. The efficacy of droplet formation and evaporation is susceptible to alteration by the presence of less volatile compounds. This affects the abundance of charged ions in the gaseous phase, thereby influencing

the detection of protein abundance [194]. It is therefore important to normalise the protein abundance parameters. One such normalisation method is the protein abundance index (PAI), in which the number of peptides per protein is normalised with the theoretical amount of peptides of a protein [195, 196].

$$PAI = \frac{N\_observed}{N\_theoretical}$$

Where the $N\_observed$ refers to the number of peptides per protein observed and $N\_theoretical$ refers to the theoretical number of peptides of a protein.

To refine the absolute quantification and incorporate the molar fraction expression, rather than merely relative abundance between proteins, the exponentially modified PAI (emPAI) was developed. The emPAI values are proportional to the protein content in the mixture. This is estimated by the following formula [197]:

$$emPAI = 10^{PAI} - 1$$

In this study, the emPAI values were used to estimate the percentage of protein content in molar fractions using:

$$Protein\ content\ (mole\%) = \frac{emPAI}{\sum(emPAI)} \times 100$$

Another measure of protein abundance employed in the study is referred to as spectral abundance. It represents the number of spectra identified by the spectrometer for a specific peptide or protein [198].

The two abundance values, emPAI and spectral counts were employed for qualitative and quantitative analysis, including similarity testing, fold change estimation, and the investigation of physio-chemical properties.

4.2.2 Bioinformatic analysis

Qualitative analysis

The Pandas and Numpy Python libraries were employed for the purpose of sorting and structuring the proteomics data generated by nano-mass spectrometry [199, 200]. The samples were classified and distinguished using these libraries according to the analyses. In a group or volunteer, a protein is considered to be expressed if its emPAI value or spectral count is above zero. Additionally, the Venn Python library was used to create Venn diagrams for visualising the grouped clusters. This diagram facilitated comprehension of the number of shared proteins amongst various categories following analysis.

Statistical analysis

In order to ascertain whether the protein abundance data of the samples can be assumed to follow a normal distribution, thus rendering the computations more straightforward and universally applicable, the Shapiro-Wilk test [85] was employed with the null hypothesis that the data is normally distributed.

The results of the Shapiro-Wilk test indicated that the data from all samples were not normally distributed. Consequently, to ascertain whether the replicate data from the samples were comparable, the Wilcoxon signed-rank [201] test was employed. To evaluate the similarity of the proteins adsorbed in all volunteers, Mann Whitney U test [202] was conducted.

Fold change analysis

Fold change analysis is a quantitative method that is utilised for the purpose of evaluating the relative alterations in protein abundance that occur between two distinct conditions or time points. The ratio of protein abundance values in the two conditions is used to estimate the fold change. For instance, in a study comparing protein adsorption between HAP and enamel, the fold change of a protein can be estimated by calculating the ratio of the protein abundance adsorbed on HAP to the abundance adsorbed on enamel.

$$Fold\ change\ = \frac{abundace\_a}{abundace\_b}$$

where $abundace\_a$ represents the protein abundance in condition a (for the above example in HAP) and $abundace\_b$ represents the protein abundance in condition b (in enamel).

If the $foldchange$ is greater than zero, it indicates that the protein in condition a is upregulated compared to condition b. Similarly, the $fold\ change$ is less than zero signifies that the protein in condition a is downregulated in comparison to condition b. In general, in the context of biological references, a $fold\ change$ of equal to or above two is considered to be significantly upregulated, where this signifies that the proteins have doubled in condition A as compared to condition B. Conversely, a $fold\ change$ of equal or less than 0.5 is considered to be significantly downregulated, which signifies that the proteins have halved in condition A as compared to condition B.

Given that these significance threshold values fall within a disproportionate range, it is more straightforward to utilise $logfoldchange$ values.

$$Log\ fold\ change\ = log(\frac{abundace\_a}{abundance\_b})$$

$$= log(abundance\_a) - log(abundance\_b)$$

The log transformation of the $fold\ change$ allows for the thresholding of significantly up- and downregulated values to be symmetrical. In other words, a $log\ fold\ change$ of 1 or greater indicates that the abundance of a protein in condition A is twice that of condition B and is therefore significantly upregulated. A $log\ fold\ change$ of less than or equal to -1 indicates that the protein abundance at condition A is half that of condition B, and thus represents a significant downregulation.

Accordingly, a $log\ fold\ change$ analysis was conducted in the present study to ascertain the discrepancy in protein abundance on disparate surfaces (HAP vs enamel), at varying time points of the pellicle (proteins adsorbed at 3 minutes vs 30 minutes) and under disparate conditions (active caries condition vs healthy conditions).

Molecular function enrichment analysis

The Ensembl BioMarts database [203] was employed to derive gene names and identifiers from UniProt accession numbers. The Gene Ontology Molecular Function terms for the proteins undergoing enrichment analysis were conducted via the following `http://geneontology.org/`. It is linked to the PANTHER classification system [125], which is a comprehensive, annotated library of gene families and protein-coding genes in the human and mouse genomes.

The enriched molecular functions represent a summary of the subset of proteins that are responsible for major molecular-level bio-processes, such as catalysis, transport, and cellular organisation. In this study, we observed the differences in enriched molecular functions between upregulated and downregulated proteins in different conditions, as well as between proteins expressed on different surfaces and at different time points.

Physiochemical properties

In order to gain further insight into the differences in protein adsorption under varying conditions, surfaces and time points, an investigation was conducted into the physicochemical properties of the proteins in question. The molecular weights of the proteins adsorbed were subjected to thorough analysis with the objective of identifying patterns of specific adsorption on surfaces, conditions or time points that correlate to molecular weights. Similarly, the patterns of adsorption based on the protein's isoelectric points (using the Proteome Isoelectric Point Database [204]) were also analysed to identify any differences. The significance of the observed differences was determined by the Mann-Whitney U test. The rank-biserial correlation [205] was used to determine the effect size and the direction of the differences.

4.3 Results and discussion

4.3.1 Proteins adsorbed on HAP and enamel

Qualitative analysis

The pellicles from the HAP and enamel surfaces were permitted to form for a period of three minutes. Subsequently, the proteome from both surfaces was extracted and considered. The combined proteome data from the two independent replicates of the oral exposure rounds on HAP and enamel were tabulated in Table 4.1 and 4.2 represented as a Venn diagram in Fig 4.1 and 4.2.

The term "diversity" is used to describe the number of proteins that have been identified on the material, at least in one of the volunteers. The term "overlap" refers to the number of proteins deposited on the material, which is a common occurrence across all subjects. The comparison of the common proteins in HAP and enamel under diversity and overlap conditions are represented as Venn diagrams in Fig 4.3.

	HAP				
Subjects	1	2	3	4	5
No. of proteins	283	98	216	371	337
Diversity			490		
Overlap			84		

Table 4.1: Number of proteins in the HAP biofilms from two independent replicates pooled of five subjects identified by protein mass spectrometry (nanoLC-ESI-MS2).

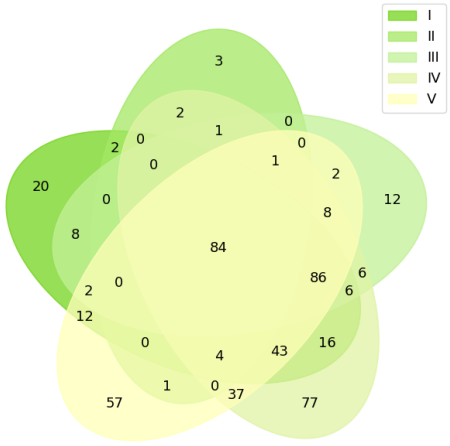


Figure 4.1: Number of proteins in HAP biofilm. Numbers inside the ellipses representing single subjects (I-V) indicate proteins identified within one or commonly within two to five subjects

	Enamel				
Subjects	1	2	3	4	5
No. of proteins	191	91	208	171	306
Diversity			381		
Overlap			71		

Table 4.2: Number of proteins in the enamel biofilms from two independent replicates pooled of five subjects identified by protein mass spectrometry (nanoLC-ESI-MS2).

A total of 567 distinct proteins were identified on both materials for all five subjects, with 490 identified on HAP and 381 on enamel. A total of 490 proteins were adsorbed on HAP surfaces by at least one of the volunteers, while 84 proteins were adsorbed by all volunteers. In the case of enamel surfaces, 381 proteins were adsorbed in at least one of the volunteers, while 71 proteins were adsorbed commonly in all volunteers. A total of 304 proteins were identified on both surfaces in at least one of the volunteers of each group. 61 proteins were identified on both surfaces and commonly in all volunteers.

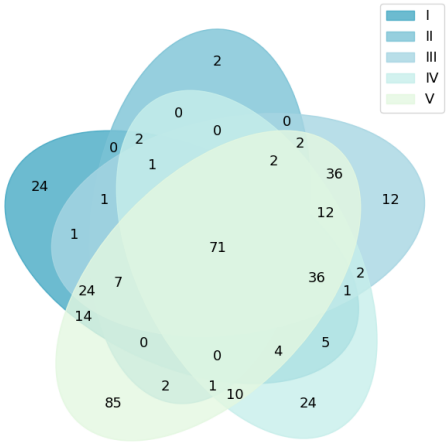


Figure 4.2: Number of proteins in enamel biofilm. Numbers inside the ellipses representing single subjects (I-V) indicate proteins identified within one or commonly within two to five subjects

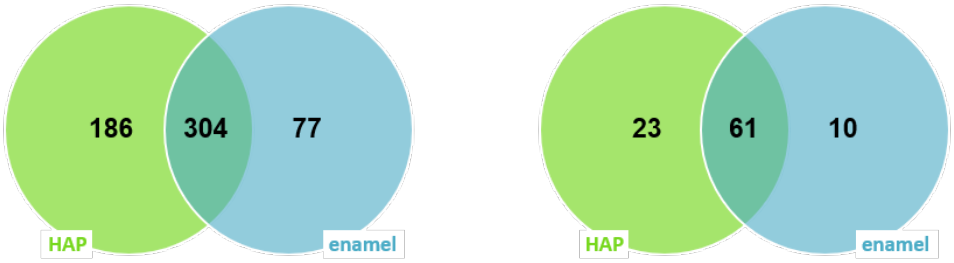
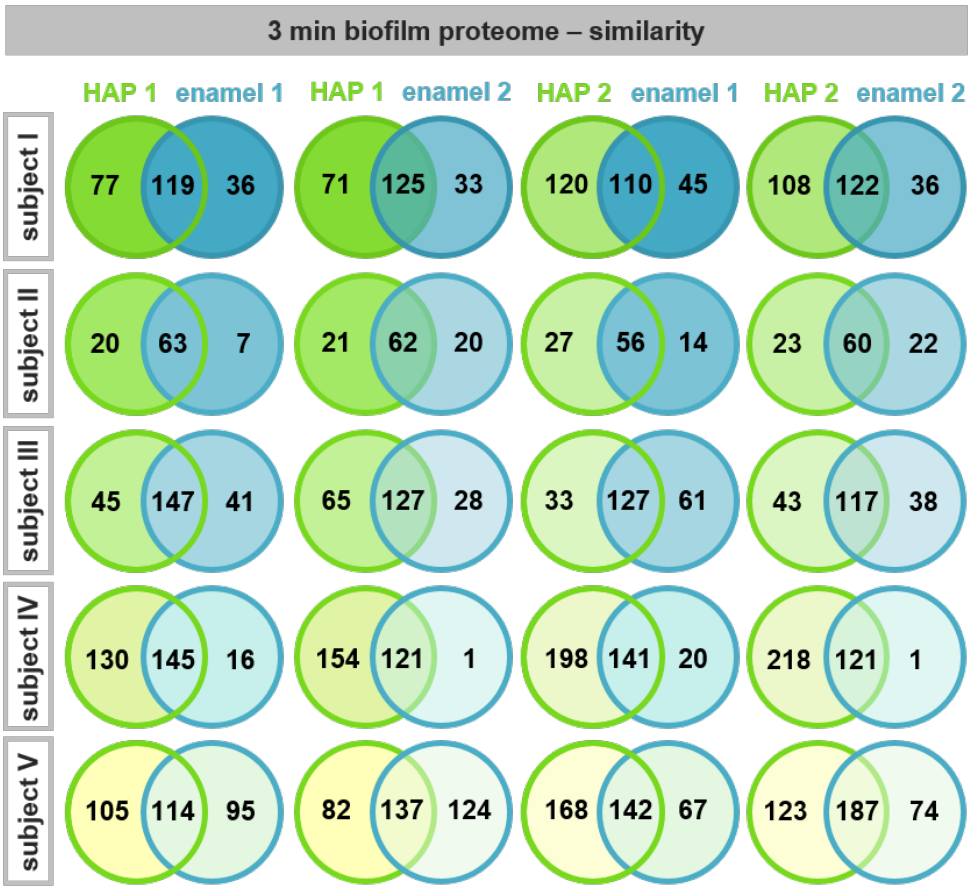


Figure 4.3: Total amount of identified individual proteins (diversity) in the biofilms of at least one volunteer formed on HAP, enamel and both materials (left). The number of identified common proteins (overlap) in the biofilms of all subjects formed on HAP, enamel and both materials (right.

Statistical analysis

The Wilcoxon signed-rank test was used to characterize the level of similarity of proteins adsorbed on HAP and enamel. We applied the null hypothesis that the number of shared proteins between the two HAP probes (or between two enamel probes) is equal to the number of shared proteins between one HAP and one enamel probe. Then, the number of proteins shared between HAP1 and HAP2 (or enamel 1 and enamel 2) shown in the Venn diagrams was compared to the number of proteins shared between HAP and enamel. The p-value generated by the test determines whether or not the null hypothesis should be rejected, based on a threshold of 0.05. A representation of this analysis can be found in Fig 4.4



all subjects	p-values			
	HAP 1-enamel 1	HAP 1-enamel 2	HAP 2-enamel 1	HAP 2-enamel 2
	compared to HAP 1-HAP 2			
	0.1875	0.0625	0.0625	0.0625
	compared to enamel 1-enamel 2			
	1	1	0.625	0.715

Figure 4.4: Similarity of proteins commonly found in 3 min biofilms formed on HAP and enamel. Wilcoxon signed-rank tests were used to compare the intersection values between HAP 1 and HAP 2 (or enamel 1 and enamel 2) with the intersection values of HAP and enamel replicates

4.3.2 Proteins in saliva vs proteins adsorbed on pellicle analysis

Quantitative analysis

The objective of this analysis was to identify the patterns of protein adsorption on enamel and protein retention in saliva at the individual volunteer level. Therefore, the pellicle formed on the enamel at different time points (10 seconds, 3 minutes and 30 minutes) and the corresponding saliva samples were analysed and compared. The combined proteome data from the pellicle of different time points and saliva are presented in Table 4.3. In this table, the proteins are considered to be adsorbed if at least one of the replicates of a given volunteer has a protein abundance value above zero.

		Vol 1	Vol 2	Vol 3	Vol 4	Vol 5	Diversity	Overlap
Saliva		377	776	315	819	598	1055	221
Pellicle	10-s	342	624	376	841	527	1108	197
	3-min	450	697	481	768	561	1080	274
	30-min	553	808	607	659	520	1161	286
Overlap time-points pellicle		286	477	287	460	402	770	174
Overlap saliva and pellicle		179	358	173	322	293	553	108
Diversity pellicle		640	986	728	1124	672	1539	339
Diversity saliva and pellicle		767	1237	823	1409	875	1816	778

Table 4.3: Number of identified proteins and overlaps in the individual pellicles at 10 seconds, 3 minutes and 30 minutes on pellicle and saliva

In total, 1,055 proteins were eluted and expressed in at least one of the volunteer readings. Furthermore, 221 proteins were eluted from all volunteers in saliva. A total of 1,539 proteins were adsorbed on the pellicles in at least one of the volunteer readings at all time points, while 174 proteins were eluted from all volunteers at all time points. The initial biofilm formation process, occurring within a 10-second timeframe, resulted in the adsorption of 197 proteins across all volunteers. Similarly, 274 and 286 proteins were identified as adsorbed on the pellicle formed after 3 and 30 minutes, respectively, with a high degree of consistency across all volunteers. A comparison of the 174 proteins that were expressed in all volunteer and time points pellicle readings and the saliva proteins of all volunteers revealed that 108 of the former were also present in the latter.

Similarity test

In this analysis, the hypothesis that the adsorption patterns of proteins were similar between the replicates of volunteers in saliva and pellicle data was tested. The Shapiro-Wilk test was employed to ascertain whether the data for each sample exhibited a normal distribution. The results demonstrated that none of the samples exhibited a normal distribution. Consequently, the Wilcoxon signed-rank test was employed to ascertain whether a discrepancy existed in the distribution of the replicate data of volunteers. The null hypothesis of this test states that there is no difference between the distributions of the replicate data.

It should be noted that not all volunteers had replicate data available for analysis. Only volunteers 2 to 4 had this information for the saliva data set, while volunteers 2 and 3 had this information for the pellicle data set. The Tables 4.4 and 4.5 illustrate the outcomes of the Wilcoxon signed-rank test, accompanied by P-values and an interpretation for the data pertaining to saliva and pellicle, respectively.

Samples being tested			P-value	Inference
Volunteer 2	Vs	Volunteer 2 replicate	4.4697×10^{-13}	The two samples do not follow a similar distribution of protein abundance
Volunteer 3	Vs	Volunteer 3 replicate	6.2328×10^{-9}	The two samples do not follow a similar distribution of protein abundance
Volunteer 4	Vs	Volunteer 4 replicate	1.0756×10^{-44}	The two samples do not follow a similar distribution of protein abundance
Volunteer 4	Vs	Volunteer 5 replicate	1.2467×10^{-38}	The two samples do not follow a similar distribution of protein abundance

Table 4.4: Results from Wilcoxon signed-rank test to test the similarity of replicates from proteome data collected from saliva

Time point: 10-s				
Samples being tested			P-value	Inference
Volunteer 2	Vs	Volunteer 2 replicate	1.0952×10^{-15}	The two samples do not follow a similar distribution of protein abundance
Volunteer 3	Vs	Volunteer 3 replicate	0.00344	The two samples do not follow a similar distribution of protein abundance
Time point: 3-min				
Samples being tested			P-value	Inference
Volunteer 2	Vs	Volunteer 2 replicate	0.1269	The two samples could potentially follow a similar distribution of protein abundance
Volunteer 3	Vs	Volunteer 3 replicate	0.0008	The two samples do not follow a similar distribution of protein abundance
Time point: 30-min				
Samples being tested			P-value	Inference
Volunteer 2	Vs	Volunteer 2 replicate	3.6187×10^{-16}	The two samples do not follow a similar distribution of protein abundance
Volunteer 3	Vs	Volunteer 3 replicate	0.0108	The two samples do not follow a similar distribution of protein abundance

Table 4.5: Results from Wilcoxon signed-rank test to test the similarity of replicates from proteome data collected from pellicle

In proteome data obtained from saliva samples, all p-values are less

than 0.05, indicating that the null hypothesis is rejected. It can therefore be concluded that the replicates are not similar. In the proteome data from the pellicle, only at time point 3 minutes is there no evidence to reject the null hypothesis between the replicates of volunteer 2. In contrast, the remaining replicates at all other time points provide evidence to reject the null hypothesis, indicating that they do not follow a similar protein abundance distribution.

In order to ascertain the degree of similarity in protein expression between volunteers, a Mann-Whitney U test is employed in the analysis of proteome data derived from saliva and pellicle. This test is employed due to the fact that the preceding analysis demonstrated that the data set in question does not adhere to a normal distribution. The results of the Mann-Whitney U test are presented in tabular form in Tables 4.6 and 4.7 to 4.9 for the saliva and pellicle data sets, respectively, together with the associated P-values and interpretation.

	Vol I	Vol II	Vol III	Vol IV
Vol I				
Vol II	6.1078×10^{-7}			
Vol III	4.1898×10^{-7}	5.4543×10^{-11}		
Vol IV	1.1982×10^{-9}	0.6884	3.5209×10^{-14}	
Vol V	0.0028	0.0003	1.0856×10^{-9}	9.6848×10^{-6}

Table 4.6: P-values from Mann Whitney U test to test the similarity between volunteers from proteome data collected from saliva

10- sec	Vol I	Vol II	Vol III	Vol IV
Vol I				
Vol II	2.2391×10^{-7}			
Vol III	0.5101	5.2181×10^{-5}		
Vol IV	1.5826×10^{-15}	$1.7592 \times 10-5$	$1.2855 \times 10-12$	
Vol V	$4.76414 \times 10-7$	0.4794	0.0008	$5.0302 \times 10-6$

Table 4.7: P-values from Mann Whitney U test to test the similarity between volunteers from proteome data collected from pellicle eluted after 10 seconds

3-min	Vol I	Vol II	Vol III	Vol IV
Vol I				
Vol II	0.0682			
Vol III	0.1286	7.2016×10^{-5}		
Vol IV	8.3623×10^{-6}	0.0011	6.1015×10^{-12}	
Vol V	0.0005	0.1967	1.9160×10^{-6}	0.0699

Table 4.8: P-values from Mann Whitney U test to test the similarity between volunteers from proteome data collected from pellicle eluted after 3 minutes

The proteome data obtained from saliva samples demonstrated a comparable distribution of protein abundance in Volunteers 2 and 4. The remaining

30-min	Vol I	Vol II	Vol III	Vol IV
Vol I				
Vol II	$7.7875 \text{ X } 10^{-5}$			
Vol III	0.1836	$1.7886 \text{ X } 10^{-5}$		
Vol IV	0.1185	0.0141	0.0161	
Vol V	0.1760	0.0531	0.3392	0.3346

Table 4.9: P-values from Mann Whitney U test to test the similarity between volunteers from proteome data collected from pellicle eluted after 30 minutes

comparison was tested with a P-value less than 0.05, thereby rejecting the null hypothesis that the distributions are similar.

In the proteome collected from the pellicle eluted at 10 seconds, there was a notable similarity between the profiles of Volunteer 1 and Volunteer 3, while Volunteer 5 exhibited a resemblance to Volunteer 2. With regard to the pellicle eluted at 3 minutes, volunteer 1 exhibited similarities to Volunteers 2 and 3, while Volunteer 5 displayed similarities to Volunteers 2 and 4. With regard to the pellicle eluted at 30 minutes, volunteer 5 exhibited similarities to all the other volunteers. Similarly, volunteer 1 displayed similarities to volunteers 3 and 4. The remaining comparison yielded P-values less than 0.05, thereby rejecting the null hypothesis that the distributions are similar.

Fold change analysis

A total of 553 proteins exhibited differential abundance in saliva and pellicle, out of the 778 proteins that were common between the proteome data from saliva and pellicle and adsorbed in at least one of the volunteers. Of these, 90 proteins exhibited significant differential expression (i.e., logarithmic fold change value ≥ 1, logarithmic fold change value $\leq$ -1, and p-value < 0.05) in all time points of pellicle proteome with respect to saliva proteome data (illustrated in Fig. 4.5).

The term "upregulated" is used to describe proteins that are expressed at higher levels in the proteome data of pellicles in comparison to the proteome data of saliva. Similarly, the term "downregulated" is used to describe the proteins that are expressed in lower abundances in the proteome data of pellicles as compared to that of saliva. The significance of the differential expression is evaluated through the application of the F-test, which compares the entire set of volunteer data from the pellicle against the data from saliva. A comparison of the proteome data of the pellicle eluted at all time points with the saliva data revealed that 32 proteins were significantly upregulated and 58 proteins were significantly downregulated in saliva.

The differential abundance of the 553 common proteins in the pellicle eluted at different times was estimated. Fig 4.6 illustrates the differential abundance of the 553 proteins present in the pellicle at different time points.

In this context, the term "upregulated" is employed to describe the proteins that are adsorbed at a higher level at a subsequent time point in

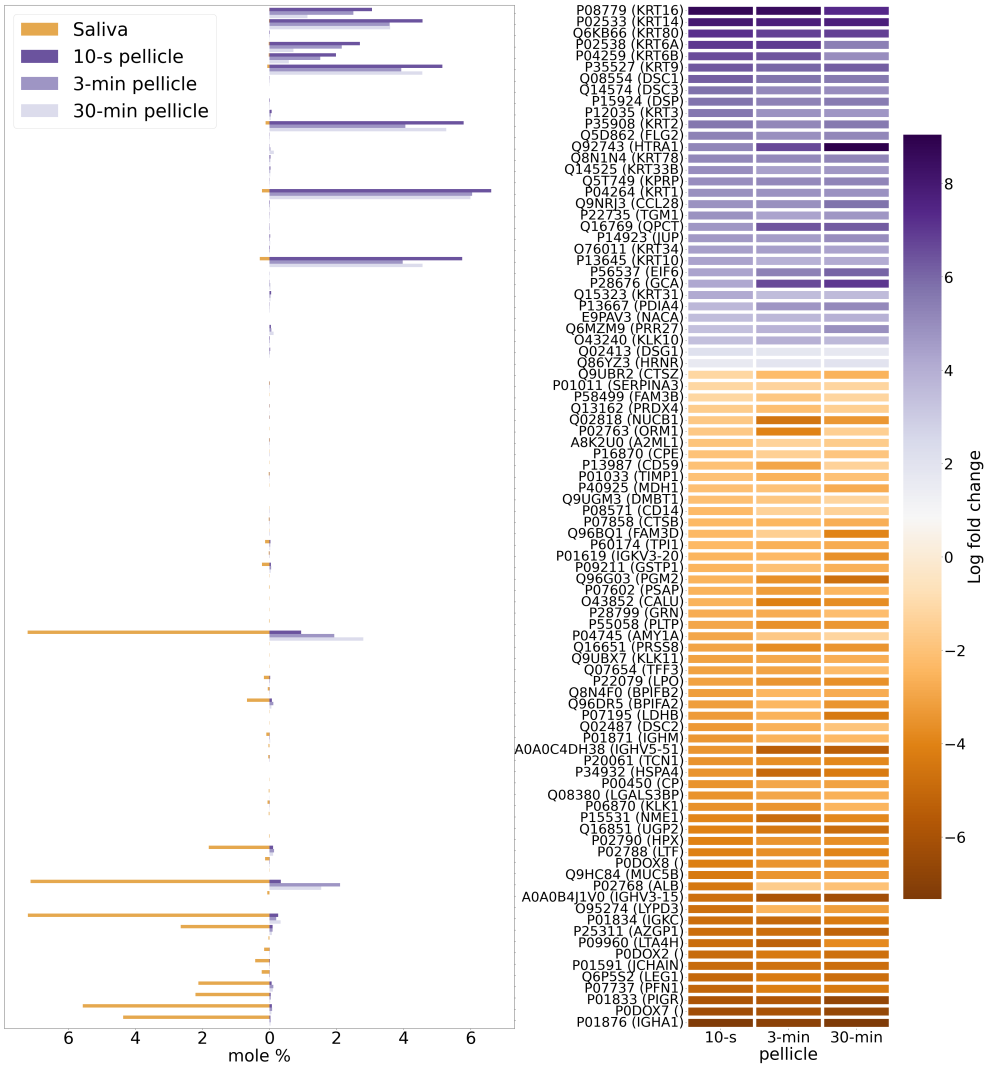


Figure 4.5: Bar plot and heatmap representation of the 90 proteins showing significantly different abundance in saliva and pellicle at 10 seconds, 3 minutes and 30 minutes.

comparison to the proteins adsorbed at an earlier time point. Similarly, the term "downregulated" is employed to describe the proteins that are adsorbed at a lower level at a later time point in comparison to the proteins adsorbed at an earlier time point. To illustrate, in the left subplot of Fig. 4.6, the upregulated proteins are those adsorbed at higher levels in the pellicle eluted at 3 minutes with respect to the pellicle eluted at 10 seconds.

A comparison of the protein abundance in the pellicle eluted at 3 minutes and 10 seconds revealed that 9 proteins exhibited a significant upregulation, while none demonstrated a downregulation. In the comparison between the protein abundances in the pellicle eluted at 30 minutes and 10 seconds, 27 proteins demonstrated increased expression, while 1 protein exhibited decreased expression. Similarly, in the comparison between the protein abun-

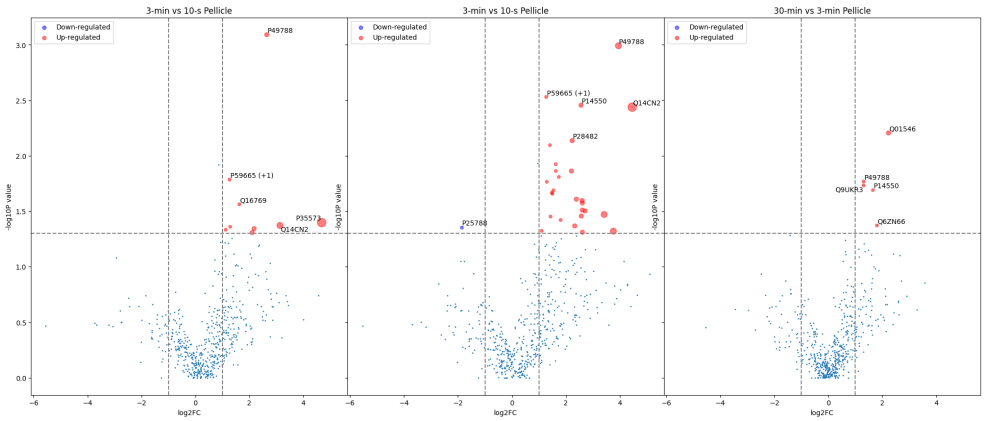


Figure 4.6: Volcano plot representation of the 553 differentially abundant proteins in pellicle at different time points. (left) depicts the differential abundance of proteins adsorbed at 3 minutes in comparison to 10 seconds. (middle) presents the differential abundance of proteins adsorbed at 30 minutes in comparison to 10 seconds. (right) illustrates the differential abundance of proteins adsorbed at 30 minutes in comparison to 3 minutes.

dances in the pellicle eluted at 30 minutes and 3 minutes, 5 proteins exhibited increased expression, while no proteins exhibited decreased expression.

Molecular function enrichment analysis

In order to gain insight into the underlying mechanisms governing the selective adsorption of proteins to the pellicle and their subsequent retention in saliva, a functional analysis is conducted. In order to achieve this, the proteins that are exclusively present in saliva and absent from the pellicle in any given volunteer are estimated. Similarly, the proteins that are only adsorbed in the pellicle and not expressed in saliva are identified. The proportion of these proteins is illustrated in a Venn diagram in Fig. 4.7

A total of 778 proteins were identified in at least one volunteer and at least one time point of the eluted pellicle. However, none of these proteins were expressed in saliva. Similarly, 277 proteins were identified in at least one of the volunteers' saliva samples, but none were identified in the proteome pellicle data. A total of 761 proteins were identified as being expressed in both the saliva and pellicle data sets for at least one of the volunteers.

The 761 and 277 proteins, which were exclusively identified in the pellicle and saliva, respectively, were subsequently subjected to molecular function enrichment analysis. The results of this analysis demonstrate the functions of these proteins, thereby elucidating why the salivary protein functions are not required in the pellicle and identifying the essential functions that are necessary for adsorption on surfaces. The top 20 enriched molecular function terms of salivary and pellicle proteins are depicted in Figs 4.8 and 4.9.

The most highly enriched molecular function term was 'structural

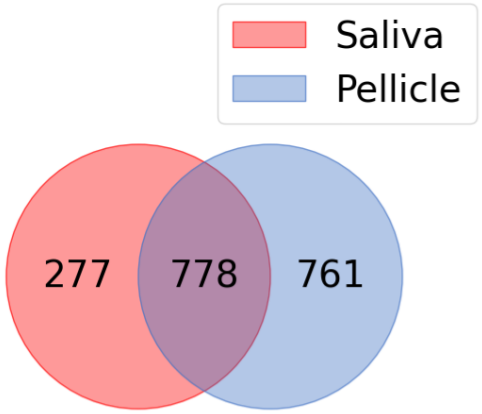


Figure 4.7: Venn diagram illustrates the number of proteins identified in saliva, pellicles, and those present in both saliva and pellicles in at least one volunteer.

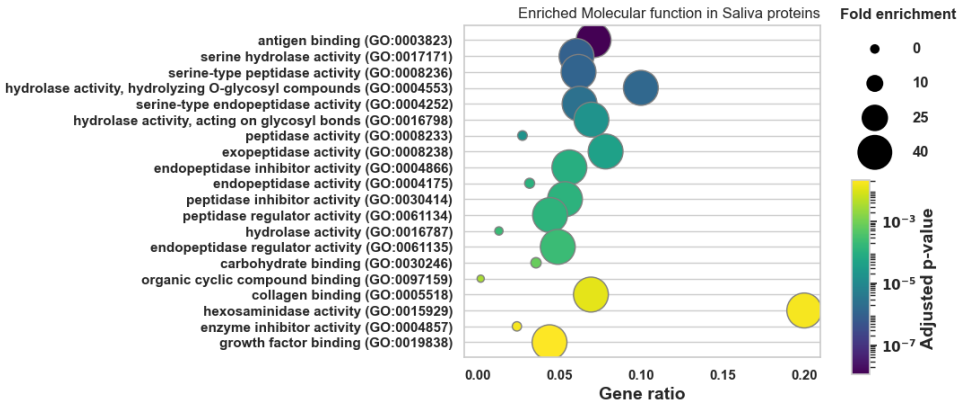


Figure 4.8: Representation of the top 20 enriched molecular functions for the 277 salivary proteins.

molecule activity', which is defined in QuickGO [206] as an action of a molecule that works on the structural integrity of a complex. This was a crucial function for the stable formation of pellicles on oral surfaces. The other enriched molecular functions, such as cadherin binding and cell adhesion molecule binding, also indicate a role in maintaining the structural integrity of the biofilm. In contrast, a multitude of catalytic and protein processing activities were observed in saliva. It would appear that minimal involvement was observed with regard to complex formation and structural management.

In order to gain further insight into the distinction between the molecular functions observed in saliva and the pellicle, a distribution of the parent molecular function terms was plotted, with the objective of determining the percentage of genes that have been annotated to a given term in relation to the total number of function hits. This would explain the proportion of proteins involved in each molecular function, thereby allowing the variation in importance attributed to each molecular function category to be concluded.

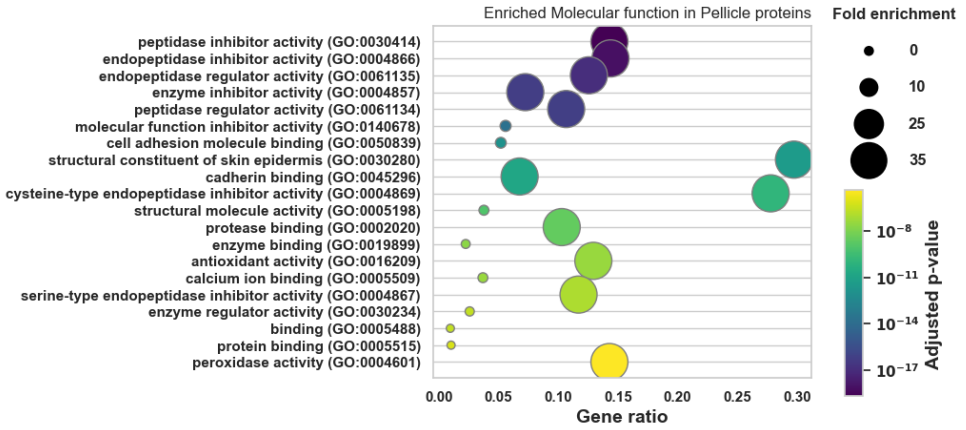


Figure 4.9: Representation of the top 20 enriched molecular functions for the 761 pellicle proteins.

Fig. 4.10 presents a graphical representation of the importance assigned to each molecular function category in saliva and the pellicle.

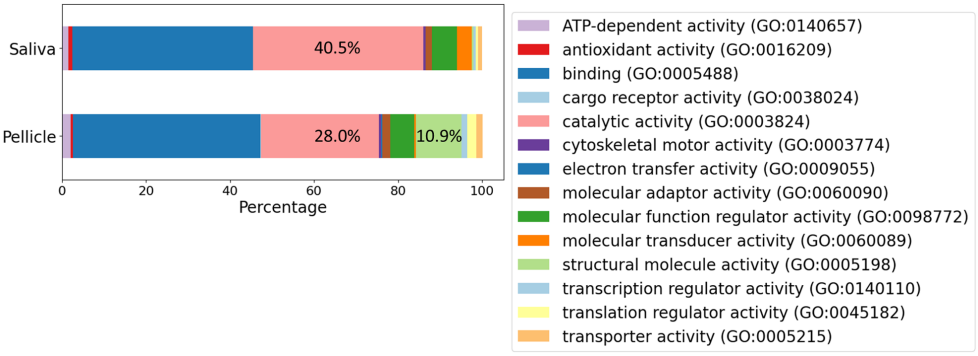


Figure 4.10: Distribution of percentage of molecular function hit in pellicle and saliva

It can be observed that proteins in saliva exhibit a greater propensity for molecular functions pertaining to catalytic activity in comparison to those present in the pellicle. In contrast, the proteins in the pellicle are more inclined to engage in structural molecule activity than those found in saliva.

Physiochemical properties

In order to gain a deeper understanding of the factors that contribute to the inability of certain proteins to adsorb onto biofilms developed on oral surfaces, an investigation was conducted into the physicochemical characteristics of the 304 proteins exclusively present in saliva, in addition to those of the 746 proteins that are exclusively present in pellicles. Firstly, the distribution of molecular weights of these proteins present in saliva and pellicle was subjected to a comprehensive and systematic analysis. It was observed that proteins with

a molecular weight of less than 100 kDa were more prevalent in the proteome data of saliva and pellicle. The Mann-Whitney U test also demonstrated that the overall distribution of salivary proteins was significantly smaller than that of pellicle proteins. The rank-biserial correlation coefficient revealed that the distribution of pellicle proteins is 0.354 times higher than that of salivary proteins. Fig. 4.11 illustrates the distribution of salivary and pellicle proteins based on their respective molecular weights.

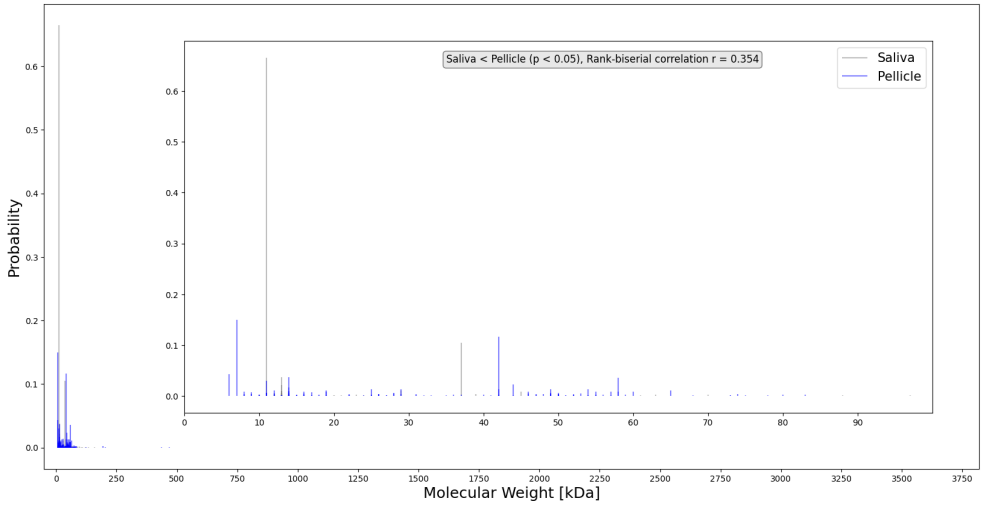


Figure 4.11: Distribution of salivary and pellicle proteins based on their respective molecular weights

Finally, the isoelectric points of the salivary and pellicle proteins were investigated. The proteins were classified into three groups based on their isoelectric point ranges. The salivary proteins were categorised into three groups based on their isoelectric points: 1) negatively charged proteins in the oral cavity (pH below 6.3), 2) neutral proteins (pH between 6.3 and 7.6, which corresponds to the physiological pH range of the oral cavity), and 3) positively charged proteins in the oral cavity (pH above 7.6). The proportion of proteins identified within each category is illustrated in Fig. 4.12.

As anticipated, the majority of salivary proteins were identified within the pH range characteristic of the oral cavity. The proportion of salivary proteins found below the oral pH range was 16.9%, while 8.3% were found above this range. In the case of pellicle proteins, the majority were found to be present at pH values above those typically observed in the oral cavity. A total of 33.3% of the pellicle proteins were identified within the pH range below that of the oral cavity, while only 16.7% were found within the pH range corresponding to that of the oral cavity. This suggests that the adsorption of positively charged proteins was favoured for the formation of the pellicle.

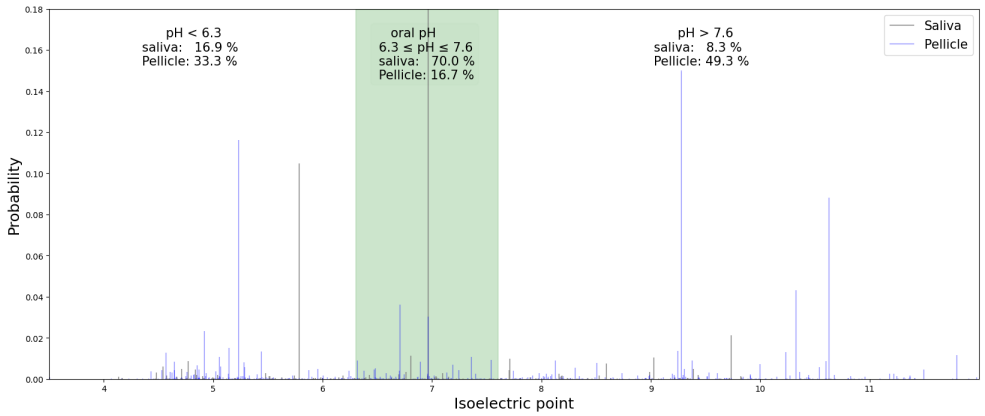


Figure 4.12: Distribution of salivary and pellicle proteins based on their respective isoelectric points

4.3.3 Proteins in pellicle and saliva from active caries, treated caries and healthy conditions

Qualitative analysis

The objective of this study was to identify the patterns of protein adsorption in pellicle and protein expressed in saliva in individuals with active caries, no caries (healthy) and treated caries conditions. This was done on 8 to 11 volunteers. Therefore, the pellicle formed under the various conditions was analysed and compared to the corresponding saliva samples. The overall statistics of the protein expression in the pellicle and saliva of the different conditions are presented in Table 4.10. The proteins included in the table exhibited an abundance value above zero in the samples.

	Pellicle active	Pellicle non-active	Pellicle treated	Saliva active	Saliva non-active	Saliva treated
No. of proteins	279	523	567	823	714	636
Diversity of proteins		694			1001	
Overlap of proteins		26			85	
Diversity of proteins found only in the group	19	91	147	197	91	47
Proteins found only in the group in all volunteers	0	3	18	72	7	3

Table 4.10: Overall statistics of the number of identified, diversity and overlap of proteins in pellicle and saliva samples from independent volunteers with active caries, no caries (healthy) and treated caries conditions

A total of 1,001 proteins were identified in the proteome data of saliva from at least one of the volunteers under all experimental conditions. Moreover, 694 proteins were adsorbed in the pellicle in at least one of the volunteers under all conditions. A total of 823 proteins were identified in the saliva samples from the volunteers with active caries, with 72 proteins being present

in all volunteers. Similarly, 714 and 636 proteins were expressed in saliva under non-active and treated caries conditions, respectively. Of these, 7 and 3 proteins were expressed in all volunteers in the respective conditions. Similarly, in the pellicle proteome data, 279 proteins were adsorbed on the pellicle under active caries conditions, yet none of them were found to be present in all the volunteers. Similarly, 523 and 567 proteins were identified in the pellicle under non-active and treated caries conditions, respectively. Of these, 3 and 18 proteins were identified as common to all volunteers in the respective conditions.

Log fold change analysis

A total of 506 out of the 1,001 proteins were expressed in all conditions in at least one of the volunteers of each group, as determined by analysis of saliva samples. Similarly, 238 out of 694 proteins were expressed in all conditions in at least one of the volunteers of each group adsorbed in the pellicle. A fold change analysis was conducted using the 506 common proteins in saliva data to elucidate the differences in protein abundances across various conditions. Similarly, the 238 common pellicle proteins were employed for fold change analysis, with a view to examining the change in protein abundances in different conditions. Fig. 4.13 and 4.14 illustrate the differential abundance of the common saliva and pellicle proteins in different conditions.

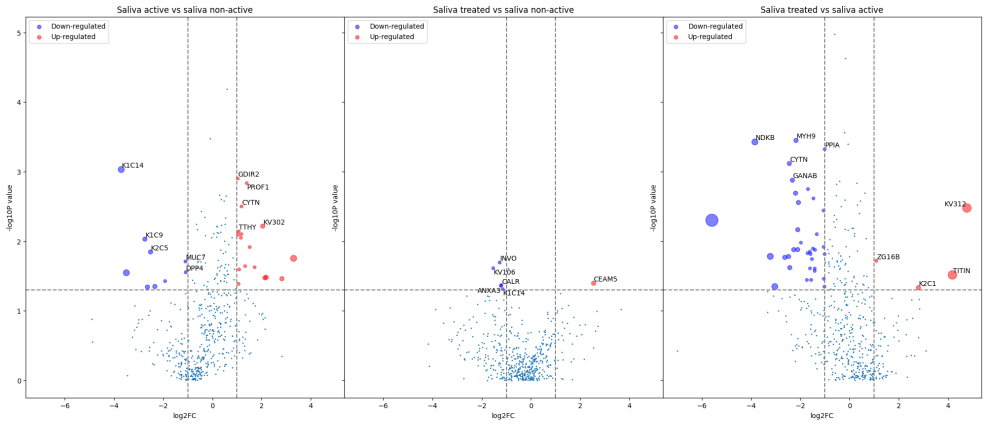


Figure 4.13: Volcano plot representation of the 506 differential abundance of proteins in different conditions from saliva samples.

In the three comparisons, the term "upregulated" is employed to describe elevated protein expression levels in active and treated caries conditions relative to non-active caries conditions, and higher in treated caries conditions in comparison to active caries conditions, as observed in both saliva and pellicle proteome data. Similarly, the term "downregulated" is employed to describe the proteins that are expressed in lower levels in active and treated caries conditions in comparison to non-active caries conditions, and in treated caries conditions in comparison to active caries conditions, in both saliva and pellicle proteome data.

