Doctoral Dissertation
Doctoral Program in Bioengineering and Medical Surgical Sciences (37th Cycle)

Bioinformatics and Machine Learning Approaches for Biomarker Discovery and Predictive Modelling

Applications in Healthcare

Konstantinos PANAGIOTOPOULOS

* *** * **

Supervisor(s):
Prof. Marco Agostino Deriu, Supervisor
Prof. Jacek Adam Tuszynski, Co-Supervisor

Doctoral Examination Committee:
Prof. Dr. Franca Fraternali, University College London (UCL), UK
Prof. Dr. Andrea Danani, Università dell Svizzera italiana (USI), Switzerland

Politecnico di Torino - Università degli Studi di Torino
July 17, 2025

I hereby declare that the contents and organisation of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

This doctoral dissertation is based on material that was previously published and for which the author holds the right for inclusion. Original publications are acknowledged in the pertinent chapters.

____

Konstantinos Panagiotopoulos
Turin, July 17, 2025

Dedication

With the completion of the Ph.D. programme, I would like to thank all the people that stood by my side during this trip. In strange and uncertain times, I began this path in the heart of a global pandemic, in a new country, and feeling a bit lost, but they were always there for me. When I was doubting myself, they were standing by me with unwavering support.

So, I dedicate this work to all of them.

Και ας μη νικήσουμε ποτέ, θα πολεμάμε πάντα! Λευτεριά στην Παλαιστίνη!

Acknowledgment

I would like to acknowledge the support I received from the European Union's Marie Sklodowska-Curie Innovative Training Network 2020 project PARENT (G.A.: N° 956394) (https://parenth2020.com/), both financially but mostly for the training events and the collaborations I had the opportunity to have during this journey.

Through the ITN training sessions and the mobility opportunities, I had the chance to meet and work in inspiring environments and alongside brilliant scientists from all over the world. The result of this networking is the present thesis, since most - if not all - of the scientific work presented below is the fruit of these collaborations.

One of the strongest supporters of this thesis I would like to acknowledge is the company InSyBio PC (https://insybio.com) located in Patras, Greece. Through close collaboration with people from InSyBio, I found inspiration and guidance that significantly contributed to the development and improvement of many of my ongoing projects.

The company very generously hosted me for two mobility periods at its offices, as part of the secondments for the PARENT project, as well as for a three-month period in the King's College London, where Dr. Konstantinos Theofilatos – a founding member of InSyBio – is based. During this time, I had the opportunity to work on machine learning applications, bioinformatics, statistics, and the development of a database, which is kindly hosted on the company's servers.

Summary

Biomarkers play a crucial role in modern medicine, serving as measurable indicators of biological processes, disease states, or responses to treatments. Their importance spans from early disease detection and diagnosis to personalized treatment strategies and the identification of predisposition characteristics, significantly improving patients' outcomes. Advancements in molecular biology and technology have led to the discovery of a wide range of biomarkers, transforming fields such as oncology, neurology, and infectious diseases. Furthermore, beyond molecular biomarkers, other types of markers such as clinical variables (e.g., blood pressure, LDL cholesterol, BMI, assessment scores), signal-based (e.g., EEG and MRI characteristics), and sensor biomarkers (e.g., glucose measuring sensors) further enhance diagnostic and monitoring capabilities in medicine.

Bioinformatics has become essential for analyzing and interpreting complex conditions with the emergence and increasing availability of high-throughput data. Computational tools enable researchers to identify new biomarkers, assess their reliability, and integrate them into clinical practice. Moreover, new machine learning algorithms have enhanced biomarker research by avoiding statistical assumptions, thus allowing for pattern recognition and the development of predictive models. Leveraging large datasets, like omics, machine learning algorithms can identify subtle correlations that may be overlooked by traditional statistical methods, ultimately leading to more accurate clinical diagnostics and improved treatment planning.

Despite these advancements, several challenges persist that this thesis seeks to address. In bioinformatics, new tools and methods are necessary in biomarker discovery which is often hindered by biological variability, limited sample sizes, and issues with reproducibility. In addition, data heterogeneity makes it more difficult for retrospective analysis which can reduce the need of repeating studies. In clinical practice, it is important to foresee the early evolution of diseases to prevent or cure them in time and create better treatments. The present thesis directly tackles these limitations by providing novel applications and methodologies rooted in data integration processes and advanced machine learning models, significantly aiding clinical decision-making and biomarker discovery.

The initial chapters of this thesis explore the critical role and advanced identification methods for both coding and non-coding molecular biomarkers, highlighting the challenges inherent in their discovery.. A major part of this work is dedicated to the harmonization of microarray datasets to increase the sample sizes of conditions and improve the statistical power of biomarkers' findings. A method that can help in cases of rare conditions and valuable past studies combination. Emphasis was given on neurodevelopmental disorders (NDDs) and blood samples to face both the problems of minimal invasiveness and data scarcity. For this, six distinct studies with four different microarray designs were used to demonstrate the challenges of data harmonization. That effort yields a homogenized pool of transcriptomic dataset for NDDs.

Beyond data integration, an application of a multi-objective evolutionary algorithm implementing machine learning modelling on clinical data is presented, proving that it is beneficial in clinical decision making by identifying important markers that play a role in disease manifestation . This method effectively integrates both omics and clinical data, achieving robust classification performance while simultaneously reducing the redundancy of important features. This is extremely useful in prognostics, precision medicine, and biomarker discovery. Furthermore, a machine learning application for early risk prediction of cardiovascular disease in people with thalassemia is proposed as a clinical decision support system. This tool demonstrated a high predictive efficiency, achieving a ROC-AUC of over 90%.