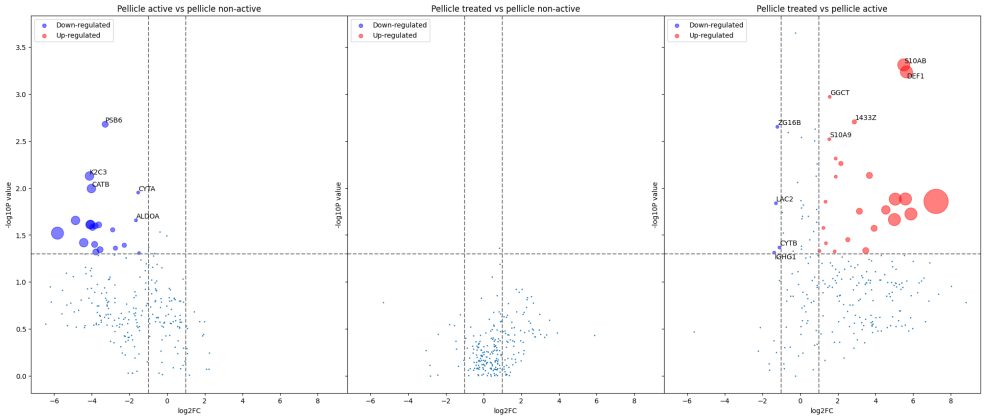


Figure 4.14: Volcano plot representation of the 238 differential abundance of proteins in different conditions from pellicle samples.

A comparative proteomic analysis of saliva samples revealed significant upregulation of 17 proteins and downregulation of 9 proteins in individuals with active caries, as compared to those with non-active caries. In the comparison between the treated caries condition and the non-active caries condition, only 1 protein was found to be significantly upregulated, while 5 proteins were significantly downregulated. In the comparison between the treated caries condition and the active caries condition, 4 proteins were found to be significantly upregulated, while 36 proteins were significantly downregulated. In the proteome data obtained from pellicle samples, no proteins were found to be significantly upregulated in the comparison between active and non-active caries conditions. Conversely, 20 proteins were identified as being significantly downregulated. In the comparison between the treated caries condition and the non-active caries condition, no proteins exhibited significant upregulation or downregulation. In the comparison between the treated caries condition and the active caries condition, 24 proteins were found to be significantly upregulated, while 4 were significantly downregulated.

A further fold change analysis was conducted to ascertain the disparity in protein abundance between the data obtained from the pellicle and saliva in a specific condition. A total of 166 proteins were identified as being commonly expressed in at least one volunteer in both the saliva and pellicle samples, irrespective of the condition under investigation. The 166 proteins were subsequently subjected to a fold change analysis, the results of which are represented in Fig. 4.15. This figure illustrates the differential abundance observed between the pellicle and saliva for each condition.

The term "upregulated" is used to describe a situation in which the level of protein expression in the pellicle is higher than that observed in saliva proteome data across all experimental conditions. Similarly, the term "downregulated" is used to describe a reduction in the levels of protein expression data observed in the pellicle proteome in comparison to the saliva proteome data across all conditions.

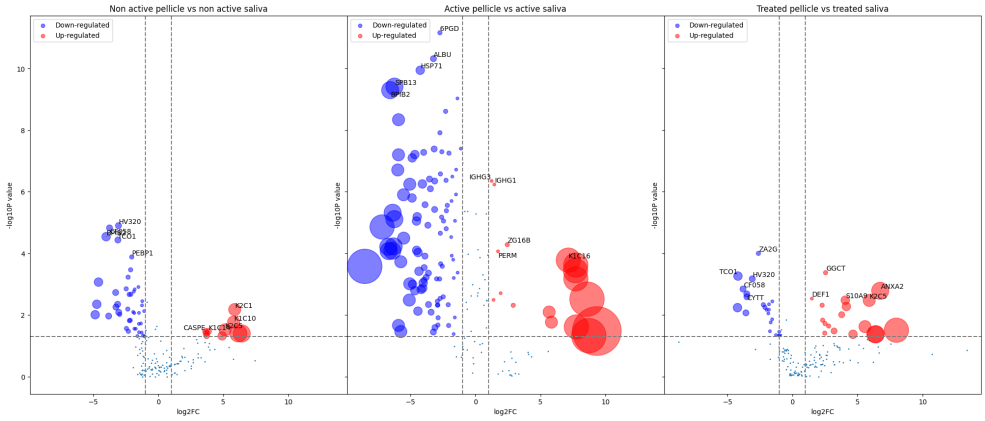


Figure 4.15: Volcano plot representation of the 166 differential abundance of proteins pellicle and saliva in each condition.

A comparison of the pellicle proteome data with that of the saliva proteome in non-active caries conditions revealed a significant upregulation of 10 proteins and a significant downregulation of 42 proteins. In the comparison between pellicle proteome data and saliva proteome data in active caries conditions, 17 proteins were found to be significantly upregulated, while 95 proteins were significantly downregulated. In the comparison between pellicle proteome data and saliva proteome data in treated caries conditions, 18 proteins were found to be significantly upregulated, while 22 proteins were significantly downregulated. It would appear that there is a significant disparity between the pellicle and saliva proteome data in the context of active caries, with 57% of the proteins exhibiting a notable decrease in expression.

Molecular function enrichment analysis

In order to gain further insight into the mechanisms that are distinct in saliva and pellicle under the influence of all the conditions, a molecular function enrichment analysis was conducted. The proteins of each condition of the saliva and pellicle proteome data were loaded into the `https://geneontology.org/` link for the enrichment analysis, and the results are presented in Fig. 4.16.

It was evident that the proteins present in the pellicle samples played a role in maintaining the structural integrity and molecular activities of the samples, whereas the proteins present in the saliva samples demonstrated a tendency to engage in catalytic activities within the oral cavity.

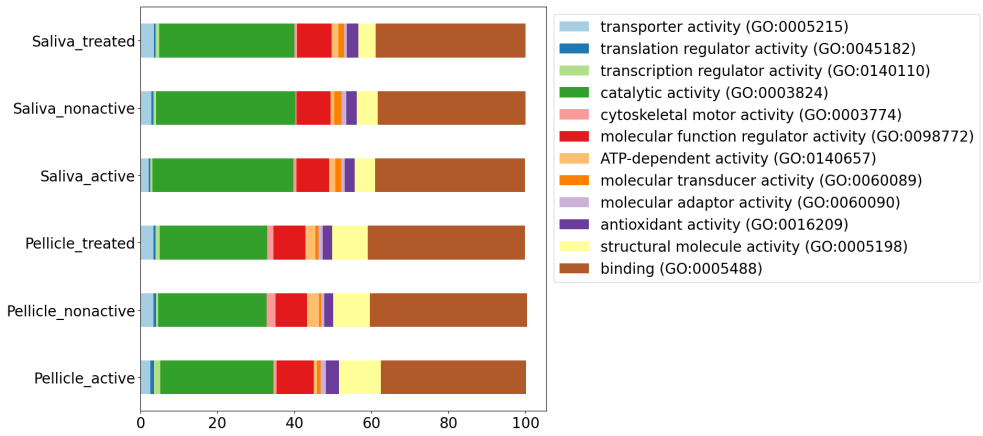


Figure 4.16: Distribution of percentage of molecular function hit in pellicle and saliva proteome data in all conditions

Principal Component Analysis

Principal Component Analysis (PCA) was further executed to determine if a clear distinction could be made between the condition of active caries, treated caries and the healthy state, using pellicle and salivary proteome data.To do so, the list of proteins that were expressed in all conditions in at least one volunteer of each condition of pellicle and salivary proteome data were considered. The PCA was performed on a total of 506 common proteins from the 1001 salivary proteins, and likewise on 238 common proteins from the 694 pellicle proteins.The results of these two PCAs are illustrated in figures 4.17 and 4.18

It is intriguing to observe that the proteome data from the pellicle of cases treated for caries (represented in fig 4.18 by blue dots) and those from healthy conditions (represented in fig 4.18 by green dots) exhibit a closer proximity in their clustering. This finding suggests that the proteome data of treated caries conditions may exhibit analogous trends to those observed in healthy conditions. This finding serves to validate the healing effects of caries.

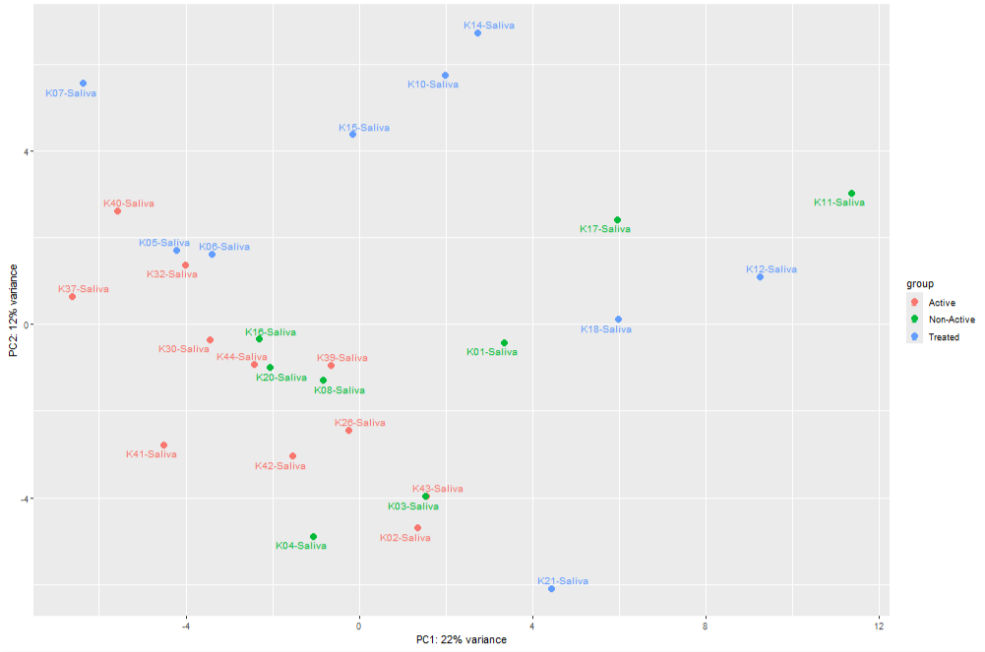


Figure 4.17: Results from Principal Component Analysis using salivary proteome data

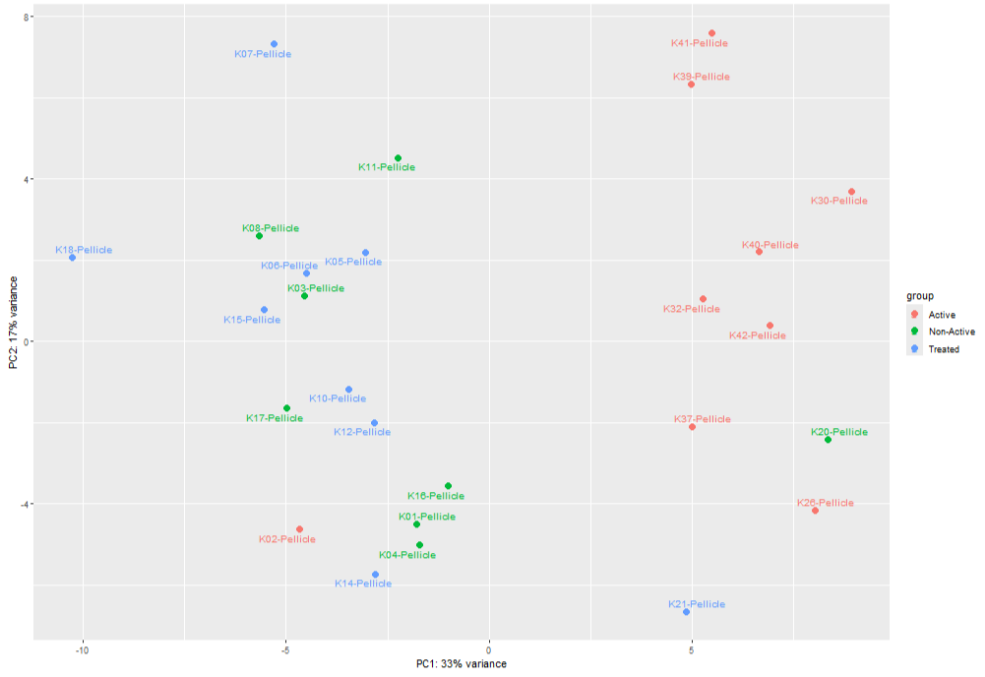


Figure 4.18: Results from Principal Component Analysis using pellicle proteome data

4.3.4 Conclusion

The findings of the comparative proteomic analysis of HAP and enamel suggest that there are minimal discrepancies in protein adsorption and biofilm formation between these two surfaces. This finding was supported by statistical similarity tests, which revealed comparable outcomes between replicates and materials. In summary, key processes such as pellicle formation, bacterial colonization, and individual-specific responses occur similarly on synthetic HAP and natural enamel. These findings underscore the viability of HAP as a substitute for enamel in future biofilm research endeavours.

A comparative proteomic analysis of saliva and pellicle was conducted, revealing discrepancies in the similarity between replicates and between volunteer data. However, it is anticipated that the incorporation of a more substantial number of replicate and volunteer data in future studies would facilitate the conclusion of statistical results with greater confidence. A further observation from the study revealed that certain proteins were found to be expressed at all time points in pellicle data, and these were also found to be significantly differentially expressed in saliva. It was also observed that certain proteins do not adsorb in the pellicles but rather remain in saliva. Further enrichment analysis of these proteins revealed that those which remain unabsorbed exhibit higher catalytic activity compared to the predominantly structural molecule activities observed in pellicle proteins. This finding suggests a potential role for these proteins in the formation of pellicle structures on oral surfaces. In addition, an analysis of molecular weight distribution patterns indicated that salivary proteins generally have smaller molecular weights than pellicle proteins. However, there was no significant correlation between their distributions. In addition, an examination of the isoelectric point distribution patterns revealed that the majority of salivary proteins fall within the pH range of the oral cavity, whereas the majority of pellicle proteins are found in a higher pH range.

The comparative analysis of the proteome of the pellicle and saliva samples collected from subjects in active caries, treated caries and healthy conditions yielded valuable insights into the dynamic changes in protein composition. It is noteworthy that the protein profile of the pellicle in treated conditions closely resembled that of healthy conditions, as evidenced by high similarity in PCA plots. Furthermore, a detailed investigation into the differential expression of proteins in saliva from caries-active conditions revealed significant up- and downregulation of proteins in comparison to healthy conditions. Notably, a substantial proportion of proteins exhibited unique downregulation in pellicle data during caries activity. In treated samples, saliva demonstrated substantial shifts in protein expression, with numerous proteins showing altered expression patterns compared to caries-active conditions. Conversely, the pellicle data from treated samples exhibited a prevalence of proteins that were found to be overexpressed, indicating a recovery process and functional adaptation. A detailed comparison of salivary and pellicle data sets across the conditions revealed a large number of differentially expressed proteins unique to each context, underscoring the distinct roles of saliva and

pellicle in oral health and disease. Molecular function enrichment analysis provided further insights, showing that pellicle proteins are predominantly involved in structural molecule activity, which is critical for forming and stabilising the pellicle layer on oral surfaces. Conversely, salivary proteins demonstrated a stronger association with catalytic activities, likely reflecting their involvement in biochemical processes essential for oral homeostasis. These findings underscore the intricate interplay between saliva and pellicle in maintaining oral health and adapting to pathological conditions, offering potential targets for therapeutic interventions.

Chapter 5
Putative protein complexes of peroxiredoxins

5.1 Introduction

This chapter describes unpublished research work carried out by myself in collaboration with Prof. Bruce Morgan. The chapter was written by me.

Peroxiredoxins (Prxs) constitute a ubiquitous family of antioxidant enzymes, with multiple isoforms expressed in a wide range of organisms, including *Arabidopsis thaliana*, *Saccharomyces cerevisiae* and *Homo sapiens* [207, 208]. They are involved in the detoxification of peroxides, aliphatic and aromatic hydroperoxides, and peroxynitrite [209, 210]. Moreover, they play a significant role in the defence against oxidative stress [211]. Cysteine (Cys) is a requisite residue for all catalytic processes involved in peroxide reduction by peroxiredoxins (Prxs). Peroxiredoxins are classified into two groups based on the number of cysteine residues present at the C-terminal: 2-Cys and 1-Cys. The cysteine residue that is responsible for the catalytic processes is considered to be the active site and is referred to as the peroxidatic cysteine (Cp or Sp). The catalytic cycle of Prx is illustrated in the Fig 5.1.

Prxs have a globular structure comprising five dimers that homo-oligomerise to form a decamer. It has been observed that certain Prxs exist in a redox-dependent equilibrium between a decameric and a dimeric state [212]. In general, hyperoxidised dimers have a propensity to form decamers, while oxidised dimers have a propensity to remain as dimers. The assembly of dimers is initiated by the interaction of two monomers at the B-type interface (beta sheets), while the formation of decamers involves the interaction of two dimers at the A-type interface (alpha helices).These interactions result in the formation of a toroid-like or doughnut-like structure, which subsequently

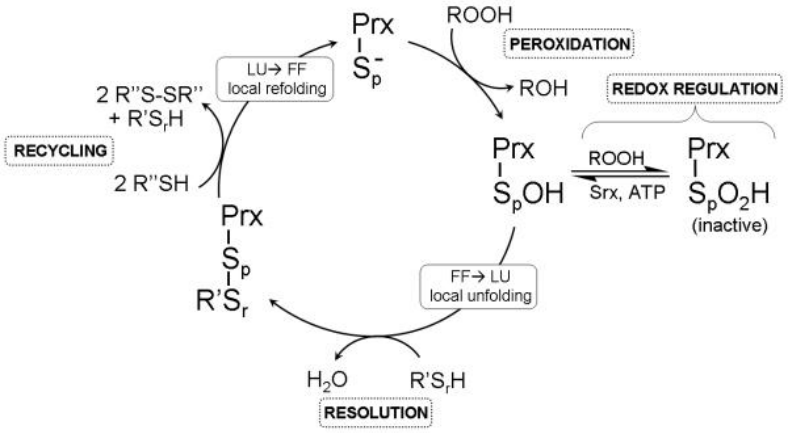


Figure 5.1: Representation of catalytic cycle of Peroxiredoxins adapted from [209]

gives rise to the assembly of decameric structures [212]. Hyperoxidised dimers have been observed to incorporate and form decamers, while oxidised dimers remain in dimer structures. Figure 5.2 illustrates the decamer, dimer with A-type and B-type interactions, and the monomer structure of a Prx from *Saccharomyces cerevisiae*.

Furthermore, it is a common observation that multiple isoforms of the Prxs are present in a number of organisms [213, 214]. For example, six subclasses of Prxs are observed in mammals, which have been demonstrated to influence a range of functions, including cell apoptosis, proliferation and differentiation, by regulating the cytokine-induced hydrogen peroxide levels [215, 216]. A number of studies have demonstrated that the monomers of the various isoforms of Prxs are capable of forming hetero-oligomers at dimeric or decameric stages in a multitude of organisms [217, 218, 219]. The reason for hetero-oligomerisation remains unclear. However, it is evident that these oligomers exhibit distinct enzymatic and mechanistic properties when compared to their homo counterparts. In light of this, an in silico modelling of Prxs hetero-oligomerisation was conducted using HADDOCK and Alphafold. These approaches were employed to gain insights and analyse the interface site, with the aim of verifying the possibility of hetero-oligomerisation.

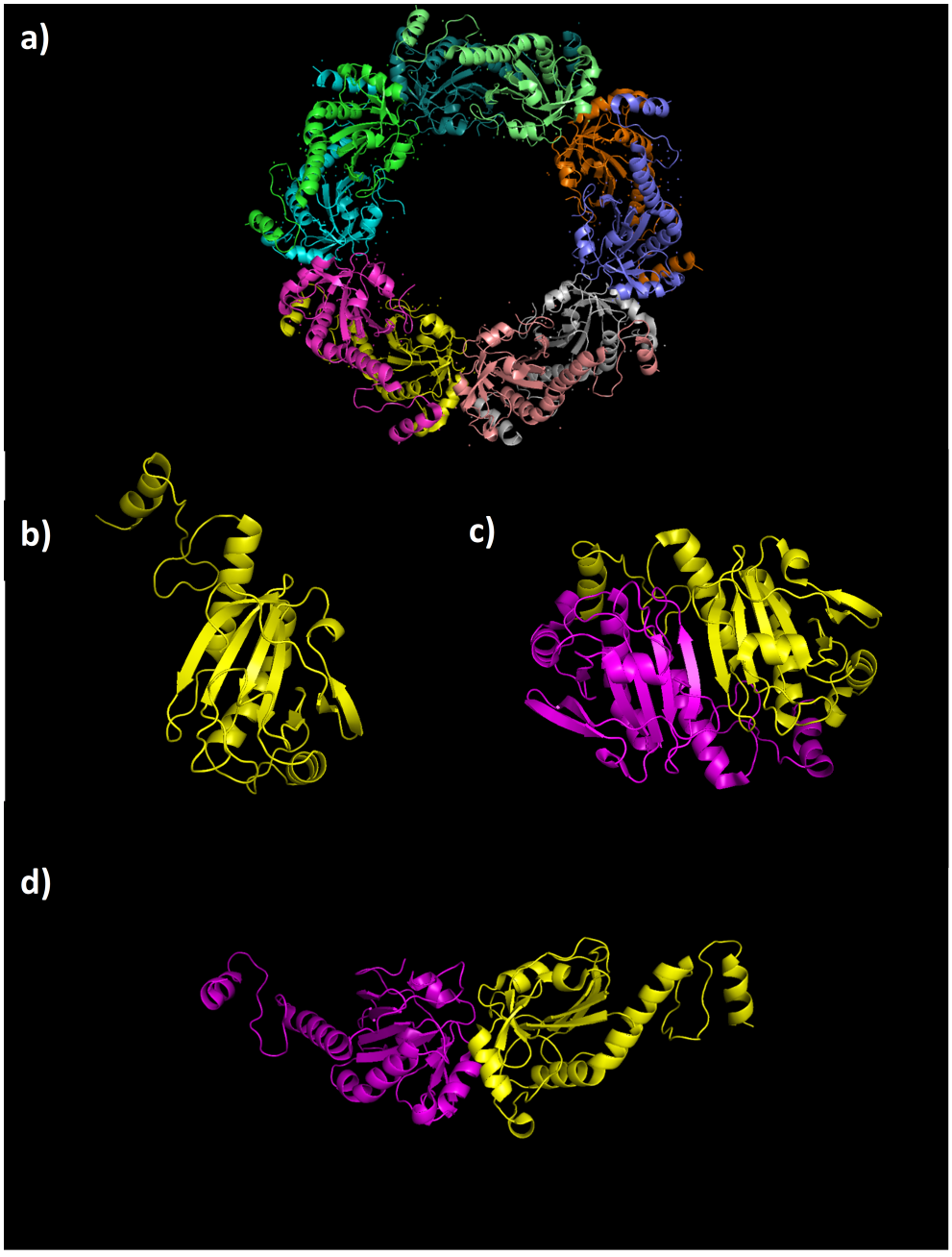


Figure 5.2: Representation of the crystal structure of TSA1 peroxiredoxin (PDB ID: 3SBC) from *Saccharomyces cerevisiae* depicts the doughnut-shaped decameric structure (a), the monomer (b), the dimers formed by interaction at beta sheets (c), and the dimers formed by interaction at alpha helices (d).

5.2 Materials and method

5.2.1 Datasets

Peroxiredoxins of *Saccharomyces cerevisiae*

Saccharomyces cerevisiae expresses five subtypes of Peroxiredoxins, TSA1, TSA2, Dot5, Aph1 and Prx1 [220, 221]. Of the five subtypes, TSA1 and TSA2 exhibit the largest homology [220, 222]. Accordingly, these two Prxs were selected for in silico modelling. The decameric crystal structure of TSA1 *Saccharomyces cerevisiae* from X-ray diffraction was deposited in the RCSB Protein Data Bank (PDB [223]) with the following PDB ID: 3SBC. The decameric crystal structure of TSA2 *Saccharomyces cerevisiae* from X-ray diffraction was submitted to the RCSB PDB, where it was assigned the PDB ID: 5DVB. In order to perform the docking, the crystal structures and the corresponding sequences were downloaded and used as input in HADDOCK and Alphafold.

Peroxiredoxins of *Arabidopsis thaliana*

The genome of *Arabidopsis thaliana* encodes ten Peroxiredoxins [224]. In this study, PrxA was selected for in silico modelling due to its classification within the 2-Cys Prx family, which is similar to that of TSA1 and TSA2 [225]. The decameric crystal structure of PrxA from *Arabidopsis thaliana*, derived from X-ray diffraction, was submitted to the RCSB PDB, where it was assigned the PDB ID: 5ZTE. Consequently, the crystal structures and the corresponding sequence were downloaded and employed as input in HADDOCK for the purpose of docking.

5.2.2 HADDOCK

HADDOCK is an integrative modelling platform that incorporates a variety of data types, including ambiguous and low-resolution data, information on the symmetry and flexibility of molecules, and information on residues that are important for binding to aid docking. Despite the fact that the docking process is based on rigid body theory, it permits a certain degree of flexibility in the structures under consideration, including those at the binding interface. Given that HADDOCK handles data of low resolution, which may be inherently inaccurate, the tool randomly eliminates half of the information in order to arrive at the optimal solutions. However, there is a risk of deleting accurate information and also producing false positive results.

The initial step in the HADDOCK process is to prepare the input structures. A maximum of 20 structures of the molecules can be inputted. For the input, the most optimal results would be obtained from experimental structures derived from X-ray, NMR crystallography or cryo-electron microscope structures that have been deposited in the Protein Data Bank (PDB). In the absence of experimental data for proteins in the PDB, homologous protein structures may be employed as templates. Homologous structures can be identified in the UniProt database [149] and the HMMER database [226].

Homology modelling can be conducted using the SWISS-MODEL database [227]and the MODELLER database [228].It should be noted, however, that the tool has a few rules regarding the PDB files that are to be input. It is imperative that all PDB files conclude with an "END" statement. In the event of structural discontinuities, it is advisable to utilise the "TER" statement to delineate the respective chains or sub-molecules. In the event that the molecule contains multiple chains, but the numbers of these chains overlap, it is of utmost importance to re-number the chains. It is possible for high-resolution crystal structures to contain multiple occupancy site conformations. Nevertheless, only one conformation is to be retained. It is imperative that the input file does not contain the name of the segment as the segment ID (SEGID). It is essential that the numbering of ions is consistent. Fortunately, the experimental crystal structures of the TSA1 and TSA2 Prxs of *Saccharomyces cerevisiae* and PrxA of *Arabidopsis thaliana* were deposited in the Protein Data Bank (PDB), and thus these structures were used as input. The PDB files were modified to comply with the aforementioned rules.

Initially, dimers were constructed and evaluated at the A-type and B-type interfaces from corresponding monomers of TSA1 and TSA 2 Prxs of *Saccharomyces cerevisiae*, using HADDOCK to ascertain whether the outcomes exhibited similarities to those observed in the wild-type deposited structure at the interaction sites. Subsequently, trimers were constructed, whereby two monomers interact at the B-type interface and a third monomer interacts with the dimer at the A-type interface. And finally, ten monomers are given as input with A and B types of interface to construct the doughnut shaped decamer. To this end, the HADDOCK server is used to input the requisite data, with the number of molecules selected as two for the construction of dimer, three for the construction of a trimer, and ten for the construction of decamer. For each molecule, the relevant PDB file is uploaded, and a chain is selected as the monomer. As part of the input parameters, the active and passive residues at the A and B-type interfaces are specified. The active residues at B-type interface (for example, the β sheet interchain interaction sites of PrxA *A. thaliana* are Q207, H208, S209, I211, and N213) and A-Type interface (the α helix interchain interaction sites of PrxA *A. thaliana* are S150, V175, and K177) were provided as input. The remaining parameters were set to their default values and the process was initiated to obtain the results.

Once the results have been generated, the top clusters of conformations with the highest scores are presented in a list. Each cluster comprises the four most optimal conformations. The conformations from the ten most prominent clusters were aligned with the previously deposited structures in order to estimate the root-mean-square deviation (RMSD) values. A lower RMSD value indicates a greater degree of similarity between the predicted conformation and the experimental structure.

In order to determine whether the Prxs form heterodimers or heterotrimers, one molecule was derived from the monomer of TSA1, and the other was derived from the monomer of TSA2. The active and passive residues were employed in conjunction with the HADDOCK algorithm to generate

docked conformation clusters. The resulting conformations were aligned with the experimental dimer and trimer structures of TSA1 and TSA2 in order to estimate the RMSD values and ascertain the degree of similarity. Subsequently, the hypothesis was tested that the Prxs of monomers from one species dimerise or trimerise with Prxs of monomers from other species. Therefore, the monomers from TSA1 and TSA2 of *Saccharomyces cerevisiae* were docked against the monomers from PrxA of *Arabidopsis thaliana*. The resulting conformations from the clusters were aligned with those of the Prxs of both species in order to estimate the similarity by RMSD values.

5.2.3 AlphaFold

AlphaFold 3 is a model designed by researchers at Google DeepMind with the objective of predicting complexes from all types of molecules, with a particular focus on protein-protein interaction complexes, deposited in the Protein Data Bank (PDB) with high accuracy [15]. The model has demonstrated the capacity to predict novel protein structures and complexes that are not currently represented in the Protein Data Bank (PDB) [229, 230].

A variety of sequences, including those derived from proteins, DNA, RNA, ligand molecules, or ions, can be entered as input by specifying the appropriate entity and the number of copies of each molecule. However, for certain molecules, such as ligands, ions, and chemical modifications, only a specific molecule can be selected from the available list. Subsequently, AlphaFold seeks to identify pertinent MSA data for RNA chains and proteins. In the case of protein inputs, the protein template structure information is also employed. Based on the aforementioned three categories of input, five conformations are predicted. The predicted structures and complexes comprise information regarding the coordinates of atoms belonging to the conformation, as well as five confidence scores designed to assess the conformations.

In order to establish whether the Prxs form heterodimers or heterotrimers, the overall decameric structures of the Prxs, including TSA1, TSA2 and PrxA, were initially replicated in order to ascertain whether they exhibited a close structural similarity to the corresponding PDB structures that had been deposited. To this end, ten monomers of the corresponding TSA1, TSA2 and PrxA were loaded in individual runs. Subsequently, the predicted complexes were compared and aligned with the experimental structures in order to confirm the accuracy of the predictions. Once the results are accurate and favourable, the hetero-dimerisation with alternating monomers will be tested to ascertain the plausibility of the resulting structures.

5.3 Results and discussion

5.3.1 Results from HADDOCK

B-type TSA1 dimers from *Saccharomyces cerevisiae*

The objective was to replicate the B-type dimers from the monomers of TSA1 *Saccharomyces cerevisiae*. The monomer and the dimer at the B-type interface is represented in Fig. 5.3

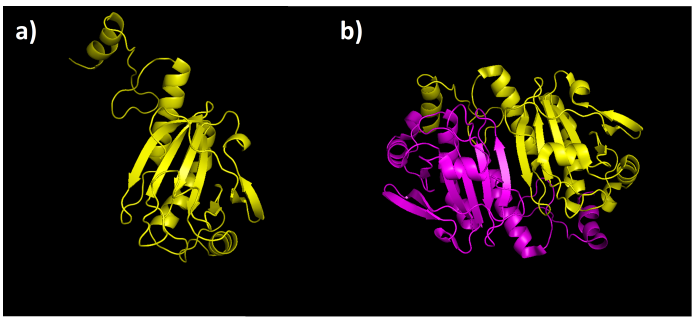


Figure 5.3: Representation of the monomer of TSA1 *Saccharomyces cerevisiae* (a), and the dimers formed by interaction at beta sheets (b).

HADDOCK yielded 13 clusters, comprising a total of 89 conformations. The ten most favourable conformations are presented in Table 5.1.

Cluster No.	HADDOCK score	Cluster size	RMSD from overall lowest energy (A°)	Z- score	RMSD aligning TSA1 dimer (A°)
2	-138.5 +/- 26.0	15	2.8 +/- 0.5	-2.5	2.5
6	-104.3 +/- 4.6	6	3.3 +/- 0.3	-0.5	3.8
1	-104.3 +/- 9.4	18	6.1 +/- 0.8	-0.5	6.7
4	-98.8 +/- 3.3	7	20.3 +/- 0.1	-0.2	20.7
13	-96.2 +/- 13.2	4	18.1 +/- 0.4	-0.1	18.3
3	-90.4 +/- 11.7	7	19.1 +/- 0.3	0.3	20.1
5	-83.5 +/- 18.9	6	18.8 +/- 0.8	0.7	20.9
8	-81.4 +/- 7.4	5	3.5 +/- 0.7	0.8	3.9
10	-79.8 +/- 12.2	4	21.7 +/- 0.0	0.9	21.8
12	-75.0 +/- 6.8	4	19.5 +/- 0.5	1.2	20.6

Table 5.1: The table depicts the ten most similar conformations to the TSA1 B-type dimer. The table contains information regarding the cluster, the HADDOCK score, the cluster size, the RMSD of the lowest energy, the z-score and the RMSD while aligning to the TSA1 B-type dimer.

In this study, a threshold of 10 A° is set as the threshold for the estimated RMSD values. The structures of clusters 1, 2, 6 and 8 exhibit a high degree of alignment with the reference dimer of TSA1. Furthermore, the low root-mean-square deviation (RMSD) of atomic position value calculated lends additional support to this conclusion. Therefore, the B-type dimers of yeast TSA1 were

successfully replicated using only one monomer sequence.

A-type TSA1 dimers from *Saccharomyces cerevisiae*

The objective was to replicate the A-type dimers from the monomers of TSA1 *Saccharomyces cerevisiae*. The monomer and the dimer at the A-type interface is represented in Fig. 5.4

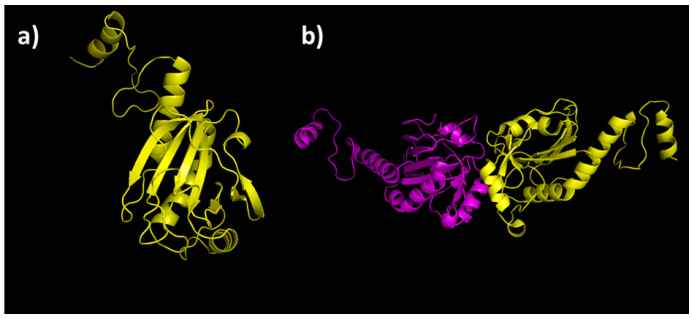


Figure 5.4: Representation of the monomer of TSA1 *Saccharomyces cerevisiae* (a), and the dimers formed by interaction at alpha helices (b).

HADDOCK yielded 18 clusters, comprising a total of 356 conformations. The ten most favourable conformations are presented in Table 5.2.

Cluster No.	HADDOCK score	Cluster size	RMSD from overall lowest energy (A°)	Z- score	RMSD aligning TSA1 dimer (A°)
1	-117.3 +/- 0.7	170	1.1 +/- 0.9	-2.1	1.4
9	-104.9 +/- 3.7	7	14.6 +/- 0.3	-1.0	14.1
2	-96.7 +/- 4.5	48	13.7 +/- 0.1	-0.3	13.5
4	-94.8 +/- 7.3	22	2.2 +/- 0.3	-0.1	1.8
7	-93.6 +/- 3.5	9	11.2 +/- 0.7	-0.0	11.9
3	-93.0 +/- 6.8	31	2.5 +/- 0.3	0.0	2.4
18	-91.1 +/- 8.4	4	2.8 +/- 0.6	0.2	1.9
17	-86.1 +/- 6.1	4	14.5 +/- 0.4	0.7	14.8
5	-82.7 +/- 7.7	13	14.2 +/- 0.3	1.0	14.2
6	-74.7 +/- 3.3	9	14.7 +/- 0.3	1.7	14.3

Table 5.2: The table depicts the ten most similar conformations to the TSA1 A-type dimer. The table contains information regarding the cluster, the HADDOCK score, the cluster size, the RMSD of the lowest energy, the z-score and the RMSD while aligning to the TSA1 B-type dimer.

The structures of clusters 1, 3, 4, and 18 exhibit a high degree of alignment with the reference A-type dimer of TSA1. This conclusion is further supported by the low root-mean-square deviation (RMSD) of atomic position values calculated. The successful replication of the A-type dimers of TSA1 *Saccharomyces cerevisiae* using a single monomer sequence is therefore demonstrated.

A-type and B-type interface interaction in TSA1 trimers from *Saccharomyces cerevisiae*

The objective was to replicate the A-type and B-type interface interactions observed in the monomers of *Saccharomyces cerevisiae* TSA1. The monomer and the trimer formed at the A-type and B-type interfaces are illustrated in the Fig. 5.5

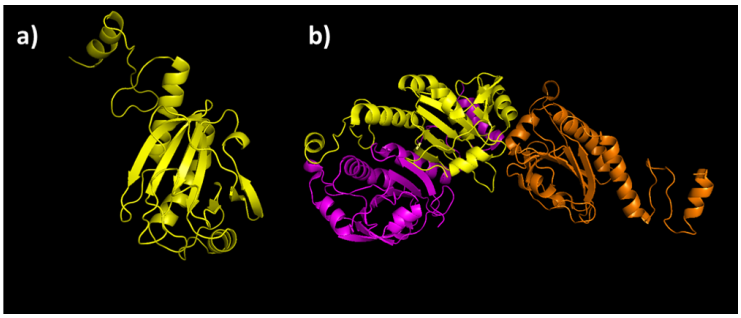


Figure 5.5: The monomer of TSA1 is represented in (a), while the trimer formed by interaction at alpha helices (between yellow and orange chains) and beta sheets (between yellow and magenta chains) is shown in (b).

HADDOCK yielded 7 clusters, comprising a total of 18 conformations. The seven conformations are presented in Table 5.3.

Cluster No.	HADDOCK score	Cluster size	RMSD from overall lowest energy (A°)	Z- score	RMSD aligning TSA1 trimer (A°)
1	-366.1 +/- 12.2	35	8.8 +/- 6.5	-2.3	2.7
2	-217.5 +/- 10.8	27	6.5 +/- 3.6	-0.3	12.6
4	-185.9 +/- 14.6	6	20.7 +/- 1.1	0.1	19.6
3	-152.7 +/- 9.8	11	14.3 +/- 6.0	0.6	5.6
7	-150.5 +/- 16.1	4	19.7 +/- 1.9	0.6	20.9
6	-148.4 +/- 18.2	4	21.4 +/- 1.2	0.6	20.5
5	-146.1 +/- 18.8	4	21.2 +/- 0.9	0.7	21.1

Table 5.3: The table depicts the seven most similar conformations to the TSA1 A-type and B-type trimer. The table contains information regarding the cluster, the HADDOCK score, the cluster size, the RMSD of the lowest energy, the z-score and the RMSD while aligning to the TSA1 trimer.

The structures of clusters 1 and 3 exhibit a high degree of alignment with the reference A-type and B-type interfaces of TSA1 *Saccharomyces cerevisiae*. Furthermore, the low root-mean-square deviation (RMSD) of atomic position value calculated lends additional support to this conclusion. It was therefore demonstrated that the A-type and B-type interfaces of TSA1 *Saccharomyces cerevisiae* could be successfully replicated using only one monomer sequence.

Decamer of TSA1 from *Saccharomyces cerevisiae*

The objective was to replicate the decamer doughnut-like shaped structure observed in *Saccharomyces cerevisiae* TSA1. Therefore, five molecules of B-type TSA1 dimers and the A-type interface active residues were incorporated as the initial input. An illustration of the decamer structure can be found in the Fig. 5.6.

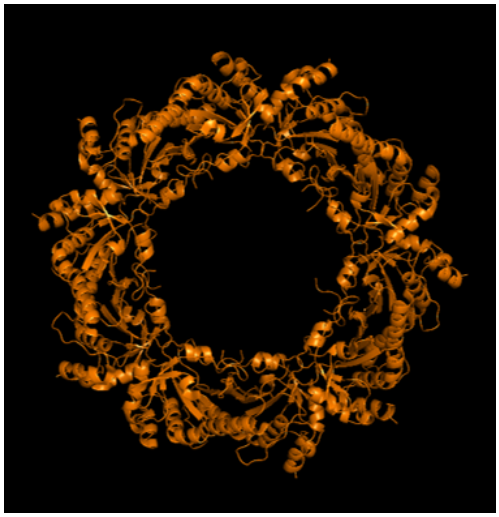


Figure 5.6: The decamer structure of TSA1 *Saccharomyces cerevisiae*.

HADDOCK yielded 200 clusters, comprising a total of 200 conformations. The ten most favourable conformations are presented in Table 5.4.

Cluster No.	HADDOCK score	Cluster size	RMSD from overall lowest energy (A°)	Z- score	RMSD aligning TSA1 decamer (A°)
1	-425.8 +/- 0.0	1	0.0 +/- 0.0	-1.7	48.6
2	-419.4 +/- 0.0	1	44.7 +/- 0.0	-1.4	23.4
3	-417.5 +/- 0.0	1	49.1 +/- 0.0	-1.3	28.6
4	-390.1 +/- 0.0	1	48.5 +/- 0.0	0.2	34.4
5	-388.3 +/- 0.0	1	48.1 +/- 0.0	0.3	29.9
6	-381.3 +/- 0.0	1	47.1 +/- 0.0	0.6	46.5
7	-380.2 +/- 0.0	1	49.7 +/- 0.0	0.7	15.1
8	-379.7 +/- 0.0	1	49.2 +/- 0.0	0.7	24.3
9	-375.2 +/- 0.0	1	42.7 +/- 0.0	0.9	42.7
10	-372.9 +/- 0.0	1	49.4 +/- 0.0	1.1	24.8

Table 5.4: The table depicts the ten most similar conformations to the TSA1 decamer. The table contains information regarding the cluster, the HADDOCK score, the cluster size, the RMSD of the lowest energy, the z-score and the RMSD while aligning to the TSA1 decamer.

It is evident that none of the predicted structures demonstrate a high degree of similarity to the reference decamer *Saccharomyces cerevisiae* TSA1.

RMSD of atomic position value calculated lends additional support to this conclusion. The overall 3D orientation of the predicted structures did not resemble a doughnut structure. Consequently, an attempt was made to dock with other beta dimers, including chains C and D, chains E and F, chains G and H, and chains I and J. However, docking with all these dimers, in individual runs, did not yielded successful results. The alignment with TSA1 was not optimal.

B-type TSA2 dimers from *Saccharomyces cerevisiae*

The objective was to replicate the B-type dimers from the monomers of TSA2 *Saccharomyces cerevisiae*.

HADDOCK yielded 29 clusters, comprising a total of 308 conformations. The ten most favourable conformations are presented in Table 5.5.

Cluster No.	HADDOCK score	Cluster size	RMSD from overall lowest energy (A°)	Z- score	RMSD aligning TSA2 dimer (A°)
1	-262.9 +/- 9.1	80	0.8 +/- 0.5	-3.0	1.3
5	-96.5 +/- 2.3	12	15.0 +/- 0.5	0.1	13.8
11	-95.0 +/- 6.5	8	5.2 +/- 0.6	0.2	5.5
3	-88.5 +/- 3.5	19	5.5 +/- 0.3	0.3	5.8
6	-85.8 +/- 4.7	11	21.7 +/- 0.2	0.3	19.8
12	-84.0 +/- 17.3	8	14.7 +/- 0.9	0.4	15.0
2	-82.5 +/- 8.2	30	5.4 +/- 0.4	0.4	5.9
14	-82.2 +/- 15.2	7	16.8 +/- 0.9	0.4	13.5
16	-81.0 +/- 10.9	6	22.0 +/- 0.2	0.4	20.5
4	-80.8 +/- 11.5	16	5.3 +/- 0.2	0.4	5.3

Table 5.5: The table depicts the ten most similar conformations to the TSA2 B-type dimer. The table contains information regarding the cluster, the HADDOCK score, the cluster size, the RMSD of the lowest energy, the z-score and the RMSD while aligning to the TSA2 B-type dimer.

The structures of clusters 1, 2, 3, 4 and 11 exhibit a high degree of alignment with the reference dimer of TSA2. Furthermore, the low RMSD of atomic position value calculated lends additional support to this conclusion. It was thus demonstrated that the B-type dimers of *Saccharomyces cerevisiae* TSA2 could be successfully replicated using only a single monomer sequence.

A-type TSA2 dimers from *Saccharomyces cerevisiae*

The objective was to replicate the A-type dimers from the monomers of TSA2 *Saccharomyces cerevisiae*.

HADDOCK yielded 7 clusters, comprising a total of 390 conformations. The seven favourable conformations are presented in Table 5.6.

All the clusters exhibit a close alignment with the reference A-type dimer of TSA2, a finding that is also corroborated by the low RMSD of atomic position value calculated. Consequently, the A-type dimers of *Saccharomyces*

Cluster No.	HADDOCK score	Cluster size	RMSD from overall lowest energy (A°)	Z- score	RMSD aligning TSA2 dimer (A°)
1	-114.4 +/- 4.9	245	1.2 +/- 1.0	-1.7	1.6
3	-97.3 +/- 5.6	58	3.0 +/- 0.2	-0.9	3.6
2	-87.3 +/- 3.2	61	4.5 +/- 0.6	-0.5	4.3
5	-66.5 +/- 15.6	7	4.3 +/- 0.4	0.5	3.8
6	-64.6 +/- 6.6	6	2.4 +/- 0.4	0.6	3.3
4	-59.4 +/- 16.9	8	5.0 +/- 0.3	0.8	3.8
7	-50.7 +/- 14.8	5	6.9 +/- 0.9	1.2	5.9

Table 5.6: The table depicts the seven most similar conformations to the TSA2 A-type dimer. The table contains information regarding the cluster, the HADDOCK score, the cluster size, the RMSD of the lowest energy, the z-score and the RMSD while aligning to the TSA2 A-type dimer.

cerevisiae TSA2 have been successfully replicated using only a single monomer sequence.

A-type and B-type interface interactions in TSA2 trimers from *Saccharomyces cerevisiae*

The objective was to replicate the A-type and B-type interface interactions observed in the monomers of *Saccharomyces cerevisiae* TSA2.

HADDOCK yielded 9 clusters, comprising a total of 46 conformations. The nine conformations are presented in Table 5.7.

Cluster No.	HADDOCK score	Cluster size	RMSD from overall lowest energy (A°)	Z- score	RMSD aligning TSA2 trimer (A°)
1	-309.1 +/- 18.3	9	10.3 +/- 0.7	-1.6	14.7
7	-303.7 +/- 13.1	4	13.3 +/- 0.3	-1.3	4.1
4	-295.4 +/- 23.2	5	2.4 +/- 1.9	-0.9	13.2
3	-273.9 +/- 14.0	5	9.6 +/- 1.4	0.3	14.0
8	-272.6 +/- 10.3	4	21.5 +/- 0.9	0.3	26.1
2	-272.1 +/- 27.7	5	22.3 +/- 0.3	0.4	23.4
6	-270.2 +/- 17.5	5	23.0 +/- 0.3	0.5	23.7
5	-265.7 +/- 12.9	5	24.0 +/- 0.5	0.7	23.0
9	-248.5 +/- 36.4	4	14.4 +/- 0.7	1.6	14.1

Table 5.7: The table depicts the nine most similar conformations to the TSA2 A-type and B-type trimer. The table contains information regarding the cluster, the HADDOCK score, the cluster size, the RMSD of the lowest energy, the z-score and the RMSD while aligning to the TSA2 trimer.

The structures of cluster 7 exhibit a close alignment to the reference A-type and B-type interfaces of *Saccharomyces cerevisiae* TSA2, a finding that is also supported by the low RMSD of atomic position value calculated. This demonstrates that the A-type and B-type interfaces of TSA2 can be successfully replicated using only one monomer sequence.

Decamers of TSA2 from *Saccharomyces cerevisiae*

The objective was to replicate the decamer doughnut-like shaped structure observed in *Saccharomyces cerevisiae* TSA1. Therefore, five molecules of B-type TSA2 dimers and the A-type interface active residues were incorporated as the initial input.

HADDOCK yielded 200 clusters, comprising a total of 200 conformations. The ten most favourable conformations are presented in Table 5.8.

Cluster No.	HADDOCK score	Cluster size	RMSD from overall lowest energy (A°)	Z- score	RMSD aligning TSA2 decamer (A°)
1	-495.7 +/- 0.0	1	0.0 +/- 0.0	-2.2	45.4
2	-427.8 +/- 0.0	1	29.9 +/- 0.0	-0.6	43.2
3	-424.9 +/- 0.0	1	40.2 +/- 0.0	-0.5	40.3
4	-422.3 +/- 0.0	1	48.7 +/- 0.0	-0.5	52.4
5	-422.2 +/- 0.0	1	36.0 +/- 0.0	-0.5	38.5
6	-397.3 +/- 0.0	1	33.7 +/- 0.0	0.1	43.8
7	-371.4 +/- 0.0	1	33.4 +/- 0.0	0.7	38.9
8	-359.0 +/- 0.0	1	28.8 +/- 0.0	1	36.0
9	-355.1 +/- 0.0	1	40.9 +/- 0.0	1.1	38.1
10	-355.1 +/- 0.0	1	31.7 +/- 0.0	1.1	37.5

Table 5.8: The table depicts the ten most similar conformations to the TSA2 decamer. The table contains information regarding the cluster, the HADDOCK score, the cluster size, the RMSD of the lowest energy, the z-score and the RMSD while aligning to the TSA2 decamer.

The doughnut-shaped decameric structure, which bears resemblance to the TSA2, is not formed. Consequently, an attempt was made to dock with other beta dimers, including chains C and D, chains E and F, chains G and H, and chains I and J. However, docking with all these dimers, in individual runs, did not yielded successful results. The alignment with TSA2 was not optimal.

Hetero-dimerisation at B-type interface of Prxs from *Saccharomyces cerevisiae*

The objective was to dock a monomer derived from TSA1 and a monomer derived from TSA2 of *Saccharomyces cerevisiae* at B-type interfaces. Accordingly, a molecule of a chain from TSA1, another molecule of a chain from TSA2, and the B-type interface active residues in TSA1 and TSA2 were loaded as the input.

HADDOCK yielded 23 clusters, comprising a total of 281 conformations. The ten most favourable conformations are presented in Table 5.9.

The structures of clusters 1, 2, 4, 12 and 15 exhibit a high degree of alignment with the reference B-type interfaces of TSA1 and TSA2. This conclusion is further supported by the low RMSD of atomic position values calculated. Consequently, successfully created heterodimer from monomers of TSA1 an TSA2 at B-type interface.

Cluster No.	HADDOCK score	Cluster size	Z- score	RMSD aligning TSA1 dimer (A°)	RMSD aligning TSA2 dimer (A°)
1	-270.2 +/- 6.2	78	-3	1.654	1.9
4	-108.3 +/- 11.6	15	0	3.754	3.9
7	-99.2 +/- 14.2	12	0.2	20.125	20.0
6	-98.7 +/- 10.5	12	0.2	22.224	21.0
3	-98.1 +/- 4.7	21	0.2	21.564	19.3
2	-90.5 +/- 8.3	24	0.4	6.015	6.0
12	-85.5 +/- 8.5	9	0.4	6.72	6.3
15	-82.4 +/- 16.4	6	0.5	4.853	5.0
5	-82.4 +/- 11.6	13	0.5	22.813	20.6
9	-80.8 +/- 1.4	11	0.5	22.072	20.3

Table 5.9: The table depicts the ten most similar conformations to the B-type TSA1 and TSA2 dimers. The table contains information regarding the cluster, the HADDOCK score, the cluster size, the RMSD of the lowest energy, the z-score and the RMSD while aligning to the TSA1 and TSA2 dimer.

Hetero-dimerisation at A-type interface of Prxs from *S. cerevisiae*

The objective was to dock a monomer derived from TSA1 and a monomer derived from TSA2 of *Saccharomyces cerevisiae* at A-type interfaces. Accordingly, a molecule of a chain from TSA1, another molecule of a chain from TSA2, and the A-type interface active residues in TSA1 and TSA2 were loaded as the input.

HADDOCK yielded 14 clusters, comprising a total of 161 conformations. The ten most favourable conformations are presented in Table 5.10.

Cluster No.	HADDOCK score	Cluster size	Z- score	RMSD aligning TSA1 dimer (A°)	RMSD aligning TSA2 dimer (A°)
5	-96.7 +/- 17.4	14	-2.1	3.849	3.7
2	-85.3 +/- 6.9	19	-1.2	1.715	1.6
3	-79.3 +/- 2.8	17	-0.7	7.629	6.9
1	-70.7 +/- 0.8	28	0	12.382	11.8
4	-66.9 +/- 5.2	17	0.2	11.408	11.6
9	-66.6 +/- 8.6	7	0.3	11.989	12.1
7	-62.7 +/- 5.4	1	0.6	5.059	4.8
8	-59.6 +/- 4.6	10	0.8	5.869	5.8
10	-59.6 +/- 7.8	6	0.8	15.567	15.4
6	-53.4 +/- 3.6	14	1.3	15.918	15.7

Table 5.10: The table depicts the ten most similar conformations to the A-type TSA1 and TSA2 dimers. The table contains information regarding the cluster, the HADDOCK score, the cluster size, the RMSD of the lowest energy, the z-score and the RMSD while aligning to the TSA1 and TSA2 dimer.

The structures of clusters 2, 3, 5, 7 and 8 exhibit a high degree of alignment with the reference A-type interfaces of TSA1 and TSA2. This is

supported by the low RMSD of atomic position values calculated. Therefore, successfully created heterodimer from monomers of TSA1 and TSA2 at A-type interface.

Hetero-trimerisation of Prxs from *Saccharomyces cerevisiae*

The objective was to dock a three-monomer derived from TSA1 and TSA2 of *Saccharomyces cerevisiae* at A-type and B-type interfaces with the intention of forming trimers that are similar in structure to those observed in the experimental data. A total of six combinations exist in which the three molecules derived from TSA1 and TSA2 can be inputted at the A-type and B-type interfaces for docking purposes. The combinations are presented in the table 5.11. In order to ascertain which combination yields the most favourable results, all six combinations have been subjected to docking.

Combination	Monomers interacting in B-type interface		Monomer interacting at A-type interface
Combination 1	TSA1	TSA1	TSA2
Combination 2	TSA1	TSA2	TSA1
Combination 3	TSA2	TSA1	TSA1
Combination 4	TSA1	TSA2	TSA2
Combination 5	TSA2	TSA1	TSA2
Combination 6	TSA2	TSA2	TSA1

Table 5.11: The six possible combinations for docking hetero-trimers from monomers of TSA1 and TSA2

Combination 1 of hetero-trimerisation

In this configuration, two monomers of TSA1 were loaded with B-type interface residues, while one monomer from TSA2 is loaded with A-type interface residues.

HADDOCK yielded 12 clusters, comprising a total of 26 conformations. The ten most favourable conformations are presented in Table 5.12.

Using the combination 1 of monomers, the structures of cluster 4 exhibit a high degree of similarity to the reference A- and B-type interfaces of TSA1 and TSA2. Furthermore, the low RMSD of atomic position value calculated lends additional support to this conclusion.

Cluster No.	HADDOCK score	Cluster size	Z- score	RMSD aligning TSA1 trimer (A°)	RMSD aligning TSA2 trimer (A°)
3	-312.0 +/- 19.8	2	-2.2	23.925	22.5
4	-251.2 +/- 79.2	2	-1.2	5.45	5.3
1	-199.4 +/- 21.5	3	-0.4	11.084	11.7
5	-191.5 +/- 4.5	2	-0.3	12.495	13.2
8	-151.8 +/- 10.6	2	0.3	24.089	23.5
6	-142.7 +/- 34.4	2	0.5	10.295	11.1
7	-133.8 +/- 29.9	2	0.6	13.921	14.1
9	-131.1 +/- 1.9	2	0.6	16.809	25.1
11	-105.9 +/- 5.4	2	1	22.295	21.8
2	-101.2 +/- 13.5	3	1.1	13.075	13.0

Table 5.12: The table depicts the ten most similar conformations to the TSA1 and TSA2 trimers using combination 1 of monomers. The table contains information regarding the cluster, the HADDOCK score, the cluster size, the RMSD of the lowest energy, the z-score and the RMSD while aligning to the TSA1 and TSA2 trimers.

Combination 2 of hetero-trimerisation

In this configuration, at the B-type interface, one monomer of TSA1 and one monomer from TSA2 with the corresponding B-type interacting residues were loaded. At the A-type interface, the hetero-dimer is expected to interact with a monomer from TSA1 and the A-type interacting residues of TSA1 were loaded accordingly.

HADDOCK yielded 4 clusters, comprising a total of 23 conformations. The four conformations are presented in Table 5.13.

Cluster No.	HADDOCK score	Cluster size	Z- score	RMSD aligning TSA1 trimer (A°)	RMSD aligning TSA2 trimer (A°)
2	-333.8 +/- 20.7	7	-1	13.018	13.7
1	-332.8 +/- 16.2	8	-0.9	9.867	11.7
4	-302.8 +/- 8.0	4	0.6	8.951	10.5
3	-289.4 +/- 16.5	4	1.3	13.749	14.8

Table 5.13: The table depicts the four conformations of the TSA1 and TSA2 trimers using combination 2 of monomers. The table contains information regarding the cluster, the HADDOCK score, the cluster size, the RMSD of the lowest energy, the z-score and the RMSD while aligning to the TSA1 and TSA2 trimers.

Using the combination 2 of monomers, the structural predictions generated by HADDOCK do not appear to exhibit similarities to TSA1 or TSA2.

Combination 3 of hetero-trimerisation

In this configuration, at the B-type interface, one monomer of TSA2 and one monomer from TSA1 with the corresponding B-type interacting residues were loaded. At the A-type interface, the hetero-dimer is expected to interact with a monomer from TSA1 and the A-type interacting residues of TSA1 were loaded accordingly.

HADDOCK yielded 3 clusters, comprising a total of 15 conformations. The three conformations are presented in Table 5.14.

Cluster No.	HADDOCK score	Cluster size	Z- score	RMSD aligning TSA1 trimer (A°)	RMSD aligning TSA2 trimer (A°)
2	-331.0 +/- 53.7	5	-1.4	1.513	6.4
1	-301.6 +/- 6.2	6	0.6	27.160	26.8
3	-297.4 +/- 12.0	4	0.8	26.056	25.8

Table 5.14: The table depicts the three conformations of the TSA1 and TSA2 trimers using combination 3 of monomers. The table contains information regarding the cluster, the HADDOCK score, the cluster size, the RMSD of the lowest energy, the z-score and the RMSD while aligning to the TSA1 and TSA2 trimers.

Using the combination 3 of monomers, the structures of cluster 2 exhibit a high degree of similarity to the reference A-type and B-type interfaces of TSA1 and TSA2. Furthermore, the low RMSD of atomic position value calculated lends additional support to this conclusion.

Combination 4 of hetero-trimerisation

In this configuration, at the B-type interface, one monomer of TSA1 and one monomer from TSA2 with the corresponding B-type interacting residues were loaded. At the A-type interface, the hetero-dimer is expected to interact with a monomer from TSA2 and the A-type interacting residues of TSA2 were loaded accordingly.

HADDOCK yielded 4 clusters, comprising a total of 23 conformations. The four conformations are presented in Table 5.15.

Using combination 4 of monomers, the structures of clusters 1 and 3 exhibit a high degree of similarity to the reference A- and B-type interfaces of TSA1 and TSA2. Furthermore, the low RMSD of atomic position value calculated lends additional support to this conclusion.

Cluster No.	HADDOCK score	Cluster size	Z- score	RMSD aligning TSA1 trimer (A°)	RMSD aligning TSA2 trimer (A°)
3	-312.5 +/- 41.1	4	-1.2	3.022	1.9
4	-306.4 +/- 14.1	4	-0.3	12.055	12.3
1	-304.4 +/- 8.8	9	0	7.541	7.0
2	-294.0 +/- 4.8	6	1.5	24.817	23.8

Table 5.15: The table depicts the four conformations of the TSA1 and TSA2 trimers using combination 4 of monomers. The table contains information regarding the cluster, the HADDOCK score, the cluster size, the RMSD of the lowest energy, the z-score and the RMSD while aligning to the TSA1 and TSA2 trimers.

Combination 5 of hetero-trimerisation

In this configuration, at the B-type interface, one monomer of TSA2 and one monomer from TSA1 with the corresponding B-type interacting residues were loaded. At the A-type interface, the hetero-dimer is expected to interact with a monomer from TSA2 and the A-type interacting residues of TSA2 were loaded accordingly.

HADDOCK yielded 3 clusters, comprising a total of 15 conformations. The three conformations are presented in Table 5.16.

Cluster No.	HADDOCK score	Cluster size	Z- score	RMSD aligning TSA1 trimer (A°)	RMSD aligning TSA2 trimer (A°)
1	-319.0 +/- 13.8	7	-1.4	13.841	13.6
2	-253.0 +/- 32.2	4	0.4	26.325	25.6
3	-228.9 +/- 63.1	4	1.0	5.478	5.3

Table 5.16: The table depicts the three conformations of the TSA1 and TSA2 trimers using combination 5 of monomers. The table contains information regarding the cluster, the HADDOCK score, the cluster size, the RMSD of the lowest energy, the z-score and the RMSD while aligning to the TSA1 and TSA2 trimers.

Using combination 5 of monomers, the structures of cluster 3 exhibit a high degree of similarity to the reference A- and B-type interfaces of TSA1 and TSA2. Furthermore, the low RMSD of atomic position value calculated lends additional support to this conclusion.

Combination 6 of hetero-trimerisation

In this configuration, at the B-type interface, two monomer of TSA2 with the corresponding B-type interacting residues were loaded. At the A-type interface, the hetero-dimer is expected to interact with a monomer from TSA1 and the A-type interacting residues of TSA1 were loaded accordingly.

HADDOCK yielded 8 clusters, comprising a total of 38 conformations. The eight conformations are presented in Table 5.17.

Cluster No.	HADDOCK score	Cluster size	Z- score	RMSD aligning TSA1 trimer (A°)	RMSD aligning TSA2 trimer (A°)
1	-307.6 +/- 11.8	7	-1.5	9.33	11.1
2	-301.0 +/- 8.5	6	-1.1	7.745	9.6
5	-283.0 +/- 24.9	4	-0.2	24.717	24.8
3	-281.8 +/- 33.8	5	-0.2	3.519	6.8
6	-277.3 +/- 26.4	4	0	21.907	22.2
8	-267.8 +/- 20.8	4	0.5	5.535	8.0
4	-267.4 +/- 34.4	4	0.5	7.293	9.4
7	-238.0 +/- 41.4	4	2	25.21	25.7

Table 5.17: The table depicts the eight conformations of the TSA1 and TSA2 trimers using combination 6 of monomers. The table contains information regarding the cluster, the HADDOCK score, the cluster size, the RMSD of the lowest energy, the z-score and the RMSD while aligning to the TSA1 and TSA2 trimers.

Using combination 6 of monomers, the structures of clusters 2, 3, 4 and 8 exhibit a high degree of similarity to the reference A- and B-type interfaces of TSA1 and TSA2. Furthermore, the low RMSD of atomic position value calculated lends additional support to this conclusion.

B-type PrxA dimers from *Arabidopsis thaliana*

The objective was to replicate the B-type dimers from the monomers of PrxA *Arabidopsis thaliana*.

HADDOCK yielded 11 clusters, comprising a total of 377 conformations. The ten most favourable conformations are presented in Table 5.18.

The structures of clusters 1, 3, 4, 5, and 6 exhibit a high degree of alignment with the reference dimer of PrxA. Furthermore, the RMSD of atomic position value calculated lends additional support to this conclusion. Therefore, the B-type dimers of PrxA were successfully replicated using only one monomer sequence.

Cluster No.	HADDOCK score	Cluster size	RMSD from overall lowest energy (A°)	Z- score	RMSD aligning PrxA dimer (A°)
1	-281.3 +/- 4.7	224	0.4 +/- 0.3	-3.0	3.8
2	-100.0 +/- 3.4	32	16.8 +/- 0.5	-0.1	15.9
3	-85.8 +/- 1.8	31	4.5 +/- 1.4	0.1	3.9
8	-73.3 +/- 13.1	10	10.2 +/- 0.3	0.3	10.1
4	-71.3 +/- 9.7	20	6.6 +/- 0.9	0.4	7.1
9	-71.2 +/- 2.8	7	18.3 +/- 0.3	0.4	17.1
6	-68.0 +/- 7.6	13	4.6 +/- 0.4	0.4	6.1
10	-63.2 +/- 9.8	6	18.2 +/- 0.3	0.5	17.3
7	-62.4 +/- 4.7	11	21.1 +/- 0.2	0.5	22.1
5	-61.8 +/- 2.1	17	9.4 +/- 0.7	0.5	9.7

Table 5.18: The table depicts the ten most similar conformations to the PrxA B-type dimer. The table contains information regarding the cluster, the HADDOCK score, the cluster size, the RMSD of the lowest energy, the z-score and the RMSD while aligning to the PrxA B-type dimer.

A-type PrxA dimers from *Arabidopsis thaliana*

The objective was to replicate the A-type dimers from the monomers of PrxA *Arabidopsis thaliana*.

HADDOCK yielded 7 clusters, comprising a total of 177 conformations. The seven conformations are presented in Table 5.19.

Cluster No.	HADDOCK score	Cluster size	RMSD from overall lowest energy (A°)	Z- score	RMSD aligning PrxA dimer (A°)
1	-108.3 +/- 3.8	90	7.2 +/- 0.2	-1.8	5.1
2	-98.2 +/- 6.0	51	6.3 +/- 0.4	-1.1	14.4
5	-82.1 +/- 10.5	7	9.7 +/- 1.5	0.0	3.8
7	-78.5 +/- 26.3	4	1.9 +/- 1.2	0.3	10.9
3	-77.9 +/- 5.1	10	7.5 +/- 0.3	0.3	5.9
6	-70.6 +/- 7.1	6	14.3 +/- 0.9	0.8	6.7
4	-62.9 +/- 4.1	9	16.4 +/- 0.8	1.4	12.3

Table 5.19: The table depicts the seven most similar conformations to the PrxA A-type dimer. The table contains information regarding the cluster, the HADDOCK score, the cluster size, the RMSD of the lowest energy, the z-score and the RMSD while aligning to the PrxA A-type dimer.

The structures of clusters 1, 3, 5 and 6 exhibit a high degree of alignment with the reference dimer of PrxA. Furthermore, the RMSD of atomic position value calculated lends additional support to this conclusion. Therefore, the B-type dimers of PrxA were successfully replicated using only one monomer sequence.

A-type and B-type interface interactions in PrxA from *Arabidopsis thaliana*

The objective was to replicate the A-type and B-type interface interactions observed in the monomers of *Arabidopsis thaliana* PrxA.

HADDOCK yielded 4 clusters, comprising a total of 17 conformations. The four conformations are presented in Table 5.20.

Cluster No.	HADDOCK score	Cluster size	RMSD from overall lowest energy (A°)	Z- score	RMSD aligning PrxA trimer (A°)
1	-298.4 +/- 25.7	5	3.0 +/- 0.9	-1.3	12.6
2	-271.8 +/- 17.1	4	12.1 +/- 0.6	-0.6	4.8
3	-223.0 +/- 46.0	4	6.7 +/- 0.4	0.6	11.1
4	-191.8 +/- 59.4	4	6.2 +/- 1.0	1.3	13.4

Table 5.20: The table depicts the four conformations to the PrxA A-type and B-type trimer. The table contains information regarding the cluster, the HADDOCK score, the cluster size, the RMSD of the lowest energy, the z-score and the RMSD while aligning to the PrxA trimer.

The structure of cluster 2 exhibits a high degree of alignment with the reference A-type and B-type interfaces of PrxA *Arabidopsis thaliana*. Furthermore, the RMSD of atomic position value calculated lends additional support to this conclusion. It was therefore demonstrated that the A-type and B-type interfaces of PrxA could be successfully replicated using only one monomer sequence.

Cross species hetero-dimerisation at B-type interface

The objective was to dock a monomer derived from peroxiredoxins of *Saccharomyces cerevisiae* and a monomer from PrxA of *Arabidopsis thaliana* at B-type interfaces. Therefore, for the input molecules, one chain from TSA1 or TSA2 of *S. cerevisiae* was loaded as one molecule and a chain from PrxA of *A. thaliana* was loaded as the other molecule, with the active and passive residues at the B-type interface of the corresponding monomers.

Upon docking a monomer from TSA1 and a monomer from PrxA at B-type interface, the HADDOCK yielded 8 clusters comprising a total of 183 conformations. The eight conformations are presented in Table 5.21. Figures B.1 and B.2 illustrate the eight aligned complexes of the predicted structures and TSA1 and PrxA dimers, which are aligned in accordance with the eight predicted structures.

The structures of cluster 1, 4 and 8 exhibit a high degree of alignment with the reference B-type interfaces of TSA1 and PrxA. Furthermore, the RMSD of atomic position value calculated lends additional support to this conclusion.

Upon docking a monomer from TSA2 and a monomer from PrxA at B-type interface, HADDOCK yielded 7 clusters comprising a total of 178 conformations. The seven conformations are presented in Table 5.22. Figure

Cluster No.	HADDOCK score	Cluster size	Z- score	RMSD aligning TSA1 dimer (A°)	RMSD aligning PrxA dimer (A°)
1	-254.1 +/- 6.2	52	-2.5	1.650	3.8
5	-117.6 +/- 4.6	18	0.0	20.225	20.8
2	-114.0 +/- 1ß.7	41	0.1	21.032	21.0
7	-111.2 +/- 10.7	10	0.1	14.475	13.8
4	-96.7 +/- 11.6	19	0.4	7.949	7.9
6	-95.7 +/- 6.5	16	0.4	20.521	21.4
3	-89.3 +/- 5.7	21	0.5	9.505	10.4
8	-59.8 +/- 11.3	6	1.1	7.433	8.0

Table 5.21: The table depicts the eight of the hetero-dimer at B-type interface. The table contains information regarding the cluster, the HADDOCK score, the cluster size, the RMSD of the lowest energy, the z-score and the RMSD while aligning to the TSA1 and PrxA dimer.

B.3 and B.4 illustrates the seven aligned complexes of the predicted structures and TSA2 and PrxA dimers, which are aligned in accordance with the seven predicted structures.

Cluster No.	HADDOCK score	Cluster size	Z- score	RMSD aligning TSA2 dimer (A°)	RMSD aligning PrxA dimer (A°)
3	-201.1 +/- 16.5	26	-2.4	1.862	4.0
4	-97.4 +/- 17.0	14	0.2	7.987	10.6
2	-91.7 +/- 6.5	32	0.3	19.011	21.6
7	-91.5 +/- 7.5	4	0.3	17.976	21.7
5	-90.7 +/- 15.3	8	0.3	5.800	5.3
1	-90.4 +/- 1.5	88	0.3	17.844	21.5
6	-62.8 +/- 5.1	6	1.0	14.183	14.0

Table 5.22: The table depicts the seven of the hetero-dimer at B-type interface. The table contains information regarding the cluster, the HADDOCK score, the cluster size, the RMSD of the lowest energy, the z-score and the RMSD while aligning to the TSA2 and PrxA dimer.

The structures of cluster 3 and 5 exhibit a high degree of alignment with the reference B-type interfaces of TSA2 and PrxA. Furthermore, the RMSD of atomic position value calculated lends additional support to this conclusion.

Cross species hetero-dimerisation at A-type interface

The objective was to dock a monomer derived from peroxiredoxins of *Saccharomyces cerevisiae* and a monomer from PrxA of *Arabidopsis thaliana* at A-type interfaces. Therefore, for the input molecules, one chain from TSA1 or TSA2 of *S. cerevisiae* was loaded as one molecule and a chain from PrxA of *A. thaliana* was loaded as the other molecule, with the active and passive residues at the A-type interface of the corresponding monomers.

Upon docking a monomer from TSA1 and a monomer from PrxA at A-type interface, the HADDOCK yielded 6 clusters comprising a total of 191 conformations. The six conformations are presented in Table 5.23. Figures B.5 and B.6 illustrates the six aligned complexes of the predicted structures and TSA1 and PrxA dimers, which are aligned in accordance with the six predicted structures.

Cluster No.	HADDOCK score	Cluster size	Z- score	RMSD aligning TSA1 dimer (A°)	RMSD aligning PrxA dimer (A°)
1	-118.0 +/- 0.9	144	-1.7	3.504	4.5
2	-94.6 +/- 3.2	18	-0.7	1.554	3.5
6	-73.9 +/- 20.3	4	0.2	6.799	7.2
3	-71.9 +/- 9.8	13	0.3	12.872	13.3
4	-67.3 +/- 16.6	7	0.5	7.603	8.4
5	-46.0 +/- 16.0	5	1.4	5.590	6.8

Table 5.23: The table depicts the six of the hetero-dimer at A-type interface. The table contains information regarding the cluster, the HADDOCK score, the cluster size, the RMSD of the lowest energy, the z-score and the RMSD while aligning to the TSA1 and PrxA dimer.

The structure of clusters 1, 2, 4, 5 and 6 exhibit a high degree of alignment with the reference A-type interfaces of TSA1 and PrxA. Furthermore, the RMSD of atomic position value calculated lends additional support to this conclusion.

Upon docking a monomer from TSA2 and a monomer from PrxA at A-type interface, HADDOCK yielded 10 clusters comprising a total of 154 conformations. The ten conformations are presented in Table 5.24. Figures B.7 and B.8 illustrate the ten aligned complexes of the predicted structures and TSA2 and PrxA dimers, which are aligned in accordance with the ten predicted structures.

The structure of clusters 1, 4, 6, 7, 8 and 9 exhibit a high degree of alignment with the reference A-type interfaces of TSA2 and PrxA. Furthermore, the RMSD of atomic position value calculated lends additional support to this conclusion.

Cluster No.	HADDOCK score	Cluster size	Z- score	RMSD aligning TSA2 dimer (A°)	RMSD aligning PrxA dimer (A°)
1	-76.2 +/- 3.5	39	-2.0	10.410	7.4
2	-67.4 +/- 3.0	33	-1.2	14.918	13.5
6	-61.9 +/- 7.0	9	-0.6	10.355	6.4
5	-56.5 +/- 9.1	13	-0.1	15.295	13.6
4	-54.4 +/- 1.2	14	0.1	9.927	6.3
9	-51.4 +/- 10.0	5	0.3	10.480	8.2
3	-50.8 +/- 11.8	22	0.4	16.514	15.4
8	-48.8 +/- 4.9	2	0.6	10.440	7.4
7	-45.4 +/- 10.7	8	0.9	11.028	8.4
10	-37.6 +/- 14.4	4	1.6	14.761	12.1

Table 5.24: The table depicts the ten of the hetero-dimer at A-type interface. The table contains information regarding the cluster, the HADDOCK score, the cluster size, the RMSD of the lowest energy, the z-score and the RMSD while aligning to the TSA2 and PrxA dimer.

5.3.2 Results from AlphaFold

The objective was to replicate the decamer structure of TSA1 and TSA2 from *S. cerevisiae* and PrxA from *A. thaliana* using AlphaFold 3. Ten corresponding monomers were input into AlphaFold 3 in order to predict the complex structure. The AlphaFold 3 algorithm yielded five top-predicted structures for each peroxiredoxin.

Decamers of TSA1 from *Saccharomyces cerevisiae*

The decamer structure of TSA1 was replicated through the loading of ten monomers in AlphaFold, which yielded five predicted structures. The scores and RMSD values obtained by comparing the predicted structure with the experimental structure are presented in Table 5.25. Figure B.9 illustrates the five aligned complexes of the predicted structures and the experimental decamers of TSA1.

Predicted structure	Fraction disordered	Has_clash	iptm	ptm	Ranking score	RMSD align to TSA1 (A°)
1	0.01	0.0	0.84	0.86	0.85	1.5
2	0.0	0.0	0.84	0.85	0.84	1.5
3	0.0	0.0	0.84	0.85	0.84	1.6
4	0.01	0.0	0.83	0.85	0.84	1.7
5	0.0	0.0	0.81	0.82	0.81	1.5

Table 5.25: The table presents the scores of the five predicted structures using the AlphaFold algorithm and the RMSD values while aligning to the TSA1 decamer. The table presents a range of data regarding various scores, including those pertaining to fraction disordered, has_clash, Iptm, Ptm, and ranking score.

As indicated by the RMSD values, all predicted structures exhibit an average degree of similarity. The results of the ptm and iptm scores also

indicate that the accuracy of the predicted complex and interface structures is high.

Decamers of TSA2 from *Saccharomyces cerevisiae*

The decamer structure of TSA2 was replicated through the loading of ten monomers in AlphaFold, which yielded five predicted structures. The scores and RMSD values obtained by comparing the predicted structure with the experimental structure are presented in Table 5.26. Figure B.10 illustrates the five aligned complexes of the predicted structures and the experimental decamers of TSA2.

Predicted structure	Fraction disordered	Has_clash	iptm	ptm	Ranking score	RMSD align to TSA2 (A°)
1	0.03	0.0	0.85	0.86	0.86	47.9
2	0.03	0.0	0.85	0.86	0.86	47.8
3	0.03	0.0	0.85	0.86	0.86	47.9
4	0.02	0.0	0.84	0.86	0.85	47.8
5	0.01	0.0	0.83	0.85	0.84	48.0

Table 5.26: The table presents the scores of the five predicted structures using the AlphaFold algorithm and the RMSD values while aligning to the TSA2 decamer. The table presents a range of data regarding various scores, including those pertaining to fraction disordered, has_clash, Iptm, Ptm, and ranking score.

As determined by RMSD values, no structures demonstrated a high degree of similarity. However, the outcomes of the ptm and iptm scores suggest that the accuracy of all the predicted complex and interface structures is high.

Decamers of PrxA from *Arabidopsis thaliana*

The decamer structure of PrxA was replicated through the loading of ten monomers in AlphaFold, which yielded five predicted structures. The scores and RMSD values obtained by comparing the predicted structure with the experimental structure are presented in Table 5.27. Figure B.11 illustrates the five aligned complexes of the predicted structures and the experimental decamers of PrxA.

Predicted structure	Fraction disordered	Has_clash	iptm	ptm	Ranking score	RMSD align to PrxA (A°)
1	0.0	0.0	0.79	0.81	0.79	39.8
2	0.0	0.0	0.79	0.81	0.79	39.9
3	0.0	0.0	0.79	0.81	0.79	39.9
4	0.0	0.0	0.79	0.8	0.79	39.9
5	0.0	0.0	0.79	0.8	0.79	40.0

Table 5.27: The table presents the scores of the five predicted structures using the AlphaFold algorithm and the RMSD values while aligning to the PrxA decamer. The table presents a range of data of various scores, including those pertaining to fraction disordered, has_clash, Iptm, Ptm, and ranking score.

As determined by RMSD values, no structures demonstrated a high degree of similarity. However, the ptm and iptm scores suggest that the accuracy of all the predicted complex and interface structures is considerable.

5.4 Conclusion

The HADDOCK method has been demonstrated to be an effective means of replicating the A-type and B-type homo dimerisation and trimerisation of the peroxiredoxins TSA1 and TSA2 of *S. cerevisiae* and PrxA of *A. thaliana*. Furthermore, HADDOCK was able to successfully predict heterodimer complexes from monomers of TSA1 and TSA2, which exhibited close similarity in RMSD values to the homodimer structures of TSA1 and TSA2. However, HADDOCK was unsuccessful in predicting the decameric structures of TSA1, and TSA2 from the ten corresponding monomers. In the attempt to create hetero-trimers at the A-type and B-type interfaces with six combinations of monomers of TSA1 and TSA2, only one combination (Combination 2, a monomer from TSA2 and a monomer from TSA2 at the B-type interface and a monomer from TSA1 at the A-type interface) did not yield structures similar to those of the trimers of TSA1 or TSA2. The predicted structures of the remaining combinations exhibited at least one cluster of trimer structures with notable similarities. In the cross-species dimerisation of docking a monomer from TSA1 or TSA2 and another from PrxA at the B-type interface, HADDOCK was successful in identifying a cluster that showed similarity to both peroxiredoxin from *S. cerevisiae* and PrxA of *A. thaliana*. While cross-species dimerisation of docking a monomer from TSA1 or TSA2 and another from PrxA at the A-type interface, HADDOCK was successful in identifying a cluster that showed similarity to both peroxiredoxin from *S. cerevisiae* and PrxA of *A. thaliana*. In conclusion, while HADDOCK predicted a number of structures, the majority of these exhibited RMSD values higher than 2 A° when compared to the experimental structures. Consequently, these structures may not be regarded as optimal for further investigation.

The replication of the dimer and trimer complex was not feasible using AlphaFold, as there was no option to provide the active binding sites for the B-type and A-type interfaces. Consequently, an attempt was made to replicate the decamer complex for TSA1, TSA2 and PrxA. RMSD scores were used to calculate the average distance between the atoms of the predicted and the experimentally deposited protein complex [231]. In addition, Alphafold generated further scores such as predicted Template Modelling (pTM) and Interface Predicted Template Modelling (ipTM) for each prediction. These scores provide the global topological structure and are insensitive to local variations in the orientations of the flexible loop or tail structures [232]. The pTM score ranges from 0 to 1, with scores above 0.5 indicating a structure that is close to the hypothetical true structure and scores below 0.5 indicating less confident predictions. Similarly, the ipTM score is a metric that quantifies the accuracy of the predicted relative positions of subunits. This score ranges from 0 to 1, with scores in the 0.6 to 0.8 range falling within what is termed the 'grey zone'.

In this zone, the predicted relative positions of the structures may or may not be correct. Scores above 0.8 are considered to be high quality predictions, while scores below 0.5 are considered to be low quality predictions.

In the present study, the predicted structures were compared to the corresponding TSA1, TSA2 and PrxA protein decameric structures using Alphafold. The RMSD, pTM and ipTM scores were estimated. A notable observation was that, in the comparison of the predicted structures with those of TSA1, the pTM and iPTM scores were sufficiently high to substantiate confident predictions, while the RMSD scores were sufficiently low to affirm the similarity of the predicted structures to the experimental ones. Conversely, when the experimental structures of TSA2 and PrxA were compared with their respective predicted structures, it was observed that the pTM and the ipTM scores were sufficiently high to substantiate the reliability of the predicted structures. However, a high deviation was observed in the RMSD scores when compared to the experimental structures. To further analyse this in detail, a comparison was made of the dimers at the A-type and B-type interfaces of the predicted structures and the experimental structures. The RMSD scores indicated that the dimers exhibited a high degree of similarity to the experimental structures. Subsequently, the trimer structures encompassing one A-type and one B-type interface of the predicted and the experimental structures were compared. The RMSD scores indicated a significant disparity between the predicted and the experimental structures. The analysis revealed significant variations in the folding of protein structures at the tails. In the predicted structures, the tails of the complex folded as helixes, but in the experimental structures, no helix formation was observed at the tails. These local variations, leading to larger distances between atoms, could be the reason for the high RMSD scores. The incorporation of helixes by Alphafold is indicative of enhanced tail stability, consequently leading to elevated pTM and ipTM scores. The absence of tail helices in the experimental structures may be attributable to external factors, such as the buffer utilized. Consequently, it can be deduced that, while Alphafold claimed being able to predict stable decameric structures of peroxiredoxins, significant local variations were observed between its structures and the experimental ones.

Chapter 6
Conclusion and Outlook

The central objective of this thesis was to further advance the understanding of protein-protein interactions, with a particular focus on their collective functions as clusters within protein-protein protein interaction networks that drive cellular state transitions. The thesis also encompassed protein abundance analyses derived from technologies such as mass spectrometry. Such analyses are of relevance to interdisciplinary fields, including dental research, where they are employed to identify materials that can substitute for enamel. Furthermore, the analyses are utilised to study protein adsorption patterns on pellicle deposited in the oral cavity, and to examine the differences in protein adsorption patterns in various active, non-active, and treated caries conditions. The thesis concludes with a computational study on peroxiredoxin decameric protein complex predictions by docking monomers using HADDOCK and Alphafold.

6.1 Analysis of transcriptomics data

In chapter 3, six pipelines were constructed for the analysis of melanoma and nevi transcriptomics sample data, with the objective of determining the potential drivers of a healthy state to a diseased state. Pipelines 0, I and II involved the processing of differential analyses prior to the construction of the protein-protein interaction network. In contrast, pipelines IV and V involved the processing of data where network construction preceded differential analyses. Pipeline III considered only rewiring events during network construction. The results from all the pipelines were diverse, with the pipelines 0, I and II yielding larger and more numerous clusters. This may be due to the fact that network construction was carried out using available databases such as STRING and IntAct, where protein interaction evidence was supported not only by literature evidence but also by experimental and predicted evidence.The enrichment analysis performed on these predicted

clusters also showed results that were not specific to the initiation or progression of melanoma. In contrast, the predicted clusters from pipelines III, IV and V were comparatively smaller in size. This was due to the fact that condition-specific networks were constructed in the initial phase, at which stage only the nodes of the networks were dependent on the expression of the corresponding genes of that sample.The enrichment analysis on these clusters resulted in pathways that were more specific to melanoma development and progression. The clusters identified in pipelines II, IV and V were found to be enriched with significant Reactome pathways that have demonstrated evidence of melanoma development in the extant literature. These pathways include the MAPK pathway [170, 233], PI3K-AKt pathway [172, 234], RAS pathway [235, 236], BRAF signalling pathway [237, 238], PTEN signalling pathway [239, 240], WNT signalling pathway [241, 242], amongst others.

In the subnetwork analysis, the clusters that were found to be overexpressed, underexpressed, or absent from nevi and melanoma samples were more clearly identified and better interpretable. The enrichment analyses of these clusters provided a more profound insight into the reactome pathways that were significantly enriched in melanoma samples, exclusively enriched in nevi samples, or enriched in melanoma samples. The findings of these differential clusters will inform future research on improving the diagnosis of changes in cell state and even targeted therapy for diseased conditions.

The significant clusters predicted in the networks and subnetworks exhibited a high degree of overlap with the existing complexes deposited in Complex Portal. This finding serves to validate the prediction of clusters. However, no gold standard reference existed to validate the predicted clusters for specific or diseased conditions. Consequently, there is currently no existing knowledgebase that provides the list of complexes that drive a healthy cell to a melanoma condition, against which our predicted clusters can be compared and ranked.

6.2 Analysis of proteomics data

In the comparative proteomic analysis of HAP and enamel to determine if HAP can be used as a standard material for enamel studies and research, it can be concluded that there exists no statistically significant deviations in the absorption of proteins or the formation of biofilms on these two substrates. Overall, the fundamental processes, encompassing pellicle formation, bacterial colonization and subject-specific responses, are observed to occur in a comparable manner on synthetic HAP compared to natural enamel surfaces. This finding thus demonstrates the aptitude of HAP for use in biofilm investigations. Consequently, it can be hypothesised that HAP may serve as a suitable replacement for enamel in subsequent research studies.

In the comparative proteomic analysis of saliva and pellicle, it can be concluded that some proteins do not absorb in the oral cavity to form pellicles. Further investigation revealed that these proteins that remain in saliva exhibited higher catalytic activities than the structural molecule activities

predominantly seen in pellicle proteins, suggesting a role in the formation of pellicle structures on oral surfaces. Furthermore, an analysis of the molecular weight distribution patterns of these salivary and pellicle proteins reveals that the overall distribution of molecular weights of the salivary proteins were significantly smaller than that of pellicle proteins, with no significant correlation values also. Furthermore, an analysis of the isoelectric points of these proteins reveals that the majority of salivary proteins are found within the pH range of the oral cavity, while the majority of pellicle proteins are found in the pH range higher than that of the oral cavity (i.e. above 7.6).

In the comparative proteomic analysis of pellicle and saliva data from active caries, treated caries and healthy conditions, it was concluded that the protein composition, especially in pellicle, of the treated condition reverts back and expresses high similarity to the healthy conditions, as seen in the PCA plots. Fold change analysis reveals a significant upregulation of 17 proteins and downregulation of 9 proteins in saliva from caries-active conditions compared to healthy conditions, while a 20 proteins were exclusively downregulated in pellicle data. In the case of treated samples, only 4 proteins were significantly upregulated and 36 proteins were significantly downregulated compared to caries active conditions in salivary data. However, a contrasting pattern was noted in pellicle data, where a greater number of proteins (24 proteins) were found to be upregulated while 4 proteins were downregulated. A further comparison of the salivary and pellicle data sets reveals a significant number of proteins with differential expression in the respective conditions. The molecular function enrichment analysis reveals a higher degree of structural molecule activity in pellicle proteins than in salivary proteins. Conversely, salivary proteins demonstrate a higher propensity for catalytic activity in comparison to pellicle proteins.

The present project encountered several drawbacks. For instance, the samples obtained from the volunteers lacked sufficient replicate data to facilitate a reliable statistical analysis. There was considerable variability both between the available replicate data and between the inter-volunteer data. In order to achieve a higher level of confidence in the results, it is desirable increase the number of volunteers and the number of replicate measurements.

6.3 Protein complex predictions of peroxiredoxins

Peroxiredoxin, an antioxidant protein of a ubiquitous family with multiple isoforms, is expressed in many organisms. However, they have a propensity to oligomerise, forming a characteristic globular doughnut-shaped structure comprising five homodimers. Several recent studies have suggested the possibility of hetero-oligomerisation at dimeric or decameric structures with monomers of close isoforms. The present study aims to determine the plausibility of such hetero-oligomerisation by employing in silico modelling tools such as HADDOCK and Alphafold. The HADDOCK approach was employed to replicate the A-type and B-type homo dimerisation, trimerisation and decamer formation of the peroxiredoxins TSA1 and TSA2 of *S. cerevisiae* and PrxA of

A. thaliana, with the objective of validating the docked results. The tool successfully docked the dimers, trimers and decamers at the A-type and B-type active site interfaces, showing high similarity to the experimental structures. Subsequently, monomers from disparate isomers were introduced into the tool in an effort to induce hetero-oligomerisation in a range of combinations. The HADDOCK program demonstrated success in docking hetero-oligomers with a comparable degree of similarity to the corresponding experimental structures, utilising the majority of the combination of monomers. Furthermore, HADDOCK was also successful in docking monomers in a cross-species manner, with an average degree of similarity to the corresponding species. While it is evident that the resulting conformers bear a strong resemblance to the experimental structures and the HADDOCK scores are indicative of efficient binding energies, it is crucial to acknowledge that HADDOCK is a data-driven tool. Consequently, it is highly sensitive to inaccuracies in the experimental data, which can significantly impact the outcomes, particularly at interfaces [243]. While the hetero-oligomer structures were predicted, there was a lack of deposited experimental hetero-oligomer structures to facilitate a comparison between the predicted structure and the experimental data. This would have provided a more robust validation of the predicted structures.

Utilising AlphaFold, the replication of the dimers and trimer structures proved unfeasible, as the functionality to feed active interface information of specific A-type and B-type interfaces was not available. Consequently, an attempt was made to replicate the complete decameric structures. While AlphaFold yielded highly confident predicted structures, a comparison of these with the corresponding experimental structures revealed high RMSD scores, thus indicating that the predicted structures did not resemble the experimental structures. Upon closer observation, it was noted that there was a high degree of local variations in the tail regions from the monomer sequences. It was also observed that Alphafold incorporated helices in instances where the experimental structures did not exhibit any helices. The tail regions were found to be enriched with alanine and leucine. These amino acids have been identified as favourable in the formation of helices. Consequently, it is quite plausible why Alphafold scored its predictions as "accurate". However, external factors, such as the buffer used to store the peroxiredoxins prior to the X-ray diffraction snapshot, may have influenced the folding process, resulting in the tails not folding into helices. However, given the observed discrepancy between the RMSD and TM scores, the hetero-oligomerisation process was not pursued.

APPENDIX

Appendix A
Supplementary Material for chapter 3

A.1 N vs M samples

Table A.1: Listed are the significant reactome pathways enriched in clusters predicted by ClusterONE from the differential network comparing all N samples with all M samples using pipeline 0.