Finally, this thesis presents a functional and interactive unified platform that integrates the above-mentioned tools into an easy-to-use online application. This open-access platform allows for the practical use of the methods developed on the user's datasets seamlessly. The platform's main role is the demonstration of the developed tools as a set of methods for diverse biomarker analysis.

In summary, the present work is presented as an innovation in the field of finding potential biomarkers of different types through data-driven approaches while also creating clinical decision support systems, following the dictates of personalized medicine and risk prediction methods.

Contents

List of Figures

List of Tables

List of equations

List of Acronyms

3’ (prime) untranslated region: **3’UTR**
Akaike information criterion: **AIC**
Alzheimer’s Disease: **AD**
Analysis of variance: **ANOVA**
Autism spectrum disorder: **ASD**
Cerebral Palsy: **CP**
Competitive endogenous RNA: **ceRNA**
Copy Number Variations: **CNV**
Differential Expression: **DE**
Effective Degrees of Freedom: **EDF**
Evolutionary Algorithms: **EA**
False discovery rate: **FDR**
Gene Expression Omnibus: **GEO**
Gene Set Enrichment Analysis: **GSEA**
Generalized additive model: **GAM**
Genetic Algorithms: **GA**
Genome-Wide Association Studies: **GWAS**
K-nearest neighbors: **KNN**
Long non-coding RNA: **lncRNA**
Magnetic resonance imaging: **MRI**
Mass Spectrometry: **MS**
Messenger RNA: **mRNA**
Micro RNA: **miRNA**
miRNA response element: **MRE**
Neonatal intensive care unit: **NICU**
Neurodevelopmental disorder: **NDD**
Next generation sequencing: **NGS**
Non-coding RNA: **ncRNA**
Normalized Enrichment Score: **NES**
P-element induced wimpy testis-interacting RNA: **piRNA**
Principal Component Analysis: **PCA**
Protein-Protein interaction: **PPI**
Ribonucleotide acid: **RNA**
RNA interference: **RNAi**
Robust Multichip Average: **RMA**
Small interference RNA: **siRNA**
Small nucleic RNA: **snRNA**
Small/short non-coding RNA: **sncRNA**
Structured Query Language: **SQL**

Tourette Syndrome: **TS**
White matter injury: **WMI**
William’s Syndrome: **WS**
False Positive Rate: **FPR**
False Negative Rate: **FNR**
Receiver operating characteristics: **ROC**
Area Under the Curve: **AUC**

Chapter 1

Introduction

1.1 Thesis field of interest

The present thesis revolves around the **discovery of biomarkers** in healthcare, and their prognostic and diagnostic role in various diseases. It focuses on the development of tools for biomarkers discovery in the field of **bioinformatics**, the application of **Machine Learning** (ML) models for risk prediction, and methods for data harmonization to enable merging for transcriptomics datasets.

One of the primary objectives of this work is to analyse expression data from preterm infants to identify potential indicators of neurodevelopmental disorders. In the realm of bioinformatics and omics research, a prevalent challenge is the disproportionate number of features relative to the limited number of available samples. Addressing this issue, the thesis emphasises the importance of harmonising datasets to increase sample sizes, thus enhancing the statistical power and reliability of findings.

The thesis also investigates different types of biomarkers, clinical and molecular, and explores how ML models can facilitate the identification of key features critical for disease prognosis and diagnosis. By leveraging the strengths of bioinformatics and ML, this work aims to propose some methods for identifying characteristics in the data which can potentially give us clinical insights. These clinical insights can be a treatment plan for a patient in the framework of precision medicine, a risk assessment that can alert clinicians early to the trajectory of a patient's health, or even a group of genes as factors of the manifestation of a disease.

This thesis presents a unified platform to integrate the diverse tools developed, ranging from bioinformatics pipelines to clinical decision-support systems and databases. This platform simplifies access to the developed tools, making them accessible and practical for users to explore and apply in their respective domains.

1.2 Thesis objectives

The current work's central aim is to combine ML algorithms with structured clinical data and transcriptomics data for biomarker identification. In particular, the transcriptomics datasets used, mined from online sources under the work frame of the PARENT project and are related to neurodevelopmental disorders with focus on prematurely born children. Under the PARENT specification, the data should also come from blood samples in order to be as less invasive as possible for the infants. These constraints lead to several challenges that must be considered and overcome, beginning with the fact that the samples are derived from blood while the affected tissue in neurodevelopmental disorders involves the brain. Other difficulties are the variability in the samples that is the effect of different factors such as ethnicity, sex, prematurity, and factors related to the study itself such as the method used, the laboratory, the technician and others. On top of those challenges, one has to face a common problem of omics data, often referred as the curse of dimensionality. This is when the number of features is much greater than the sample sizes, creating a large features to samples ratio. This characteristic imbalance causes traditional distance metrics to lose their significance and diminish the effectiveness of machine learning algorithms. In the case the studies are conducted on infants, one more clinical and technical problem is the volume of blood that can be sampled because of the babies' fragility and low (or very low in some cases) birth weight, that lead to insufficient material to conduct the studies properly, or to strategy change and selection of simpler methods. All these obstacles require solutions that are not trivial and require data harmonization and merging and removal of non-biological variations while retaining the meaningful differences between samples.

On the other hand, in clinical data of children and adult patients, the nature of the challenges differs. Here, the population is, most of the time, able to communicate and express their needs, give feedback, and cooperate with medics. A continual issue is the fragmentation and sparsity of data across healthcare systems, stemming from the manual curation of records. Additionally, the conditions in which the aid of ML is required are frequently cases with high imbalance, which makes generalization harder. Thus, the challenge is for an ML model to find underlying patterns through imbalanced datasets, provide a reliable prediction and ideally an explanation for the classification in a way that is easy to interpret by clinicians.