Cluster	Size	Reactome pathways	Adjusted P-value
1	1466	Unclassified (UNCLASSIFIED) Translation (R-HSA-72766) Cellular responses to stress (R-HSA-2262752)	7.790e-71 8.380e-58 1.660e-53
2	232	Transport of small molecules (R-HSA-382551) SLC-mediated transmembrane transport (R-HSA-425407) Transport of bile salts and organic acids, metal ions and amine compounds (R-HSA-425366)	5.630e-04 8.730e-04 1.020e-03
3	128	Keratinization (R-HSA-6805567) Developmental Biology (R-HSA-1266738)	1.670e-43 3.950e-18
4	15	Chemokine receptors bind chemokines (R-HSA-380108) Peptide ligand-binding receptors (R-HSA-375276) Class A/1 (Rhodopsin-like receptors) (R-HSA-373076)	6.610e-28 1.090e-21 5.460e-19
5	36	Chemokine receptors bind chemokines (R-HSA-380108) Peptide ligand-binding receptors (R-HSA-375276) Class A/1 (Rhodopsin-like receptors) (R-HSA-373076)	4.730e-23 2.050e-16 1.500e-13
6	20	Chemokine receptors bind chemokines (R-HSA-380108) Peptide ligand-binding receptors (R-HSA-375276) Class A/1 (Rhodopsin-like receptors) (R-HSA-373076)	4.800e-25 7.710e-19 3.730e-16
7	278	Signaling by Rho GTPases , Miro GTPases and RHOBTB3 (R-HSA-9716542) Signaling by Rho GTPases (R-HSA-194315) RHO GTPase cycle (R-HSA-9012999)	9.470e-10 9.960e-10 6.550e-08
8	19		
9	10		
10	16	Intrinsic Pathway for Apoptosis (R-HSA-109606) BH3-only proteins associate with and inactivate anti-apoptotic BCL-2 members (R-HSA-111453) Apoptosis (R-HSA-109581)	3.150e-11 2.330e-10 3.170e-08

11	7	Complement cascade (R-HSA-166658) Lectin pathway of complement activation (R-HSA-166662) Surfactant metabolism (R-HSA-5683826)	5.280e-03 6.550e-03 3.000e-02

Table A.2: Listed are the significant reactome pathways enriched in clusters predicted by ClusterONE from the network constructed by significantly downregulated genes in M samples with respect to N samples using pipeline 0.

Cluster	Size	Reactome pathways	Adjusted P-value
1	448	GTP hydrolysis and joining of the 60S ribosomal subunit (R-HSA-72706) L13a-mediated translational silencing of Ceruloplasmin expression (R-HSA-156827) Eukaryotic Translation Initiation (R-HSA-72613)	6.100e-66 7.670e-66 7.150e-65
2	72	Keratinization (R-HSA-6805567) Developmental Biology (R-HSA-1266738)	9.060e-51 6.110e-24
3	14		
4	20		
5	20		
6	39	Prolactin receptor signaling (R-HSA-1170546) Growth hormone receptor signaling (R-HSA-982772) STAT5 Activation (R-HSA-9645135)	5.800e-03 1.400e-02 3.360e-02
7	21		
8	18	Signaling by PDGFRA transmembrane, juxtamembrane and kinase domain mutants (R-HSA-9673767) Erythropoietin activates Phosphoinositide-3-kinase (PI3K) (R-HSA-9027276) Signaling by PDGFRA extracellular domain mutants (R-HSA-9673770)	1.090e-04 1.490e-04 1.630e-04
9	42	Formation of the cornified envelope (R-HSA-6809371) Keratinization (R-HSA-6805567) Developmental Biology (R-HSA-1266738)	6.800e-24 3.360e-20 2.340e-08
10	6	Intraflagellar transport (R-HSA-5620924) Cilium Assembly (R-HSA-5617833) Organelle biogenesis and maintenance (R-HSA-1852241)	1.580e-09 6.300e-07 2.820e-06
11	12	BBSome-mediated cargo-targeting to cilium (R-HSA-5620922) Cilium Assembly (R-HSA-5617833) Cargo trafficking to the periciliary membrane (R-HSA-5620920)	6.850e-04 3.450e-03 3.990e-03
12	25		
13	26		
14	18	Regulation of FOXO transcriptional activity by acetylation (R-HSA-9617629) FOXO-mediated transcription of cell cycle genes (R-HSA-9617828) FOXO-mediated transcription (R-HSA-9614085)	2.200e-07 1.240e-06 1.910e-06
15	38	Formation of the cornified envelope (R-HSA-6809371) Keratinization (R-HSA-6805567) Developmental Biology (R-HSA-1266738)	7.840e-25 3.940e-21 3.270e-09
16	5	Trafficking of myristoylated proteins to the cilium (R-HSA-5624138) Cargo trafficking to the periciliary membrane (R-HSA-5620920) Cilium Assembly (R-HSA-5617833)	1.770e-07 1.840e-04 7.640e-03
17	5	Complement cascade (R-HSA-166658) Regulation of Complement cascade (R-HSA-977606) Terminal pathway of complement (R-HSA-166665)	7.960e-09 1.020e-08 3.300e-07
18	21		

Table A.3: Listed are the significant reactome pathways enriched in clusters predicted by ClusterONE from the network constructed by significantly upregulated genes in M samples with respect to N samples using pipeline 0.

Cluster	Size	Reactome pathways	Adjusted P-value
1	23	Chemokine receptors bind chemokines (R-HSA-380108) Peptide ligand-binding receptors (R-HSA-375276) Class A/1 (Rhodopsin-like receptors) (R-HSA-373076)	3.690e-25 4.770e-18 3.220e-15
2	20	Chemokine receptors bind chemokines (R-HSA-380108) Peptide ligand-binding receptors (R-HSA-375276) Class A/1 (Rhodopsin-like receptors) (R-HSA-373076)	2.700e-29 1.490e-21 1.780e-18
3	149	Chk1/Chk2(Cds1) mediated inactivation of Cyclin B:Cdk1 complex (R-HSA-75035) Signaling by Rho GTPases, Miro GTPases and RHOBTB3 (R-HSA-9716542) Signaling by Rho GTPases (R-HSA-194315)	5.960e-07 1.240e-04 1.300e-04
4	13	Lectin pathway of complement activation (R-HSA-166662) Complement cascade (R-HSA-166658)	1.400e-02 3.960e-02
5	49		
6	6	Butyrophilin (BTN) family interactions (R-HSA-8851680) Adaptive Immune System (R-HSA-1280218) RUNX1 regulates transcription of genes involved in differentiation of keratinocytes (R-HSA-8939242)	7.780e-06 7.860e-04 1.700e-03
7	11	Complement cascade (R-HSA-166658) Lectin pathway of complement activation (R-HSA-166662) Initial triggering of complement (R-HSA-166663)	5.770e-04 5.730e-03 6.850e-03
8	24		
9	26	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell (R-HSA-198933) Endosomal/Vacuolar pathway (R-HSA-1236977) Immune System (R-HSA-168256)	8.030e-14 6.640e-12 1.580e-09
10	22	Metabolism of polyamines (R-HSA-351202) Regulation of ornithine decarboxylase (ODC) (R-HSA-350562)	3.840e-02 5.170e-02
11	11		
12	11	Intrinsic Pathway for Apoptosis (R-HSA-109606) BH3-only proteins associate with and inactivate anti-apoptotic BCL-2 members (R-HSA-111453) Apoptosis (R-HSA-109581)	2.270e-10 3.970e-09 1.230e-07
13	12		
14	36	G1/S Transition (R-HSA-69206) Mitotic G1 phase and G1/S transition (R-HSA-453279) Cell Cycle, Mitotic (R-HSA-69278)	9.960e-15 2.270e-14 2.860e-12
15	13		
16	104	Respiratory electron transport (R-HSA-611105) Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins. (R-HSA-163200) The citric acid (TCA) cycle and respiratory electron transport (R-HSA-1428517)	6.250e-27 4.650e-25 1.810e-23
17	7	Interferon alpha/beta signaling (R-HSA-909733) Interferon Signaling (R-HSA-913531) SARS-CoV-2 activates/modulates innate and adaptive immune responses (R-HSA-9705671)	3.220e-08 2.240e-06 3.360e-05
18	17	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell (R-HSA-198933) Immune System (R-HSA-168256) Endosomal/Vacuolar pathway (R-HSA-1236977)	1.130e-12 3.430e-09 2.190e-07

A.1.1 Enriched reactome pathways using pipeline 1

Table A.4: Listed are the significant reactome pathways enriched in clusters predicted by ClusterONE from the differential network comparing all N samples with all M samples using pipeline I.

Cluster	Size	Reactome pathways	Adjusted P-value

1	487	Immune System (R-HSA-168256) Cytokine Signaling in Immune system (R-HSA-1280215) Innate Immune System (R-HSA-168249)	5.430e-123 3.280e-69 2.280e-43
2	86	Keratinization (R-HSA-6805567) Developmental Biology (R-HSA-1266738) Formation of the cornified envelope (R-HSA-6809371)	1.140e-115 1.130e-71 2.080e-40
3	361	Cell Cycle (R-HSA-1640170) Cell Cycle, Mitotic (R-HSA-69278) Cell Cycle Checkpoints (R-HSA-69620)	9.490e-105 6.380e-95 2.360e-62
4	469	Translation (R-HSA-72766) Metabolism of RNA (R-HSA-8953854) rRNA processing (R-HSA-72312)	8.040e-119 2.660e-75 4.700e-75
5	60	Keratinization (R-HSA-6805567) Developmental Biology (R-HSA-1266738) Formation of the cornified envelope (R-HSA-6809371)	9.210e-96 2.260e-59 1.400e-20
6	517	Metabolism (R-HSA-1430728) The citric acid (TCA) cycle and respiratory electron transport (R-HSA-1428517) Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins. (R-HSA-163200)	4.960e-146 1.800e-40 1.530e-30
7	1084	Immune System (R-HSA-168256) Unclassified (UNCLASSIFIED) Signal Transduction (R-HSA-162582)	1.150e-94 5.000e-70 7.470e-64
8	45		
9	288	Signaling by WNT (R-HSA-195721) TCF dependent signaling in response to WNT (R-HSA-201681) Developmental Biology (R-HSA-1266738)	1.120e-14 4.510e-14 1.260e-09
10	347	Signaling by WNT (R-HSA-195721) TCF dependent signaling in response to WNT (R-HSA-201681) Developmental Biology (R-HSA-1266738)	5.790e-11 1.610e-10 1.340e-08
11	73		
12	36	Formation of the cornified envelope (R-HSA-6809371) Keratinization (R-HSA-6805567) Developmental Biology (R-HSA-1266738)	3.090e-15 3.680e-13 1.000e-05
13	40		
14	14	Keratinization (R-HSA-6805567) Developmental Biology (R-HSA-1266738)	1.060e-21 8.750e-13
15	125	Extracellular matrix organization (R-HSA-1474244) Collagen formation (R-HSA-1474290) Collagen biosynthesis and modifying enzymes (R-HSA-1650814)	9.320e-68 1.020e-37 3.760e-37
16	199	Developmental Biology (R-HSA-1266738) Transcriptional regulation of pluripotent stem cells (R-HSA-452723) POU5F1 (OCT4), SOX2, NANOG activate genes related to proliferation (R-HSA-2892247)	7.070e-08 3.920e-06 4.200e-06
17	75		
18	6		
19	17	Amino acid transport across the plasma membrane (R-HSA-352230) Transport of inorganic cations/anions and amino acids/ oligopeptides (R-HSA-425393) SLC-mediated transmembrane transport (R-HSA-425407)	3.480e-19 3.180e-15 3.350e-12
20	253	Neuronal System (R-HSA-112316) Transmission across Chemical Synapses (R-HSA-112315) Protein-protein interactions at synapses (R-HSA-6794362)	5.560e-42 1.330e-30 3.400e-24

Table A.5: Listed are the significant reactome pathways enriched in clusters predicted by ClusterONE from the network constructed by significantly downregulated genes in M samples with respect to N samples using pipeline I.

Cluster	Size	Reactome pathways	Adjusted P-value

1	147	L13a-mediated translational silencing of Ceruloplasmin expression (R-HSA-156827) GTP hydrolysis and joining of the 60S ribosomal subunit (R-HSA-72706) Formation of a pool of free 40S subunits (R-HSA-72689)	1.630e-104 1.740e-104 2.540e-103
2	43	Keratinization (R-HSA-6805567) Developmental Biology (R-HSA-1266738) Formation of the cornified envelope (R-HSA-6809371)	4.780e-71 9.670e-43 3.510e-29
3	53		
4	60		
5	31		
6	39		
7	108	Metabolism of RNA (R-HSA-8953854) Processing of Capped Intron-Containing Pre-mRNA (R-HSA-72203) mRNA Splicing (R-HSA-72172)	3.980e-17 7.380e-12 5.210e-10
8	37		
9	48	Post-translational modification: synthesis of GPI-anchored proteins (R-HSA-163125) NOTCH2 Activation and Transmission of Signal to the Nucleus (R-HSA-2979096) Signaling by NOTCH2 (R-HSA-1980145)	4.270e-05 1.120e-02 1.960e-02
10	45		
11	70	GPCR downstream signalling (R-HSA-388396) Signaling by GPCR (R-HSA-372790) GPCR ligand binding (R-HSA-500792)	6.290e-25 1.060e-24 1.110e-19
12	14	Keratinization (R-HSA-6805567) Developmental Biology (R-HSA-1266738)	4.910e-22 7.620e-13
13	72	Rab regulation of trafficking (R-HSA-9007101) Membrane Trafficking (R-HSA-199991) RAB GEFs exchange GTP for GDP on RABs (R-HSA-8876198)	3.250e-06 1.040e-03 2.780e-03
14	56	Transport of small molecules (R-HSA-382551) ABC-family proteins mediated transport (R-HSA-382556) ABC transporters in lipid homeostasis (R-HSA-1369062)	4.900e-06 6.630e-05 6.990e-05
15	21		
16	55	Metabolism (R-HSA-1430728) Metabolism of amino acids and derivatives (R-HSA-71291) Branched-chain amino acid catabolism (R-HSA-70895)	5.940e-16 2.410e-07 5.320e-07
17	27		
18	56		
19	18		
20	58		
21	25	Cilium Assembly (R-HSA-5617833) Organelle biogenesis and maintenance (R-HSA-1852241) Cargo trafficking to the periciliary membrane (R-HSA-5620920)	1.670e-15 5.710e-14 1.770e-09
22	23	Biological oxidations (R-HSA-211859) Phase II - Conjugation of compounds (R-HSA-156580) Metabolism (R-HSA-1430728)	8.450e-33 4.900e-15 1.750e-14
23	31	COPI-dependent Golgi-to-ER retrograde traffic (R-HSA-6811434) Golgi-to-ER retrograde transport (R-HSA-8856688) Intra-Golgi and retrograde Golgi-to-ER traffic (R-HSA-6811442)	1.070e-05 3.060e-05 2.300e-04
24	5		
25	11		
26	30	ABC transporters in lipid homeostasis (R-HSA-1369062) ABC-family proteins mediated transport (R-HSA-382556)	3.280e-03 8.530e-03
27	71	Hyaluronan uptake and degradation (R-HSA-2160916) Integrin cell surface interactions (R-HSA-216083) Hyaluronan metabolism (R-HSA-2142845)	8.160e-05 1.390e-04 1.940e-04
28	41		
29	18		
30	36	Glycosaminoglycan metabolism (R-HSA-1630316) Metabolism of carbohydrates (R-HSA-71387) Chondroitin sulfate/dermatan sulfate metabolism (R-HSA-1793185)	6.330e-10 6.690e-07 1.200e-05
31	28		
32	58		

33	37	Keratinization (R-HSA-6805567)	2.370e-03
34	9		
35	37		
36	18		
37	16		
38	36	cGMP effects (R-HSA-418457) Physiological factors (R-HSA-5578768) Nitric oxide stimulates guanylate cyclase (R-HSA-392154)	1.460e-03 2.330e-03 3.260e-03
39	70	Metabolism of lipids (R-HSA-556833) Sphingolipid metabolism (R-HSA-428157) Metabolism (R-HSA-1430728)	4.400e-09 8.450e-06 1.230e-04
40	50	Signaling by WNT (R-HSA-195721) TCF dependent signaling in response to WNT (R-HSA-201681) Regulation of FZD by ubiquitination (R-HSA-4641263)	8.740e-10 6.620e-09 4.150e-07
41	10		
42	50		
43	10	O-linked glycosylation of mucins (R-HSA-913709) O-linked glycosylation (R-HSA-5173105)	6.410e-03 1.750e-02
44	39	Gene Silencing by RNA (R-HSA-211000) Gene expression (Transcription) (R-HSA-74160) Pre-NOTCH Transcription and Translation (R-HSA-1912408)	6.370e-05 7.510e-05 8.730e-05
45	19		
46	7		
47	8		
48	62	Pre-NOTCH Transcription and Translation (R-HSA-1912408) Pre-NOTCH Expression and Processing (R-HSA-1912422) Signaling by NOTCH (R-HSA-157118)	2.790e-05 3.840e-05 5.400e-05
49	35	Transcriptional regulation of white adipocyte differentiation (R-HSA-381340) Metabolism of lipids (R-HSA-556833) Regulation of lipid metabolism by PPARalpha (R-HSA-400206)	2.740e-18 9.730e-10 2.700e-09
50	6		
51	6		
52	11		

Table A.6: Listed are the significant reactome pathways enriched in clusters predicted by ClusterONE from the network constructed by significantly upregulated genes in M samples with respect to N samples using pipeline I.

Cluster	Size	Reactome pathways	Adjusted P-value
1	300	Cell Cycle (R-HSA-1640170) Cell Cycle, Mitotic (R-HSA-69278) Cell Cycle Checkpoints (R-HSA-69620)	1.160e-111 5.540e-101 2.250e-75
2	409	Immune System (R-HSA-168256) Cytokine Signaling in Immune system (R-HSA-1280215) Innate Immune System (R-HSA-168249)	9.680e-121 1.490e-67 1.320e-42
3	250	Translation (R-HSA-72766) Mitochondrial translation (R-HSA-5368287) Mitochondrial translation initiation (R-HSA-5368286)	9.200e-51 1.040e-26 3.150e-26
4	223	Developmental Biology (R-HSA-1266738) Myogenesis (R-HSA-525793) Cardiogenesis (R-HSA-9733709)	1.500e-09 4.060e-09 9.530e-08
5	36		
6	29	Formation of the cornified envelope (R-HSA-6809371) Keratinization (R-HSA-6805567) Developmental Biology (R-HSA-1266738)	2.040e-10 6.880e-09 2.000e-03
7	143	Transcriptional regulation of pluripotent stem cells (R-HSA-452723) POU5F1 (OCT4), SOX2, NANOG activate genes related to proliferation (R-HSA-2892247) Developmental Biology (R-HSA-1266738)	5.520e-08 1.080e-07 9.230e-07

8	95	Metabolism of RNA (R-HSA-8953854) rRNA processing (R-HSA-72312) rRNA processing in the nucleus and cytosol (R-HSA-8868773)	1.900e-24 5.280e-24 4.600e-23
9	56		
10	47	Striated Muscle Contraction (R-HSA-390522) Muscle contraction (R-HSA-397014) Smooth Muscle Contraction (R-HSA-445355)	7.950e-17 5.650e-16 1.090e-03
11	217	Asparagine N-linked glycosylation (R-HSA-446203) Metabolism of proteins (R-HSA-392499) Post-translational protein modification (R-HSA-597592)	1.990e-46 1.600e-25 3.980e-25
12	269	Gene expression (Transcription) (R-HSA-74160) Chromatin organization (R-HSA-4839726) Chromatin modifying enzymes (R-HSA-3247509)	7.920e-35 3.140e-34 4.700e-34
13	160	Antigen processing: Ubiquitination & Proteasome degradation (R-HSA-983168) Class I MHC mediated antigen processing & presentation (R-HSA-983169) Neddylation (R-HSA-8951664)	1.900e-68 7.970e-63 1.780e-62
14	6		
15	157	Metabolism (R-HSA-1430728) Metabolism of carbohydrates (R-HSA-71387) Glucose metabolism (R-HSA-70326)	5.420e-64 5.570e-24 1.360e-21
16	60	Extracellular matrix organization (R-HSA-1474244) Collagen formation (R-HSA-1474290) Collagen biosynthesis and modifying enzymes (R-HSA-1650814)	2.590e-57 3.770e-43 7.410e-42
17	26		
18	82	Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins. (R-HSA-163200) The citric acid (TCA) cycle and respiratory electron transport (R-HSA-1428517) Respiratory electron transport (R-HSA-611105)	2.090e-37 1.200e-33 6.270e-32
19	149	Striated Muscle Contraction (R-HSA-390522) Muscle contraction (R-HSA-397014) RHO GTPases Activate WASPs and WAVEs (R-HSA-5663213)	9.580e-12 2.920e-11 6.320e-07

A.1.2 Enriched reactome pathways using pipeline 2

Table A.7: Listed are the significant reactome pathways enriched in clusters predicted by ClusterONE from the differential network comparing all N samples with all M samples using pipeline II.

Cluster	Size	Reactome pathways	Adjusted P-value
1	144	Cell Cycle (R-HSA-1640170) Cell Cycle, Mitotic (R-HSA-69278) Cell Cycle Checkpoints (R-HSA-69620)	5.460e-59 4.780e-56 3.720e-42
2	190	Immune System (R-HSA-168256) Peptide ligand-binding receptors (R-HSA-375276) Class A/1 (Rhodopsin-like receptors) (R-HSA-373076)	9.190e-29 1.030e-19 4.920e-19
3	81	Complement cascade (R-HSA-166658) Extracellular matrix organization (R-HSA-1474244) Regulation of Complement cascade (R-HSA-977606)	1.910e-05 6.100e-05 7.910e-05
4	54	GPCR ligand binding (R-HSA-500792) Signaling by GPCR (R-HSA-372790) GPCR downstream signalling (R-HSA-388396)	1.150e-21 1.450e-21 1.520e-21
5	54	Passive transport by Aquaporins (R-HSA-432047) Extracellular matrix organization (R-HSA-1474244)	7.820e-03 1.400e-02
6	14	ABC-family proteins mediated transport (R-HSA-382556) ABC transporters in lipid homeostasis (R-HSA-1369062) Transport of small molecules (R-HSA-382551)	2.740e-04 3.860e-04 1.730e-02
7	38	Passive transport by Aquaporins (R-HSA-432047)	4.830e-03

8	28		
9	51	Passive transport by Aquaporins (R-HSA-432047)	1.110e-02
10	8		
11	14	Metabolism (R-HSA-1430728) Metabolism of carbohydrates (R-HSA-71387) Glycolysis (R-HSA-70171)	3.210e-05 3.260e-05 9.800e-03
12	22		
13	37	Transport of small molecules (R-HSA-382551) Transport of inorganic cations/anions and amino acids/ oligopeptides (R-HSA-425393) SLC-mediated transmembrane transport (R-HSA-425407)	7.510e-06 1.460e-05 3.860e-05
14	9		
15	11		
16	5		
17	17	Biological oxidations (R-HSA-211859) Glutathione conjugation (R-HSA-156590) Highly calcium permeable nicotinic acetylcholine receptors (R-HSA-629597)	1.140e-08 4.050e-05 8.770e-05
18	15		
19	14	Biological oxidations (R-HSA-211859) Glutathione conjugation (R-HSA-156590) Ethanol oxidation (R-HSA-71384)	1.450e-09 1.710e-05 9.710e-05
20	7		
21	23		
22	12	Transport of inorganic cations/anions and amino acids/oligopeptides (R-HSA-425393)	2.960e-02
23	32		
24	15	Vitamin B1 (thiamin) metabolism (R-HSA-196819)	1.680e-02
25	12	Cytochrome P450 - arranged by substrate type (R-HSA-211897) Phase I - Functionalization of compounds (R-HSA-211945) Metabolism (R-HSA-1430728)	6.140e-05 1.890e-04 3.600e-04
26	12		
27	9		
28	10	Cytochrome P450 - arranged by substrate type (R-HSA-211897) Phase I - Functionalization of compounds (R-HSA-211945) Biological oxidations (R-HSA-211859)	3.690e-05 1.140e-04 1.410e-03
29	18	Vitamin B1 (thiamin) metabolism (R-HSA-196819)	2.670e-02
30	7		
31	9		
32	22		

Table A.8: Listed are the significant reactome pathways enriched in clusters predicted by ClusterONE from the network constructed by significantly downregulated genes in M samples with respect to N samples using pipeline II.

Cluster	Size	Reactome pathways	Adjusted P-value
1	140	Cell Cycle (R-HSA-1640170) Cell Cycle, Mitotic (R-HSA-69278) Cell Cycle Checkpoints (R-HSA-69620)	2.580e-61 1.020e-56 7.550e-44
2	150	Immune System (R-HSA-168256) Cytokine Signaling in Immune system (R-HSA-1280215) Innate Immune System (R-HSA-168249)	6.340e-28 2.230e-16 1.650e-12
3	39	Neurotransmitter receptors and postsynaptic signal transmission (R-HSA-112314) Glucagon-type ligand receptors (R-HSA-420092) Opioid Signalling (R-HSA-111885)	9.360e-03 1.020e-02 1.040e-02
4	38	G alpha (s) signalling events (R-HSA-418555) Neurotransmitter receptors and postsynaptic signal transmission (R-HSA-112314) Glucagon-type ligand receptors (R-HSA-420092)	5.180e-04 1.090e-02 1.120e-02

5	16	G alpha (s) signalling events (R-HSA-418555) Neurotransmitter receptors and postsynaptic signal transmission (R-HSA-112314) Platelet homeostasis (R-HSA-418346)	1.180e-04 1.830e-04 2.360e-04
6	13		
7	23	Signaling by GPCR (R-HSA-372790) GPCR downstream signalling (R-HSA-388396) G alpha (q) signalling events (R-HSA-416476)	1.250e-06 5.800e-06 4.940e-05
8	10		
9	25	G alpha (s) signalling events (R-HSA-418555) GPCR downstream signalling (R-HSA-388396) Platelet homeostasis (R-HSA-418346)	1.150e-03 1.210e-03 1.380e-03
10	38	Neurotransmitter receptors and postsynaptic signal transmission (R-HSA-112314) Transmission across Chemical Synapses (R-HSA-112315) Meiotic synapsis (R-HSA-1221632)	1.770e-06 7.890e-06 2.610e-05
11	7		
12	9		
13	5		
14	5		
15	7		
16	28	G alpha (s) signalling events (R-HSA-418555) GPCR downstream signalling (R-HSA-388396) Signaling by GPCR (R-HSA-372790)	7.850e-04 8.700e-04 1.000e-03
17	11		
18	7		

Table A.9: Listed are the significant reactome pathways enriched in clusters predicted by ClusterONE from the network constructed by significantly upregulated genes in M samples with respect to N samples using pipeline II.

Cluster	Size	Reactome pathways	Adjusted P-value
1	15		
2	17		
3	16	ABC transporters in lipid homeostasis (R-HSA-1369062) ABC-family proteins mediated transport (R-HSA-382556) Transport of small molecules (R-HSA-382551)	3.160e-04 3.510e-04 2.800e-02
4	8		
5	8		
6	10	Biological oxidations (R-HSA-211859) Ethanol oxidation (R-HSA-71384) Phase I - Functionalization of compounds (R-HSA-211945)	1.320e-09 1.630e-05 6.510e-05
7	11		
8	13	GPCR ligand binding (R-HSA-500792) Class A/1 (Rhodopsin-like receptors) (R-HSA-373076) GPCR downstream signalling (R-HSA-388396)	4.890e-16 1.530e-15 6.220e-15
9	10	Cytochrome P450 - arranged by substrate type (R-HSA-211897) Phase I - Functionalization of compounds (R-HSA-211945) Biological oxidations (R-HSA-211859)	2.900e-05 9.770e-05 1.320e-03
10	12	ABC transporters in lipid homeostasis (R-HSA-1369062) Transport of small molecules (R-HSA-382551) ABC-family proteins mediated transport (R-HSA-382556)	3.160e-04 2.800e-02 3.240e-02
11	13		
12	5		
13	12	Passive transport by Aquaporins (R-HSA-432047) Transport of small molecules (R-HSA-382551)	2.130e-02 2.830e-02
14	8		
15	6		
16	5		
17	5		

18	8	Complement cascade (R-HSA-166658) Regulation of Complement cascade (R-HSA-977606) Terminal pathway of complement (R-HSA-166665)	1.230e-09 1.450e-09 1.850e-06
19	7	HSP90 chaperone cycle for steroid hormone receptors (SHR) in the presence of ligand (R-HSA-3371497) SUMOylation of intracellular receptors (R-HSA-4090294)	1.030e-03 2.510e-02
20	9		

A.1.3 Enriched reactome pathways using pipeline 3

Table A.10: Listed are the significant reactome pathways enriched in clusters predicted by ClusterONE from the differential network comparing all N samples with all M samples using pipeline III.

Cluster	Size	Reactome pathways	Adjusted P-value
1	40	Nucleotide biosynthesis (R-HSA-8956320) Purine ribonucleoside monophosphate biosynthesis (R-HSA-73817) Metabolism of nucleotides (R-HSA-15869)	5.800e-03 6.240e-03 3.040e-02
2	40	Prolonged ERK activation events (R-HSA-169893) Signaling by Rho GTPases, Miro GTPases and RHOBTB3 (R-HSA-9716542) RHO GTPase cycle (R-HSA-9012999)	1.160e-02 2.940e-02 3.100e-02
3	46	Signaling by BRAF and RAF1 fusions (R-HSA-6802952) Signaling by moderate kinase activity BRAF mutants(R-HSA-6802946) Oncogenic MAPK signaling (R-HSA-6802957)	3.23e-02 3.97e-02 4.32e-02
4	33	Chemokine receptors bind chemokines (R-HSA-380108) Peptide ligand-binding receptors (R-HSA-375276) Class A/1 (Rhodopsin-like receptors) (R-HSA-373076)	2.440e-46 1.320e-35 8.260e-31
5	23	Transcriptional regulation of white adipocyte differentiation (R-HSA-381340) Regulation of lipid metabolism by PPARalpha (R-HSA-400206) PPARA activates gene expression (R-HSA-1989781)	5.960e-15 4.280e-14 5.910e-14
6	30	Interleukin-23 signaling (R-HSA-9020933) Interleukin-12 family signaling (R-HSA-447115) Interleukin-4 and Interleukin-13 signaling (R-HSA-6785807)	6.640e-06 3.320e-05 1.930e-02
7	16		
8	24		
9	22	Lectin pathway of complement activation (R-HSA-166662) Complement cascade (R-HSA-166658) Creation of C4 and C2 activators (R-HSA-166786)	3.080e-12 5.800e-08 2.160e-07
10	16		
11	36		
12	30		
13	6	Glucuronidation (R-HSA-156588) Phase II - Conjugation of compounds (R-HSA-156580) Biological oxidations (R-HSA-211859)	7.120e-11 3.660e-08 7.600e-07
14	23	Signaling by Nuclear Receptors (R-HSA-9006931) Estrogen-stimulated signalling through PRKCZ (R-HSA-9634635) Extra-nuclear estrogen signaling (R-HSA-9009391)	2.420e-02 3.210e-02 3.470e-02
15	29		
16	32	Recruitment of mitotic centrosome proteins and complexes (R-HSA-380270) Regulation of PLK1 Activity at G2/M Transition (R-HSA-2565942) AURKA Activation by TPX2 (R-HSA-8854518)	4.350e-03 4.560e-03 4.610e-03
17	11	Trafficking and processing of endosomal TLR (R-HSA-1679131) Innate Immune System (R-HSA-168249) Immune System (R-HSA-168256)	2.330e-02 2.390e-02 3.070e-02

18	18	Interferon alpha/beta signaling (R-HSA-909733) Interferon Signaling (R-HSA-913531) Regulation of IFNA/IFNB signalling (R-HSA-912694)	5.900e-12 5.050e-09 2.310e-08
19	11	Metal sequestration by antimicrobial proteins (R-HSA-6799990)	1.870e-02
20	14	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell (R-HSA-198933)	2.100e-02
21	29		
22	8		
23	31		
24	16	Class A/1 (Rhodopsin-like receptors) (R-HSA-373076) Adenosine P1 receptors (R-HSA-417973) GPCR ligand binding (R-HSA-500792)	7.570e-03 9.540e-03 1.280e-02
25	19	WNT ligand biogenesis and trafficking (R-HSA-3238698) Class B/2 (Secretin family receptors) (R-HSA-373080) Signaling by WNT (R-HSA-195721)	1.570e-20 7.110e-16 3.170e-11
26	12	Metabolism of amine-derived hormones (R-HSA-209776) Serotonin and melatonin biosynthesis (R-HSA-209931)	5.140e-04 8.400e-03
27	36		
28	24		
29	12	GPCR ligand binding (R-HSA-500792) GPCR downstream signalling (R-HSA-388396) Signaling by GPCR (R-HSA-372790)	5.770e-06 2.250e-05 3.180e-05
30	10		
31	14		
32	9	Synthesis of bile acids and bile salts via 27-hydroxycholesterol (R-HSA-193807) Synthesis of bile acids and bile salts via 24-hydroxycholesterol (R-HSA-193775) Synthesis of bile acids and bile salts via 7alpha-hydroxycholesterol (R-HSA-193368)	8.120e-11 1.220e-10 3.430e-10
33	5	Interleukin-20 family signaling (R-HSA-8854691) Signaling by Interleukins (R-HSA-449147) Cytokine Signaling in Immune system (R-HSA-1280215)	1.190e-11 7.320e-06 4.390e-05
34	15	Thromboxane signalling through TP receptor (R-HSA-428930) Signal amplification (R-HSA-392518) Thrombin signalling through proteinase activated receptors (PARs) (R-HSA-456926)	7.380e-14 1.440e-13 1.770e-13
35	9	Creatine metabolism (R-HSA-71288) Metabolism of amino acids and derivatives (R-HSA-71291)	4.320e-08 2.970e-04
36	16	Ligand-receptor interactions (R-HSA-5632681) Hedgehog 'on' state (R-HSA-5632684) Activation of SMO (R-HSA-5635838)	3.080e-07 4.660e-04 6.520e-04
37	26		
38	7		
39	14		
40	11		
41	13	G alpha (s) signalling events (R-HSA-418555) Peptide ligand-binding receptors (R-HSA-375276) Class A/1 (Rhodopsin-like receptors) (R-HSA-373076)	8.900e-05 1.310e-04 1.110e-03
42	24		
43	16		
44	38		
45	6	Autophagy (R-HSA-9612973) Macroautophagy (R-HSA-1632852) Translation of Replicase and Assembly of the Replication Transcription Complex (R-HSA-9694676)	5.740e-05 7.590e-05 5.450e-03
46	22		
47	16		
48	24	Highly calcium permeable nicotinic acetylcholine receptors (R-HSA-629597) Presynaptic nicotinic acetylcholine receptors (R-HSA-622323) Highly calcium permeable postsynaptic nicotinic acetylcholine receptors (R-HSA-629594)	7.020e-09 7.210e-09 7.640e-09
49	15		

50	38		
51	12	Creation of C4 and C2 activators (R-HSA-166786) Role of LAT2/NTAL/LAB on calcium mobilization (R-HSA-2730905) FCERI mediated Ca+2 mobilization (R-HSA-2871809)	2.930e-07 3.100e-07 3.530e-07
52	20		
53	20		
54	32	Clathrin-mediated endocytosis (R-HSA-8856828) Signaling by NOTCH1 t(7;9)(NOTCH1:M1580_K2555) Translocation Mutant (R-HSA-2660825) Diseases of signal transduction by growth factor receptors and second messengers (R-HSA-5663202)	6.920e-03 4.030e-02 4.470e-02
55	8		
56	29		
57	13	Toll Like Receptor 4 (TLR4) Cascade (R-HSA-166016) Toll-like Receptor Cascades (R-HSA-168898) Innate Immune System (R-HSA-168249)	5.650e-05 5.660e-05 6.520e-05
58	23		
59	37		
60	13		
61	12		
62	15		
63	14		
64	12	Cooperation of Prefoldin and TriC/CCT in actin and tubulin folding (R-HSA-389958) Protein folding (R-HSA-391251) Chaperonin-mediated protein folding (R-HSA-390466)	1.110e-05 2.240e-04 2.630e-04
65	37	Condensation of Prometaphase Chromosomes (R-HSA-2514853)	3.040e-05
66	8		
67	15		
68	16	Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs) (R-HSA-381426) Post-translational protein phosphorylation (R-HSA-8957275) Extracellular matrix organization (R-HSA-1474244)	7.260e-11 1.860e-09 1.270e-06
69	33	Mitotic Prometaphase (R-HSA-68877) Anchoring of the basal body to the plasma membrane (R-HSA-5620912) Loss of Nlp from mitotic centrosomes (R-HSA-380259)	7.310e-08 3.120e-07 1.150e-06
70	10	Interferon alpha/beta signaling (R-HSA-909733) SARS-CoV-2 activates/modulates innate and adaptive immune responses (R-HSA-9705671) Regulation of IFNA/IFNB signaling (R-HSA-912694)	1.120e-04 2.260e-04 2.740e-04
71	30	Generic Transcription Pathway (R-HSA-212436) RNA Polymerase II Transcription (R-HSA-73857) Gene expression (Transcription) (R-HSA-74160)	1.250e-05 1.960e-05 5.100e-05
72	15	Inhibition of the proteolytic activity of APC/C required for the onset of anaphase by mitotic spindle checkpoint components (R-HSA-141405) Inactivation of APC/C via direct inhibition of the APC/C complex (R-HSA-141430) Conversion from APC/C:Cdc20 to APC/C:Cdh1 in late anaphase (R-HSA-176407)	3.260e-09 4.080e-09 4.390e-09
73	11		
74	24	MET activates PI3K/AKT signaling (R-HSA-8851907) MET activates PTPN11 (R-HSA-8865999) Signaling by MET (R-HSA-6806834)	1.950e-02 2.100e-02 2.910e-02
75	28		
76	10		
77	10	Fertilization (R-HSA-1187000) Interaction With Cumulus Cells And The Zona Pellucida (R-HSA-2534343) Reproduction (R-HSA-1474165)	3.550e-09 4.910e-08 1.810e-06
78	22		

79	16		
80	10		
81	17		
82	23		
83	31		
84	11	Complement cascade (R-HSA-166658) Regulation of Complement cascade (R-HSA-977606) Innate Immune System (R-HSA-168249)	1.980e-18 3.690e-16 2.020e-09
85	11		
86	8		
87	9		
88	5		
89	17	Striated Muscle Contraction (R-HSA-390522)	1.080e-02
90	27		
91	5		
92	20	Anti-inflammatory response favouring Leishmania parasite infection (R-HSA-9662851) Interleukin-3, Interleukin-5 and GM-CSF signaling (R-HSA-512988) Constitutive Signaling by EGFRvIII (R-HSA-5637810)	1.270e-02 1.280e-02 2.400e-02
93	13		
94	11		
95	7	Long-term potentiation (R-HSA-9620244) Synaptic adhesion-like molecules (R-HSA-8849932) Unblocking of NMDA receptors, glutamate binding and activation (R-HSA-438066)	3.550e-05 4.160e-05 6.250e-05
96	7		
97	7	Tachykinin receptors bind tachykinins (R-HSA-380095)	5.350e-03
98	25	Dectin-1 mediated noncanonical NF-kB signaling (R-HSA-5607761) Defective CFTR causes cystic fibrosis (R-HSA-5678895) Degradation of GLI1 by the proteasome (R-HSA-5610780)	9.100e-05 9.360e-05 9.450e-05
99	21	Response of EIF2AK1 (HRI) to heme deficiency (R-HSA-9648895) ATF4 activates genes in response to endoplasmic reticulum stress (R-HSA-380994) PERK regulates gene expression (R-HSA-381042)	7.870e-06 4.680e-03 5.020e-03
100	13	PRC2 methylates histones and DNA (R-HSA-212300) ERCC6 (CSB) and EHMT2 (G9a) positively regulate rRNA expression (R-HSA-427389) Defective pyroptosis (R-HSA-9710421)	8.240e-04 8.960e-04 9.420e-04
101	9		
102	27	Signal Transduction (R-HSA-162582) G alpha (i) signalling events (R-HSA-418594) Negative regulators of DDX58/IFIH1 signaling (R-HSA-936440)	6.830e-03 7.720e-03 1.410e-02
103	13	Phase 0 - rapid depolarisation (R-HSA-5576892) Cardiac conduction (R-HSA-5576891) Muscle contraction (R-HSA-397014)	4.620e-08 1.780e-05 1.060e-04
104	28	Myogenesis (R-HSA-525793)	1.550e-06
105	32	Hormone ligand-binding receptors (R-HSA-375281) Peptide hormone biosynthesis (R-HSA-209952) Glycoprotein hormones (R-HSA-209822)	9.610e-06 1.100e-05 1.220e-05
106	8		
107	31	SRP-dependent cotranslational protein targeting to membrane (R-HSA-1799339) Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC) (R-HSA-975956) Viral mRNA Translation (R-HSA-192823)	2.390e-22 5.790e-22 5.880e-22

Table A.11: Listed are the significant reactome pathways enriched in clusters predicted by ClusterONE from the interactions only found in M samples and not in N samples using pipeline III.

Cluster	Size	Reactome pathways	Adjusted P-value

1	48		
2	29		
3	40		
4	37	Chemokine receptors bind chemokines (R-HSA-380108) Peptide ligand-binding receptors (R-HSA-375276) Class A/1 (Rhodopsin-like receptors) (R-HSA-373076)	1.300e-48 2.950e-37 9.680e-32
5	39	Condensation of Prometaphase Chromosomes (R-HSA-2514853)	9.250e-06
6	28	Transcriptional regulation of white adipocyte differentiation (R-HSA-381340) Regulation of lipid metabolism by PPARalpha (R-HSA-400206) PPARA activates gene expression (R-HSA-1989781)	1.200e-16 1.720e-15 2.330e-15
7	40	Creation of C4 and C2 activators (R-HSA-166786) Role of LAT2/NTAL/LAB on calcium mobilization (R-HSA-2730905) FCGR activation (R-HSA-2029481)	2.420e-06 2.470e-06 2.840e-06
8	40	Metabolism of RNA (R-HSA-8953854) Eukaryotic Translation Initiation (R-HSA-72613) GTP hydrolysis and joining of the 60S ribosomal subunit (R-HSA-72706)	6.350e-03 2.060e-02 2.280e-02
9	26	Integrin cell surface interactions (R-HSA-216083) Laminin interactions (R-HSA-3000157) Extracellular matrix organization (R-HSA-1474244)	1.590e-04 9.130e-03 2.700e-02
10	21	Response of EIF2AK1 (HRI) to heme deficiency (R-HSA-9648895) ATF4 activates genes in response to endoplasmic reticulum stress (R-HSA-380994) PERK regulates gene expression (R-HSA-381042)	2.790e-06 3.390e-03 3.820e-03
11	17		
12	37		
13	23		
14	35	Clathrin-mediated endocytosis (R-HSA-8856828) Signaling by EGFR (R-HSA-177929) Diseases of signal transduction by growth factor receptors and second messengers (R-HSA-5663202)	9.500e-03 3.480e-02 3.610e-02
15	25		
16	17		
17	30	Diseases of Mismatch Repair (MMR) (R-HSA-5423599)	4.920e-02
18	15	Adenosine P1 receptors (R-HSA-417973) Class A/1 (Rhodopsin-like receptors) (R-HSA-373076) GPCR ligand binding (R-HSA-500792)	3.310e-03 4.770e-03 8.260e-03
19	22	Muscarinic acetylcholine receptors (R-HSA-390648)	2.800e-02
20	21		
21	16	Gastrulation (R-HSA-9758941) Regulation of beta-cell development (R-HSA-186712) Developmental Biology (R-HSA-1266738)	2.750e-03 4.450e-03 1.010e-02
22	39		
23	22	Highly calcium permeable nicotinic acetylcholine receptors (R-HSA-629597) Presynaptic nicotinic acetylcholine receptors (R-HSA-622323) Highly calcium permeable postsynaptic nicotinic acetylcholine receptors (R-HSA-629594)	3.160e-07 4.130e-07 4.130e-07
24	36	Mitotic Prometaphase (R-HSA-68877) Anchoring of the basal body to the plasma membrane (R-HSA-5620912) Loss of Nlp from mitotic centrosomes (R-HSA-380259)	1.190e-07 3.690e-07 1.200e-06
25	12	Metabolism of amine-derived hormones (R-HSA-209776) Serotonin and melatonin biosynthesis (R-HSA-209931)	3.160e-04 4.010e-03
26	16	Signal amplification (R-HSA-392518) Thromboxane signalling through TP receptor (R-HSA-428930) Thrombin signalling through proteinase activated receptors (PARs) (R-HSA-456926)	2.160e-16 9.170e-15 4.630e-14
27	44	Transcriptional Regulation by TP53 (R-HSA-3700989) Diseases of Mismatch Repair (MMR) (R-HSA-5423599) Disease (R-HSA-1643685)	1.860e-02 3.470e-02 3.620e-02
28	14		

29	34	Synthesis of very long-chain fatty acyl-CoAs (R-HSA-75876) Fatty acyl-CoA biosynthesis (R-HSA-75105)	2.090e-02 4.300e-02
30	15	trans-Golgi Network Vesicle Budding (R-HSA-199992)	4.660e-02
31	40		
32	39	Interleukin-2 signaling (R-HSA-9020558) Interleukin receptor SHC signaling (R-HSA-912526) Interleukin-4 and Interleukin-13 signaling (R-HSA-6785807)	2.990e-03 1.740e-02 3.950e-02
33	6		
34	23		
35	6	Cell surface interactions at the vascular wall (R-HSA-202733)	4.350e-02
36	17	FCERI mediated MAPK activation (R-HSA-2871796) Creation of C4 and C2 activators (R-HSA-166786) Role of LAT2/NTAL/LAB on calcium mobilization (R-HSA-2730905)	1.690e-07 1.410e-06 1.440e-06
37	22		
38	9		
39	18		
40	19	HSP90 chaperone cycle for steroid hormone receptors (SHR) in the presence of ligand (R-HSA-3371497) COPI-independent Golgi-to-ER retrograde traffic (R-HSA-6811436) MHC class II antigen presentation (R-HSA-2132295)	1.230e-11 1.430e-11 1.660e-11
41	18	Activation of the pre-replicative complex (R-HSA-68962) Formation of Incision Complex in GG-NER (R-HSA-5696395) Translesion synthesis by POLI (R-HSA-5656121)	7.750e-03 8.720e-03 1.570e-02
42	22	Estrogen-stimulated signalling through PRKCZ (R-HSA-9634635) Extra-nuclear estrogen signaling (R-HSA-9009391) ESR-mediated signaling (R-HSA-8939211)	2.800e-02 3.230e-02 3.560e-02
43	14		
44	14		
45	15	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell (R-HSA-198933)	2.630e-02
46	15		
47	12	Fertilization (R-HSA-1187000) Interaction With Cumulus Cells And The Zona Pellucida (R-HSA-2534343) Reproduction (R-HSA-1474165)	4.330e-09 2.800e-08 4.610e-06
48	14	Downregulation of TGF-beta receptor signaling (R-HSA-2173788) Triglyceride catabolism (R-HSA-163560) Triglyceride metabolism (R-HSA-8979227)	5.780e-04 1.020e-03 1.410e-03
49	16		
50	18	IFNG signaling activates MAPKs (R-HSA-9732724)	4.610e-02
51	15		
52	13	Toll Like Receptor 4 (TLR4) Cascade (R-HSA-166016) Toll-like Receptor Cascades (R-HSA-168898) Innate Immune System (R-HSA-168249)	4.770e-05 4.890e-05 6.200e-05
53	26	MET promotes cell motility (R-HSA-8875878) MET activates PI3K/AKT signaling (R-HSA-8851907) MET activates RAP1 and RAC1 (R-HSA-8875555)	1.150e-02 1.670e-02 1.880e-02
54	9	Interleukin-2 signaling (R-HSA-9020558)	2.240e-02
55	37		
56	5	Interleukin-20 family signaling (R-HSA-8854691) Signaling by Interleukins (R-HSA-449147) Cytokine Signaling in Immune system (R-HSA-1280215)	4.440e-12 6.940e-06 4.250e-05
57	14	Creation of C4 and C2 activators (R-HSA-166786) Role of LAT2/NTAL/LAB on calcium mobilization (R-HSA-2730905) FCGR activation (R-HSA-2029481)	2.820e-09 2.880e-09 3.310e-09
58	8	G alpha (s) signalling events (R-HSA-418555) Peptide ligand-binding receptors (R-HSA-375276) Class A/1 (Rhodopsin-like receptors) (R-HSA-373076)	3.420e-06 5.230e-06 4.810e-05
59	8		
60	8	Sensing of DNA Double Strand Breaks (R-HSA-5693548) HDR through MMEJ (alt-NHEJ) (R-HSA-5685939) Defective HDR through Homologous Recombination Repair (HRR) due to PALB2 loss of BRCA1 binding function (R-HSA-9704331)	5.100e-03 1.120e-02 1.340e-02

61	22		
62	61	Keratinization (R-HSA-6805567)	2.230e-06
63	13	Lectin pathway of complement activation (R-HSA-166662) Ficolins bind to repetitive carbohydrate structures on the target cell surface (R-HSA-2855086) Complement cascade (R-HSA-166658)	8.600e-09 1.010e-06 6.580e-06
64	9		
65	10	Retinoid cycle disease events (R-HSA-2453864) Diseases of the neuronal system (R-HSA-9675143) The canonical retinoid cycle in rods (twilight vision) (R-HSA-2453902)	1.420e-02 2.130e-02 3.440e-02
66	17		
67	5	Ligand-receptor interactions (R-HSA-5632681) Hedgehog 'on' state (R-HSA-5632684) Activation of SMO (R-HSA-5635838)	1.200e-10 1.650e-06 4.810e-06
68	5		
69	14		
70	11		
71	5		
72	18	Synthesis of active ubiquitin: roles of E1 and E2 enzymes (R-HSA-8866652) Protein ubiquitination (R-HSA-8852135) Antigen processing: Ubiquitination & Proteasome degradation (R-HSA-983168)	5.030e-08 8.090e-08 3.710e-06
73	13	Creation of C4 and C2 activators (R-HSA-166786) Role of LAT2/NTAL/LAB on calcium mobilization (R-HSA-2730905) FCGR activation (R-HSA-2029481)	1.620e-09 1.650e-09 1.900e-09
74	26		
75	9		
76	40	Cellular Senescence (R-HSA-2559583) Oxidative Stress Induced Senescence (R-HSA-2559580) PRC2 methylates histones and DNA (R-HSA-212300)	1.040e-06 1.110e-06 9.790e-06
77	6	Factors involved in megakaryocyte development and platelet production (R-HSA-983231)	2.250e-02
78	6	mRNA Splicing (R-HSA-72172) Metabolism of RNA (R-HSA-8953854) Processing of Capped Intron-Containing Pre-mRNA (R-HSA-72203)	1.810e-02 2.460e-02 3.040e-02
79	11	Downregulation of ERBB4 signaling (R-HSA-1253288) Signaling by ERBB4 (R-HSA-1236394) Antigen processing: Ubiquitination & Proteasome degradation (R-HSA-983168)	1.630e-05 3.570e-03 1.280e-02
80	30		
81	7	Long-term potentiation (R-HSA-9620244) Synaptic adhesion-like molecules (R-HSA-8849932) Unblocking of NMDA receptors, glutamate binding and activation (R-HSA-438066)	2.380e-05 2.740e-05 4.110e-05
82	7	Defective ABCC9 causes CMD10, ATFB12 and Cantu syndrome (R-HSA-5678420) ATP sensitive Potassium channels (R-HSA-1296025)	2.550e-04 7.650e-04
83	10	Creatine metabolism (R-HSA-71288) Metabolism of amino acids and derivatives (R-HSA-71291)	1.510e-08 5.490e-04
84	19		
85	27		
86	12		
87	9		
88	26	Cellular response to heat stress (R-HSA-3371556) HSP90 chaperone cycle for steroid hormone receptors (SHR) in the presence of ligand (R-HSA-3371497) Attenuation phase (R-HSA-3371568)	2.010e-04 6.470e-04 8.300e-04
89	7	Synthesis of bile acids and bile salts via 27-hydroxycholesterol (R-HSA-193807) Synthesis of bile acids and bile salts via 24-hydroxycholesterol (R-HSA-193775) Synthesis of bile acids and bile salts via 7alpha-hydroxycholesterol (R-HSA-193368)	1.410e-05 2.250e-05 3.650e-05

90	19		
91	5	VEGFR2 mediated cell proliferation (R-HSA-5218921) RHO GTPases Activate NADPH Oxidases (R-HSA-5668599)	1.460e-02 1.670e-02
92	5	Serotonin receptors (R-HSA-390666) Amine ligand-binding receptors (R-HSA-375280) Class A/1 (Rhodopsin-like receptors) (R-HSA-373076)	3.890e-06 1.090e-04 2.780e-04
93	5	Elevation of cytosolic Ca2+ levels (R-HSA-139853) Platelet calcium homeostasis (R-HSA-418360) Platelet homeostasis (R-HSA-418346)	9.900e-06 2.890e-05 6.000e-04
94	12	GABA receptor activation (R-HSA-977443) Signaling by ERBB4 (R-HSA-1236394) Neurotransmitter receptors and postsynaptic signal transmission (R-HSA-112314)	4.080e-05 5.300e-05 3.250e-03
95	37	Generic Transcription Pathway (R-HSA-212436) RNA Polymerase II Transcription (R-HSA-73857) Gene expression (Transcription) (R-HSA-74160)	2.260e-04 3.450e-04 8.610e-04
96	14	Nervous system development (R-HSA-9675108) Axon guidance (R-HSA-422475) RET signaling (R-HSA-8853659)	5.760e-04 6.660e-04 1.140e-03
97	8		
98	27	SARS-CoV-2 modulates autophagy (R-HSA-9754560)	8.480e-04
99	14		
100	14		
101	11	Peptide ligand-binding receptors (R-HSA-375276) Class A/1 (Rhodopsin-like receptors) (R-HSA-373076) G alpha (i) signalling events (R-HSA-418594)	6.420e-03 1.700e-02 2.070e-02
102	5		
103	5		
104	25		
105	12	Creation of C4 and C2 activators (R-HSA-166786) FCGR activation (R-HSA-2029481) Role of phospholipids in phagocytosis (R-HSA-2029485)	5.630e-05 6.390e-05 6.450e-05
106	14		
107	15	Nuclear Receptor transcription pathway (R-HSA-383280)	8.450e-03
108	8		
109	15		
110	38		
111	11		
112	16		
113	6		
114	7		
115	7		
116	19	tRNA modification in the nucleus and cytosol (R-HSA-6782315)	1.930e-02
117	8		

Table A.12: Listed are the significant reactome pathways enriched in clusters predicted by ClusterONE from the interactions only found in N samples and not in M samples using pipeline III

Cluster	Size	Reactome pathways	Adjusted P-value
1	37		
2	30	Interleukin-23 signaling (R-HSA-9020933)	6.000e-04
3	40	Recruitment of mitotic centrosome proteins and complexes (R-HSA-380270) AURKA Activation by TPX2 (R-HSA-8854518) Regulation of PLK1 Activity at G2/M Transition (R-HSA-2565942)	8.940e-03 9.270e-03 9.470e-03
4	45	IKBKG deficiency causes anhidrotic ectodermal dysplasia with immunodeficiency (EDA-ID) (via TLR) (R-HSA-5603027) IKBKB deficiency causes SCID (R-HSA-5602636)	1.800e-02 3.600e-02
5	33		
6	28		

7	31		
8	35		
9	20		
10	39	O-linked glycosylation (R-HSA-5173105)	6.080e-03
11	26		
12	40	Transport of small molecules (R-HSA-382551) Cation-coupled Chloride cotransporters (R-HSA-426117) Transport of inorganic cations/anions and amino acids/ oligopeptides (R-HSA-425393)	1.460e-11 3.040e-04 1.150e-03
13	40	RNA Polymerase I Transcription Initiation (R-HSA-73762) RNA Polymerase I Transcription (R-HSA-73864) Regulation of TP53 Activity through Acetylation (R-HSA-6804758)	4.360e-03 1.220e-02 1.590e-02
14	40	rRNA modification in the nucleus and cytosol (R-HSA-6790901) rRNA processing (R-HSA-72312) rRNA processing in the nucleus and cytosol (R-HSA-8868773)	2.780e-04 2.840e-02 2.970e-02
15	29		
16	38		
17	35		
18	30		
19	18		
20	28		
21	28	Cooperation of Prefoldin and TriC/CCT in actin and tubulin folding (R-HSA-389958) Prefoldin mediated transfer of substrate to CCT/TriC (R-HSA-389957) Chaperonin-mediated protein folding (R-HSA-390466)	1.900e-06 7.050e-05 1.170e-04
22	31		
23	12	Striated Muscle Contraction (R-HSA-390522)	2.750e-03
24	16	Processive synthesis on the C-strand of the telomere (R-HSA-174414) Telomere C-strand (Lagging Strand) Synthesis (R-HSA-174417) Extension of Telomeres (R-HSA-180786)	2.410e-06 1.430e-05 5.100e-05
25	44		
26	13	WNT ligand biogenesis and trafficking (R-HSA-3238698) Class B/2 (Secretin family receptors) (R-HSA-373080) Signaling by WNT (R-HSA-195721)	6.440e-18 2.070e-13 1.720e-09
27	11		
28	19	Inhibition of the proteolytic activity of APC/C required for the onset of anaphase by mitotic spindle checkpoint components (R-HSA-141405) Inactivation of APC/C via direct inhibition of the APC/C complex (R-HSA-141430) Conversion from APC/C:Cdc20 to APC/C:Cdh1 in late anaphase (R-HSA-176407)	2.890e-09 3.610e-09 3.670e-09
29	37	Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC) (R-HSA-975956) SARS-CoV-2 modulates host translation machinery (R-HSA-9754678) Eukaryotic Translation Elongation (R-HSA-156842)	6.670e-06 7.280e-06 7.310e-06
30	24	Collagen biosynthesis and modifying enzymes (R-HSA-1650814) Collagen formation (R-HSA-1474290) Extracellular matrix organization (R-HSA-1474244)	2.670e-03 4.180e-03 1.070e-02
31	16		
32	24	IKBKG deficiency causes anhidrotic ectodermal dysplasia with immunodeficiency (EDA-ID) (via TLR) (R-HSA-5603027) IKBKB deficiency causes SCID (R-HSA-5602636) IkBA variant leads to EDA-ID (R-HSA-5603029)	5.030e-03 1.010e-02 2.340e-02
33	17	mRNA Splicing (R-HSA-72172) Processing of Capped Intron-Containing Pre-mRNA (R-HSA-72203) mRNA Splicing - Major Pathway (R-HSA-72163)	6.390e-11 3.730e-10 1.560e-09
34	10		
35	26	HSP90 chaperone cycle for steroid hormone receptors (SHR) in the presence of ligand (R-HSA-3371497) Attenuation phase (R-HSA-3371568) HSF1-dependent transactivation (R-HSA-3371571)	9.710e-04 1.660e-03 3.050e-03
36	14		

37	12	Maturation of spike protein (R-HSA-9694548) Translation of Structural Proteins (R-HSA-9694635) Late SARS-CoV-2 Infection Events (R-HSA-9772573)	2.990e-03 5.600e-03 6.950e-03
38	29		
39	26		
40	23		
41	11		
42	23		
43	9	Interferon alpha/beta signaling (R-HSA-909733) Interferon Signaling (R-HSA-913531) Regulation of IFNA/IFNB signaling (R-HSA-912694)	1.920e-07 1.330e-05 8.740e-05
44	27	Platelet activation, signaling and aggregation (R-HSA-76002) Response to elevated platelet cytosolic Ca2+ (R-HSA-76005) Platelet degranulation (R-HSA-114608)	5.840e-04 6.210e-04 7.690e-04
45	12		
46	13		
47	8		
48	14	Phase 0 - rapid depolarisation (R-HSA-5576892) Cardiac conduction (R-HSA-5576891) Muscle contraction (R-HSA-397014)	3.330e-08 2.290e-05 1.450e-04
49	30	CRMPs in Sema3A signaling (R-HSA-399956) Semaphorin interactions (R-HSA-373755)	1.690e-05 2.820e-03
50	14	Metal sequestration by antimicrobial proteins (R-HSA-6799990)	1.660e-02
51	10	Fibronectin matrix formation (R-HSA-1566977) Cell surface interactions at the vascular wall (R-HSA-202733) Neutrophil degranulation (R-HSA-6798695)	4.250e-06 1.400e-03 1.950e-03
52	10	Neutrophil degranulation (R-HSA-6798695)	3.890e-03
53	6	Glucuronidation (R-HSA-156588) Phase II - Conjugation of compounds (R-HSA-156580) Biological oxidations (R-HSA-211859)	2.660e-11 2.920e-08 6.790e-07
54	14		
55	13		
56	17	Activation of BAD and translocation to mitochondria (R-HSA-111447) Chk1/Chk2(Cds1) mediated inactivation of Cyclin B:Cdk1 complex (R-HSA-75035) SARS-CoV-1 targets host intracellular signalling and regulatory pathways (R-HSA-9735871)	1.360e-04 1.710e-04 1.810e-04
57	8	Tachykinin receptors bind tachykinins (R-HSA-380095)	3.400e-03
58	8	COPII-mediated vesicle transport (R-HSA-204005) RAB GEFs exchange GTP for GDP on RABs (R-HSA-8876198) Rab regulation of trafficking (R-HSA-9007101)	7.420e-11 1.970e-10 8.650e-10
59	13		
60	12		
61	12	Clathrin-mediated endocytosis (R-HSA-8856828) InlB-mediated entry of Listeria monocytogenes into host cell (R-HSA-8875360) Negative regulation of MET activity (R-HSA-6807004)	2.730e-02 3.030e-02 3.160e-02
62	13	RNA Polymerase II Transcription Initiation And Promoter Clearance (R-HSA-76042) Inhibition of DNA recombination at telomere (R-HSA-9670095) HIV Transcription Initiation (R-HSA-167161)	4.580e-02 4.800e-02 5.040e-02
63	16		
64	10		
65	14	Cooperation of Prefoldin and TriC/CCT in actin and tubulin folding (R-HSA-389958) Chaperonin-mediated protein folding (R-HSA-390466) Protein folding (R-HSA-391251)	3.930e-08 2.510e-06 2.610e-06
66	18	GPCR ligand binding (R-HSA-500792) GPCR downstream signalling (R-HSA-388396) Signaling by GPCR (R-HSA-372790)	1.930e-06 2.080e-06 2.740e-06
67	8		
68	9		

69	9	Formation of the HIV-1 Early Elongation Complex (R-HSA-167158) HIV Transcription Elongation (R-HSA-167169) HIV elongation arrest and recovery (R-HSA-167287)	2.230e-02 2.540e-02 2.550e-02
70	6	Class A/1 (Rhodopsin-like receptors) (R-HSA-373076) GPCR ligand binding (R-HSA-500792) G alpha (q) signalling events (R-HSA-416476)	2.470e-03 4.740e-03 9.420e-03
71	6	Autophagy (R-HSA-9612973) Macroautophagy (R-HSA-1632852) Translation of Replicase and Assembly of the Replication Transcription Complex (R-HSA-9694676)	5.170e-05 6.760e-05 4.140e-03
72	6		
73	7		
74	16		
75	7		
76	17	Integrin cell surface interactions (R-HSA-216083) Signal transduction by L1 (R-HSA-445144) Extracellular matrix organization (R-HSA-1474244)	1.540e-05 6.800e-04 2.830e-03
77	17	Signaling by Rho GTPases, Miro GTPases and RHOBTB3 (R-HSA-9716542) Signaling by Rho GTPases (R-HSA-194315) Striated Muscle Contraction (R-HSA-390522)	8.110e-04 1.400e-03 2.820e-03
78	12		
79	17	Folding of actin by CCT/TriC (R-HSA-390450) Formation of tubulin folding intermediates by CCT/TriC (R-HSA-389960) Prefoldin mediated transfer of substrate to CCT/TriC (R-HSA-389957)	3.520e-21 6.070e-17 8.040e-17
80	8	Cilium Assembly (R-HSA-5617833) Intraflagellar transport (R-HSA-5620924) Organelle biogenesis and maintenance (R-HSA-1852241)	5.790e-06 7.560e-06 2.570e-05
81	18	Metabolism of vitamin K (R-HSA-6806664) Fatty acyl-CoA biosynthesis (R-HSA-75105)	5.570e-03 5.980e-03
82	10		
83	8		
84	11	SARS-CoV-1-host interactions (R-HSA-9692914)	4.060e-02
85	6		
86	17	Signaling by activated point mutants of FGFR1 (R-HSA-1839122)	3.090e-02
87	9		
88	8	Terminal pathway of complement (R-HSA-166665)	9.520e-03
89	9		
90	20		
91	31		
92	11		
93	28	WNT 5:FZD7-mediated leishmania damping (R-HSA-9673324) Killing mechanisms (R-HSA-9664420) RHO GTPases Activate NADPH Oxidases (R-HSA-5668599)	6.330e-04 1.270e-03 3.840e-03
94	20	Interleukin receptor SHC signaling (R-HSA-912526) Interleukin-3, Interleukin-5 and GM-CSF signaling (R-HSA-512988) Interleukin-2 family signaling (R-HSA-451927)	5.170e-03 1.060e-02 1.300e-02
95	10		
96	11		
97	5		
98	6		
99	6		
100	30		
101	8	RHO GTPases activate PAKs (R-HSA-5627123) RHO GTPases Activate ROCKs (R-HSA-5627117) RHO GTPases activate CIT (R-HSA-5625900)	2.370e-02 2.900e-02 5.800e-02
102	16		
103	10		
104	8		

105	9	Complement cascade (R-HSA-166658) Regulation of Complement cascade (R-HSA-977606) Activation of C3 and C5 (R-HSA-174577)	1.470e-17 5.910e-15 5.050e-10
106	11	Butyrophilin (BTN) family interactions (R-HSA-8851680)	4.400e-02
107	28		
108	12		
109	12		
110	7	Platelet Adhesion to exposed collagen (R-HSA-75892) Regulation of signaling by CBL (R-HSA-912631) GPVI-mediated activation cascade (R-HSA-114604)	2.810e-05 4.110e-05 1.350e-04
111	15		
112	7		
113	9	Signaling by NTRK3 (TRKC) (R-HSA-9034015) Signaling by NTRK2 (TRKB) (R-HSA-9006115) Downstream signal transduction (R-HSA-186763)	5.400e-05 9.850e-05 1.280e-04
114	7		
115	10		
116	31	mRNA Splicing (R-HSA-72172) mRNA Splicing - Major Pathway (R-HSA-72163) Processing of Capped Intron-Containing Pre-mRNA (R-HSA-72203)	1.670e-08 2.370e-08 1.250e-07
117	65	Formation of the cornified envelope (R-HSA-6809371) Keratinization (R-HSA-6805567) Developmental Biology (R-HSA-1266738)	6.020e-36 3.250e-30 1.160e-13
118	20		
119	12		
120	6		
121	13	Incretin synthesis, secretion, and inactivation (R-HSA-400508) Glucagon-type ligand receptors (R-HSA-420092) GPCR ligand binding (R-HSA-500792)	7.830e-04 1.050e-03 3.390e-03
122	9	Postsynaptic nicotinic acetylcholine receptors (R-HSA-622327) Presynaptic nicotinic acetylcholine receptors (R-HSA-622323) Highly calcium permeable nicotinic acetylcholine receptors (R-HSA-629597)	6.620e-03 7.200e-03 7.860e-03
123	23		
124	25		
125	5		
126	5	Sensory perception of salty taste (R-HSA-9730628) Sensory perception of taste (R-HSA-9717189) Stimuli-sensing channels (R-HSA-2672351)	3.540e-07 1.430e-04 1.130e-03
127	21		
128	16		
129	16		
130	7		
131	6		
132	24	SHOC2 M1731 mutant abolishes MRAS complex function (R-HSA-9726840) Gain-of-function MRAS complexes activate RAF signaling (R-HSA-9726842) Signaling by MRAS-complex mutants (R-HSA-9660537)	3.120e-02 4.670e-02 9.350e-02
133	14	Myogenesis (R-HSA-525793)	2.090e-03
134	6		
135	12		
136	5	XAV939 stabilizes AXIN (R-HSA-5545619) Signaling by WNT in cancer (R-HSA-4791275)	1.220e-04 3.400e-02
137	53	Neddylation (R-HSA-8951664) Macroautophagy (R-HSA-1632852) Formation of the cornified envelope (R-HSA-6809371)	6.540e-08 1.090e-03 1.250e-03

A.1.4 Enriched reactome pathways using pipeline 4

Table A.13: Listed are the significant reactome pathways enriched in clusters predicted by ClusterONE from the differential network comparing all N samples with all M samples using pipeline IV.

Cluster	Size	Reactome pathways	Adjusted P-value
1	37		
2	42	Chemokine receptors bind chemokines (R-HSA-380108) Peptide ligand-binding receptors (R-HSA-375276) Class A/1 (Rhodopsin-like receptors) (R-HSA-373076)	9.460e-47 4.190e-33 7.430e-28
3	32	Chemokine receptors bind chemokines (R-HSA-380108) Peptide ligand-binding receptors (R-HSA-375276) Class A/1 (Rhodopsin-like receptors) (R-HSA-373076)	9.010e-42 3.630e-30 1.210e-25
4	35		
5	27	CREB phosphorylation (R-HSA-199920) Inhibition of DNA recombination at telomere (R-HSA-9670095) Cellular Senescence (R-HSA-2559583)	1.780e-02 2.020e-02 2.120e-02
6	24		
7	39	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell (R-HSA-198933) Antigen Presentation: Folding, assembly and peptide loading of class I MHC (R-HSA-983170)	4.590e-03 3.170e-02
8	26		
9	33		
10	23		
11	26		
12	21	trans-Golgi Network Vesicle Budding (R-HSA-199992) Formation of annular gap junctions (R-HSA-196025) Gap junction degradation (R-HSA-190873)	1.980e-07 3.380e-07 3.380e-07
13	20		
14	23		
15	29		
16	25		
17	25	Transferrin endocytosis and recycling (R-HSA-917977) ROS and RNS production in phagocytes (R-HSA-1222556) Insulin receptor recycling (R-HSA-77387)	3.690e-07 5.440e-07 6.200e-07
18	21	Gastrulation (R-HSA-9758941) Developmental Biology (R-HSA-1266738) Regulation of beta-cell development (R-HSA-186712)	1.890e-06 8.020e-03 1.050e-02
19	30	RUNX2 regulates bone development (R-HSA-8941326)	2.560e-02
20	32		
21	32	Ovarian tumor domain proteases (R-HSA-5689896) Regulation of TNFR1 signaling (R-HSA-5357905) Negative regulators of DDX58/IFIH1 signaling (R-HSA-936440)	3.860e-02 4.250e-02 5.560e-02
22	14	RAB geranylgeranylation (R-HSA-8873719) RAB GEFs exchange GTP for GDP on RABs (R-HSA-8876198) Rab regulation of trafficking (R-HSA-9007101)	1.350e-06 3.370e-06 1.060e-05
23	28	Platelet activation, signaling and aggregation (R-HSA-76002) GPVI-mediated activation cascade (R-HSA-114604) Anti-inflammatory response favouring Leishmania parasite infection (R-HSA-9662851)	2.940e-03 1.840e-02 1.850e-02
24	25		
25	12	Tachykinin receptors bind tachykinins (R-HSA-380095) Class A/1 (Rhodopsin-like receptors) (R-HSA-373076) GPCR ligand binding (R-HSA-500792)	3.890e-06 1.810e-05 6.720e-05
26	19	Signaling by CTNNB1 phospho-site mutants (R-HSA-4839743) CTNNB1 S33 mutants aren't phosphorylated (R-HSA-5358747) Platelet sensitization by LDL (R-HSA-432142)	1.670e-02 1.810e-02 1.870e-02
27	27	Infectious disease (R-HSA-5663205) Disease (R-HSA-1643685) SARS-CoV Infections (R-HSA-9679506)	3.210e-03 3.770e-03 4.550e-03
28	32	Keratinization (R-HSA-6805567) Developmental Biology (R-HSA-1266738)	2.550e-10 1.220e-03

29	21	rRNA processing (R-HSA-72312) rRNA processing in the nucleus and cytosol (R-HSA-8868773) Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC) (R-HSA-975956)	5.270e-05 7.800e-05 7.150e-04
30	18		
31	20	FGFR3 mutant receptor activation (R-HSA-2033514) Signaling by activated point mutants of FGFR3 (R-HSA-1839130)	3.730e-02 7.450e-02
32	14	Cargo recognition for clathrin-mediated endocytosis (R-HSA-8856825) Negative regulation of MET activity (R-HSA-6807004) Clathrin-mediated endocytosis (R-HSA-8856828)	4.440e-02 4.470e-02 4.690e-02
33	14		
34	20	Telomere Extension By Telomerase (R-HSA-171319) Extension of Telomeres (R-HSA-180786) Telomere Maintenance (R-HSA-157579)	1.560e-13 3.630e-11 1.790e-09
35	19		
36	13	Class A/1 (Rhodopsin-like receptors) (R-HSA-373076) GPCR ligand binding (R-HSA-500792) GPCR downstream signalling (R-HSA-388396)	6.630e-05 2.450e-04 9.550e-04
37	14		
38	18		
39	10	IL-6-type cytokine receptor ligand interactions (R-HSA-6788467) Interleukin-6 family signaling (R-HSA-6783589)	1.440e-04 2.140e-04
40	13		
41	13	Activation of Matrix Metalloproteinases (R-HSA-1592389) Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs) (R-HSA-381426) Extracellular matrix organization (R-HSA-1474244)	6.830e-04 7.370e-04 9.250e-04
42	7	DDX58/IFIH1-mediated induction of interferon-alpha/beta (R-HSA-168928) Interferon alpha/beta signaling (R-HSA-909733) Regulation of IFNA/IFNB signaling (R-HSA-912694)	2.230e-08 3.220e-08 3.550e-08
43	7		
44	11		
45	14		
46	14	Germ layer formation at gastrulation (R-HSA-9754189) Formation of definitive endoderm (R-HSA-9823730) Condensation of Prometaphase Chromosomes (R-HSA-2514853)	4.980e-02 5.780e-02 6.060e-02
47	16	Nervous system development (R-HSA-9675108) Axon guidance (R-HSA-422475) RET signaling (R-HSA-8853659)	9.840e-04 1.140e-03 1.160e-03
48	29	IRE1alpha activates chaperones (R-HSA-381070) XBP1(S) activates chaperone genes (R-HSA-381038) Unfolded Protein Response (UPR) (R-HSA-381119)	8.280e-04 1.400e-03 6.980e-03
49	13		
50	15		
51	10		
52	10		
53	6		
54	6		
55	6		
56	6		
57	13		
58	12	Condensation of Prometaphase Chromosomes (R-HSA-2514853) Formation of Senescence-Associated Heterochromatin Foci (SAHF) (R-HSA-2559584)	4.400e-02 4.790e-02
59	18		
60	7		
61	7		

62	11	Lectin pathway of complement activation (R-HSA-166662) Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs) (R-HSA-381426) Complement cascade (R-HSA-166658)	1.150e-02 2.130e-02 2.900e-02
63	12		
64	8	Cilium Assembly (R-HSA-5617833) Intraflagellar transport (R-HSA-5620924) Organelle biogenesis and maintenance (R-HSA-1852241)	5.790e-06 7.560e-06 2.570e-05
65	9		
66	9	Diseases associated with the TLR signaling cascade (R-HSA-5602358) Diseases of Immune System (R-HSA-5260271) Toll-like Receptor Cascades (R-HSA-168898)	3.320e-04 6.640e-04 3.520e-02
67	7	Prolactin receptor signaling (R-HSA-1170546) Growth hormone receptor signaling (R-HSA-982772)	2.320e-02 3.220e-02
68	6		
69	14		
70	13		
71	5		
72	8		
73	9		
74	9		
75	10		
76	7		
77	7		
78	28		
79	11		
80	10	Postsynaptic nicotinic acetylcholine receptors (R-HSA-622327) Presynaptic nicotinic acetylcholine receptors (R-HSA-622323) Highly calcium permeable postsynaptic nicotinic acetylcholine receptors (R-HSA-629594)	1.540e-05 1.550e-05 1.750e-05
81	10	Signaling by Hippo (R-HSA-2028269)	2.050e-04
82	6	ADP signalling through P2Y purinoceptor 12 (R-HSA-392170) Adrenaline, noradrenaline inhibits insulin secretion (R-HSA-400042) Signal amplification (R-HSA-392518)	5.440e-05 5.780e-05 5.840e-05
83	6		
84	11		
85	7	Neutrophil degranulation (R-HSA-6798695)	2.450e-02
86	7		
87	11		
88	6		
89	7	Interleukin-1 processing (R-HSA-448706) Signaling by Interleukins (R-HSA-449147) Interleukin-1 family signaling (R-HSA-446652)	9.180e-03 1.040e-02 1.110e-02
90	7	RNA Polymerase II Transcription Initiation And Promoter Clearance (R-HSA-76042) HIV Transcription Initiation (R-HSA-167161) RNA Polymerase II HIV Promoter Escape (R-HSA-167162)	3.040e-02 3.420e-02 3.910e-02
91	8		
92	9		
93	8		
94	10		
95	11		
96	7	PRC2 methylates histones and DNA (R-HSA-212300) ERCC6 (CSB) and EHMT2 (G9a) positively regulate rRNA expression (R-HSA-427389) Defective pyroptosis (R-HSA-9710421)	8.210e-05 9.060e-05 9.380e-05
97	10	Initial triggering of complement (R-HSA-166663) Creation of C4 and C2 activators (R-HSA-166786) Classical antibody-mediated complement activation (R-HSA-173623)	4.780e-05 4.800e-05 6.520e-05
98	5		
99	8		
100	9	Terminal pathway of complement (R-HSA-166665)	1.220e-02

101	7		
102	6		
103	9		
104	12	Signaling by Overexpressed Wild-Type EGFR in Cancer (R-HSA-5638302) EGFR interacts with phospholipase C-gamma (R-HSA-212718) GRB2 events in EGFR signaling (R-HSA-179812)	1.090e-05 1.090e-05 1.280e-05
105	5		
106	5	Meiotic synapsis (R-HSA-1221632) Meiosis (R-HSA-1500620) Reproduction (R-HSA-1474165)	5.160e-04 8.690e-04 1.770e-03
107	7		
108	9		
109	6		
110	10		
111	13	PTEN Regulation (R-HSA-6807070) Regulation of PTEN gene transcription (R-HSA-8943724) PIP3 activates AKT signaling (R-HSA-1257604)	3.300e-03 8.060e-03 1.230e-02
112	7		
113	6		
114	9		
115	12		
116	8		
117	7	Resolution of Sister Chromatid Cohesion (R-HSA-2500257) EML4 and NUDC in mitotic spindle formation (R-HSA-9648025) Amplification of signal from the kinetochores (R-HSA-141424)	2.120e-03 2.160e-03 2.200e-03
118	7		
119	9		
120	7		
121	7		
122	5		
123	9	Negative regulation of TCF-dependent signaling by WNT ligand antagonists (R-HSA-3772470)	4.580e-02
124	6		
125	15		
126	13		
127	9		
128	6	DAP12 signaling (R-HSA-2424491) Nuclear signaling by ERBB4 (R-HSA-1251985) Other semaphorin interactions (R-HSA-416700)	2.730e-02 2.810e-02 3.110e-02
129	6		
130	6		
131	9		
132	5		
133	5	Sensory perception of salty taste (R-HSA-9730628) Sensory perception of taste (R-HSA-9717189) Stimuli-sensing channels (R-HSA-2672351)	3.540e-07 1.430e-04 1.130e-03
134	5	Complement cascade (R-HSA-166658)	4.930e-03
135	8		
136	5		
137	5		
138	8	Eukaryotic Translation Initiation (R-HSA-72613) Selenoamino acid metabolism (R-HSA-2408522) Cap-dependent Translation Initiation (R-HSA-72737)	1.700e-03 1.800e-03 1.820e-03
139	7		
140	8	DNA Damage/Telomere Stress Induced Senescence (R-HSA-2559586) HCMV Late Events (R-HSA-9610379) HDACs deacetylate histones (R-HSA-3214815)	1.680e-03 2.580e-03 3.360e-03
141	7		
142	9		
143	5		
144	5		

145	5	Negative regulation of the PI3K/AKT network (R-HSA-199418) PI5P, PP2A and IER3 Regulate PI3K/AKT Signaling (R-HSA-6811558) Estrogen-dependent nuclear events downstream of ESR-membrane signaling (R-HSA-9634638)	3.220e-06 4.790e-06 1.190e-05
146	6		
147	6		
148	8		
149	17	eNOS activation (R-HSA-203615) Tetrahydrobiopterin (BH4) synthesis, recycling, salvage and regulation (R-HSA-1474151)	4.520e-02 7.410e-02
150	8		
151	8		
152	5		

Table A.14: Listed are the significant reactome pathways enriched in clusters predicted by ClusterONE from the interactions only found in M samples and not in N samples using pipeline IV.

Cluster	Size	Reactome pathways	Adjusted P-value
1	39	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell (R-HSA-198933) Antigen Presentation: Folding, assembly and peptide loading of class I MHC (R-HSA-983170)	4.590e-03 3.170e-02
2	34		
3	37	Chemokine receptors bind chemokines (R-HSA-380108) Peptide ligand-binding receptors (R-HSA-375276) Class A/1 (Rhodopsin-like receptors) (R-HSA-373076)	9.180e-43 1.720e-30 9.880e-26
4	36		
5	25	Transferrin endocytosis and recycling (R-HSA-917977) ROS and RNS production in phagocytes (R-HSA-1222556) Insulin receptor recycling (R-HSA-77387)	3.690e-07 5.440e-07 6.200e-07
6	24		
7	26		
8	35		
9	36		
10	22	Gastrulation (R-HSA-9758941) Regulation of beta-cell development (R-HSA-186712) Developmental Biology (R-HSA-1266738)	2.580e-06 1.010e-04 8.050e-04
11	27		
12	23		
13	24	HCMV Infection (R-HSA-9609646) CREB phosphorylation (R-HSA-199920) Generic Transcription Pathway (R-HSA-212436)	1.070e-02 1.170e-02 1.250e-02
14	21	trans-Golgi Network Vesicle Budding (R-HSA-199992) Formation of annular gap junctions (R-HSA-196025) Gap junction degradation (R-HSA-190873)	1.980e-07 3.380e-07 3.380e-07
15	25		
16	22		
17	34		
18	17		
19	31	RUNX2 regulates bone development (R-HSA-8941326)	2.840e-02
20	28	Infectious disease (R-HSA-5663205) Disease (R-HSA-1643685) Antigen activates B Cell Receptor (BCR) leading to generation of second messengers (R-HSA-983695)	4.540e-03 5.740e-03 5.780e-03
21	29	Platelet activation, signaling and aggregation (R-HSA-76002) GPVI-mediated activation cascade (R-HSA-114604) Anti-inflammatory response favouring Leishmania parasite infection (R-HSA-9662851)	3.670e-03 2.050e-02 2.130e-02

22	18	Signaling by CTNNB1 phospho-site mutants (R-HSA-4839743) Metabolism of carbohydrates (R-HSA-71387) CTNNB1 S33 mutants aren't phosphorylated (R-HSA-5358747)	1.490e-02 1.580e-02 1.620e-02
23	18		
24	22	Telomere Extension By Telomerase (R-HSA-171319) Extension of Telomeres (R-HSA-180786) Chromosome Maintenance (R-HSA-73886)	3.430e-13 7.950e-11 2.510e-10
25	13	RAB geranylgeranylation (R-HSA-8873719) RAB GEFs exchange GTP for GDP on RABs (R-HSA-8876198) Rab regulation of trafficking (R-HSA-9007101)	8.710e-07 2.170e-06 6.860e-06
26	12		
27	12		
28	18	Downstream signal transduction (R-HSA-186763) Axon guidance (R-HSA-422475) Nervous system development (R-HSA-9675108)	9.260e-06 7.440e-05 7.550e-05
29	29	IRE1alpha activates chaperones (R-HSA-381070) XBP1(S) activates chaperone genes (R-HSA-381038) Unfolded Protein Response (UPR) (R-HSA-381119)	8.280e-04 1.400e-03 6.980e-03
30	17		
31	9		
32	14		
33	17		
34	10		
35	10	IL-6-type cytokine receptor ligand interactions (R-HSA-6788467) Interleukin-6 family signaling (R-HSA-6783589)	1.440e-04 2.140e-04
36	13	Activation of Matrix Metalloproteinases (R-HSA-1592389) Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs) (R-HSA-381426) Extracellular matrix organization (R-HSA-1474244)	6.830e-04 7.370e-04 9.250e-04
37	13		
38	7		
39	12		
40	9		
41	10		
42	15		
43	14		
44	6		
45	12	Metal sequestration by antimicrobial proteins (R-HSA-6799990) IRAK4 deficiency (TLR2/4) (R-HSA-5603041) Regulation of TLR by endogenous ligand (R-HSA-5686938)	1.200e-02 4.070e-02 4.180e-02
46	9		
47	9		
48	7		
49	10		
50	12	Condensation of Prometaphase Chromosomes (R-HSA-2514853) Formation of Senescence-Associated Heterochromatin Foci (SAHF) (R-HSA-2559584)	4.400e-02 4.790e-02
51	12		
52	12		
53	10		
54	9		
55	9	Diseases associated with the TLR signaling cascade (R-HSA-5602358) Diseases of Immune System (R-HSA-5260271) Toll-like Receptor Cascades (R-HSA-168898)	3.320e-04 6.640e-04 3.520e-02
56	11		
57	14		
58	5	DDX58/IFIH1-mediated induction of interferon-alpha/beta (R-HSA-168928) Interferon alpha/beta signaling (R-HSA-909733) SARS-CoV-2 activates/modulates innate and adaptive immune responses (R-HSA-9705671)	1.430e-06 2.200e-06 4.850e-06
59	5		

60	9		
61	8	Lectin pathway of complement activation (R-HSA-166662) Complement cascade (R-HSA-166658)	5.350e-03 8.550e-03
62	7	PRC2 methylates histones and DNA (R-HSA-212300) ERCC6 (CSB) and EHMT2 (G9a) positively regulate rRNA expression (R-HSA-427389) Defective pyroptosis (R-HSA-9710421)	8.210e-05 9.060e-05 9.380e-05
63	12		
64	10	Postsynaptic nicotinic acetylcholine receptors (R-HSA-622327) Presynaptic nicotinic acetylcholine receptors (R-HSA-622323) Highly calcium permeable postsynaptic nicotinic acetylcholine receptors (R-HSA-629594)	1.540e-05 1.550e-05 1.750e-05
65	10	Signaling by Hippo (R-HSA-2028269)	2.050e-04
66	6	ADP signalling through P2Y purinoceptor 12 (R-HSA-392170) Adrenaline, noradrenaline inhibits insulin secretion (R-HSA-400042) Signal amplification (R-HSA-392518)	5.440e-05 5.780e-05 5.840e-05
67	6		
68	7	Complement cascade (R-HSA-166658) Regulation of Complement cascade (R-HSA-977606) Innate Immune System (R-HSA-168249)	3.310e-07 3.430e-05 1.420e-04
69	9		
70	7	Interleukin-1 processing (R-HSA-448706) Signaling by Interleukins (R-HSA-449147) Interleukin-1 family signaling (R-HSA-446652)	9.180e-03 1.040e-02 1.110e-02
71	7	RNA Polymerase II Transcription Initiation And Promoter Clearance (R-HSA-76042) HIV Transcription Initiation (R-HSA-167161) RNA Polymerase II HIV Promoter Escape (R-HSA-167162)	3.040e-02 3.420e-02 3.910e-02
72	14		
73	7		
74	8	Formation of the cornified envelope (R-HSA-6809371)	3.390e-02
75	6		
76	9	DNA Damage/Telomere Stress Induced Senescence (R-HSA-2559586) HCMV Late Events (R-HSA-9610379) HDACs deacetylate histones (R-HSA-3214815)	2.510e-03 3.860e-03 5.030e-03
77	14		
78	5		
79	9		
80	9		
81	7		
82	6	Insertion of tail-anchored proteins into the endoplasmic reticulum membrane (R-HSA-9609523)	3.110e-02
83	10	RUNX2 regulates osteoblast differentiation (R-HSA-8940973) Transcriptional regulation by RUNX2 (R-HSA-8878166) Signaling by CSF1 (M-CSF) in myeloid cells (R-HSA-9680350)	3.670e-02 3.770e-02 3.780e-02
84	12	Signaling by Overexpressed Wild-Type EGFR in Cancer (R-HSA-5638302) EGFR interacts with phospholipase C-gamma (R-HSA-212718) GRB2 events in EGFR signaling (R-HSA-179812)	1.090e-05 1.090e-05 1.280e-05
85	5	Meiotic synapsis (R-HSA-1221632) Meiosis (R-HSA-1500620) Reproduction (R-HSA-1474165)	5.160e-04 8.690e-04 1.770e-03
86	5		
87	8	Role of phospholipids in phagocytosis (R-HSA-2029485) Creation of C4 and C2 activators (R-HSA-166786) Initial triggering of complement (R-HSA-166663)	2.140e-03 2.380e-03 2.410e-03
88	6		
89	7		
90	7		
91	9	Negative regulation of the PI3K/AKT network (R-HSA-199418) PI5P, PP2A and IER3 Regulate PI3K/AKT Signaling (R-HSA-6811558) Diseases of signal transduction by growth factor receptors and second messengers (R-HSA-5663202)	3.710e-07 5.100e-07 4.620e-06
92	6		

93	8		
94	7		
95	6		
96	7		
97	7		
98	7		
99	7		
100	7		
101	25		
102	5		
103	9		
104	5		
105	5		
106	5		
107	7		
108	6	Germ layer formation at gastrulation (R-HSA-9754189) Formation of definitive endoderm (R-HSA-9823730)	1.240e-02 1.910e-02
109	6	Resolution of Sister Chromatid Cohesion (R-HSA-2500257) EML4 and NUDC in mitotic spindle formation (R-HSA-9648025) Amplification of signal from the kinetochores (R-HSA-141424)	2.120e-03 2.160e-03 2.200e-03
110	12		
111	6		
112	6		
113	6	ABC transporters in lipid homeostasis (R-HSA-1369062)	2.780e-02
114	6	DAP12 signaling (R-HSA-2424491) Nuclear signaling by ERBB4 (R-HSA-1251985) Other semaphorin interactions (R-HSA-416700)	2.730e-02 2.810e-02 3.110e-02
115	6		
116	6		
117	9		
118	11	Signaling by ERBB4 (R-HSA-1236394) Defective POMT2 causes MDDGA2, MDDGB2 and MDDGC2 (R-HSA-5083629) Defective POMT1 causes MDDGA1, MDDGB1 and MDDGC1 (R-HSA-5083633)	3.540e-05 6.680e-04 1.000e-03
119	5		
120	9		

Table A.15: Listed are the significant reactome pathways enriched in clusters predicted by ClusterONE from the interactions only found in N samples and not in M samples using pipeline IV.