Consequently, this research focused on these two main challenges: (C1) developing methods integrating ML for biomarkers' discovery in omics data with focus on neurodevelopmental disorders (NDDs), and (C2) the development of ML models for prediction of outcomes and identification of characteristics important in clinical decision support systems. Both research lines come with their own objectives (denoted as "O" followed by a number) which were set as follows:

C1) Biomarker discovery in transcriptomics for NDDs

O1. Gathering of transcriptomics datasets related to neurodevelopmental disorders from online open-source databanks. This will be the foundation for Objectives O2 and O4 in which the data will be analyzed and stored in the custom MySQL database.

O2. Analysis of the datasets collected in objective O1 with classical bioinformatics methods and building a harmonized reference pool for increasing sample sizes and statistical power in the findings.

O3. Development of a method that efficiently searches the feature space of omics data for addressing the high-dimensionality with ML modelling. This method will provide a non-redundant feature set and models that perform well both on classification and minimization of redundant biomarkers.

O4. Creation of a MySQL database for three serving purposes: storing the data used for this research in one place, creation of a homogenized pool of transcriptomics data for NDDs, and serving as a job management platform for the objective O7

C2) development of ML models for biomarker discovery and clinical decision support systems

O5. Development of a sophisticated algorithm that leverages the benefits of machine learning to counteract the effects of large features-to-samples ratios in the datasets typical for -omics data, in line with objective 3 (O3). This method is a multi-objective evolutionary algorithm that enables both feature-space search, and hyper-parameter optimization. In its core it uses an XGBoost algorithm and a Pareto frontier method to avoid premature convergence while returning multiple models which perform equally well depending on the objective of the user.

O6. Creation of a risk predictor of cardiovascular disease in β-thalassemia patients in a mid-term future, for the clinical decision-making support and lifestyle adaptation. Two models predict the CVD risk within one and three years respectively after a patient is presented to the hospital.

O7. Implementing of a user-friendly platform to unify the models and tool developed in the research. The platform will allow users to have their personal accounts and submit job queries through the tool-dedicated pages within the platform. For the creation of the platform, the database's functionality is increased to handle job requests and prioritize them.

Concluding the challenges that this work tries to tackle, they span in the two main fields, that of bioinformatics and that of machine learning. These fields are traditionally distinct, but increasingly integrated. On the bioinformatics domain, this thesis contributes to elucidating the diversity and complexity of biology driving pathogenesis with a dedicated section to biomarkers in clinical applications. Additionally, it describes a common problem related to low sample sizes and methodological heterogeneity in studies by proposing a data harmonization pipeline. This pipeline features a custom normalization step for microarray datasets before merging which considers sample-derived differences in gene expression.

Integrating advanced ML with bioinformatics, this thesis introduces an evolutionary algorithm that optimizes an ensemble classifier. This innovative method is designed to simultaneously enhance classification performance and significantly reduce the number of redundant biomarkers, providing a more efficient and interpretable feature set. Benchmarked against both omics and clinical dataset, this method demonstrates comparable performances to state-of-the-art algorithms. Considering biomarker discovery and clinical decision support systems, this thesis provides an example in hematology through the development of a cardiovascular disease risk predictor for thalassemia patients. This model achieved a significantly high performance predicting a cardiovascular disease within a mid-term period. Finally, all developed tools are unified within a Python-based platform, laying the groundwork for a robust and integrated suite of bioinformatics and ML tools for future research.

A concise visual summary of the present thesis can be found below in Figure 1. The machine learning side (left) is dedicated to tools developed for the objectives of C2 (biomarkers discovery with ML), and the bioinformatics side (right) is dedicated to classical analysis of microarray data and a method for data harmonization as described on the objectives of C1. All the developed technologies and tools are linked to the platform and the database, which are the final unified result of this thesis. The role of the platform is information exchange and the calling of the proper functions to execute job queries of the user.

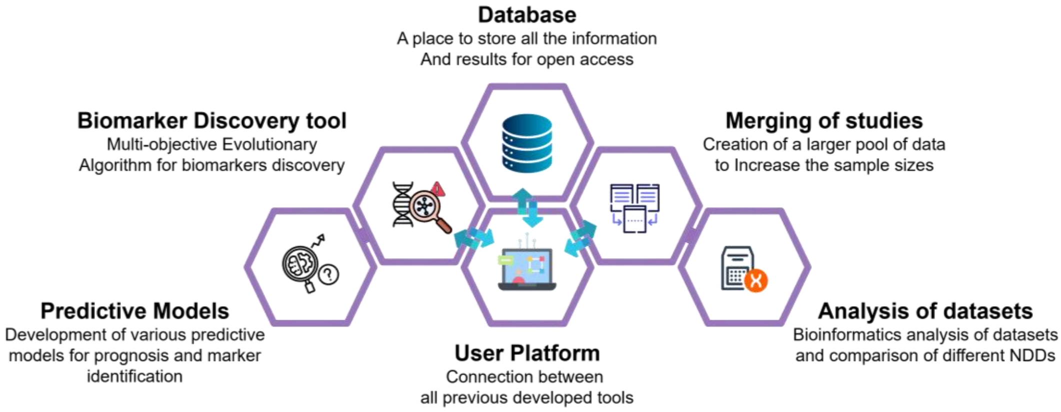


Figure 1. Graphical representation of the developed work for this thesis.

Chapter 2

Biomarkers in Clinical Applications

In this chapter a general overview of biomarkers used in clinical practice will be given with a special focus on coding and non-coding RNA molecules. On these, we will explore their role in neurological and other disorders, and their mechanisms to unravel some of their complexity and regulatory role in physiology.