Cluster	Size	Reactome pathways	Adjusted P-value
1	21	rRNA processing (R-HSA-72312) rRNA processing in the nucleus and cytosol (R-HSA-8868773) Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC) (R-HSA-975956)	5.270e-05 7.800e-05 7.150e-04
2	27	Chk1/Chk2(Cds1) mediated inactivation of Cyclin B:Cdk1 complex (R-HSA-75035) Activation of BAD and translocation to mitochondria (R-HSA-111447) SARS-CoV-2 targets host intracellular signalling and regulatory pathways (R-HSA-9755779)	4.310e-09 5.020e-09 5.310e-09
3	20		
4	14	Cargo recognition for clathrin-mediated endocytosis (R-HSA-8856825) Negative regulation of MET activity (R-HSA-6807004) Clathrin-mediated endocytosis (R-HSA-8856828)	4.440e-02 4.470e-02 4.690e-02
5	21		
6	18		
7	15		

8	12		
9	15	PTEN Regulation (R-HSA-6807070) Regulation of PTEN gene transcription (R-HSA-8943724) Chromatin organization (R-HSA-4839726)	6.230e-03 1.280e-02 1.380e-02
10	12	Terminal pathway of complement (R-HSA-166665) Parasite infection (R-HSA-9664407) Nephrin family interactions (R-HSA-373753)	2.240e-02 2.630e-02 3.360e-02
11	19		
12	23	Keratinization (R-HSA-6805567)	2.330e-06
13	10		
14	14		
15	18	Keratinization (R-HSA-6805567)	3.700e-05
16	9	Cilium Assembly (R-HSA-5617833) Intraflagellar transport (R-HSA-5620924) Organelle biogenesis and maintenance (R-HSA-1852241)	1.290e-05 1.360e-05 5.720e-05
17	12		
18	10		
19	11		
20	7		
21	7		
22	8	Serotonin receptors (R-HSA-390666)	2.240e-02
23	8		
24	9		
25	15		
26	17	Keratinization (R-HSA-6805567)	1.430e-02
27	6		
28	6		
29	6		
30	6	Sensory perception of salty taste (R-HSA-9730628) Sensory perception of taste (R-HSA-9717189) Stimuli-sensing channels (R-HSA-2672351)	7.080e-07 2.860e-04 2.250e-03
31	6	Neutrophil degranulation (R-HSA-6798695)	1.070e-02
32	6		
33	6		
34	10		
35	6		
36	6	Prolactin receptor signaling (R-HSA-1170546) Growth hormone receptor signaling (R-HSA-982772)	1.660e-02 2.300e-02
37	10		
38	10		
39	9		
40	13		
41	5		
42	9		
43	13		
44	11		
45	7	Cell Cycle, Mitotic (R-HSA-69278) Cell Cycle (R-HSA-1640170) Cell Cycle Checkpoints (R-HSA-69620)	5.030e-04 7.740e-04 8.420e-04
46	8	Influenza Infection (R-HSA-168255) Eukaryotic Translation Initiation (R-HSA-72613) Selenoamino acid metabolism (R-HSA-2408522)	5.300e-04 1.510e-03 1.580e-03
47	12	FGFR3 mutant receptor activation (R-HSA-2033514) Signaling by activated point mutants of FGFR3 (R-HSA-1839130) Signaling by FGFR3 in disease (R-HSA-5655332)	1.440e-02 2.880e-02 3.620e-02
48	9	FGFR3 mutant receptor activation (R-HSA-2033514) Signaling by activated point mutants of FGFR3 (R-HSA-1839130) Signaling by FGFR3 in disease (R-HSA-5655332)	7.860e-03 1.570e-02 1.980e-02
49	11		
50	5	Common Pathway of Fibrin Clot Formation (R-HSA-140875) Formation of Fibrin Clot (Clotting Cascade) (R-HSA-140877)	2.800e-02 4.490e-02
51	5		
52	5		

53	6		
54	6		
55	8		
56	5		
57	6		
58	5	Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs) (R-HSA-381426)	5.440e-03
59	6		
60	6		
61	5		

A.1.5 Enriched reactome pathways using pipeline 5

Table A.16: Listed are the significant reactome pathways enriched in clusters predicted by ClusterONE from the differential network comparing all N samples with all M samples using pipeline V.

Cluster	Size	Reactome pathways	Adjusted P-value
1	40	Gene expression (Transcription) (R-HSA-74160) Generic Transcription Pathway (R-HSA-212436) Formation of WDR5-containing histone-modifying complexes (R-HSA-9772755)	1.660e-05 1.640e-03 1.670e-03
2	40		
3	50	RAB geranylgeranylation (R-HSA-8873719) Neutrophil degranulation (R-HSA-6798695) Rab regulation of trafficking (R-HSA-9007101)	4.760e-17 8.880e-17 2.430e-10
4	40		
5	29		
6	36	Chemokine receptors bind chemokines (R-HSA-380108) Peptide ligand-binding receptors (R-HSA-375276) Class A/1 (Rhodopsin-like receptors) (R-HSA-373076)	1.850e-46 1.720e-33 1.780e-28
7	27	Integrin cell surface interactions (R-HSA-216083) Laminin interactions (R-HSA-3000157) Extracellular matrix organization (R-HSA-1474244)	1.940e-04 1.030e-02 3.280e-02
8	23		
9	30		
10	21		
11	18		
12	31	Hemostasis (R-HSA-109582) Estrogen-stimulated signaling through PRKCZ (R-HSA-9634635) Platelet activation, signaling and aggregation (R-HSA-76002)	1.260e-02 1.880e-02 2.070e-02
13	31		
14	40	FCGR3A-mediated IL10 synthesis (R-HSA-9664323) Anti-inflammatory response favouring Leishmania parasite infection (R-HSA-9662851) Leishmania parasite growth and survival (R-HSA-9664433)	1.480e-06 1.660e-06 1.870e-06
15	18		
16	34		
17	40		
18	23	MET activates PI3K/AKT signaling (R-HSA-8851907) MET promotes cell motility (R-HSA-8875878) MET activates PTPN11 (R-HSA-8865999)	8.490e-03 1.140e-02 1.530e-02
19	35		
20	11		
21	28		
22	19		
23	19		
24	33		
25	9	Interleukin-2 signaling (R-HSA-9020558)	2.240e-02

26	12	GPCR ligand binding (R-HSA-500792) GPCR downstream signalling (R-HSA-388396) Signaling by GPCR (R-HSA-372790)	7.380e-08 4.050e-07 6.460e-07
27	6		
28	32		
29	17		
30	13	Lectin pathway of complement activation (R-HSA-166662) Ficolins bind to repetitive carbohydrate structures on the target cell surface (R-HSA-2855086) Complement cascade (R-HSA-166658)	8.600e-09 1.010e-06 6.580e-06
31	19		
32	36	Mitotic Prometaphase (R-HSA-68877) Anchoring of the basal body to the plasma membrane (R-HSA-5620912) Loss of Nlp from mitotic centrosomes (R-HSA-380259)	1.190e-07 3.690e-07 1.200e-06
33	12		
34	13	Metabolism of amine-derived hormones (R-HSA-209776) Serotonin and melatonin biosynthesis (R-HSA-209931)	4.110e-04 4.730e-03
35	40	Viral mRNA Translation (R-HSA-192823) SRP-dependent cotranslational protein targeting to membrane (R-HSA-1799339) Selenocysteine synthesis (R-HSA-2408557)	1.100e-27 1.300e-27 1.480e-27
36	40		
37	34	Clathrin-mediated endocytosis (R-HSA-8856828) Signaling by EGFR (R-HSA-177929) Signaling by NOTCH1 t(7;9)(NOTCH1:M1580_K2555) Translocation Mutant (R-HSA-2660825)	8.140e-03 3.180e-02 3.350e-02
38	10		
39	14		
40	28	Highly calcium permeable nicotinic acetylcholine receptors (R-HSA-629597) Highly calcium permeable postsynaptic nicotinic acetylcholine receptors (R-HSA-629594) Presynaptic nicotinic acetylcholine receptors (R-HSA-622323)	1.030e-09 1.890e-09 2.150e-09
41	14	Signaling by CTNNB1 phospho-site mutants (R-HSA-4839743) CTNNB1 S33 mutants aren't phosphorylated (R-HSA-5358747) Platelet sensitization by LDL (R-HSA-432142)	9.630e-03 1.050e-02 1.070e-02
42	8	VEGFR2 mediated cell proliferation (R-HSA-5218921) Ethanol oxidation (R-HSA-71384) RHO GTPases Activate NADPH Oxidases (R-HSA-5668599)	2.040e-02 2.240e-02 3.120e-02
43	9	Sensing of DNA Double Strand Breaks (R-HSA-5693548) HDR through MMEJ (alt-NHEJ) (R-HSA-5685939) Defective HDR through Homologous Recombination Repair (HRR) due to PALB2 loss of BRCA1 binding function (R-HSA-9704331)	6.560e-03 1.440e-02 1.720e-02
44	9		
45	5	Interleukin-12 family signaling (R-HSA-447115) Interleukin-23 signaling (R-HSA-9020933) Interleukin-4 and Interleukin-13 signaling (R-HSA-6785807)	1.600e-10 2.170e-10 3.420e-06
46	71	Deactivation of the beta-catenin transactivating complex (R-HSA-3769402) Defects in biotin (Btn) metabolism (R-HSA-3323169) Formation of the beta-catenin:TCF transactivating complex (R-HSA-201722)	7.800e-04 1.340e-03 1.460e-03
47	12	Retinoid cycle disease events (R-HSA-2453864) Diseases of the neuronal system (R-HSA-9675143) Diseases associated with visual transduction (R-HSA-2474795)	2.080e-02 3.120e-02 6.230e-02
48	14		
49	16		
50	10	Long-term potentiation (R-HSA-9620244) Synaptic adhesion-like molecules (R-HSA-8849932) CREB1 phosphorylation through NMDA receptor-mediated activation of RAS signaling (R-HSA-442742)	2.510e-02 2.860e-02 3.180e-02

51	10	Intraflagellar transport (R-HSA-5620924) Cilium Assembly (R-HSA-5617833) Organelle biogenesis and maintenance (R-HSA-1852241)	2.260e-05 2.570e-05 1.130e-04
52	16		
53	13		
54	16	Interferon alpha/beta signaling (R-HSA-909733) Interferon Signaling (R-HSA-913531) Regulation of IFNA/IFNB signaling (R-HSA-912694)	2.020e-10 1.040e-07 1.830e-06
55	32		
56	8		
57	37		
58	10		
59	32	HATs acetylate histones (R-HSA-3214847) Chromatin organization (R-HSA-4839726) Chromatin modifying enzymes (R-HSA-3247509)	8.870e-19 1.160e-14 1.740e-14
60	39	Generic Transcription Pathway (R-HSA-212436) RNA Polymerase II Transcription (R-HSA-73857) Gene expression (Transcription) (R-HSA-74160)	4.610e-04 6.960e-04 1.710e-03
61	5	CREB3 factors activate genes (R-HSA-8874211)	4.370e-03
62	5		
63	15		
64	16	Signal amplification (R-HSA-392518) Thromboxane signalling through TP receptor (R-HSA-428930) Thrombin signalling through proteinase activated receptors (PARs) (R-HSA-456926)	2.160e-16 9.170e-15 4.630e-14
65	10		
66	5		
67	30		
68	13	Complement cascade (R-HSA-166658) Activation of C3 and C5 (R-HSA-174577) Initial triggering of complement (R-HSA-166663)	9.870e-06 1.770e-05 1.610e-04
69	7		
70	6		
71	29	ERBB2 Regulates Cell Motility (R-HSA-6785631) EGRB2 events in ERBB2 signaling (R-HSA-1963640) GRB2 events in EGFR signaling (R-HSA-179812)	4.230e-02 4.760e-02 5.380e-02
72	11	WNT ligand biogenesis and trafficking (R-HSA-3238698) Class B/2 (Secretin family receptors) (R-HSA-373080) Signaling by WNT (R-HSA-195721)	2.530e-09 9.820e-07 2.190e-04
73	9		
74	33	Defective CFTR causes cystic fibrosis (R-HSA-5678895) Hedgehog ligand biogenesis (R-HSA-5358346) Hh mutants abrogate ligand secretion (R-HSA-5387390)	2.540e-05 2.830e-05 2.850e-05
75	32		
76	5		
77	14		
78	14		
79	17	MET activates RAS signaling (R-HSA-8851805)	4.610e-02
80	16		
81	17		
82	16		
83	22		
84	18		
85	6		
86	17	WNT 5A-dependent internalization of FZD4 (R-HSA-5099900) WNT 5A-dependent internalization of FZD2, FZD5 and ROR2 (R-HSA-5140745) trans-Golgi Network Vesicle Budding (R-HSA-199992)	5.560e-07 5.830e-07 2.330e-06
87	31		
88	10		
89	15		

90	10	GABA receptor activation (R-HSA-977443) Signaling by ERBB4 (R-HSA-1236394) Neurotransmitter receptors and postsynaptic signal transmission (R-HSA-112314)	1.740e-05 2.260e-05 1.400e-03
91	14	Innate Immune System (R-HSA-168249) Interleukin-1 family signaling (R-HSA-446652) Immune System (R-HSA-168256)	1.550e-05 3.130e-05 3.610e-05
92	12		
93	10	MHC class II antigen presentation (R-HSA-2132295) HSP90 chaperone cycle for steroid hormone receptors (SHR) in the presence of ligand (R-HSA-3371497) COPI-independent Golgi-to-ER retrograde traffic (R-HSA-6811436)	1.130e-06 3.140e-06 3.500e-06
94	7	mRNA decay by 5′ to 3′ exoribonuclease (R-HSA-430039) mRNA Splicing (R-HSA-72172) mRNA Splicing - Major Pathway (R-HSA-72163)	2.670e-02 3.140e-02 4.190e-02
95	7		
96	11	Creation of C4 and C2 activators (R-HSA-166786) Role of LAT2/NTAL/LAB on calcium mobilization (R-HSA-2730905) FCGR activation (R-HSA-2029481)	1.250e-07 1.310e-07 1.530e-07
97	6		
98	17		
99	11		
100	24		
101	8	Interferon alpha/beta signaling (R-HSA-909733) SARS-CoV-2 activates/modulates innate and adaptive immune responses (R-HSA-9705671) Regulation of IFNA/IFNB signaling (R-HSA-912694)	3.050e-05 6.690e-05 8.750e-05
102	13		
103	29	Factors involved in megakaryocyte development and platelet production (R-HSA-983231)	6.580e-03
104	9	Nuclear Receptor transcription pathway (R-HSA-383280)	1.570e-03
105	9		
106	17	Platelet Adhesion to exposed collagen (R-HSA-75892) Dectin-2 family (R-HSA-5621480) GPVI-mediated activation cascade (R-HSA-114604)	5.440e-04 2.170e-03 2.580e-03
107	28	SARS-CoV-2 modulates autophagy (R-HSA-9754560)	9.500e-04
108	24	Anchoring of the basal body to the plasma membrane (R-HSA-5620912) Recruitment of mitotic centrosome proteins and complexes (R-HSA-380270) Cilium Assembly (R-HSA-5617833)	1.130e-02 3.970e-02 4.060e-02
109	15	Factors involved in megakaryocyte development and platelet production (R-HSA-983231)	1.100e-02
110	35		
111	9		
112	20	Elevation of cytosolic Ca2+ levels (R-HSA-139853) Platelet calcium homeostasis (R-HSA-418360) Role of second messengers in netrin-1 signaling (R-HSA-418890)	3.010e-06 1.680e-05 8.040e-05
113	33		
114	11		
115	8		
116	18	Complement cascade (R-HSA-166658) Regulation of Complement cascade (R-HSA-977606) Laminin interactions (R-HSA-3000157)	1.560e-06 4.140e-05 1.440e-03
117	6		
118	21	TNFs bind their physiological receptors (R-HSA-5669034) TRAF6 mediated IRF7 activation in TLR7/8 or 9 signaling (R-HSA-975110) TNFR2 non-canonical NF-kB pathway (R-HSA-5668541)	4.150e-05 3.350e-04 2.330e-03
119	21	Highly calcium permeable nicotinic acetylcholine receptors (R-HSA-629597) Highly calcium permeable postsynaptic nicotinic acetylcholine receptors (R-HSA-629594) Presynaptic nicotinic acetylcholine receptors (R-HSA-622323)	2.590e-07 3.380e-07 3.380e-07

120	10	Fertilization (R-HSA-1187000) Interaction With Cumulus Cells And The Zona Pellucida (R-HSA-2534343) Reproduction (R-HSA-1474165)	1.380e-09 1.190e-08 1.480e-06
121	13		
122	26	Ovarian tumor domain proteases (R-HSA-5689896) Regulation of TNFR1 signaling (R-HSA-5357905) TNF signaling (R-HSA-75893)	2.040e-02 2.250e-02 2.890e-02
123	18		
124	7	Phase 0 - rapid depolarisation (R-HSA-5576892) Cardiac conduction (R-HSA-5576891) Muscle contraction (R-HSA-397014)	4.310e-07 6.740e-05 2.750e-04
125	28		
126	10		
127	13		
128	16		
129	27	Neurotoxicity of clostridium toxins (R-HSA-168799) Toxicity of botulinum toxin type C (botC) (R-HSA-5250971) Uptake and actions of bacterial toxins (R-HSA-5339562)	1.260e-06 2.590e-06 2.500e-05
130	24	Antigen activates B Cell Receptor (BCR) leading to generation of second messengers (R-HSA-983695) Signaling by the B Cell Receptor (BCR) (R-HSA-983705) CD22 mediated BCR regulation (R-HSA-5690714)	2.050e-06 4.290e-05 6.670e-04
131	11	Nuclear Receptor transcription pathway (R-HSA-383280)	7.250e-06
132	7		
133	12	Highly calcium permeable nicotinic acetylcholine receptors (R-HSA-629597) Highly calcium permeable postsynaptic nicotinic acetylcholine receptors (R-HSA-629594) Presynaptic nicotinic acetylcholine receptors (R-HSA-622323)	2.140e-08 2.800e-08 2.800e-08
134	40	M Phase (R-HSA-68886) DNA Double Strand Break Response (R-HSA-5693606) Cell Cycle Checkpoints (R-HSA-69620)	5.890e-03 7.570e-03 8.170e-03
135	6		
136	6		
137	29	Chromatin organization (R-HSA-4839726) Heme signaling (R-HSA-9707616) HATs acetylate histones (R-HSA-3214847)	3.510e-04 4.020e-04 4.160e-04
138	18	CREB1 phosphorylation through the activation of Adenylate Cyclase (R-HSA-442720) PKA activation (R-HSA-163615) PKA-mediated phosphorylation of CREB (R-HSA-111931)	2.640e-04 3.250e-04 3.400e-04
139	7		
140	12	Metal sequestration by antimicrobial proteins (R-HSA-6799990)	1.200e-02
141	22		
142	12		
143	13		
144	18	Gastrulation (R-HSA-9758941) Developmental Biology (R-HSA-1266738) Formation of intermediate mesoderm (R-HSA-9761174)	6.250e-05 2.530e-02 2.590e-02
145	6		
146	10		
147	26		
148	11		
149	11		
150	9		
151	11		
152	10		
153	21		
154	7		
155	13		

Table A.17: Listed are the significant reactome pathways enriched in clusters predicted by ClusterONE from the interactions only found in M samples and not in N samples using pipeline V.

Cluster	Size	Reactome pathways	Adjusted P-value
1	39	Viral mRNA Translation (R-HSA-192823) SRP-dependent cotranslational protein targeting to membrane (R-HSA-1799339) Selenocysteine synthesis (R-HSA-2408557)	2.750e-30 3.690e-30 3.910e-30
2	51	Formation of the cornified envelope (R-HSA-6809371) Keratinization (R-HSA-6805567)	1.780e-06 5.360e-05
3	33		
4	40	Oxidative Stress Induced Senescence (R-HSA-2559580) Cellular Senescence (R-HSA-2559583) PRC2 methylates histones and DNA (R-HSA-212300)	1.130e-03 1.780e-03 3.540e-02
5	32	Anchoring of the basal body to the plasma membrane (R-HSA-5620912) Loss of Nlp from mitotic centrosomes (R-HSA-380259) AURKA Activation by TPX2 (R-HSA-8854518)	3.020e-07 5.640e-07 5.890e-07
6	29	Integrin cell surface interactions (R-HSA-216083) Laminin interactions (R-HSA-3000157) Extracellular matrix organization (R-HSA-1474244)	2.840e-04 1.280e-02 4.710e-02
7	34	Clathrin-mediated endocytosis (R-HSA-8856828) Signaling by EGFR (R-HSA-177929) Signaling by NOTCH1 t(7;9) (NOTCH1:M1580_K2555) Translocation Mutant (R-HSA-2660825)	8.140e-03 3.180e-02 3.350e-02
8	25		
9	36		
10	30	Chemokine receptors bind chemokines (R-HSA-380108) Peptide ligand-binding receptors (R-HSA-375276) Class A/1 (Rhodopsin-like receptors) (R-HSA-373076)	1.030e-45 2.020e-33 1.210e-28
11	40		
12	21		
13	45		
14	34	Translation (R-HSA-72766)	1.990e-02
15	32		
16	40		
17	25		
18	30	Creation of C4 and C2 activators (R-HSA-166786) Role of LAT2/NTAL/LAB on calcium mobilization (R-HSA-2730905) FCGR activation (R-HSA-2029481)	3.400e-07 3.470e-07 3.990e-07
19	23	MET activates PI3K/AKT signaling (R-HSA-8851907) MET promotes cell motility (R-HSA-8875878) MET activates PTPN11 (R-HSA-8865999)	8.490e-03 1.140e-02 1.530e-02
20	35		
21	28	Hemostasis (R-HSA-109582) Platelet activation, signaling and aggregation (R-HSA-76002) Estrogen-stimulated signaling through PRKCZ (R-HSA-9634635)	1.570e-02 2.050e-02 2.290e-02
22	19		
23	20		
24	32		
25	13		
26	18	Gastrulation (R-HSA-9758941) Developmental Biology (R-HSA-1266738) Regulation of beta-cell development (R-HSA-186712)	6.250e-05 3.010e-03 4.320e-03
27	22	Signal amplification (R-HSA-392518) Thromboxane signalling through TP receptor (R-HSA-428930) Thrombin signalling through proteinase activated receptors (PARs) (R-HSA-456926)	1.650e-23 2.680e-19 2.750e-18
28	28	Chromatin organization (R-HSA-4839726) HATs acetylate histones (R-HSA-3214847) Heme signaling (R-HSA-9707616)	2.790e-04 3.430e-04 3.450e-04

29	15	Lectin pathway of complement activation (R-HSA-166662) Ficolins bind to repetitive carbohydrate structures on the target cell surface (R-HSA-2855086) Complement cascade (R-HSA-166658)	1.640e-08 1.610e-06 1.520e-05
30	17	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell (R-HSA-198933)	4.520e-02
31	9	Interleukin-2 signaling (R-HSA-9020558)	2.240e-02
32	12	HATs acetylate histones (R-HSA-3214847) Chromatin organization (R-HSA-4839726) Chromatin modifying enzymes (R-HSA-3247509)	5.510e-13 1.160e-10 1.730e-10
33	21		
34	19		
35	23	TNFs bind their physiological receptors (R-HSA-5669034) TRAF6 mediated IRF7 activation in TLR7/8 or 9 signaling (R-HSA-975110) Cytokine Signaling in Immune system (R-HSA-1280215)	6.130e-05 4.450e-04 5.330e-04
36	22	Neurotoxicity of clostridium toxins (R-HSA-168799) Toxicity of botulinum toxin type C (botC) (R-HSA-5250971) Uptake and actions of bacterial toxins (R-HSA-5339562)	5.260e-07 1.360e-06 1.050e-05
37	20		
38	26	Highly calcium permeable nicotinic acetylcholine receptors (R-HSA-629597) Highly calcium permeable postsynaptic nicotinic acetylcholine receptors (R-HSA-629594) Presynaptic nicotinic acetylcholine receptors (R-HSA-622323)	6.450e-07 8.430e-07 8.430e-07
39	16	Interferon alpha/beta signaling (R-HSA-909733) Interferon Signaling (R-HSA-913531) Cytokine Signaling in Immune system (R-HSA-1280215)	4.140e-08 7.630e-06 3.280e-04
40	39		
41	13	Metabolism of amine-derived hormones (R-HSA-209776) Serotonin and melatonin biosynthesis (R-HSA-209931)	4.110e-04 4.730e-03
42	14		
43	9		
44	9	Sensing of DNA Double Strand Breaks (R-HSA-5693548) HDR through MMEJ (alt-NHEJ) (R-HSA-5685939) Defective HDR through Homologous Recombination Repair (HRR) due to PALB2 loss of BRCA1 binding function (R-HSA-9704331)	6.560e-03 1.440e-02 1.720e-02
45	15		
46	15	Signaling by CTNNB1 phospho-site mutants (R-HSA-4839743) CTNNB1 S33 mutants aren't phosphorylated (R-HSA-5358747) Platelet sensitization by LDL (R-HSA-432142)	1.110e-02 1.210e-02 1.230e-02
47	31	rRNA processing in the nucleus and cytosol (R-HSA-8868773) rRNA processing (R-HSA-72312) Major pathway of rRNA processing in the nucleolus and cytosol (R-HSA-6791226)	1.990e-10 2.210e-10 2.330e-10
48	27		
49	6		
50	58	Keratinization (R-HSA-6805567)	1.990e-03
51	19	tRNA modification in the nucleus and cytosol (R-HSA-6782315)	1.930e-02
52	15		
53	38		
54	11	Fertilization (R-HSA-1187000) Interaction With Cumulus Cells And The Zona Pellucida (R-HSA-2534343) Reproduction (R-HSA-1474165)	2.530e-09 1.870e-08 2.700e-06
55	13	Collagen biosynthesis and modifying enzymes (R-HSA-1650814) Collagen formation (R-HSA-1474290)	2.370e-02 2.790e-02
56	14	Creation of C4 and C2 activators (R-HSA-166786) Role of LAT2/NTAL/LAB on calcium mobilization (R-HSA-2730905) FCGR activation (R-HSA-2029481)	2.820e-09 2.880e-09 3.310e-09
57	12	Innate Immune System (R-HSA-168249) Toll Like Receptor 4 (TLR4) Cascade (R-HSA-166016) Toll-like Receptor Cascades (R-HSA-168898)	2.500e-05 2.950e-05 3.030e-05
58	8		

59	9		
60	10	Creation of C4 and C2 activators (R-HSA-166786) Role of LAT2/NTAL/LAB on calcium mobilization (R-HSA-2730905) FCGR activation (R-HSA-2029481)	6.830e-08 7.160e-08 8.360e-08
61	19	WNT 5A-dependent internalization of FZD4 (R-HSA-5099900) WNT 5A-dependent internalization of FZD2, FZD5 and ROR2 (R-HSA-5140745) trans-Golgi Network Vesicle Budding (R-HSA-199992)	9.040e-07 9.480e-07 4.360e-06
62	11	GABA receptor activation (R-HSA-977443) Signaling by ERBB4 (R-HSA-1236394) Neurotransmitter receptors and postsynaptic signal transmission (R-HSA-112314)	2.730e-05 3.540e-05 2.180e-03
63	13		
64	25		
65	14	Signaling by Hippo (R-HSA-2028269)	6.200e-04
66	31		
67	15		
68	11	MHC class II antigen presentation (R-HSA-2132295) HSP90 chaperone cycle for steroid hormone receptors (SHR) in the presence of ligand (R-HSA-3371497) COPI-independent Golgi-to-ER retrograde traffic (R-HSA-6811436)	9.920e-09 1.460e-08 1.500e-08
69	12		
70	17		
71	14		
72	16	rRNA processing (R-HSA-72312) Eukaryotic Translation Initiation (R-HSA-72613) Selenoamino acid metabolism (R-HSA-2408522)	1.310e-02 1.390e-02 1.440e-02
73	7	mRNA decay by 5′ to 3′ exoribonuclease (R-HSA-430039) mRNA Splicing (R-HSA-72172) mRNA Splicing - Major Pathway (R-HSA-72163)	2.670e-02 3.140e-02 4.190e-02
74	9		
75	16		
76	16		
77	8		
78	14		
79	9		
80	18		
81	19		
82	8		
83	7		
84	15	Nervous system development (R-HSA-9675108) Axon guidance (R-HSA-422475) RET signaling (R-HSA-8853659)	9.840e-04 1.140e-03 1.160e-03
85	10		
86	14	Downregulation of ERBB4 signaling (R-HSA-1253288) Signaling by ERBB4 (R-HSA-1236394) Defective POMT2 causes MDDGA2, MDDGB2 and MDDGC2 (R-HSA-5083629)	3.600e-05 5.340e-05 8.290e-04
87	34	SARS-CoV-2 modulates autophagy (R-HSA-9754560)	1.580e-03
88	5	CREB3 factors activate genes (R-HSA-8874211)	4.370e-03
89	11	Retinoid cycle disease events (R-HSA-2453864) Diseases of the neuronal system (R-HSA-9675143) The canonical retinoid cycle in rods (twilight vision) (R-HSA-2453902)	1.730e-02 2.600e-02 4.200e-02
90	11	RAB geranylgeranylation (R-HSA-8873719) RAB GEFs exchange GTP for GDP on RABs (R-HSA-8876198) Rab regulation of trafficking (R-HSA-9007101)	3.140e-07 7.850e-07 2.490e-06
91	9	STING mediated induction of host immune responses (R-HSA-1834941) IRF3-mediated induction of type I IFN (R-HSA-3270619) Regulation of lipid metabolism by PPARalpha (R-HSA-400206)	4.150e-05 4.240e-05 7.330e-05

92	6	STING mediated induction of host immune responses (R-HSA-1834941) IRF3-mediated induction of type I IFN (R-HSA-3270619) Regulation of lipid metabolism by PPARalpha (R-HSA-400206)	4.150e-05 4.240e-05 7.330e-05
93	34	Unfolded Protein Response (UPR) (R-HSA-381119) Metabolism of proteins (R-HSA-392499)	4.020e-02 4.140e-02
94	11		
95	10	IFNG signaling activates MAPKs (R-HSA-9732724) Regulation of IFNG signaling (R-HSA-877312) Potential therapeutics for SARS (R-HSA-9679191)	1.530e-02 1.800e-02 4.310e-02
96	46	Keratinization (R-HSA-6805567)	1.020e-05
97	8		
98	10	Post-translational modification: synthesis of GPI-anchored proteins (R-HSA-163125)	2.610e-02
99	5		
100	14		
101	14		
102	20	Condensation of Prometaphase Chromosomes (R-HSA-2514853) Cell Cycle (R-HSA-1640170)	3.310e-04 3.220e-02
103	6	Factors involved in megakaryocyte development and platelet production (R-HSA-983231)	2.250e-02
104	8		
105	17		
106	19	Neurotransmitter receptors and postsynaptic signal transmission (R-HSA-112314) RNA Polymerase II Transcription (R-HSA-73857) CREB phosphorylation (R-HSA-199920)	3.620e-02 4.310e-02 4.350e-02
107	36		
108	20		
109	23	Transferrin endocytosis and recycling (R-HSA-917977) ROS and RNS production in phagocytes (R-HSA-1222556) Insulin receptor recycling (R-HSA-77387)	2.340e-07 3.450e-07 3.930e-07
110	32	TGF-beta receptor signaling activates SMADs (R-HSA-2173789) Molecules associated with elastic fibres (R-HSA-2129379) Lysosomal oligosaccharide catabolism (R-HSA-8853383)	2.980e-02 3.290e-02 3.610e-02
111	5		
112	11		
113	9	IFNG signaling activates MAPKs (R-HSA-9732724) Regulation of IFNG signaling (R-HSA-877312)	1.220e-02 1.440e-02
114	8		
115	6		
116	10		
117	10	Defective ABCC9 causes CMD10, ATFB12 and Cantu syndrome (R-HSA-5678420) ATP sensitive Potassium channels (R-HSA-1296025)	5.470e-04 1.640e-03
118	13	Telomere C-strand synthesis initiation (R-HSA-174430) Removal of the Flap Intermediate from the C-strand (R-HSA-174437) Processive synthesis on the C-strand of the telomere (R-HSA-174414)	1.750e-07 2.910e-07 3.160e-07
119	5	VEGFR2 mediated cell proliferation (R-HSA-5218921) RHO GTPases Activate NADPH Oxidases (R-HSA-5668599)	1.460e-02 1.670e-02
120	5	Elevation of cytosolic Ca2+ levels (R-HSA-139853) Platelet calcium homeostasis (R-HSA-418360) Platelet homeostasis (R-HSA-418346)	9.900e-06 2.890e-05 6.000e-04
121	16		
122	10		
123	12	PRC2 methylates histones and DNA (R-HSA-212300) ERCC6 (CSB) and EHMT2 (G9a) positively regulate rRNA expression (R-HSA-427389) Defective pyroptosis (R-HSA-9710421)	4.560e-04 5.090e-04 5.130e-04
124	9		
125	7		
126	8		
127	30		

128	15	Telomere C-strand synthesis initiation (R-HSA-174430) Processive synthesis on the C-strand of the telomere (R-HSA-174414) Removal of the Flap Intermediate from the C-strand (R-HSA-174437)	1.840e-04 2.070e-04 2.180e-04
129	11		
130	21	Cell Cycle, Mitotic (R-HSA-69278) Mitotic Prometaphase (R-HSA-68877) Cell Cycle (R-HSA-1640170)	2.440e-02 3.820e-02 4.390e-02
131	30	Ovarian tumor domain proteases (R-HSA-5689896) Regulation of TNFR1 signaling (R-HSA-5357905) TNF signaling (R-HSA-75893)	3.170e-02 3.490e-02 4.480e-02
132	6	Defective AMN causes MGA1 (R-HSA-3359462) Defective CUBN causes MGA1 (R-HSA-3359463) Uptake of dietary cobalamins into enterocytes (R-HSA-9758881)	2.730e-04 5.470e-04 2.730e-03
133	21	CREB1 phosphorylation through the activation of Adenylate Cyclase (R-HSA-442720) PKA activation (R-HSA-163615) PKA activation in glucagon signalling (R-HSA-164378)	8.210e-07 1.690e-06 1.970e-06
134	37		
135	12		
136	10		
137	9		
138	25	Signaling by Receptor Tyrosine Kinases (R-HSA-9006934) Diseases of signal transduction by growth factor receptors and second messengers (R-HSA-5663202) Insulin receptor signalling cascade (R-HSA-74751)	4.470e-10 9.410e-10 8.510e-09
139	8	Response of EIF2AK1 (HRI) to heme deficiency (R-HSA-9648895)	3.560e-02
140	19	SARS-CoV-1 Infection (R-HSA-9678108) Viral Infection Pathways (R-HSA-9824446) Infectious disease (R-HSA-5663205)	9.710e-03 9.810e-03 1.370e-02
141	9		
142	10		
143	13		
144	5	Interleukin-12 family signaling (R-HSA-447115) Interleukin-12 signaling (R-HSA-9020591) Gene and protein expression by JAK-STAT signaling after Interleukin-12 stimulation (R-HSA-8950505)	3.730e-02 3.760e-02 4.840e-02
145	9		
146	7		

Table A.18: Listed are the significant reactome pathways enriched in clusters predicted by ClusterONE from the interactions only found in N samples and not in M samples using pipeline V.

Cluster	Size	Reactome pathways	Adjusted P-value
1	40	Activation of the mRNA upon binding of the cap-binding complex and eIFs, and subsequent binding to 43S (R-HSA-72662) Defective CFTR causes cystic fibrosis (R-HSA-5678895) Translation initiation complex formation (R-HSA-72649)	4.820e-07 5.420e-07 5.580e-07
2	35		
3	44		
4	39		
5	35		
6	32	NGF-stimulated transcription (R-HSA-9031628) Nuclear Events (kinase and transcription factor activation) (R-HSA-198725) Signaling by NTRK1 (TRKA) (R-HSA-187037)	9.790e-04 3.030e-03 2.280e-02
7	23		
8	39	HATs acetylate histones (R-HSA-3214847) Chromatin organization (R-HSA-4839726) Chromatin modifying enzymes (R-HSA-3247509)	1.020e-09 4.260e-07 6.390e-07
9	39		

10	29	N-glycan antennae elongation in the medial/trans-Golgi (R-HSA-975576) Glycosaminoglycan metabolism (R-HSA-1630316) Signaling by RNF43 mutants (R-HSA-5340588)	1.200e-04 3.300e-02 3.440e-02
11	31		
12	32		
13	41		
14	18	Antigen processing: Ubiquitination & Proteasome degradation (R-HSA-983168) Class I MHC mediated antigen processing & presentation (R-HSA-983169) Immune System (R-HSA-168256)	4.380e-04 7.760e-04 1.830e-02
15	29	Smooth Muscle Contraction (R-HSA-445355) Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs) (R-HSA-381426)	4.120e-02 6.820e-02
16	31	CRMPs in Sema3A signaling (R-HSA-399956) Semaphorin interactions (R-HSA-373755)	1.940e-05 5.080e-05
17	32		
18	21	Factors involved in megakaryocyte development and platelet production (R-HSA-983231)	4.650e-02
19	26		
20	24	Transport of small molecules (R-HSA-382551) ABC-family proteins mediated transport (R-HSA-382556)	1.510e-04 7.200e-03
21	11		
22	21		
23	22		
24	33	rRNA processing (R-HSA-72312) rRNA processing in the nucleus and cytosol (R-HSA-8868773) Major pathway of rRNA processing in the nucleolus and cytosol (R-HSA-6791226)	2.410e-05 2.550e-05 3.530e-05
25	13		
26	28		
27	17		
28	13		
29	20		
30	12		
31	10	Interferon alpha/beta signaling (R-HSA-909733) Interferon Signaling (R-HSA-913531) Regulation of IFNA/IFNB signaling (R-HSA-912694)	3.830e-07 2.630e-05 1.250e-04
32	26		
33	19		
34	12	Maturation of spike protein (R-HSA-9694548) Translation of Structural Proteins (R-HSA-9694635) Late SARS-CoV-2 Infection Events (R-HSA-9772573)	2.990e-03 5.600e-03 6.950e-03
35	14		
36	8		
37	14		
38	9		
39	11		
40	21	Signaling by KIT in disease (R-HSA-9669938) PI-3K cascade:FGFR2 (R-HSA-5654695) PI-3K cascade:FGFR4 (R-HSA-5654720)	1.190e-02 1.210e-02 1.230e-02
41	9		
42	9		
43	16	Signaling by Rho GTPases, Miro GTPases and RHOBTB3 (R-HSA-9716542) Signaling by Rho GTPases (R-HSA-194315) Striated Muscle Contraction (R-HSA-390522)	4.910e-04 8.450e-04 2.320e-03
44	14		
45	15	Folding of actin by CCT/TriC (R-HSA-390450) Prefoldin mediated transfer of substrate to CCT/TriC(R-HSA-389957) Formation of tubulin folding intermediates by CCT/TriC (R-HSA-389960)	6.310e-11 8.140e-09 8.180e-09

46	14		
47	14		
48	14		
49	10		
50	8		
51	22	Response of EIF2AK4 (GCN2) to amino acid deficiency (R-HSA-9633012) Signaling by ROBO receptors (R-HSA-376176) Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC) (R-HSA-975956)	1.900e-05 2.080e-05 2.100e-05
52	13		
53	9		
54	9	Cilium Assembly (R-HSA-5617833) Intraflagellar transport (R-HSA-5620924) Organelle biogenesis and maintenance (R-HSA-1852241)	1.290e-05 1.360e-05 5.720e-05
55	17		
56	8	Terminal pathway of complement (R-HSA-166665)	9.520e-03
57	8		
58	32		
59	31		
60	16		
61	5		
62	5	COPII-mediated vesicle transport (R-HSA-204005) RAB GEFs exchange GTP for GDP on RABs (R-HSA-8876198) Rab regulation of trafficking (R-HSA-9007101)	8.830e-04 9.990e-04 1.680e-03
63	14	Terminal pathway of complement (R-HSA-166665) Parasite infection (R-HSA-9664407) Fcgamma receptor (FCGR) dependent phagocytosis (R-HSA-2029480)	3.090e-02 4.320e-02 4.510e-02
64	7		
65	33		
66	10		
67	6		
68	6		
69	6		
70	17	MET activates RAS signaling (R-HSA-8851805)	4.610e-02
71	7	Processive synthesis on the C-strand of the telomere (R-HSA-174414) Base-Excision Repair, AP Site Formation (R-HSA-73929) Recognition and association of DNA glycosylase with site containing an affected purine (R-HSA-110330)	1.450e-02 1.490e-02 1.510e-02
72	7		
73	14		
74	12		
75	32		
76	17		
77	11	Long-term potentiation (R-HSA-9620244) Unblocking of NMDA receptors, glutamate binding and activation (R-HSA-438066) CREB1 phosphorylation through NMDA receptor-mediated activation of RAS signaling (R-HSA-442742)	3.070e-02 3.490e-02 3.880e-02
78	6	GPCR ligand binding (R-HSA-500792) Incretin synthesis, secretion, and inactivation (R-HSA-400508) GPCR downstream signalling (R-HSA-388396)	9.490e-03 1.250e-02 1.570e-02
79	6		
80	22		
81	5		
82	8	Defective AVP does not bind AVPR1A, B and causes neurohypophyseal diabetes insipidus (NDI) (R-HSA-5619099) Vasopressin-like receptors (R-HSA-388479)	1.020e-03 2.550e-03
83	11	WNT 5:FZD7-mediated leishmania damping (R-HSA-9673324) Killing mechanisms (R-HSA-9664420) RHO GTPases Activate NADPH Oxidases (R-HSA-5668599)	3.200e-05 6.410e-05 1.960e-04
84	27		

85	9	Postsynaptic nicotinic acetylcholine receptors (R-HSA-622327) Presynaptic nicotinic acetylcholine receptors (R-HSA-622323) Highly calcium permeable nicotinic acetylcholine receptors (R-HSA-629597)	6.620e-03 7.200e-03 7.860e-03
86	6		
87	8		
88	11		
89	5	Signaling by FGFR2 in disease (R-HSA-5655253) FGFR2 mutant receptor activation (R-HSA-1839126)	4.260e-02 5.270e-02
90	5		
91	5	WNT ligand biogenesis and trafficking (R-HSA-3238698) Class B/2 (Secretin family receptors) (R-HSA-373080) Signaling by WNT (R-HSA-195721)	4.600e-05 1.140e-03 2.470e-02
92	5		
93	5	Complement cascade (R-HSA-166658) Activation of C3 and C5 (R-HSA-174577) Initial triggering of complement (R-HSA-166663)	1.590e-08 3.100e-07 1.160e-06
94	5		
95	5		
96	12		
97	6	Sensory perception of salty taste (R-HSA-9730628) Sensory perception of taste (R-HSA-9717189) Stimuli-sensing channels (R-HSA-2672351)	7.080e-07 2.860e-04 2.250e-03
98	6		
99	7	Cooperation of Prefoldin and TriC/CCT in actin and tubulin folding (R-HSA-389958) Protein folding (R-HSA-391251) Chaperonin-mediated protein folding (R-HSA-390466)	3.370e-04 3.000e-03 3.720e-03
100	14		
101	8	Uptake and function of diphtheria toxin (R-HSA-5336415) HSP90 chaperone cycle for steroid hormone receptors (SHR) in the presence of ligand (R-HSA-3371497) HSF1 activation (R-HSA-3371511)	2.550e-03 2.870e-03 7.470e-03
102	8	Metal sequestration by antimicrobial proteins (R-HSA-6799990)	5.100e-03
103	9		
104	9		
105	11	Downstream signal transduction (R-HSA-186763) Signaling by PDGF (R-HSA-186797) Signaling by NTRK3 (TRKC) (R-HSA-9034015)	1.690e-06 1.770e-05 3.530e-05
106	7	Phase 0 - rapid depolarisation (R-HSA-5576892) Cardiac conduction (R-HSA-5576891) Muscle contraction (R-HSA-397014)	4.310e-07 6.740e-05 2.750e-04
107	6	Butyrophilin (BTN) family interactions (R-HSA-8851680)	1.200e-02
108	6		
109	21	Folding of actin by CCT/TriC (R-HSA-390450) Prefoldin mediated transfer of substrate to CCT/TriC (R-HSA-389957) Formation of tubulin folding intermediates by CCT/TriC (R-HSA-389960)	2.790e-07 8.970e-06 9.840e-06
110	11		
111	12		
112	15		
113	8		
114	8		
115	9		
116	20		
117	8	Attenuation phase (R-HSA-3371568) HSF1-dependent transactivation (R-HSA-3371571) Drug resistance in ERBB2 TMD/JMD mutants (R-HSA-9665737)	3.600e-05 9.990e-05 1.570e-04
118	45		
119	16	Formation of the cornified envelope (R-HSA-6809371) Keratinization (R-HSA-6805567) Developmental Biology (R-HSA-1266738)	9.500e-05 6.350e-04 1.010e-02
120	10		

121	7		
122	14		
123	12		
124	10		
125	19	Activation of BAD and translocation to mitochondria (R-HSA-111447) Chk1/Chk2(Cds1) mediated inactivation of Cyclin B:Cdk1 complex (R-HSA-75035) SARS-CoV-1 targets host intracellular signalling and regulatory pathways (R-HSA-9735871)	1.940e-04 2.440e-04 2.580e-04
126	16		
127	7	Integrin cell surface interactions (R-HSA-216083) IRAK4 deficiency (TLR2/4) (R-HSA-5603041) MyD88 deficiency (TLR2/4) (R-HSA-5602498)	5.830e-03 7.790e-03 8.650e-03
128	7		
129	11		
130	8		
131	10	Long-term potentiation (R-HSA-9620244) Unblocking of NMDA receptors, glutamate binding and activation (R-HSA-438066) CREB1 phosphorylation through NMDA receptor-mediated activation of RAS signaling (R-HSA-442742)	8.140e-05 9.380e-05 1.240e-04
132	14	WNT 5:FZD7-mediated leishmania damping (R-HSA-9673324) Killing mechanisms (R-HSA-9664420) RHO GTPases Activate NADPH Oxidases (R-HSA-5668599)	5.550e-05 1.110e-04 3.390e-04
133	8		
134	33		
135	13		
136	18	rRNA processing (R-HSA-72312) rRNA processing in the nucleus and cytosol (R-HSA-8868773) Metabolism of RNA (R-HSA-8953854)	8.810e-04 1.370e-03 1.300e-02

Appendix B
Supplementary Material for chapter 5

The objective is to attempt to replicate dimer, trimer structures with both interfaces and doughnut-shaped decameric structures of TSA1 and TSA2 of *Saccharomyces cerevisiae* and PrxA of *Arabidopsis thaliana* using corresponding monomers and dimers as input with the HADDOCK method. Subsequently, the results of the predicted structures were aligned with the deposited experimental structures, and the root-mean-square deviation (RMSD) was estimated. The formula for estimating the RMSD value is as follows:

$$RMSD = \sqrt{\frac{\sum(x_e - x_o)}{n}}$$

where x_e represents the experimental structure position values, x_o represents the observed structure position values and n represnts the number of equivalent atoms.

The hypothesis of hetero-dimerisation and hetero-trimerisation was tested using monomers and dimers derived from TSA1 and TSA2. The resulting predicted structure was then compared with the experimental structures of TSA1 and TSA2, and root-mean-square deviation (RMSD) values were estimated to test for similarity. Additionally, the hypothesis of cross-species dimerisation was tested with monomers from *Saccharomyces cerevisiae* and *Arabidopsis thaliana* at A-type and B-type interfaces. The degree of similarity was evaluated by calculating the root-mean-square deviation (RMSD) values through the alignment with the experimental structures.

B.1 Results from HADDOCK

B.1.1 Cross species hetero-dimerisation at B-type interface

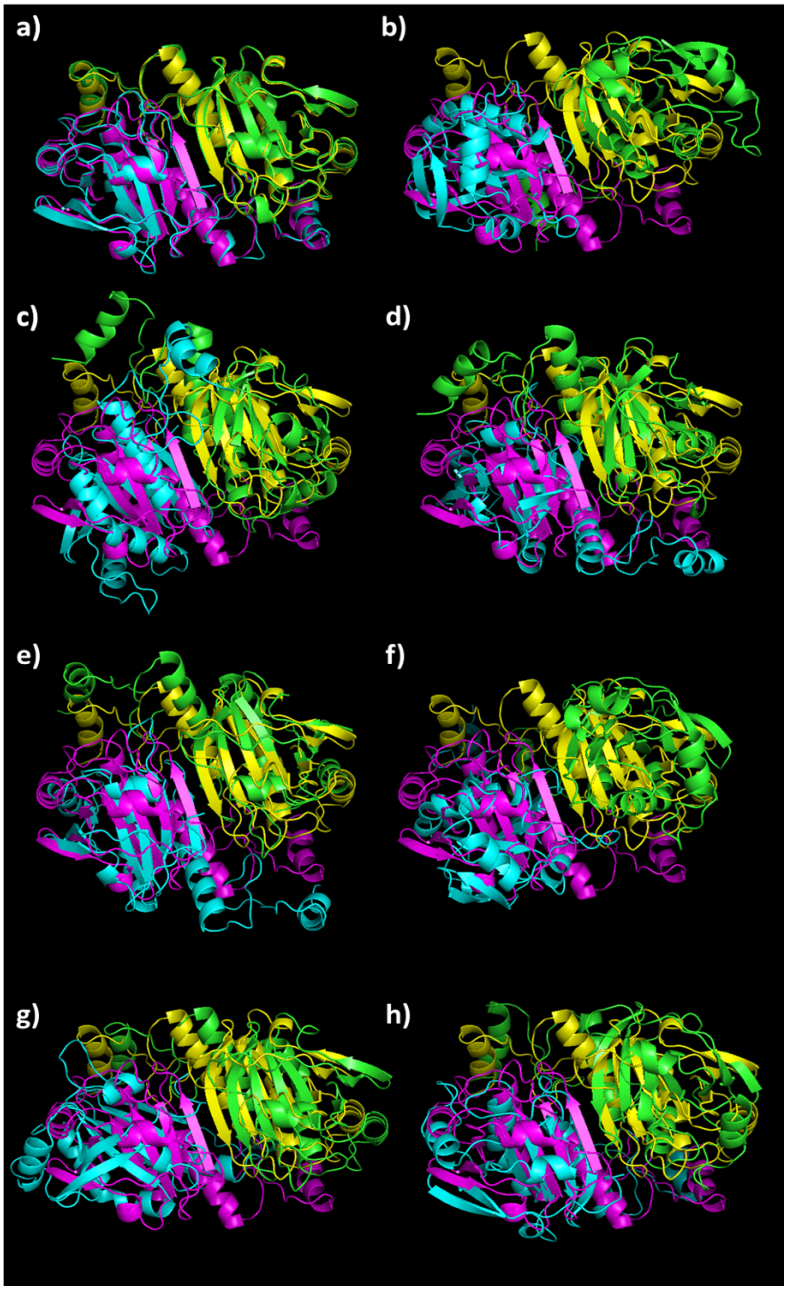


Figure B.1: The figure illustrates the eight predicted cross species hetero dimer structures of TSA1 and PrxA (coloured blue and green) and the experimental B-type homo dimer structures of TSA1 (coloured yellow and pink) of cluster 1 (a), cluster 5 (b). Cluster 2 (c), Cluster 7 (d), Cluster 4 (e), Cluster 6 (f), Cluster 3 (g), and Cluster 8 (h).

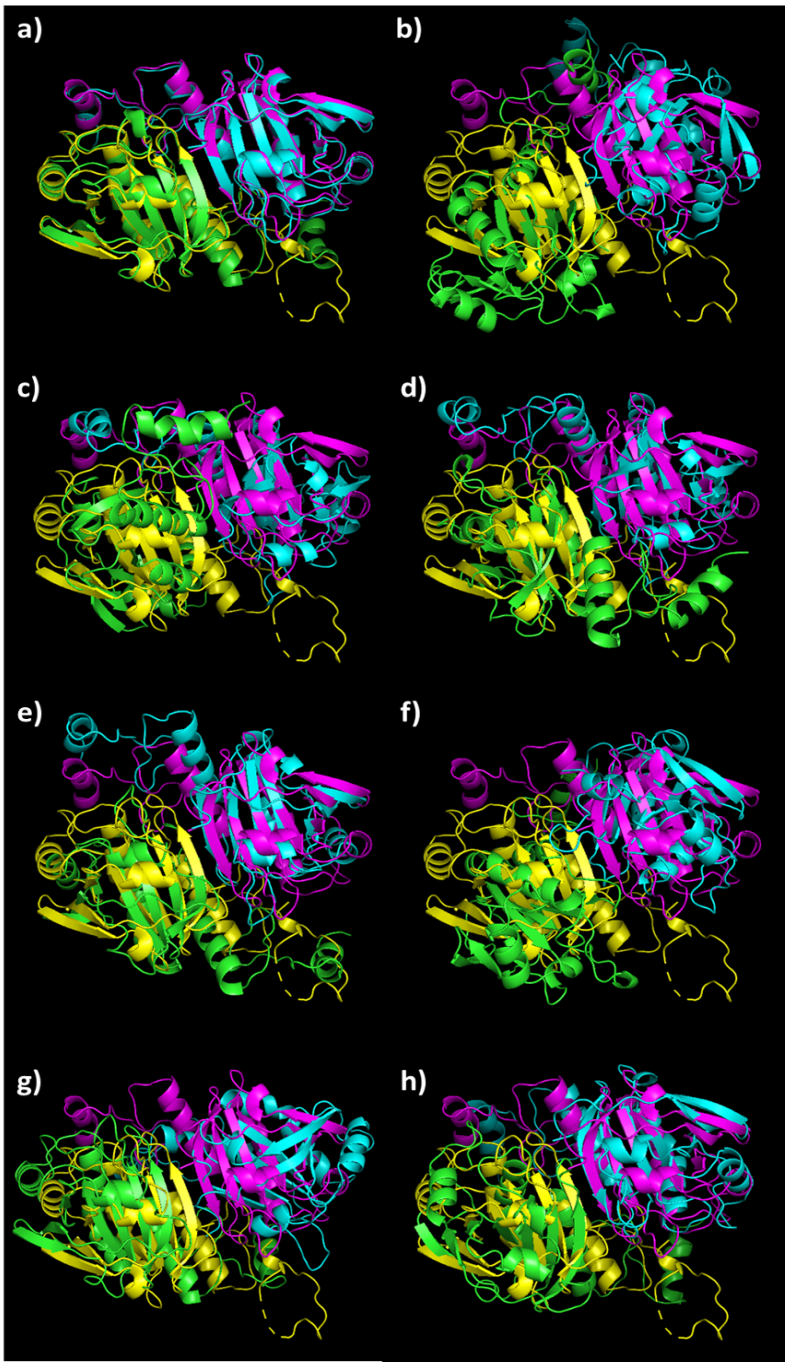


Figure B.2: The figure illustrates the eight predicted cross species hetero dimer structures of TSA1 and PrxA (coloured blue and green) and the experimental B-type homo dimer structures of PrxA (coloured yellow and pink) of cluster 1 (a), cluster 5 (b). Cluster 2 (c), Cluster 7 (d), Cluster 4 (e), Cluster 6 (f), Cluster 3 (g), and Cluster 8 (h).

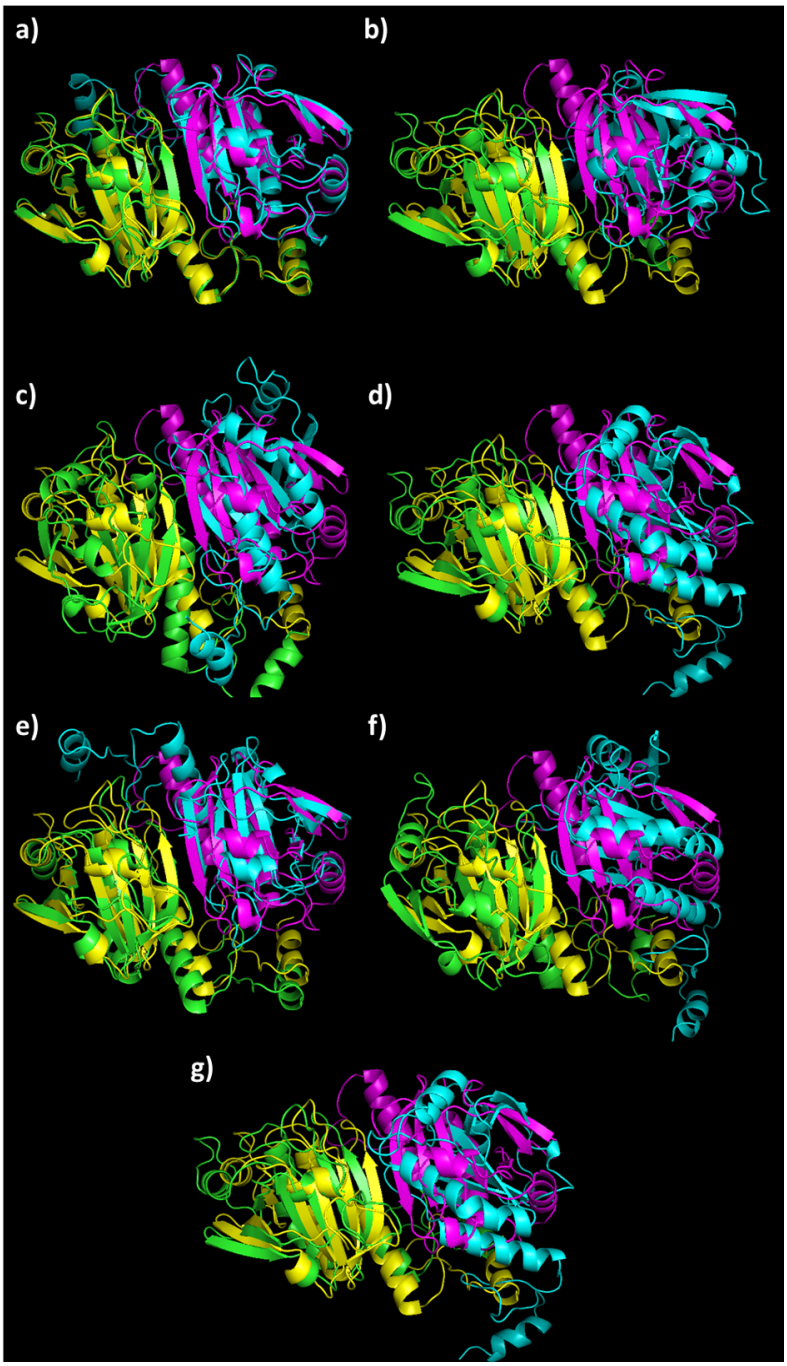


Figure B.3: The figure illustrates the seven predicted cross species hetero dimer structures of TSA2 and PrxA (coloured blue and green) and the experimental B-type homo dimer structures of TSA2 (coloured yellow and pink) of cluster 3 (a), cluster 4 (b). Cluster 2 (c), Cluster 7 (d), Cluster 5 (e), Cluster 1 (f), and Cluster 6 (g).

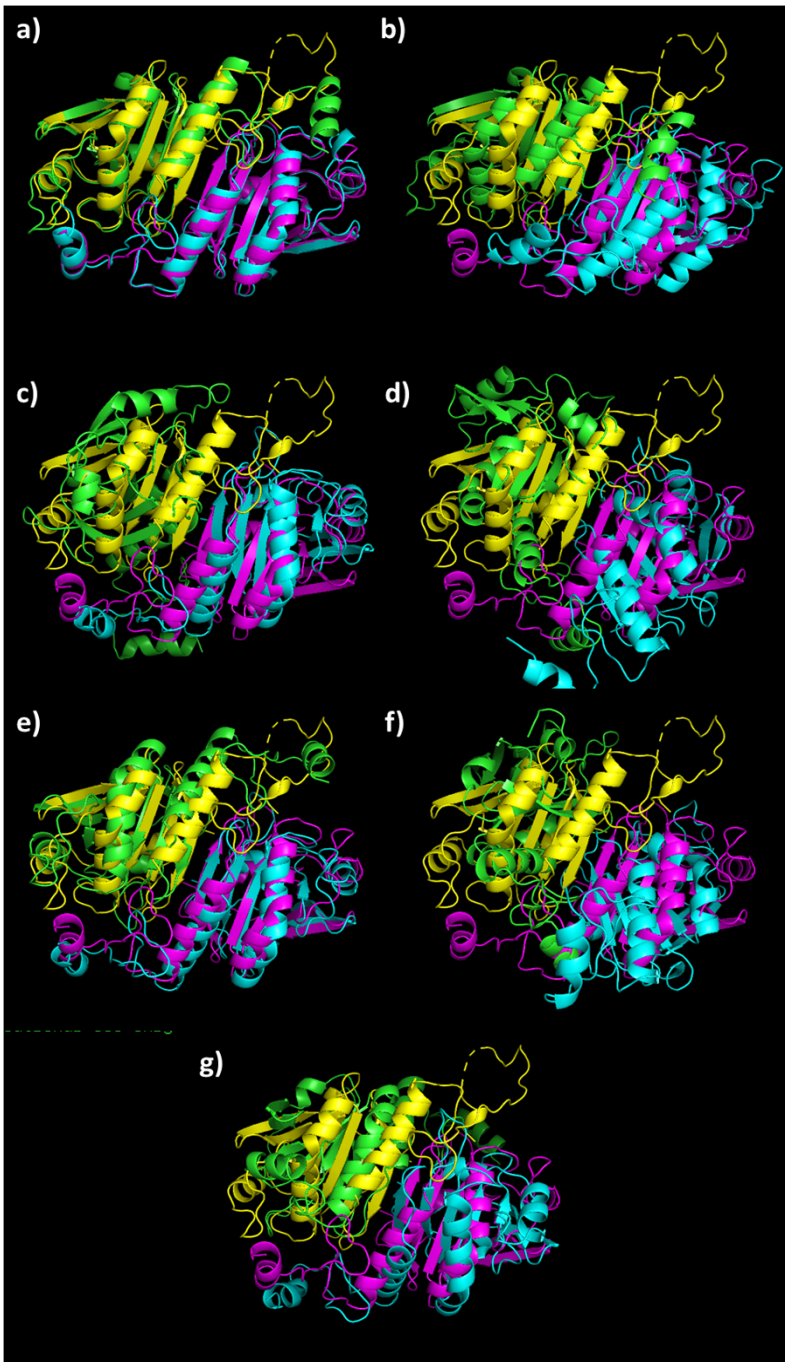


Figure B.4: The figure illustrates the eight predicted cross species hetero dimer structures of TSA2 and PrxA (coloured blue and green) and the experimental B-type homo dimer structures of PrxA (coloured yellow and pink) of cluster 3 (a), cluster 4 (b). Cluster 2 (c), Cluster 7 (d), Cluster 5 (e), Cluster 1 (f), and Cluster 6 (g).

B.1.2 Cross species hetero-dimerisation at A-type interface

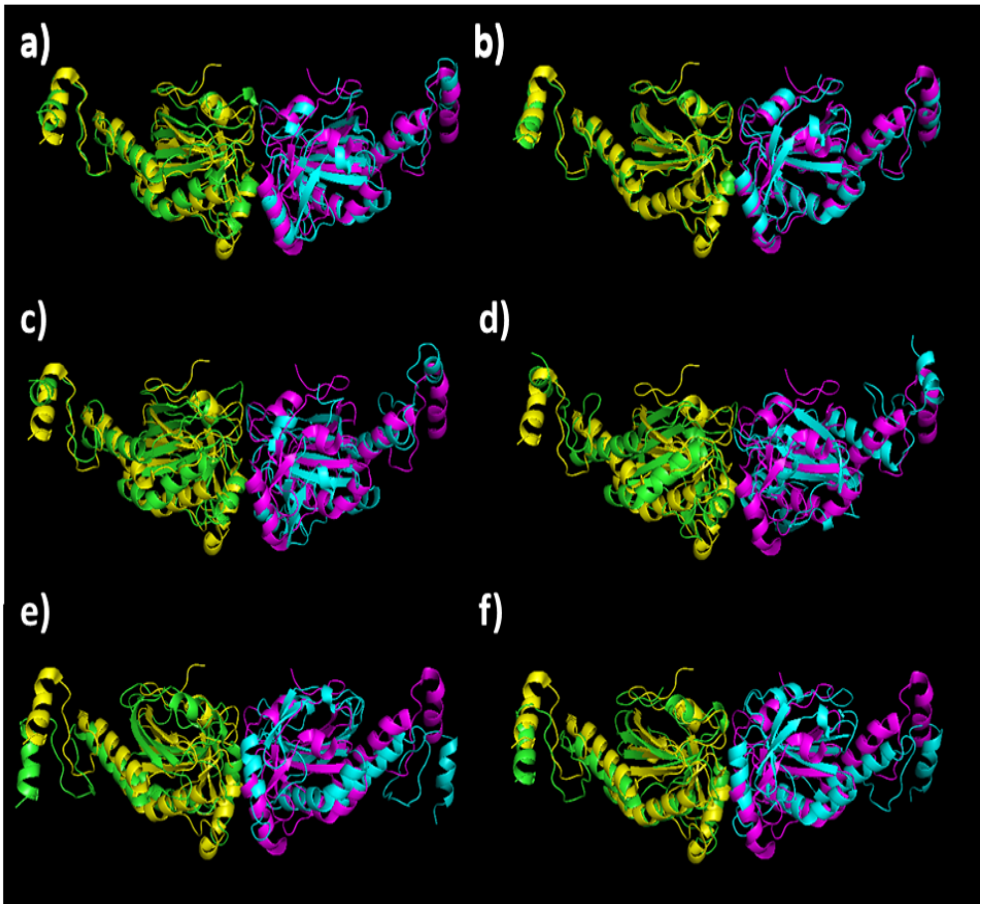


Figure B.5: The figure illustrates the six predicted cross species hetero dimer structures of TSA1 and PrxA (coloured blue and green) and the experimental A-type homo dimer structures of TSA1 (coloured yellow and pink) of cluster 1 (a), cluster 2 (b). Cluster 6 (c), Cluster 3 (d), Cluster 4 (e), and Cluster 5 (f).

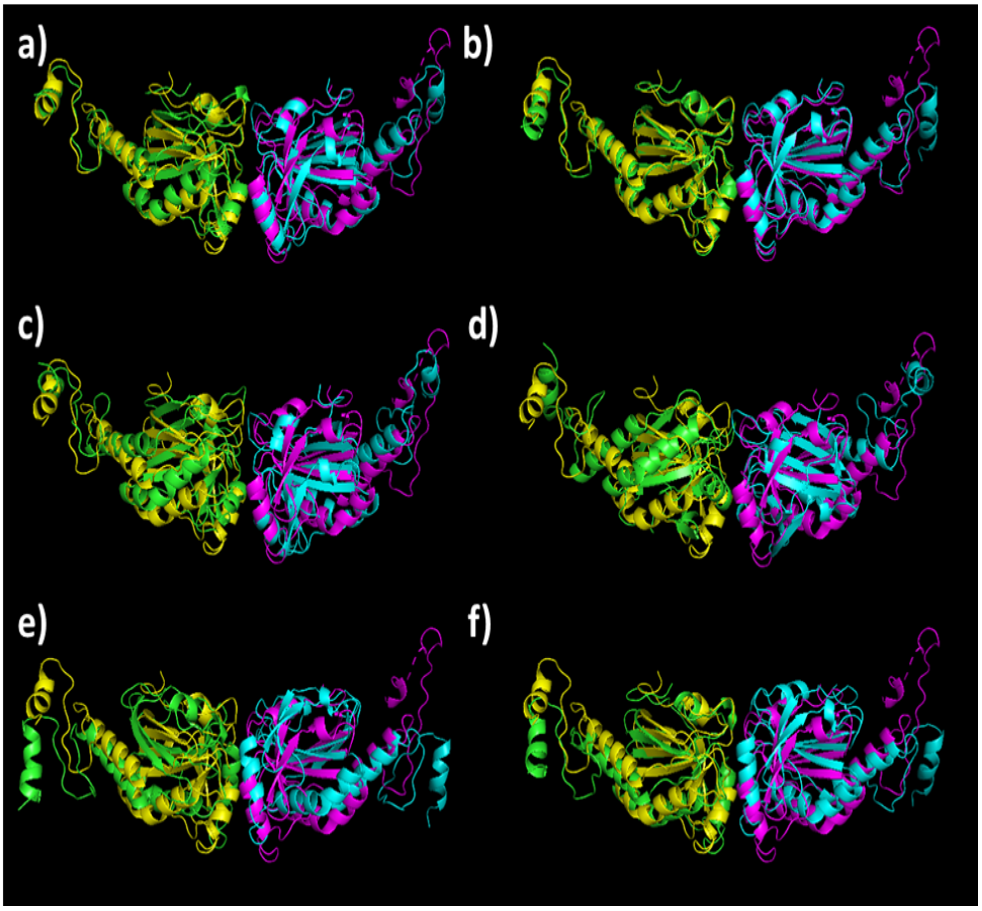


Figure B.6: The figure illustrates the six predicted cross species hetero dimer structures of TSA1 and PrxA (coloured blue and green) and the experimental A-type homo dimer structures of PrxA (coloured yellow and pink) of cluster 1 (a), cluster 2 (b). Cluster 6 (c), Cluster 3 (d), Cluster 4 (e), and Cluster 5 (f).

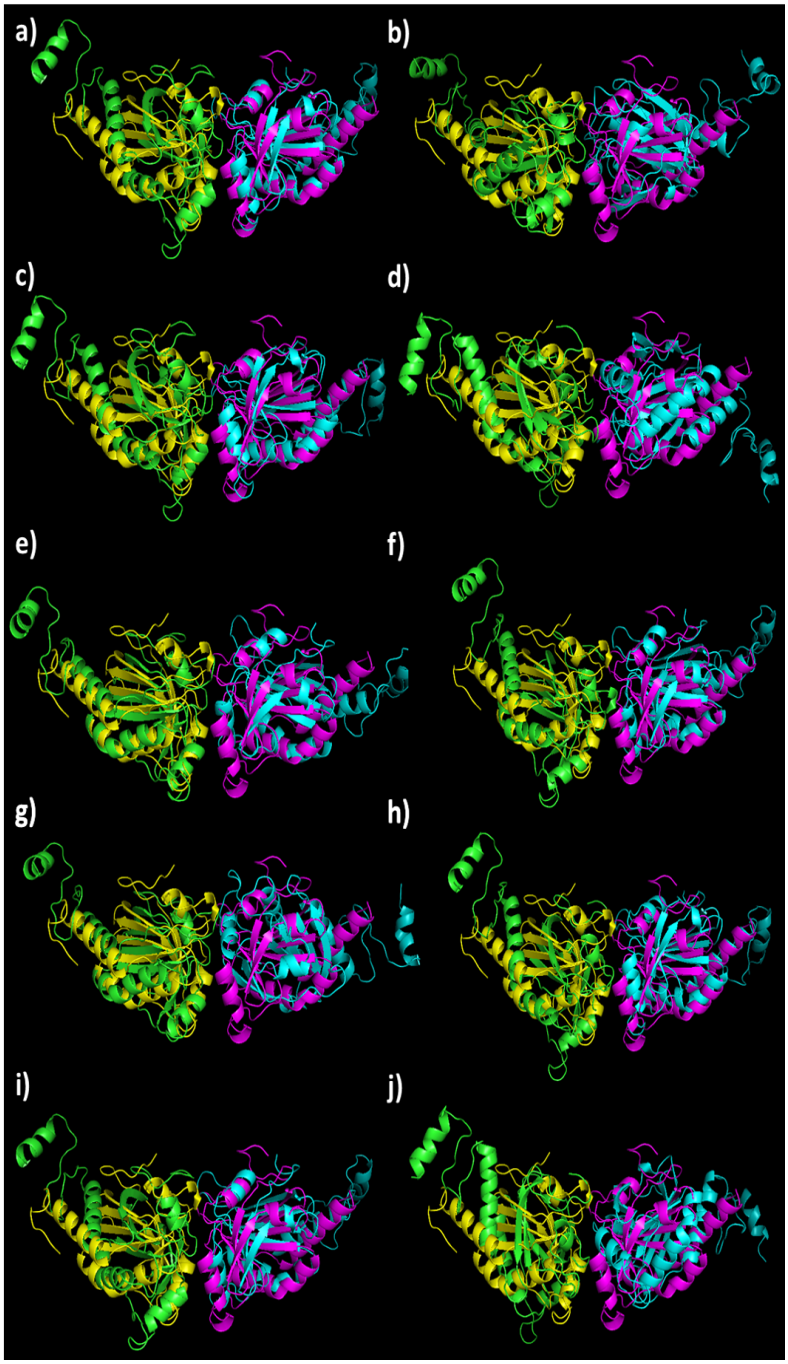


Figure B.7: The figure illustrates the ten predicted cross species hetero dimer structures of TSA2 and PrxA (coloured blue and green) and the experimental A-type homo dimer structures of TSA2 (coloured yellow and pink) of cluster 1 (a), cluster 2 (b). Cluster 6 (c), Cluster 5 (d), Cluster 4 (e), Cluster 9 (f), Cluster 3 (g), Cluster 8 (h), Cluster 7 (i) and Cluster 10 (j).

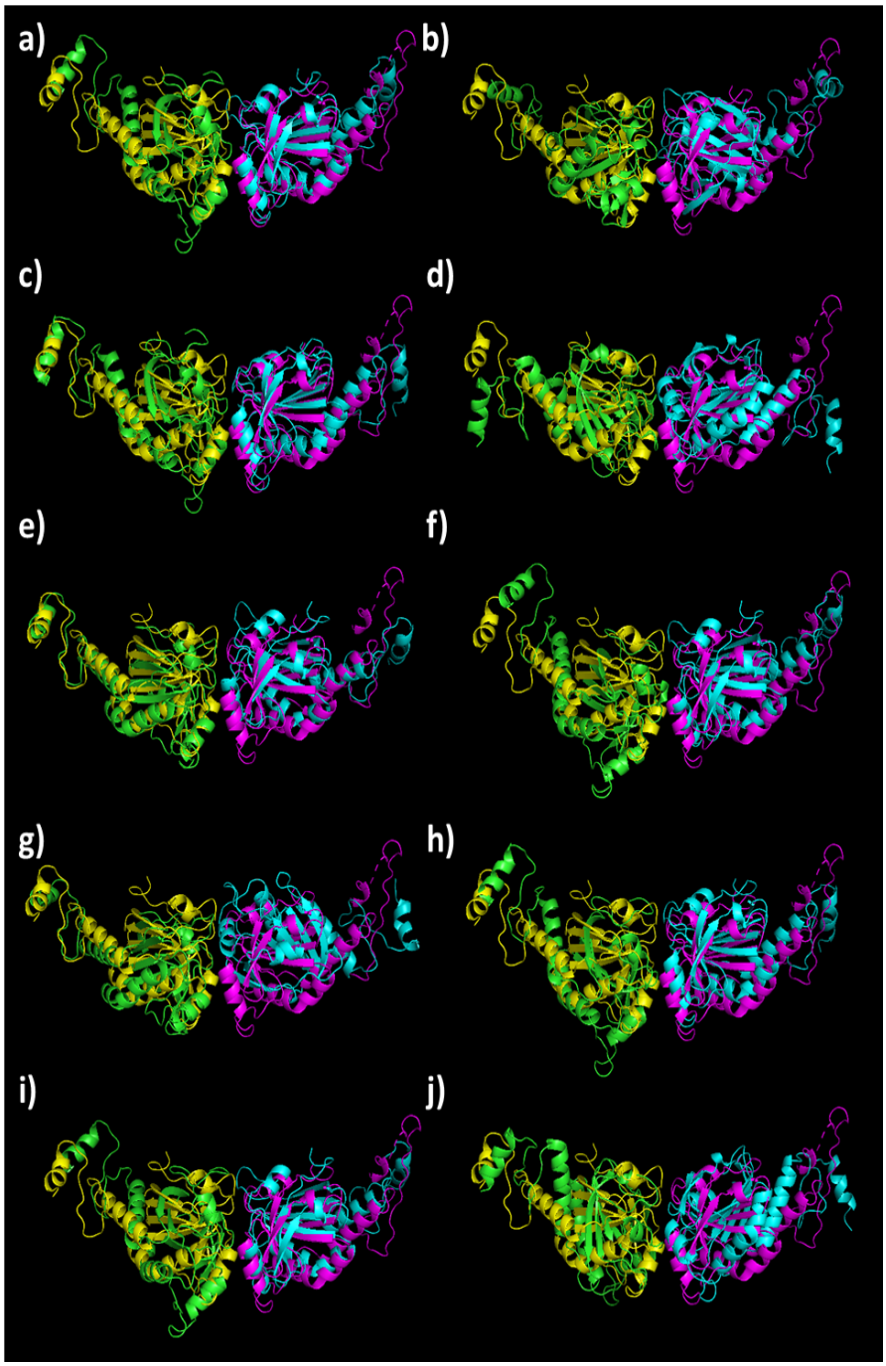


Figure B.8: The figure illustrates the eight predicted cross species hetero dimer structures of TSA2 and PrxA (coloured blue and green) and the experimental A-type homo dimer structures of PrxA (coloured yellow and pink) of cluster 1 (a), cluster 2 (b). Cluster 6 (c), Cluster 5 (d), Cluster 4 (e), Cluster 9 (f), Cluster 3 (g), Cluster 8 (h), Cluster 7 (i) and Cluster 10 (j).

B.2 Results from AlphaFold

B.2.1 Decamers of TSA1 from *Saccharomyces cerevisiae*

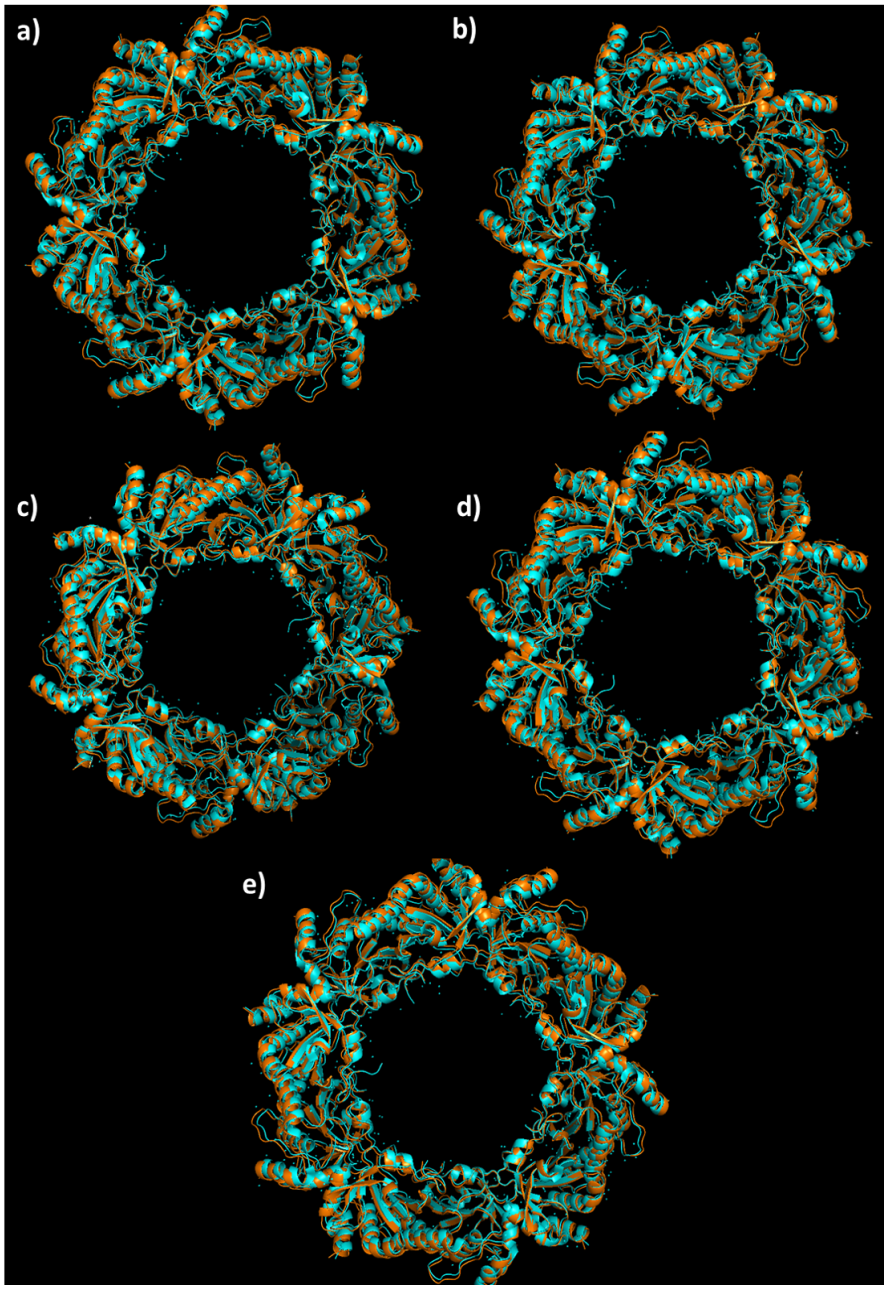


Figure B.9: The figure illustrates overlap of the five predicted decameric structures of TSA1 (coloured orange) and the experimental decameric structures of TSA1 (coloured blue) of Predicted structure 1 (a), structure 2 (b). structure 3 (c), structure 4 (d), and structure 5 (e).

B.2.2 Decamers of TSA2 from *Saccharomyces cerevisiae*

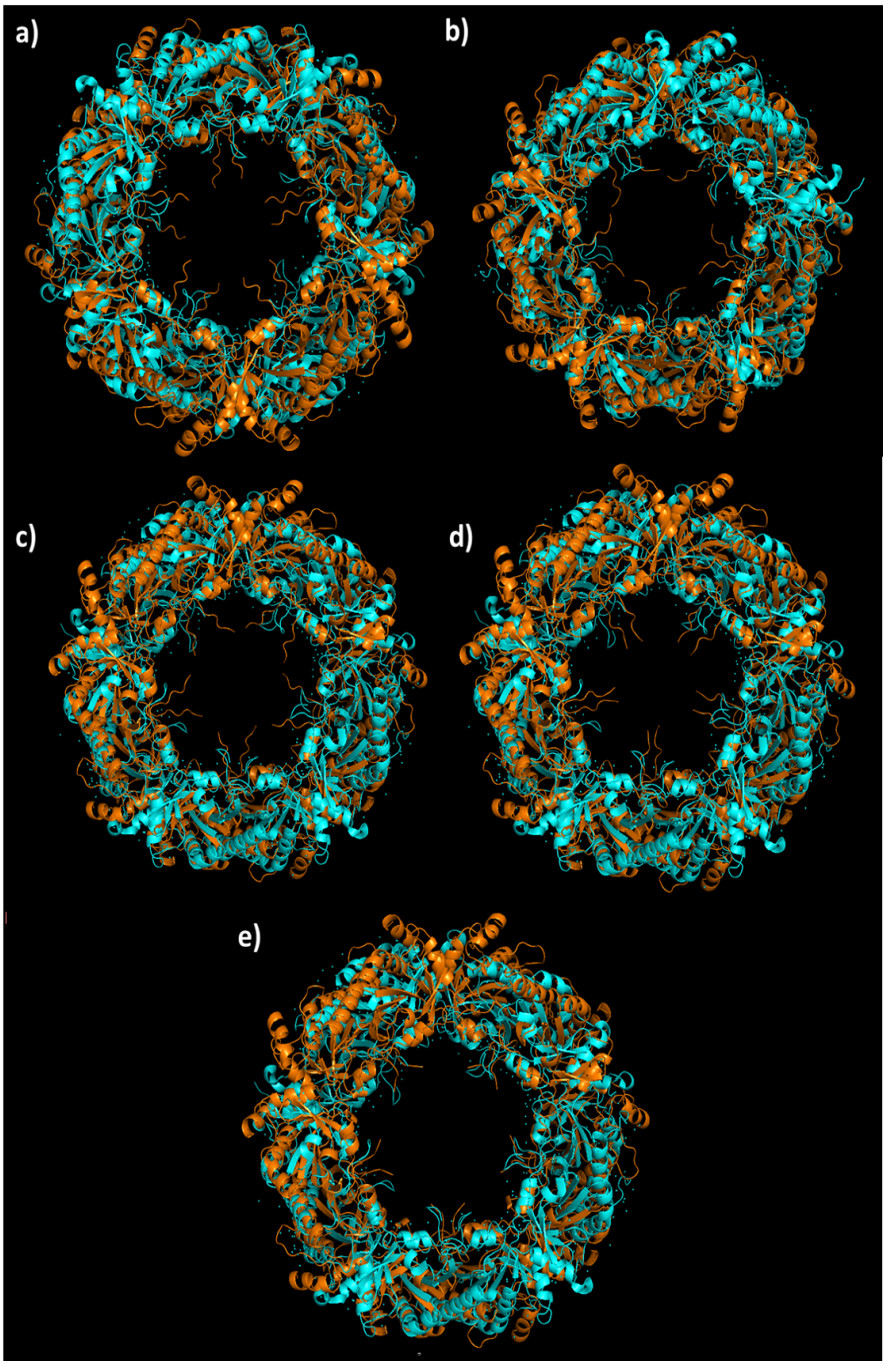


Figure B.10: The figure illustrates overlap of the five predicted decameric structures of TSA2 (coloured orange) and the experimental decameric structures of TSA2 (coloured blue) of Predicted structure 1 (a), structure 2 (b). structure 3 (c), structure 4 (d), and structure 5 (e).

B.2.3 Decamers of PrxA from *Arabidopsis thaliana*

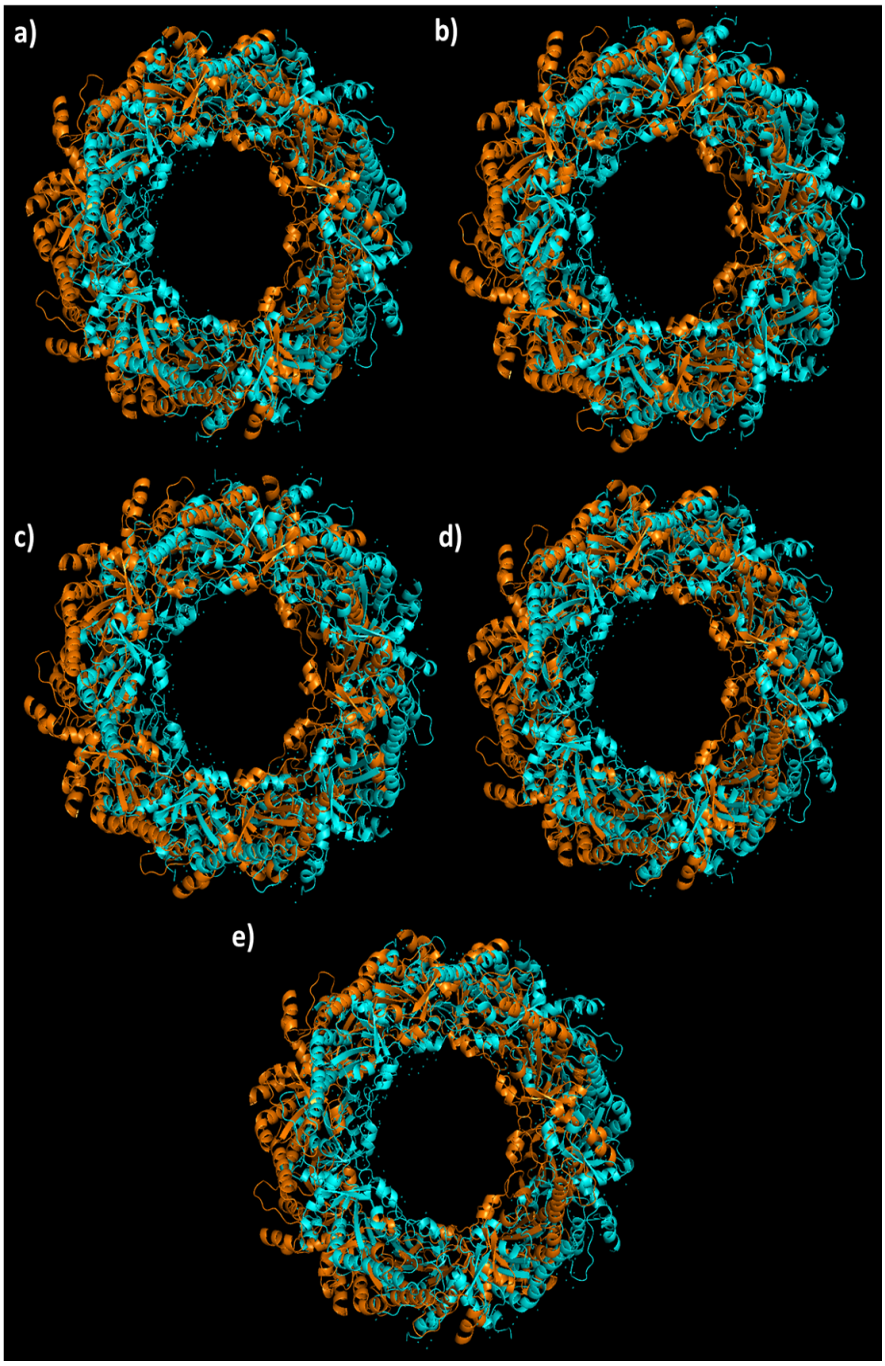


Figure B.11: The figure illustrates overlap of the five predicted decameric structures of PrxA (coloured orange) and the experimental decameric structures of PrxA (coloured blue) of Predicted structure 1 (a), structure 2 (b). structure 3 (c), structure 4 (d), and structure 5 (e).

Bibliography

[1] Neil J Smelser, Paul B Baltes, et al. *International encyclopedia of the social & behavioral sciences*, volume 11. Elsevier Amsterdam, 2001.

[2] ZA Bhutta, K Sadiq, and T Aga. Protein digestion and bioavailability. *Encyclopedia of Human Nutrition*, 4(4):116–122, 2013.

[3] Joaquín Goñi Cortés, Francisco J Esteban Ruiz, Nieves Vélez de Mendizábal, Jorge Sepulcre, Sergio Ardanza Trevijano, Ion Agirrezabal, and Pablo Villoslada. A computational analysis of protein-protein interaction networks in neurodegenerative diseases. 2008.

[4] Gavin CKW Koh, Pablo Porras, Bruno Aranda, Henning Hermjakob, and Sandra E Orchard. Analyzing protein–protein interaction networks. *Journal of proteome research*, 11(4):2014–2031, 2012.

[5] Sama Akbarzadeh, Özlem Coşkun, and Başak Günçer. Studying protein–protein interactions: Latest and most popular approaches. *Journal of Structural Biology*, page 108118, 2024.

[6] Tord Berggård, Sara Linse, and Peter James. Methods for the detection and analysis of protein–protein interactions. *Proteomics*, 7(16):2833–2842, 2007.

[7] Michael E Cusick, Haiyuan Yu, Alex Smolyar, Kavitha Venkatesan, Anne-Ruxandra Carvunis, Nicolas Simonis, Jean-François Rual, Heather Borick, Pascal Braun, Matija Dreze, et al. Literature-curated protein interaction datasets. *Nature methods*, 6(1):39–46, 2009.

[8] V Srinivasa Rao, K Srinivas, GN Sujini, and GN Sunand Kumar. Protein-protein interaction detection: methods and analysis. *International journal of proteomics*, 2014(1):147648, 2014.

[9] Christine Vogel and Edward M Marcotte. Calculating absolute and relative protein abundance from mass spectrometry-based protein expression data. *Nature protocols*, 3(9):1444–1451, 2008.

[10] Christine Vogel and Edward M Marcotte. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature reviews genetics*, 13(4):227–232, 2012.

[11] BC Searle, DL Tabb, JA Falkner, JA Kowalak, K Meyer-Arendt, PA Rudnick, SL Seymour, and WS Lane. iprg2009 study: testing for differences between complex samples in proteomics datasets. *Poster at ABRF*, 28(1):83–89, 2009.

[12] Mushan Li, Shiqi Tu, Zijia Li, Fengxiang Tan, Jian Liu, Qian Wang, Yuannyu Zhang, Jian Xu, Yijing Zhang, Feng Zhou, et al. Map: model-based analysis of proteomic data to detect proteins with significant abundance changes. *Cell Discovery*, 5(1):40, 2019.

[13] Kai Kammers, Robert N Cole, Calvin Tiengwe, and Ingo Ruczinski. Detecting significant changes in protein abundance. *EuPA open proteomics*, 7:11–19, 2015.

[14] Bruno Domon and Ruedi Aebersold. Challenges and opportunities in proteomics data analysis. *Molecular & Cellular Proteomics*, 5(10):1921–1926, 2006.

[15] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3, 2024.

[16] Cyril Dominguez, Rolf Boelens, and Alexandre MJJ Bonvin. Haddock: a protein- protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, 125(7):1731–1737, 2003.

[17] Benjamin Webb and Andrej Sali. Comparative protein structure modeling using modeller. *Current protocols in bioinformatics*, 54(1):5–6, 2016.

[18] Volkhard Helms and Olga V Kalinina. *Protein Interactions: The Molecular Basis of Interactomics*. John Wiley & Sons, 2022.

[19] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[20] Zhendong Chen, Yuxiong Gao, and Dafang Zhong. Technologies to improve the sensitivity of existing chromatographic methods used for bioanalytical studies. *Biomedical chromatography*, 34(3):e4798, 2020.

[21] Katherine L Sanders and James L Edwards. Nano-liquid chromatography-mass spectrometry and recent applications in omics investigations. *Analytical Methods*, 12(36):4404–4417, 2020.

[22] Ron Milo. What is the total number of protein molecules per cell volume? a call to rethink some published values. *Bioessays*, 35(12):1050–1055, 2013.

[23] Michal Gorka, Corné Swart, Beata Siemiatkowska, Silvia Martínez-Jaime, Aleksandra Skirycz, Sebastian Streb, and Alexander Graf. Protein complex identification and quantitative complexome by cn-page. *Scientific reports*, 9(1):11523, 2019.

[24] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. Biogrid: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl_1):D535–D539, 2006.