2.1 Biomarkers in Medicine

Biomarkers are different kinds of evidence that can be used to directly or indirectly diagnose or prognose potential diseases and impairments in one's health. The first part of the word, "bio" shows the nature of those indicators which are related to the broader medicine field and spans from signals such as electrocardiograms (ECGs), to images like X-rays or magnetic resonance imaging (MRIs), and to genetic sequences or expression of biomolecules. Biomarkers are used in almost all fields of medicine and have been involved over time into a useful tool for prognostics, diagnostics and disease progression. Their role in complex diseases is phenomenal with the most emphatic example being the cancer diagnosis and therapy [1], but also in neurodegenerative diseases which can be detected through cerebrospinal fluid analysis or PET scans as key biomarkers [2], [3].

On the field of pediatrics and preterm infants that is one of the main interests of this work, neuroimaging (i.e., MRIs, or 3D ultrasound) is the common practice to use for identifying potential brain injuries and impairments in newborns and kids in risk of neurodevelopmental disorders (i.e., Cerebral Palsy) [4]. Despite that, the correct classification based on imaging is challenging and depends on various factors such as the type of lesion, the type of imaging and others. As an example, in the case of Cerebral Palsy, studies have shown promising using structural magnetic resonance imaging (sMRI) but with no evidence of any kind of brain damage findings in up to 11.7% of cases [5]. Additionally to that, it is not yet clear if other classes of MRI is more suitable for those cases [6]. On top of these, even with some optical findings in an MR image, the prediction of a neurodevelopmental problem is not certain as S. Arulkumaran et al. stated in their work, especially focused on preterm infants [7]. Additional to neuroimaging, there are other types of indicators at a different levels such as oxidative stress markers from blood samples, urine or saliva that can be used for monitoring the stress that can be induced to a preterm neonate by mechanical ventilation, apnea and resuscitation [8].

Continuous or repeated measurements are another significant type of biomarkers used in medicine. For instance, continuous glucose monitoring in neonates provide real-time glucose levels in blood, which, when left dysregulated, is correlated with morbidity, mortality, and poor neurologic outcome [9], [10]. Wearable devices, increasingly common in recent years, track heart rate and offer potential as early indicators of stress or cardiac issues. These types of markers have the advantage of being a dynamic reflection of physiology since they measure time variant factors, but also introduce new challenges in data analysis and integrity, particularly given their automatic generation, which necessitates robust validation and quality control measures. Managing and analyzing the vast amounts of continuous data generated requires sophisticated algorithms.

Additionally, there are markers at the molecular level that are gaining significant attention in the study of neurodevelopmental disorders (NDDs), particularly in preterm neonates. Among the different types of molecular biomarkers, transcriptomic biomarkers have gained significant attention. These include coding RNA, which is transcribed and translated into proteins that carry out essential cellular functions, and non-coding RNA, which does not encode proteins but plays a key role in gene regulation. While coding biomarkers have historically been the primary focus of research, growing evidence suggests that non-coding RNAs (ncRNAs) play a crucial role in neurodevelopment by influencing processes such as neuronal differentiation, synaptic plasticity, and neuroinflammation. Given their regulatory functions, ncRNAs are emerging as promising candidates for biomarker discovery in NDDs.

2.2 Non-coding Markers

In this subsection, we provide a brief introduction to the established roles of the non-coding part of the genome in physiological processes and its relevance to NDDs. Based on the work presented at the DETERMINED2022 conference, we present a concise analysis of the different types and functions of ncRNAs, with a particular emphasis on small ncRNAs and their interactions with other biomolecules in the context of neurodevelopmental disorders and other neurological diseases.

The present content has been partially extracted from the published article:

A review on computational tools for analytical visualisation and molecular interactions of sncRNAs: prospects in NDDs

1st International Workshop in Neurodevelopmental Impairments in Preterm Children - Computational Advancements (DETERMINED 2022).

CEUR. Volume 3363, November 2023. doi: https://doi.org/10.5281/zenodo.8009881

2.2.1 Abstract

Neurodevelopmental disorders (NDDs), including cognitive impairments, motor disabilities, and psychosocial disorders, are common among infants born prematurely. However, the molecular mechanisms underlying these disorders remain unclear. Recent studies have highlighted shared molecular pathways between NDDs and neurodegenerative diseases, with small non-coding RNAs (sncRNAs) playing a significant role in their manifestation. Understanding these mechanisms is crucial for early prediction using biomarkers, enabling medical interventions during the period of heightened neuroplasticity in newborns, which allows for potential recovery. In this section, we explore the role of sncRNAs and their involvement in shared pathways in NDDs.

2.2.2 Introduction

Preterm babies are defined as those born before the 37th week of gestation, with births before the 32nd week classified as very preterm [11], [12]. Premature deliveries account for about 15% of total labors globally, with an increasing trend [13]. From the clinical point of view, preterm infants with low birth weight have higher chances of experiencing short- or long-term neurodevelopmental disorders (NDDs) and related comorbidities [14]. Common NDDs are related to motor deficits such as cerebral palsy (CP), cognitive and speech delays, visual and hearing impairments, and some psychosocial and behavioral disorders such as autism spectrum disorder (ASD) and schizophrenia. The most common methods of assessment of the developing brains in infants are using magnetic resonance imaging (MRI), ultra-sound wave imaging, and Neuropsychological battery tests [15]. But this is a way of seeing the phenotype itself or predicting the outcome rather than finding the source of the problem. Previous works have shown that genetic factors such as copy number variations (CNVs) - which are repeated segments of DNA with higher (duplications) or lower (deletions) abundance than the reference genome - are linked to both intellectual disabilities [16], motor impairments [17], and a statistically significant relationship with NDDs and psychiatric comorbidities [14], [18], [19].

It is well known that even though most of the human genome (>76%) can be transcribed into RNA products, only a small fraction (~3%) of it encodes for proteins [20]. These RNA molecules that do not follow the central dogma of molecular biology [21], are called non-coding RNAs (ncRNAs) and for many years were considered byproducts with low biological meaning. This perspective started to change, and scientists began to unravel the ncRNA mystery over the last decades with the help of advancements in sequencing methods and computational tools. Projects such as The Human Genome Project and The Encyclopedia of DNA Elements (ENCODE) [22], promoted the discovery of novel genes and shed light on functional elements encoded in the human genome, especially in non-coding areas, expanding our knowledge of their importance and their regulatory mechanisms. Studies and computational predictions suggest that even though NDDs and neuropsychiatric diseases are highly heterogenous, there are common enriched pathways and genetic factors between some of them [23], [24], [25]. As an example, ASD, Tourette syndrome (TS), and Schizophrenia share some genetic modifications that may lead to dysregulation of gene expression related to micro RNAs (miRNAs); a specific regulatory group of short non-coding RNA molecules [20].

According to their average size, ncRNAs can be categorized into two general groups: long non-coding (lncRNA) and small or short non-coding (sncRNA). LncRNAs extend to over 200 nucleotides (nt) and usually have a similar size to messenger RNAs which is more than 1000nt [26], while sncRNAs typically have a length below 200nt and they are separated into two groups based on their role in the cell; Housekeeping and Regulatory [20]. Except for other important functional roles in the cell, lncRNAs such as pseudogenes and circular RNAs can interact with some classes of sncRNAs, lowering their abundance in the free form through complementarity sequences. Housekeeping sncRNAs were discovered relatively early and are well studied, due to their abundance and their fundamental roles in the function of the cell. For example, their roles can be the amino acid transfer (tRNAs) at protein synthesis or being involved in RNA processing and splicing in the nucleus (snRNAs). Regulatory sncRNAs have drawn the attention of scientists only in the last decades when technological advancements allowed for it. Since then, their important role started to unravel and it was found that they actively interact and interfere with other molecules, regulate gene expression, and involve in important molecular pathways [27], [28], [29], [30], [31]. This control over the gene expression of the regulatory sncRNAs is important because, in many diseases dysregulation of sncRNAs sequentially causes dysregulation of functional elements that then lead to pathological phenotypes [32], [33]. Because of the great importance of these molecules, bioinformatics tools and dedicated databases have been developed in the last decades to explore their role as biomarkers and their potential in medicine. A visual taxonomy of the classification of RNA molecules can be seen in Figure 2.

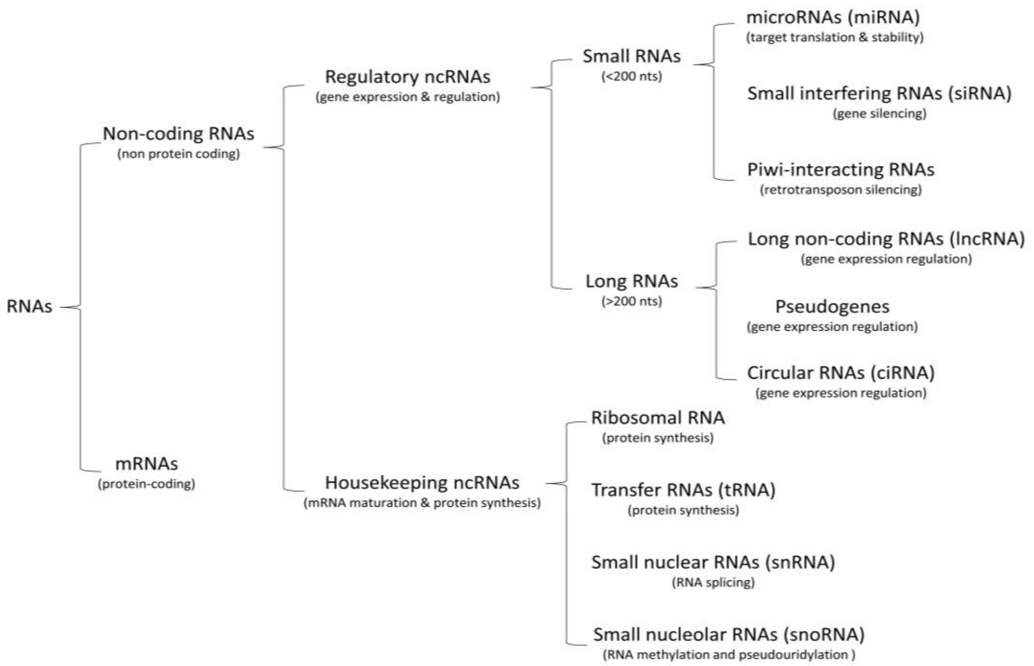


Figure 2. Classification of RNAs. The first split is based on the transcript product (coding/non-coding), followed by discrimination on the general role of the non-coding RNAs and further on the size of these molecules. This is a modified version of Figure 1 in the work of Gomes et al. (2020) [34]

After the systematic studies of sncRNAs, biologists clustered them by similarity and function with the most common ones being: microRNAs (miRNA) and small interfering RNA (siRNA) which regulate gene expression, small nuclear (snRNAs) that involve in RNA splicing, and piwi-interacting RNA (piRNA) that mainly interfere with transposable elements (or transposons) [20]. It is well established that sncRNAs hold a significant role also in many diseases in humans, and they can be used as biomarkers for diagnosis or prognosis, as drug targets, and as potential therapeutic methods [35]. Special attention has been given to miRNAs due to their high theoretical and experimental total number, the number of their interactions, and the role they have in both defending the homeostasis in the cell but also related to diseases like cancer when dysregulated [36].