[25] Alberto Calderone, Luisa Castagnoli, and Gianni Cesareni. Mentha: a resource for browsing integrated protein-interaction networks. *Nature methods*, 10(8):690–691, 2013.

[26] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, et al. String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, 47(D1):D607–D613, 2019.

[27] Andrew Chatr-Aryamontri, Arnaud Ceol, Luisa Montecchi Palazzi, Giuliano Nardelli, Maria Victoria Schneider, Luisa Castagnoli, and Gianni Cesareni. Mint: the molecular interaction database. *Nucleic acids research*, 35(suppl_1):D572–D574, 2007.

[28] Sandra Orchard, Mais Ammari, Bruno Aranda, Lionel Breuza, Leonardo Briganti, Fiona Broackes-Carter, Nancy H Campbell, Gayatri Chavali, Carol Chen, Noemi Del-Toro, et al. The mintact project—intact as a common curation platform for 11 molecular interaction databases. *Nucleic acids research*, 42(D1):D358–D363, 2014.

[29] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2):101–113, 2004.

[30] Gary D Bader, Ian Donaldson, Cheryl Wolting, BF Francis Ouellette, Tony Pawson, and Christopher WV Hogue. Bind—the biomolecular interaction network database. *Nucleic acids research*, 29(1):242–245, 2001.

[31] Ioannis Xenarios, Danny W Rice, Lukasz Salwinski, Marisa K Baron, Edward M Marcotte, and David Eisenberg. Dip: the database of interacting proteins. *Nucleic acids research*, 28(1):289–291, 2000.

[32] Max Kotlyar, Chiara Pastrello, Nicholas Sheahan, and Igor Jurisica. Integrated interactions database: tissue-specific view of the human and model organism interactomes. *Nucleic acids research*, 44(D1):D536–D541, 2016.

[33] Diego Alonso-Lopez, Miguel A Gutiérrez, Katia P Lopes, Carlos Prieto, Rodrigo Santamaría, and Javier De Las Rivas. Apid interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks. *Nucleic acids research*, 44(W1):W529–W535, 2016.

[34] Suraj Peri, J Daniel Navarro, Troels Z Kristiansen, Ramars Amanchy, Vineeth Surendranath, Babylakshmi Muthusamy, TKB Gandhi, KN Chandrika, Nandan Deshpande, Shubha Suresh, et al. Human protein reference database as a discovery resource for proteomics. *Nucleic acids research*, 32(suppl_1):D497–D501, 2004.

[35] Pablo Porras, Elisabet Barrera, Alan Bridge, Noemi Del-Toro, Gianni Cesareni, Margaret Duesbury, Henning Hermjakob, Marta Iannuccelli, Igor Jurisica, Max Kotlyar, et al. Towards a unified open access dataset of molecular interactions. *Nature communications*, 11(1):6144, 2020.

[36] Anne-Claude Gavin, Markus Bösche, Roland Krause, Paola Grandi, Martina Marzioch, Andreas Bauer, Jörg Schultz, Jens M Rick, Anne-Marie Michon, Cristina-Maria Cruciat, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147, 2002.

[37] Peter Uetz, Loic Giot, Gerard Cagney, Traci A Mansfield, Richard S Judson, James R Knight, Daniel Lockshon, Vaibhav Narayan, Maithreyan Srinivasan, Pascale Pochart, et al. A comprehensive analysis of protein–protein interactions in saccharomyces cerevisiae. *Nature*, 403(6770):623–627, 2000.

[38] Stanley Fields and Ok-kyu Song. A novel genetic system to detect protein–protein interactions. *Nature*, 340(6230):245–246, 1989.

[39] Jing-Dong J Han, Denis Dupuy, Nicolas Bertin, Michael E Cusick, and Marc Vidal. Effect of sampling on topology predictions of protein-protein interaction networks. *Nature biotechnology*, 23(7):839–844, 2005.

[40] Haidong Wang, Boyko Kakaradov, Sean R Collins, Lena Karotki, Dorothea Fiedler, Michael Shales, Kevan M Shokat, Tobias C Walther, Nevan J Krogan, and Daphne Koller. A complex-based reconstruction of the saccharomyces cerevisiae interactome. *Molecular & Cellular Proteomics*, 8(6):1361–1381, 2009.

[41] Elizabeth A Winzeler, Daniel D Shoemaker, Anna Astromoff, Hong Liang, Keith Anderson, Bruno Andre, Rhonda Bangham, Rocio Benito, Jef D Boeke, Howard Bussey, et al. Functional characterization of the s. cerevisiae genome by gene deletion and parallel analysis. *science*, 285(5429):901–906, 1999.

[42] Guri Giaever, Angela M Chu, Li Ni, Carla Connelly, Linda Riles, Steeve Véronneau, Sally Dow, Ankuta Lucau-Danila, Keith Anderson, Bruno André, et al. Functional profiling of the saccharomyces cerevisiae genome. *nature*, 418(6896):387–391, 2002.

[43] Zhaojie Zhang and Qun Ren. Why are essential genes essential?-the essentiality of saccharomyces genes. *Microbial Cell*, 2(8):280, 2015.

[44] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[45] Yuen Ho, Albrecht Gruhler, Adrian Heilbut, Gary D Bader, Lynda Moore, Sally-Lin Adams, Anna Millar, Paul Taylor, Keiryn Bennett, Kelly Boutilier, et al. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, 415(6868):180–183, 2002.

[46] Qiangfeng Cliff Zhang, Donald Petrey, José Ignacio Garzón, Lei Deng, and Barry Honig. Preppi: a structure-informed database of protein–protein interactions. *Nucleic acids research*, 41(D1):D828–D833, 2012.

[47] Christian Von Mering, Roland Krause, Berend Snel, Michael Cornell, Stephen G Oliver, Stanley Fields, and Peer Bork. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, 417(6887):399–403, 2002.

[48] Jing-Dong J Han, Nicolas Bertin, Tong Hao, Debra S Goldberg, Gabriel F Berriz, Lan V Zhang, Denis Dupuy, Albertha JM Walhout, Michael E Cusick, Frederick P Roth, et al. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature*, 430(6995):88–93, 2004.

[49] Nizar N Batada, Laurence D Hurst, and Mike Tyers. Evolutionary and physiological importance of hub proteins. *PLoS computational biology*, 2(7):e88, 2006.

[50] Sumeet Agarwal, Charlotte M Deane, Mason A Porter, and Nick S Jones. Revisiting date and party hubs: novel approaches to role assignment in protein interaction networks. *PLoS computational biology*, 6(6):e1000817, 2010.

[51] Mihaela Pertea, Alaina Shumate, Geo Pertea, Ales Varabyou, Florian P Breitwieser, Yu-Chi Chang, Anil K Madugundu, Akhilesh Pandey, and Steven L Salzberg. Chess: a new human gene catalog curated from thousands of large-scale rna sequencing experiments reveals extensive transcriptional noise. *Genome biology*, 19:1–14, 2018.

[52] Ulrich Stelzl, Uwe Worm, Maciej Lalowski, Christian Haenig, Felix H Brembeck, Heike Goehler, Martin Stroedicke, Martina Zenkner, Anke Schoenherr, Susanne Koeppen, et al. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–968, 2005.

[53] Monica Agrawal, Marinka Zitnik, and Jure Leskovec. Large-scale analysis of disease pathways in the human interactome. In *PACIFIC SYMPOSIUM on BIOCOMPUTING 2018: Proceedings of the Pacific Symposium*, pages 111–122. World Scientific, 2018.

[54] Jörg Menche, Amitabh Sharma, Maksim Kitsak, Susan Dina Ghiassian, Marc Vidal, Joseph Loscalzo, and Albert-László Barabási. Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224):1257601, 2015.

[55] Andrew Chatr-Aryamontri, Bobby-Joe Breitkreutz, Rose Oughtred, Lorrie Boucher, Sven Heinicke, Daici Chen, Chris Stark, Ashton Breitkreutz, Nadine Kolas, Lara O'Donnell, et al. The biogrid interaction database: 2015 update. *Nucleic acids research*, 43(D1):D470–D478, 2015.

[56] Woong-Hee Shin, Charles W Christoffer, and Daisuke Kihara. In silico structure-based approaches to discover protein-protein interaction-targeting drugs. *Methods*, 131:22–32, 2017.

[57] Yingnan Han, Clarence Wang, Katherine Klinger, Deepak K Rajpal, and Cheng Zhu. An integrative network-based approach for drug target indication expansion. *PloS one*, 16(7):e0253614, 2021.

[58] Michelle R Arkin, Yinyan Tang, and James A Wells. Small-molecule inhibitors of protein-protein interactions: progressing toward the reality. *Chemistry & biology*, 21(9):1102–1114, 2014.

[59] Arnout Voet, Francois Berenger, and Kam YJ Zhang. Electrostatic similarities between protein and small molecule ligands facilitate the design of protein-protein interaction inhibitors. *PLoS One*, 8(10):e75762, 2013.

[60] Arnout RD Voet, Akihiro Ito, Mikako Hirohama, Seiji Matsuoka, Naoya Tochio, Takanori Kigawa, Minoru Yoshida, and Kam YJ Zhang. Discovery of small molecule inhibitors targeting the sumo–sim interaction using a protein interface consensus approach. *MedChemComm*, 5(6):783–786, 2014.

[61] Guimei Liu, Limsoon Wong, and Hon Nian Chua. Complex discovery from weighted ppi networks. *Bioinformatics*, 25(15):1891–1897, 2009.

[62] Tamás Nepusz, Haiyuan Yu, and Alberto Paccanaro. Detecting overlapping protein complexes in protein-protein interaction networks. *Nature methods*, 9(5):471–472, 2012.

[63] Ilya A Vakser. Protein-protein docking: From interaction to interactome. *Biophysical journal*, 107(8):1785–1793, 2014.

[64] Sergey Lyskov and Jeffrey J Gray. The rosettadock server for local protein–protein docking. *Nucleic acids research*, 36(suppl_2):W233–W238, 2008.

[65] Luis Angel Rodríguez-Lumbreras, Brian Jiménez-García, Silvia Giménez-Santamarina, and Juan Fernández-Recio. pydockdna: a new web server for energy-based protein-dna docking and scoring. *Frontiers in molecular biosciences*, 9:988996, 2022.

[66] Andrey Tovchigrechko and Ilya A Vakser. Gramm-x public web server for protein–protein docking. *Nucleic acids research*, 34(suppl_2):W310–W314, 2006.

[67] Stephen R Comeau, David W Gatchell, Sandor Vajda, and Carlos J Camacho. Cluspro: a fully automated algorithm for protein–protein docking. *Nucleic acids research*, 32(suppl_2):W96–W99, 2004.

[68] Gary Macindoe, Lazaros Mavridis, Vishwesh Venkatraman, Marie-Dominique Devignes, and David W Ritchie. Hexserver: an fft-based protein docking server powered by graphics processors. *Nucleic acids research*, 38(suppl_2):W445–W449, 2010.

[69] Dina Schneidman-Duhovny, Yuval Inbar, Ruth Nussinov, and Haim J Wolfson. Patchdock and symmdock: servers for rigid and symmetric docking. *Nucleic acids research*, 33(suppl_2):W363–W367, 2005.

[70] Ezgi Karaca, Adrien SJ Melquiond, Sjoerd J De Vries, Panagiotis L Kastritis, and Alexandre MJJ Bonvin. Building macromolecular assemblies by information-driven docking. *Molecular & Cellular Proteomics*, 9(8):1784–1794, 2010.

[71] Sjoerd J De Vries, Marc Van Dijk, and Alexandre MJJ Bonvin. The haddock web server for data-driven biomolecular docking. *Nature protocols*, 5(5):883–897, 2010.

[72] Christian B Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.

[73] Robert Zwanzig, Attila Szabo, and Biman Bagchi. Levinthal's paradox. *Proceedings of the National Academy of Sciences*, 89(1):20–22, 1992.

[74] Timo Lassmann, Oliver Frings, and Erik LL Sonnhammer. Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. *Nucleic acids research*, 37(3):858–865, 2009.

[75] William B Ware, John M Ferron, and Barbara M Miller. *Introductory statistics: A conceptual approach using R*. Routledge, 2013.

[76] Frank Emmert-Streib and Matthias Dehmer. Understanding statistical hypothesis testing: The logic of statistical inference. *Machine Learning and Knowledge Extraction*, 1(3):945–962, 2019.

[77] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

[78] Ajay Kulkarni, Deri Chong, and Feras A Batarseh. Foundations of data imbalance and solutions for a data democracy. In *Data democracy*, pages 83–106. Elsevier, 2020.

[79] David J Sheskin. *Handbook of parametric and nonparametric statistical procedures*. Chapman and hall/CRC, 2003.

[80] Brown University. Hypothesis testing. `https://cs.brown.edu/courses/csci1450/lectures/lec19_hypothesis.pdf`, n.d. Accessed: 2025-01-27.

[81] Helio S Migon, Dani Gamerman, and Francisco Louzada. *Statistical inference: an integrated approach*. CRC press, 2014.

[82] Jerzy Neyman and Egon Sharpe Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.

[83] Ronald L Wasserstein and Nicole A Lazar. The asa statement on p-values: context, process, and purpose, 2016.

[84] Sheldon M Ross. *Introductory statistics*. Academic Press, 2017.

[85] S. S. SHAPIRO and M. B. WILK. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611, dec 1965.

[86] Michael A Stephens. Introduction to kolmogorov (1933) on the empirical determination of a distribution. In *Breakthroughs in statistics: Methodology and distribution*, pages 93–105. Springer, 1992.

[87] Theodore W Anderson and Donald A Darling. Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes. *The annals of mathematical statistics*, pages 193–212, 1952.

[88] Student Biometrika. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908.

[89] Bernard L Welch. The generalization of 'student's'problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35, 1947.

[90] Ellen R Girden. *ANOVA: Repeated measures*. Number 84. Sage, 1992.

[91] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[92] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

[93] K Hartmann, J Krois, and A Rudolph. Statistics and geodata analysis using r (soga-r); department of earth sciences, freie universitaet berlin, 2023.

[94] John H McDonald. *Handbook of biological statistics*, volume 2. sparky house publishing Baltimore, MD, 2009.

[95] RH Randles. Wilcoxon signed rank test. encyclopedia of statistical sciences, 1988.

[96] Kandethody M Ramachandran and Chris P Tsokos. *Mathematical statistics with applications in R*. Academic Press, 2020.

[97] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Nature Precedings*, pages 1–1, 2010.

[98] Alicia Oshlack, Mark D Robinson, and Matthew D Young. From rna-seq reads to differential expression results. *Genome biology*, 11:1–10, 2010.

[99] Charlotte Soneson and Mauro Delorenzi. A comparison of methods for differential expression analysis of rna-seq data. *BMC bioinformatics*, 14:1–18, 2013.

[100] Diletta Rosati, Maria Palmieri, Giulia Brunelli, Andrea Morrione, Francesco Iannelli, Elisa Frullanti, and Antonio Giordano. Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: A review. *Computational and structural biotechnology journal*, 2024.

[101] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15:1–21, 2014.

[102] Davis J McCarthy and Gordon K Smyth. Testing significance relative to a fold-change threshold is a treat. *Bioinformatics*, 25(6):765–771, 2009.

[103] Jun Li and Robert Tibshirani. Finding consistent patterns: a nonparametric approach for identifying differential expression in rna-seq data. *Statistical methods in medical research*, 22(5):519–536, 2013.

[104] Pierre Fermat. Varia opera mathematica d. petri de fermat.

[105] Somnath Datta and Dan Nettleton. Statistical analysis of next generation sequencing data. 2014.

[106] Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, 11:1–9, 2010.

[107] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1), 2004.

[108] Eric S Lander. Initial impact of the sequencing of the human genome. *Nature*, 470(7333):187–197, 2011.

[109] Zachary D Stephens, Skylar Y Lee, Faraz Faghri, Roy H Campbell, Chengxiang Zhai, Miles J Efron, Ravishankar Iyer, Michael C Schatz, Saurabh Sinha, and Gene E Robinson. Big data: astronomical or genomical? *PLoS biology*, 13(7):e1002195, 2015.

[110] Jüri Reimand, Ruth Isserlin, Veronique Voisin, Mike Kucera, Christian Tannus-Lopes, Asha Rostamianfar, Lina Wadi, Mona Meyer, Jeff Wong, Changjiang Xu, et al. Pathway enrichment analysis and visualization of omics data using g: Profiler, gsea, cytoscape and enrichmentmap. *Nature protocols*, 14(2):482–517, 2019.

[111] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[112] The FlyBase Consortiuma. The flybase database of the drosophila genome projects and community literature. *Nucleic Acids Research*, 30(1):106, 2002.

[113] Judith A Blake, Janan T Eppig, Joel E Richardson, Muriel T Davisson, Mouse Genome Database Group, et al. The mouse genome database (mgd): expanding genetic and genomic resources for the laboratory mouse. *Nucleic Acids Research*, 28(1):108–111, 2000.

[114] Martin Ringwald, Janan T Eppig, James A Kadin, and Joel E Richardson. Gxd: a gene expression database for the laboratory mouse: current status and recent enhancements. *Nucleic acids research*, 28(1):115–119, 2000.

[115] Catherine A Ball, Kara Dolinski, Selina S Dwight, Midori A Harris, Laurie Issel-Tarver, Andrew Kasarskis, Charles R Scafe, Gavin Sherlock, Gail Binkley, Heng Jin, et al. Integrating functional genomic information into the saccharomyces genome database. *Nucleic acids research*, 28(1):77–80, 2000.

[116] Kirk J Maurer and Fred W Quimby. Animal models in biomedical research. In *Laboratory animal medicine*, pages 1497–1534. Elsevier, 2015.

[117] Rachael P Huntley, David Binns, Emily Dimmer, Daniel Barrell, Claire O'Donovan, and Rolf Apweiler. Quickgo: a user tutorial for the web-based gene ontology browser. *Database*, 2009:bap010, 2009.

[118] Sangdun Choi. *Systems biology for signaling networks*. Springer, 2010.

[119] Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, et al. The gene ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, 2023.

[120] G Joshi-Tope, Marc Gillespie, Imre Vastrik, Peter D'Eustachio, Esther Schmidt, Bernard de Bono, Bijay Jassal, Gopal R Gopinath, GR Wu, Lisa Matthews, et al. Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, 33(suppl_1):D428–D432, 2005.

[121] Antonio Fabregat, Konstantinos Sidiropoulos, Guilherme Viteri, Pablo Marin-Garcia, Peipei Ping, Lincoln Stein, Peter D'Eustachio, and Henning Hermjakob. Reactome diagram viewer: data structures and strategies to boost performance. *Bioinformatics*, 34(7):1208–1214, 2018.

[122] Kaumadi Wijesooriya, Sameer A Jadaan, Kaushalya L Perera, Tanuveer Kaur, and Mark Ziemann. Urgent need for consistent standards in functional enrichment analysis. *PLoS computational biology*, 18(3):e1009935, 2022.

[123] Sarah Mubeen, Alpha Tom Kodamullil, Martin Hofmann-Apitius, and Daniel Domingo-Fernández. On the influence of several factors on pathway enrichment analysis. *Briefings in bioinformatics*, 23(3):bbac143, 2022.

[124] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37(1):1–13, 2009.

[125] Paul D Thomas, Dustin Ebert, Anushya Muruganujan, Tremayne Mushayahama, Laurent-Philippe Albou, and Huaiyu Mi. Panther: Making genome-scale phylogenetics accessible to all. *Protein Science*, 31(1):8–22, 2022.

[126] Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R Clark, and Avi Ma'ayan. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC bioinformatics*, 14:1–14, 2013.

[127] Liis Kolberg, Uku Raudvere, Ivan Kuzmin, Priit Adler, Jaak Vilo, and Hedi Peterson. g: Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic acids research*, 51(W1):W207–W212, 2023.

[128] Huaiyu Mi, Anushya Muruganujan, John T Casagrande, and Paul D Thomas. Large-scale gene function analysis with the panther classification system. *Nature protocols*, 8(8):1551–1566, 2013.

[129] Isabelle Rivals, Léon Personnaz, Lieng Taing, and Marie-Claude Potier. Enrichment or depletion of a go category within a class of genes: which test? *Bioinformatics*, 23(4):401–407, 2007.

[130] Nazar Zaki, Harsh Singh, and Elfadil A Mohamed. Identifying protein complexes in protein-protein interaction data using graph convolutional network. *IEEE Access*, 9:123717–123726, 2021.

[131] Xiaoxia Liu, Zhihao Yang, Shengtian Sang, Ziwei Zhou, Lei Wang, Yin Zhang, Hongfei Lin, Jian Wang, and Bo Xu. Identifying protein complexes based on node embeddings obtained from protein-protein interaction networks. *BMC bioinformatics*, 19:1–14, 2018.

[132] Kasper Lage, E Olof Karlberg, Zenia M Størling, Páll I Olason, Anders G Pedersen, Olga Rigina, Anders M Hinsby, Zeynep Tümer, Flemming Pociot, Niels Tommerup, et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature biotechnology*, 25(3):309–316, 2007.

[133] Gary D Bader and Christopher WV Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics*, 4:1–27, 2003.

[134] Stijn Van Dongen. Graph clustering via a discrete uncoupling process. *SIAM Journal on Matrix Analysis and Applications*, 30(1):121–141, 2008.

[135] Thorsten Will and Volkhard Helms. Identifying transcription factor complexes and their roles. *Bioinformatics*, 30(17):i415–i421, 2014.

[136] Mileidy W Gonzalez and Maricel G Kann. Chapter 4: Protein interactions and disease. *PLoS computational biology*, 8(12):e1002819, 2012.

[137] Clara Pizzuti, Simona E Rombo, and Elena Marchiori. Complex detection in protein-protein interaction networks: a compact overview for researchers and practitioners. In *European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, pages 211–223. Springer, 2012.

[138] Frank W Albert and Leonid Kruglyak. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4):197–212, 2015.

[139] Tuuli Lappalainen and Michael Sammeth. Friedlä nder. *MR,'t Hoen, PA, Monlong, J., Rivas, MA, Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, PG, et al*, pages 506–511, 2013.

[140] Moritz Kebschull, Melanie Julia Fittler, Ryan T Demmer, and Panos N Papapanou. Differential expression and functional analysis of high-throughput-omics data using open source tools. *Oral Biology: Molecular Techniques and Applications*, pages 327–345, 2017.

[141] Alice Bossi and Ben Lehner. Tissue specificity and the human protein interaction network. *Molecular systems biology*, 5(1):260, 2009.

[142] Tiago JS Lopes, Martin Schaefer, Jason Shoemaker, Yukiko Matsuoka, Jean-Fred Fontaine, Gabriele Neumann, Miguel A Andrade-Navarro, Yoshihiro Kawaoka, and Hiroaki Kitano. Tissue-specific subnetworks and characteristics of publicly available human protein interaction databases. *Bioinformatics*, 27(17):2414–2421, 2011.

[143] Thorsten Will and Volkhard Helms. Ppixpress: construction of condition-specific protein interaction networks based on transcript expression. *Bioinformatics*, 32(4):571–578, 2016.

[144] Evans Kataka, Jan Zaucha, Goar Frishman, Andreas Ruepp, and Dmitrij Frishman. Edgetic perturbation signatures represent known and novel cancer biomarkers. *Scientific reports*, 10(1):4350, 2020.

[145] Thorsten Will and Volkhard Helms. Rewiring of the inferred protein interactome during blood development studied with the tool ppicompare. *BMC Systems Biology*, 11:1–19, 2017.

[146] Manfred Kunz, Henry Löffler-Wirth, Michael Dannemann, Edith Willscher, Gero Doose, Janet Kelso, Tina Kottek, Birgit Nickel, Lydia Hopp, Jenny Landsberg, et al. Rna-seq analysis identifies different transcriptomic types and developmental trajectories of primary melanomas. *Oncogene*, 37(47):6136–6151, 2018.

[147] David Talavera, David L Robertson, and Simon C Lovell. Alternative splicing and protein interaction data sets. *Nature biotechnology*, 31(4):292–293, 2013.

[148] Jose Manuel Rodriguez, Paolo Maietta, Iakes Ezkurdia, Alessandro Pietrelli, Jan-Jaap Wesselink, Gonzalo Lopez, Alfonso Valencia, and Michael L Tress. Appris: annotation of principal and alternative splice isoforms. *Nucleic acids research*, 41(D1):D110–D117, 2013.

[149] UniProt Consortium. Uniprot: a hub for protein information. *Nucleic acids research*, 43(D1):D204–D212, 2015.

[150] Fergal J Martin, M Ridwan Amode, Alisha Aneja, Olanrewaju Austine-Orimoloye, Andrey G Azov, If Barnes, Arne Becker, Ruth Bennett, Andrew Berry, Jyothish Bhai, et al. Ensembl 2023. *Nucleic acids research*, 51(D1):D933–D941, 2023.

[151] Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik LL Sonnhammer, Silvio CE Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, et al. Pfam: The protein families database in 2021. *Nucleic acids research*, 49(D1):D412–D419, 2021.

[152] Philip Jones, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, et al. Interproscan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240, 2014.

[153] Sailu Yellaboina, Asba Tasneem, Dmitri V Zaykin, Balaji Raghavachari, and Raja Jothi. Domine: a comprehensive collection of known and predicted domain-domain interactions. *Nucleic acids research*, 39(suppl_1):D730–D735, 2011.

[154] Yul Kim, Bumki Min, and Gwan-Su Yi. Iddi: integrated domain-domain interaction and protein interaction analysis system. In *Proteome science*, volume 10, pages 1–9. Springer, 2012.

[155] Roberto Mosca, Arnaud Céol, Amelie Stein, Roger Olivella, and Patrick Aloy. 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic acids research*, 42(D1):D374–D379, 2014.

[156] Robert D Finn, Benjamin L Miller, Jody Clements, and Alex Bateman. ipfam: a database of protein family and domain interactions found in the protein data bank. *Nucleic acids research*, 42(D1):D364–D373, 2014.

[157] Michael Levandowsky and David Winter. Distance between sets. *Nature*, 234(5323):34–35, 1971.

[158] Vasek Chvatal. A greedy heuristic for the set-covering problem. *Mathematics of operations research*, 4(3):233–235, 1979.

[159] Birgit H M Meldal, Hema Bye-A-Jee, Lukáš Gajdoš, Zuzana Hammerová, Aneta Horáčková, Filip Melicher, Livia Perfetto, Daniel Pokornỳ, Milagros Rodriguez Lopez, Alžběta Türková, et al. Complex portal 2018: extended content and enhanced visualization tools for macromolecular complexes. *Nucleic acids research*, 47(D1):D550–D558, 2019.

[160] Wenlou Liu, Yu Lu, Xiang Yan, Quansheng Lu, Yujin Sun, Xiao Wan, Yizhi Li, Jiaqin Zhao, Yuchen Li, and Guan Jiang. Current understanding on the role of cct3 in cancer research. *Frontiers in Oncology*, 12:961733, 2022.

[161] Wenlou Liu, Xiuli Zhang, Cheng Chen, Yizhi Li, Chunsheng Yang, Zhengxiang Han, Guan Jiang, and Yanqun Liu. Suppression of cct3 inhibits melanoma cell proliferation by downregulating cdk1 expression. *Journal of Cancer*, 13(6):1958, 2022.

[162] Anna Maria Lucianò and Ada Maria Tata. Functional characterization of cholinergic receptors in melanoma cells. *Cancers*, 12(11):3141, 2020.

[163] Mohamed Nabil Bakr, Haruko Takahashi, and Yutaka Kikuchi. Chrna1 and its correlated-myogenesis/cell cycle genes are prognosis-related markers of metastatic melanoma. *Biochemistry and Biophysics Reports*, 33:101425, 2023.

[164] Cui Liu, Li-Li Zhu, Si-Guang Xu, Hong-Long Ji, and Xiu-Min Li. Enac/deg in tumor development and progression. *Journal of Cancer*, 7(13):1888, 2016.

[165] Chibuzo Sampson, Qiuping Wang, Wuxiyar Otkur, Haifeng Zhao, Yun Lu, Xiaolong Liu, and Hailong Piao. The roles of e3 ubiquitin ligases in cancer progression and targeted therapy. *Clinical and Translational Medicine*, 13(3):e1204, 2023.

[166] Kristina Bielskienė, Lida Bagdonienė, Julija Mozūraitienė, Birutė Kazbarienė, and Ernestas Janulionis. E3 ubiquitin ligases as drug targets and prognostic biomarkers in melanoma. *Medicina*, 51(1):1–9, 2015.

[167] Frédéric Soysouvanh, Serena Giuliano, Nadia Habel, Najla El-Hachem, Céline Pisibon, Corine Bertolotto, and Robert Ballotti. An update on the role of ubiquitination in melanoma development and therapies. *Journal of Clinical Medicine*, 10(5):1133, 2021.

[168] Todd D Prickett, Brad J Zerlanko, Victoria K Hill, Jared J Gartner, Nouar Qutob, Jiji Jiang, May Simaan, John Wunderlich, J Silvio Gutkind, Steven A Rosenberg, et al. Somatic mutation of grin2a in malignant melanoma results in loss of tumor suppressor activity via aberrant nmdar complex formation. *Journal of Investigative Dermatology*, 134(9):2390–2398, 2014.

[169] Yemin Wang and Gang Li. Ing3 promotes uv-induced apoptosis via fas/caspase-8 pathway in melanoma cells. *Journal of Biological Chemistry*, 281(17):11887–11893, 2006.

[170] Gajanan S Inamdar, SubbaRao V Madhunapantula, and Gavin P Robertson. Targeting the mapk pathway in melanoma: why some approaches succeed and other fail. *Biochemical pharmacology*, 80(5):624–637, 2010.

[171] Anna Gajos-Michniewicz and Malgorzata Czyz. Wnt signaling in melanoma. *International journal of molecular sciences*, 21(14):4852, 2020.

[172] Michael A Davies. The role of the pi3k-akt pathway in melanoma. *The Cancer Journal*, 18(2):142–147, 2012.

[173] Jeff S Pawlikowski, Tony McBryan, John van Tuyn, Mark E Drotar, Rachael N Hewitt, Andrea B Maier, Ayala King, Karen Blyth, Hong Wu, and Peter D Adams. Wnt signaling potentiates nevogenesis. *Proceedings of the National Academy of Sciences*, 110(40):16009–16014, 2013.

[174] Janet Piñero, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, and Laura I Furlong. The disgenet knowledge platform for disease genomics: 2019 update. *Nucleic acids research*, 48(D1):D845–D855, 2020.

[175] Darren Dhananthat Chawhuaveang, Ollie Yiru Yu, Iris Xiaoxue Yin, Walter Yu-Hang Lam, May Lei Mei, and Chun-Hung Chu. Acquired salivary pellicle and oral diseases: A literature review. *Journal of Dental Sciences*, 16(1):523–529, 2021.

[176] Anderson T Hara and Domenick T Zero. The caries environment: saliva, pellicle, diet, and hard tissue ultrastructure. *Dental Clinics*, 54(3):455–467, 2010.

[177] M Hannig and A Joiner. The structure, function and properties of the acquired pellicle. *Monographs in oral science*, 19:29, 2006.

[178] Matthias Hannig. Ultrastructural investigation of pellicle morphogenesis at two different intraoral sites during a 24-h period. *Clinical oral investigations*, 3:88–95, 1999.

[179] TRYGGVE LIE. Scanning and transmission electron microscope study of pellicle morphogenesis. *European Journal of Oral Sciences*, 85(4):217–231, 1977.

[180] Sabine Güth-Thiel, Ines Kraus-Kuleszka, Hubert Mantz, Wiebke Hoth-Hannig, Hendrik Hähl, Johanna Dudek, Karin Jacobs, and Matthias Hannig. Comprehensive measurements of salivary pellicle thickness formed at different intraoral sites on si wafers and bovine enamel. *Colloids and Surfaces B: Biointerfaces*, 174:246–251, 2019.

[181] Joachim Enax, Bernhard Ganss, Bennett T Amaechi, Erik Schulze zur Wiesche, and Frederic Meyer. The composition of the dental pellicle: an updated literature review. *Frontiers in Oral Health*, 4:1260442, 2023.

[182] Floyd E Dewhirst, Tuste Chen, Jacques Izard, Bruce J Paster, Anne CR Tanner, Wen-Han Yu, Abirami Lakshmanan, and William G Wade. The human oral microbiome. *Journal of bacteriology*, 192(19):5002–5017, 2010.

[183] Tove Larsen and Nils-Erik Fiehn. Dental biofilm infections–an update. *Apmis*, 125(4):376–384, 2017.

[184] Ghaeth H Yassen, Jeffrey A Platt, and Anderson T Hara. Bovine teeth as substitute for human teeth in dental research: a review of literature. *Journal of oral science*, 53(3):273–282, 2011.

[185] Christian Zeitz, Thomas Faidt, Samuel Grandthyll, Hendrik Hähl, Nicolas Thewes, Christian Spengler, Jörg Schmauch, Michael Johannes Deckarm, Carsten Gachot, Harald Natter, et al. Synthesis of hydroxyapatite substrates: bridging the gap between model surfaces and enamel. *ACS Applied Materials & Interfaces*, 8(39):25848–25855, 2016.

[186] J Zheng, Y Li, MY Shi, YF Zhang, LM Qian, and ZR Zhou. Microtribological behaviour of human tooth enamel and artificial hydroxyapatite. *Tribology International*, 63:177–185, 2013.

[187] Thomas Faidt, Christian Zeitz, Samuel Grandthyll, Michael Hans, Matthias Hannig, Karin Jacobs, and Frank Muüller. Time dependence of fluoride uptake in hydroxyapatite. *ACS Biomaterials Science & Engineering*, 3(8):1822–1826, 2017.

[188] Thomas Faidt, Andreas Friedrichs, Samuel Grandthyll, Christian Spengler, Karin Jacobs, and Frank Müller. Effect of fluoride treatment on the acid resistance of hydroxyapatite. *Langmuir*, 34(50):15253–15258, 2018.

[189] Johannes Mischo, Thomas Faidt, Ryan B McMillan, Johanna Dudek, Gubesh Gunaratnam, Pardis Bayenat, Anne Holtsch, Christian Spengler, Frank Müller, Hendrik Hähl, et al. Hydroxyapatite pellets as versatile model surfaces for systematic adhesion studies on enamel: A force spectroscopy case study. *ACS Biomaterials Science & Engineering*, 8(4):1476–1485, 2022.

[190] Christina PC Sim, Stuart G Dashper, and Eric C Reynolds. Oral microbial biofilm models and their application to the testing of anticariogenic agents. *Journal of Dentistry*, 50:1–11, 2016.

[191] J. PHILIP SAPP, LEWIS R. EVERSOLE, and GEORGE P. WYSOCKI. Chapter 3 - infections of teeth and bone. In J. PHILIP SAPP, LEWIS R. EVERSOLE, and GEORGE P. WYSOCKI, editors, *Contemporary Oral and Maxillofacial Pathology (Second Edition)*, pages 70–93. Mosby, Saint Louis, second edition edition, 2004.

[192] Jasmin Flemming, Christian Hannig, and Matthias Hannig. Caries management—the role of surface interactions in de-and remineralization-processes. *Journal of Clinical Medicine*, 11(23):7044, 2022.

[193] Simone Trautmann, Nicolas Künzel, Claudia Fecher-Trost, Ahmad Barghash, Pascal Schalkowsky, Johanna Dudek, Judith Delius, Volkhard Helms, and Matthias Hannig. Deep proteomic insights into the individual short-term pellicle formation on enamel—an in situ pilot study. *PROTEOMICS–Clinical Applications*, 14(3):1900090, 2020.

[194] Thomas M Annesley. Ion suppression in mass spectrometry. *Clinical chemistry*, 49(7):1041–1044, 2003.

[195] Juri Rappsilber, Ursula Ryder, Angus I Lamond, and Matthias Mann. Large-scale proteomic analysis of the human spliceosome. *Genome research*, 12(8):1231–1245, 2002.

[196] Dongyou Liu. *Handbook of Molecular Biotechnology*. CRC Press, 2024.

[197] Yasushi Ishihama, Yoshiya Oda, Tsuyoshi Tabata, Toshitaka Sato, Takeshi Nagasu, Juri Rappsilber, and Matthias Mann. Exponentially modified protein abundance index (empai) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein* s. *Molecular & Cellular Proteomics*, 4(9):1265–1272, 2005.

[198] Deborah H Lundgren, Sun-Il Hwang, Linfeng Wu, and David K Han. Role of spectral counting in quantitative proteomics. *Expert review of proteomics*, 7(1):39–53, 2010.

[199] Wes McKinney et al. Data structures for statistical computing in python. In *SciPy*, volume 445, pages 51–56, 2010.

[200] Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.

[201] Denise Rey and Markus Neuhäuser. Wilcoxon-signed-rank test. *International encyclopedia of statistical science*, pages 1658–1659, 2011.

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

[203] Rhoda J Kinsella, Andreas Kähäri, Syed Haider, Jorge Zamora, Glenn Proctor, Giulietta Spudich, Jeff Almeida-King, Daniel Staines, Paul Derwent, Arnaud Kerhornou, et al. Ensembl biomarts: a hub for data retrieval across taxonomic space. *Database*, 2011:bar030, 2011.

[204] Lukasz P Kozlowski. Proteome-pi: proteome isoelectric point database. *Nucleic acids research*, 45(D1):D1112–D1116, 2017.

[205] Edward E Cureton. Rank-biserial correlation. *Psychometrika*, 21(3):287–290, 1956.

[206] David Binns, Emily Dimmer, Rachael Huntley, Daniel Barrell, Claire O'donovan, and Rolf Apweiler. Quickgo: a web-based tool for gene ontology searching. *Bioinformatics*, 25(22):3045–3046, 2009.

[207] Karl-Josef Dietz. Peroxiredoxins in plants and cyanobacteria. *Antioxidants & redox signaling*, 15(4):1129–1159, 2011.

[208] Bernard Knoops, ElÉonore Loumaye, and ValÉrie Van Der Eecken. Evolution of the peroxiredoxins: Taxonomy, homology and characterization. *Peroxiredoxin Systems: Structures and Functions*, pages 27–40, 2007.

[209] Leslie B Poole, Andrea Hall, and Kimberly J Nelson. Overview of peroxiredoxins in oxidant defense and redox regulation. *Current protocols in toxicology*, 49(1):7–9, 2011.

[210] Marlène Dubuisson, Delphine Vander Stricht, André Clippe, Florence Etienne, Thomas Nauser, Reinhard Kissner, Willem H Koppenol, Jean-François Rees, and Bernard Knoops. Human peroxiredoxin 5 is a peroxynitrite reductase. *FEBS letters*, 571(1-3):161–165, 2004.

[211] Arden Perkins, Kimberly J Nelson, Derek Parsonage, Leslie B Poole, and P Andrew Karplus. Peroxiredoxins: guardians against oxidative stress and modulators of peroxide signaling. *Trends in biochemical sciences*, 40(8):435–445, 2015.

[212] Andrea Hall, Kimberly Nelson, Leslie B Poole, and P Andrew Karplus. Structure-based insights into the catalytic power and conformational dexterity of peroxiredoxins. *Antioxidants & redox signaling*, 15(3):795–815, 2011.

[213] Zachary A Wood, Ewald Schröder, J Robin Harris, and Leslie B Poole. Structure, mechanism and regulation of peroxiredoxins. *Trends in biochemical sciences*, 28(1):32–40, 2003.

[214] Chi-Ming Wong, Yuan Zhou, Raymond WM Ng, Hsiang-fu Kung, and Dong-Yan Jin. Cooperation of yeast peroxiredoxins tsa1p and tsa2p in the cellular defense against oxidative and nitrosative stress. *Journal of Biological Chemistry*, 277(7):5385–5394, 2002.

[215] Birgit Hofmann, H-J Hecht, and Leopold Flohé. Peroxiredoxins. 2002.

[216] Henry Jay Forman and Enrique Cadenas. *Oxidative stress and signal transduction*. Springer Science & Business Media, 2012.

[217] Daniel Pastor-Flores, Deepti Talwar, Brandán Pedre, and Tobias P Dick. Real-time monitoring of peroxiredoxin oligomerization dynamics in living cells. *Proceedings of the National Academy of Sciences*, 117(28):16313–16323, 2020.

[218] Luis A Ralat, Yefim Manevich, Aron B Fisher, and Roberta F Colman. Direct evidence for the formation of a complex between 1-cysteine peroxiredoxin and glutathione s-transferase π with activity changes in both enzymes. *Biochemistry*, 45(2):360–372, 2006.

[219] Jesalyn Bolduc, Katarina Koruza, Ting Luo, Julia Malo Pueyo, Trung Nghia Vo, Daria Ezeriņa, and Joris Messens. Peroxiredoxins wear many hats: Factors that fashion their peroxide sensing personalities. *Redox Biology*, 42:101959, 2021.

[220] Ho Zoon Chae, Sang Jin Chung, and Sue Goo Rhee. Thioredoxin-dependent peroxide reductase from yeast. *Journal of biological chemistry*, 269(44):27670–27678, 1994.

[221] Sung Goo Park, Mee-Kyung Cha, Woojin Jeong, and Il-Han Kim. Distinct physiological functions of thiol peroxidase isoenzymes in saccharomyces cerevisiae. *Journal of Biological Chemistry*, 275(8):5723–5732, 2000.

[222] Luis ES Netto, Ho Zoon Chae, Sang-Won Kang, Sue Goo Rhee, and Earl R Stadtman. Removal of hydrogen peroxide by thiol-specific antioxidant enzyme (tsa) is involved with its antioxidant properties: Tsa possesses thiol peroxidase activity. *Journal of Biological Chemistry*, 271(26):15315–15321, 1996.

[223] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[224] Karl-Josef Dietz. Plant peroxiredoxins. *Annual review of plant biology*, 54(1):93–107, 2003.

[225] Ye Yang, Wenguang Cai, Junchao Wang, Weimin Pan, Lin Liu, Mingzhu Wang, and Min Zhang. Crystal structure of arabidopsis thaliana peroxiredoxin a c119s mutant. *Acta Crystallographica Section F: Structural Biology Communications*, 74(10):625–631, 2018.

[226] Robert D Finn, Jody Clements, and Sean R Eddy. Hmmer web server: interactive sequence similarity searching. *Nucleic acids research*, 39(suppl_2):W29–W37, 2011.

[227] Andrew Waterhouse, Martino Bertoni, Stefan Bienert, Gabriel Studer, Gerardo Tauriello, Rafal Gumienny, Florian T Heer, Tjaart A P de Beer, Christine Rempfer, Lorenza Bordoli, et al. Swiss-model: homology modelling of protein structures and complexes. *Nucleic acids research*, 46(W1):W296–W303, 2018.

[228] Andrej Šali and Tom L Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology*, 234(3):779–815, 1993.

[229] Nicola Bordin, Ian Sillitoe, Vamsi Nallapareddy, Clemens Rauer, Su Datt Lam, Vaishali P Waman, Neeladri Sen, Michael Heinzinger, Maria Littmann, Stephanie Kim, et al. Alphafold2 reveals commonalities and novelties in protein structure space for 21 model organisms. *Communications biology*, 6(1):160, 2023.

[230] Janani Durairaj, Andrew M Waterhouse, Toomas Mets, Tetiana Brodiazhenko, Minhal Abdullah, Gabriel Studer, Gerardo Tauriello, Mehmet Akdel, Antonina Andreeva, Alex Bateman, et al. Uncovering new families and folds in the natural protein universe. *Nature*, 622(7983):646–653, 2023.

[231] Sebastian Raschka. Molecular docking, estimating free energies of binding, and autodock's semi empirical force field. *can be found under http://sebastianraschka. com/Articles/2014_autodo ck_energycomps. html# table-of-contents.–2014*, 2014.

[232] Jinrui Xu and Yang Zhang. How significant is a protein structure similarity with tm-score= 0.5? *Bioinformatics*, 26(7):889–895, 2010.

[233] RJ Sullivan and K Flaherty. Map kinase signaling and inhibition in melanoma. *Oncogene*, 32(19):2373–2379, 2013.

[234] Lawrence N Kwong and Michael A Davies. Navigating the therapeutic complexity of pi3k pathway inhibition in melanoma. *Clinical cancer research*, 19(19):5310–5319, 2013.

[235] Zhenyu Ji, Keith T Flaherty, and Hensin Tsao. Targeting the ras pathway in melanoma. *Trends in molecular medicine*, 18(1):27–35, 2012.

[236] Amira Al Mahi and Julien Ablain. Ras pathway regulation in melanoma. *Disease Models & Mechanisms*, 15(2):dmm049229, 2022.

[237] Giorgia Castellani, Mariachiara Buccarelli, Maria Beatrice Arasi, Stefania Rossi, Maria Elena Pisanu, Maria Bellenghi, Carla Lintas, and Claudio Tabolacci. Braf mutations in melanoma: biological aspects, therapeutic implications, and circulating biomarkers. *Cancers*, 15(16):4026, 2023.

[238] Aljawharah Alqathama. Braf in malignant melanoma progression and metastasis: potentials and challenges. *American Journal of Cancer Research*, 10(4):1103, 2020.

[239] Heng Wu, Vikas Goel, and Frank G Haluska. Pten signaling pathways in melanoma. *Oncogene*, 22(20):3113–3122, 2003.

[240] Furkan Akif Ince, Artur Shariev, and Katie Dixon. Pten as a target in melanoma. *Journal of Clinical Pathology*, 75(9):581–584, 2022.

[241] Gongda Xue, Emanuela Romano, Daniela Massi, and Mario Mandalà. Wnt/β-catenin signaling in melanoma: Preclinical rationale and novel therapeutic insights. *Cancer treatment reviews*, 49:1–12, 2016.

[242] Daniela Kovacs, Emilia Migliano, Luca Muscardin, Vitaliano Silipo, Caterina Catricalà, Mauro Picardo, and Barbara Bellei. The role of wnt/β-catenin signaling pathway in melanoma epithelial-to-mesenchymal-like switching: Evidences from patients-derived cell lines. *Oncotarget*, 7(28):43295, 2016.

[243] Sjoerd J de Vries, Adrien SJ Melquiond, Panagiotis L Kastritis, Ezgi Karaca, Annalisa Bordogna, Marc van Dijk, Joao PGLM Rodrigues, and Alexandre MJJ Bonvin. Strengths and weaknesses of data-driven docking in critical assessment of prediction of interactions. *Proteins: Structure, Function, and Bioinformatics*, 78(15):3242–3249, 2010.