Because of the numerous interactions of sncRNAs and other molecules in the manifestation of diseases, a common approach is to handle this complexity with the use of interaction networks. Since our understanding of the underlying mechanisms is still unclear for the majority of these diseases, studying individual relationships is not enough to unravel and understand the dynamic of these pathologies. Rather than this, a more holistic view is needed with the help of multi-layer networks integrating instances belonging to different levels of complexity and domains (RNAs, proteins, diseases, functions, etc.) [37]. In this context, computational modelling can help in reconciling the advancements in high throughput technologies with studies under the scope of systems and also explore the pathogenesis of diseases by understanding the molecular relations driving them, promoting treatments, drug discoveries, and precision medicine [38].

There are multiple tools nowadays that have been developed to predict the interactions of sncRNAs and especially miRNAs. Computational methods try to predict targets of these molecules [39], [40], pathways, and mechanisms involved in multiple diseases and disorders. Due to their interesting nature, ncRNAs have been systematically studied and there are multiple databases available where one can find experimental and computational information about them, based on their categorization. Most of these databases are open-access and publicly available. This makes the contained information accessible to everyone, helping scientists to build predictive models for diseases, discover potential biomarkers, and even design potential therapeutic targets.

Relevant studies were identified in PubMed, Scopus, ScienceDirect, and IEEE Xplore with no language restrictions. The first search from these databases was performed by the first author of this review and double-checked by the other corresponding authors. The following keywords were used: (sncRNAs OR miRNA OR siRNA OR piRNA OR RNAi), (neurodevelopmental comorbidities OR co-occurrence of neurodevelopmental disorders), non-coding RNA Databases, (bioinformatics tools AND target prediction of sncRNA). We included only papers from January 2000 up to August 2022. Older papers were excluded, with the exception of papers explaining concepts or statistical and mathematical techniques

In this article, we review some of the most widely used molecular biology-related databases for the characterization and functionality of sncRNAs, and the state-of-the-art for computational tools for the analysis of these RNA molecules in various comorbidities, such as NDDs observed in some preterm infants [41], [42]. The main objective is to comprehensively collect in one article information on the effectiveness and usability of biological databases and databanks, as well as some computational tools for different types of bioinformatics analysis that are considered or could be considered in the future for research in the field of neurodevelopmental disorders, also considering preterm infants.

2.2.3 Features of sncRNAs

Regulatory sncRNAs can derive usually from individual genes or introns of other genes, but it is known that by the procedure of alternative splicing they may also contain some exon sequences. The most studied categories are miRNAs and siRNAs which have been found to involve in many pathologies [43] and the developmental processes [44]. The biogenesis of miRNAs has five main steps: transcription into a primary (pri-miRNA) form of a stem-loop, cleavage into a shorter stem-loop precursor (pre-miRNA) known as hairpin, transportation of the hairpin out of the nucleus, and a second cleavage followed by the unstranding of the two counterparts to produce the mature miRNA. These molecules are typically 20-24nt long [30], [45] and bind to their targets through partial complementarity - not necessarily perfect- of their seed (nucleotides 2-8), and their mRNA target sequences in the 3' untranslated region (3'UTR) which are called miRNA response elements (MREs). This leads to degradation of the mRNA molecule, or the disruption of translation by preventing the binding of ribosomes on the mRNA. In both cases, translational inhibition results in the silencing of the target gene, a process also called RNA interference (RNAi). MREs can be found also in other types of RNAs like in lncRNAs and pseudogenes, which increases the number of targets for miRNAs since they are not strictly target-specific. This lower specificity gave rise to the idea of competitive endogenous RNAs (ceRNA). From the ceRNA point of view, miRNAs become the target of other competing molecules which based on their concentration, affinity, and the number of MREs regulate the abundance of available miRNAs resulting in an indirect regulation of their own expression.

Similarly, siRNAs regulate the gene expression of their target. These molecules by structure are almost identical to miRNAs, but with the difference of having very high specificity to their target since they usually have perfect complementarity of base pairing with them [46]. The function of siRNAs lies in the interference with gene expression by degrading their transcript targets which are far fewer targets than those of miRNAs. piRNAs on the other hand, follow a different biogenesis process which remains unclear to some extent, and also have different mechanisms of action. They are produced by a process related to the P-element induced wimpy testis or PIWI subfamily members, and recently they have been associated with cancer biology. The structure of piRNAs is single-stranded molecules of length range 26-31nt and they are known for epigenetic regulation through histone modification, but mostly for interfering with transposable elements or “genomic parasites”, protecting the genome of the host [47].

Biogenesis and the mechanism of interaction of regulatory sncRNAs are important since this knowledge is also implemented in the computational tools that predict their targets. Common features on which the majority of target prediction tools are based are the seed match, the conservation sequences, the free energy, and the site accessibility [48].

2.2.4 sncRNAs in neurological disorders

sncRNAs are crucial in the maintenance of homeostasis since they coordinate the expression of genes through the RNAi process. It has been reported by many studies that sncRNAs have a linked role in neurodegenerative diseases, various types of cancers [36], [47] where the expression of sncRNAs is heavily dysregulated due to mutations, and neurodevelopmental disorders [30], [49], [50]. Specifically, sncRNAs have been found to be part of enriched molecular pathways in numerous neurodevelopmental disorders and comorbidities like Rett syndrome, ASD, Down syndrome, and others [17], [20], [47]. In fact, a known, commonly altered pathway in neurodevelopmental and psychiatric disorders is the mTOR pathway [51], [52]. This may indicate that there are similar mechanisms between these disorders that lead to higher probabilities of comorbidity.

Another example is the dysregulation of specific miRNAs in neurodegenerative diseases. Specifically, miR-132 plays a crucial regulatory role in central nervous system disorders by influencing neuronal function, oxidative stress, neurodegeneration, immune responses, and neural stem cell behavior, highlighting its potential as a therapeutic target. mir-124, which according to Son et.al. (2024) is the most abundant miRNA in the brain, is a crucial element of neurogenesis and brain development and this is supported from other works that note the role of the molecule in cognitive decline when dysregulated [53], [54]. There are plenty of evidence and works that support the role of miRNAs in neuro-diseases and how they enhance their development, showcasing their potential as therapeutic target in drug discovery research [55], [56], [57].

In general, the role of sncRNAs in pathologies makes these molecules perfect biomarker candidates for early detection of diseases and mechanisms of diagnosis [58]. From what is known, although there are some sncRNAs that have been identified to have different expression levels and to be involved in the manifestation of NDDs, they do not show specific characteristics or significant features compared to other sncRNAs. However, mutations in the genes of the regulatory ncRNAs may be responsible for the occurrence of specific NDDs [59].

2.2.5 Conclusion

Since their discovery, the importance of non-coding transcripts has become clear, and they stop being considered as “Junk” DNA regions. With the advancements in technology allowing for the detection of these molecules and especially sncRNAs, a huge number of ncRNAs were discovered and got annotated. Even though some of their mechanisms of action have been decoded, their full functionality and role remains to be discovered, especially in complex diseases. Despite what is unknown, the focus on regulatory sncRNAs, led to significant improvements in our understanding of the molecular mechanisms driving certain diseases, and the prognosis of pathologies. Specific characteristics and features of sncRNAs involved in NDDs are not known, and thus, further studies are needed to understand and unravel the sources of these disorders. Computer science and Bioinformatics have a tremendous impact on systems biology, with the ever-improving development of tools helping scientists draw important conclusions from the massive amounts of available data. And this is why it is so important to have comprehensive, curated, open-source, and well-organized databases like the ones presented in this work in the following chapter related to tools and databanks (Chapter – 3.1.1).

2.3 Coding Molecular Biomarkers

Beyond non-coding molecular markers there are of course other types of biomarkers that do encode for proteins and have a functional role in pathologies. These can be again of various types such as single nucleotide polymorphism (SNPs), which can be studied through genome-wide association studies (GWAS) to identify genetic predispositions for conditions or phenotypes. As the name suggests, SNPs are point mutations on the DNA that depending on the position of the mutation can affect the expression of a gene, the effectiveness of the transcript and/or even the protein these genes encode for. Such mutations rarely cause a condition alone but rather impose a predisposition for it. While studying SNPs through GWAS, usually two populations are compared to finding not one single SNP but rather groups that occur more frequently in people with a certain disease or trait and synergistically lead to this predisposition. This approach has led to identifying genes related to Parkinson's, and Chron's diseases and Thalassemia among others. [60], [61], [62]. Additional to SNPs, other types of molecular biomarkers include loss of DNA segments known as deletions, insertion of nucleotide in the DNA, repetitions such as in the case of the fragile X syndrome where there is a known CGG triplet repeat [63], translocations where a segment of DNA moves from one chromosome to another and inversion of DNA within the gene region. All these alterations in DNA can occur within coding regions of the genome, potentially leading to altered molecular products. However, they do not necessarily induce a functional change, as their impact depends on their location and context. Nevertheless, they can influence molecular interactions, affect gene expression, and even alter the stability of gene products.

A huge field of study in biomarker discovery is proteomics, and as the term suggests, it revolves around proteins. Proteins have multiple roles and functions which are closely related to their structure. They are long amino acid chains folded in a specific way that determines their function and stability and can work either alone or very often in complexes. The aim of proteomics is to investigate the structure and function of proteins, to characterize the expression profiles for different conditions and life stages, and to study the interactions between them. Therefore, identifying and characterizing these protein alterations is central to discovering novel biomarkers.

Proteomics took a high throughput form with the development of mass spectrometry (MS) which enabled the identification of proteins and their quantification [64]. As an example, in clinical practice elevated levels of Prostate-Specific Antigen (PSA) are used among others as a biomarker for prostate cancer [65]. Unlike the genome where the complete set of genes in an organism is constant, the composition of the proteome is dynamic and in a continuous flux state given the environment, the development stage, and health conditions. Additionally, the proteins undergo extensive post-translational modifications which vastly increase their diversity.

Although the study of proteins can elucidate functional mechanisms of diseases, it presents some inherent challenges, such as the vast complexity, diversity and dynamic range of protein abundance in biological samples. Additionally, according to studies, the majority of proteins remain heavily understudied with their biological function unexplored [66]. Artificial intelligence has been proven extremely helpful in the field of structural analysis of proteins with the recent introduction of tools such as the Alpha-Fold2 that helped in the computational prediction of the 3D protein folding paving the way for drug targeting and designing [67].

In modern medicine, biomarkers are used routinely for prognosis, diagnosis, and personalized therapy selection. They have been a valuable tool in clinical practice for many years. With the emergence of bioinformatics and computational biology, molecular biomarkers have become even more powerful for clinical decision-making. One of the most well-established fields where molecular biomarkers are systematically applied is oncology. Since cancer is a highly heterogeneous disease, varying in type, stage, and molecular characteristics, specialized biomarkers help classify tumors, guide treatment choices, and predict patient outcomes. Some very well-known examples are the mutations on the breast cancer genes BRCA1/BRCA2 which produce repairing proteins and have a gene-suppressing role [68]. For instance, women with BRACA mutant genes are in a higher risk of breast or ovarian cancer. Being informed of this can lead them to routine examinations and thus early detection of a possible neoplasm, potentially improving their heath outcome.

Except for cancers which are the diseases with the highest number of biomarkers identified due to their mutation rate and hijacking mechanisms to bypass the control of the immune system, there are examples of significant biomarkers in neurology. As an example, the Rett syndrome, which is a rare postnatal progressive neurodevelopmental disorder affecting mostly females. It is characterized by developmental regression of spoken language, repetitive hand movements, impaired ambulation, and some autistic features. These can lead to the manifestation of severe impairments and comorbidities such as anxiety, breathing difficulties and constipation [69]. It has been shown that pathology is often linked to variations in the *MECP2* gene, with the mechanism of progression being through the MeCP2 protein which is a regulation protein binding on methylated DNA highly expressed in the brain [70]. There have been found many mutations of the gene that lead to Rett syndrome and Encephalopathy with different genotypes of the MECP2 gene effecting differently the patient's phenotype [71].

From examples like these, it becomes clear that biomarkers are essential in modern medicine either as a form of early diagnosis or even indicators of predisposition, or as indicators of condition that needs immediate treatment. Molecular biomarkers have the potential to show these and help in the tailoring of precision medicine treatments and drug therapies based on the genotype of the patient with oncology being in the frontier of this effort [72].

Discussion

In this chapter we explored briefly some clinical biomarkers used in everyday practice, as well as the nature and mechanism of molecular markers. These measurable indicators span in all scales from small molecules concentrated in the body, to visual features in clinical images. Their identification follows a general sequence of steps which starts with the well-designed clinical question, followed by the selection of the appropriate population (cases and controls), on which samples are collected and analyzed, and a careful observation of characteristics is applied. The characteristics are analyzed computationally, get interpreted clinically and then, evaluated on external validation data to ensure reproducibility and robustness (Figure 3). Regardless of the indicator's nature, those main steps are the foundation of biomarker discovery.

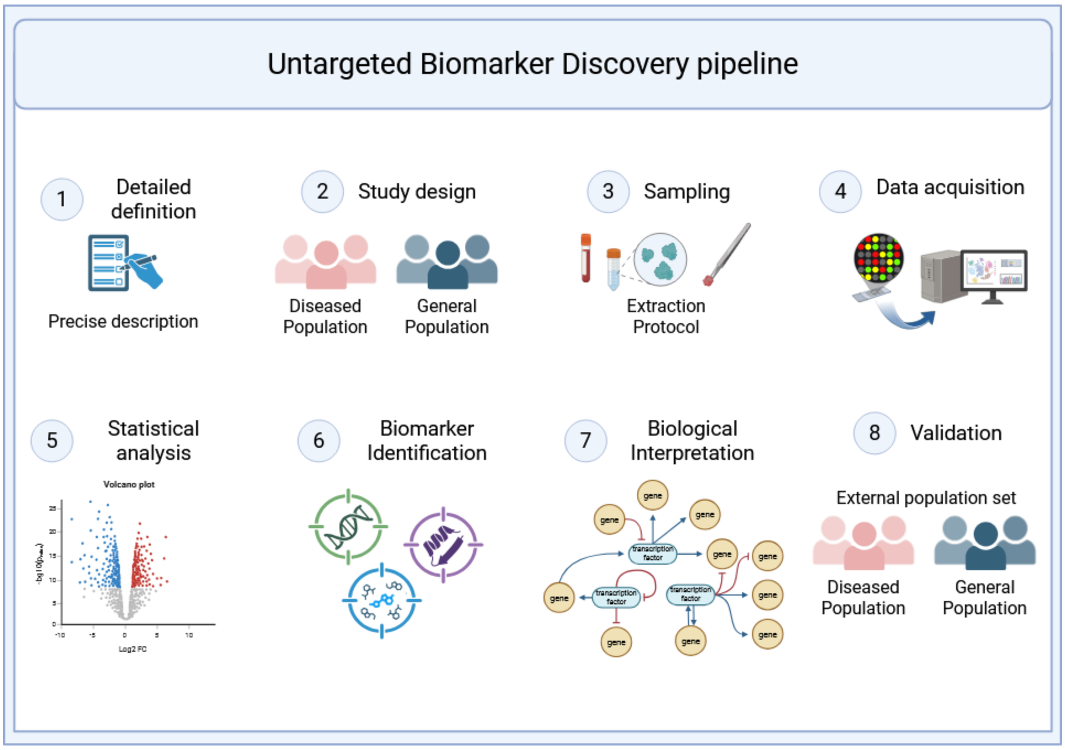


Figure 3. Fundamental steps of an untargeted biomarker discovery pipeline from the design of the experiment (step 1) to the validation phase (step 8).

Focusing on the computational aspect, the analysis of data differs for the omics level of analysis since the acquisition and laboratory methods change accordingly. Although, the rationality governing the basic steps is the same for all omics layers. As before, the design of the model is necessary to be fully understood and thoroughly explained. This step is very important for choosing an appropriate sample size, selecting the best method, and collecting the necessary meta-data for future analysis.

Data processing and quality control are necessary independent of the omics layer and method used. This step is important to remove low-quality measurements; eliminate technical noise; normalize the raw data; perform outlier detection and removal. Regarding the statistical analysis of the data, the goal is to identify dysregulated or differentially expressed molecules between the control population and the diseased one. This identification aims not only in the difference in distributions in terms of mean values and variance, but also in the quantitative change (i.e., fold change). This dual comparison allows for identifying statistical differences of distributions by using the p-value (or the adjusted p-value for multiple testing) and physical differences by using the fold change between groups since none of these alone is sufficient to describe real differences. A comparative analysis can be done with ML methods as well that are less affected by distributional differences in the dataset and exploit non-linear relations in the data (Figure 4).

The last steps of the computational analysis for biomarker discovery are related to the interpretation of the results and their evaluation. Thus, after the differential expression analysis (or the ML feature importance) a secondary functional annotation and pathway analysis helps in providing an explanation of the findings by comparing them with knowledge obtained from literature. External validation is needed for evaluating that the identified biomarkers are able to indicate a difference between groups which then must be validated in the laboratory in a targeted experiment.

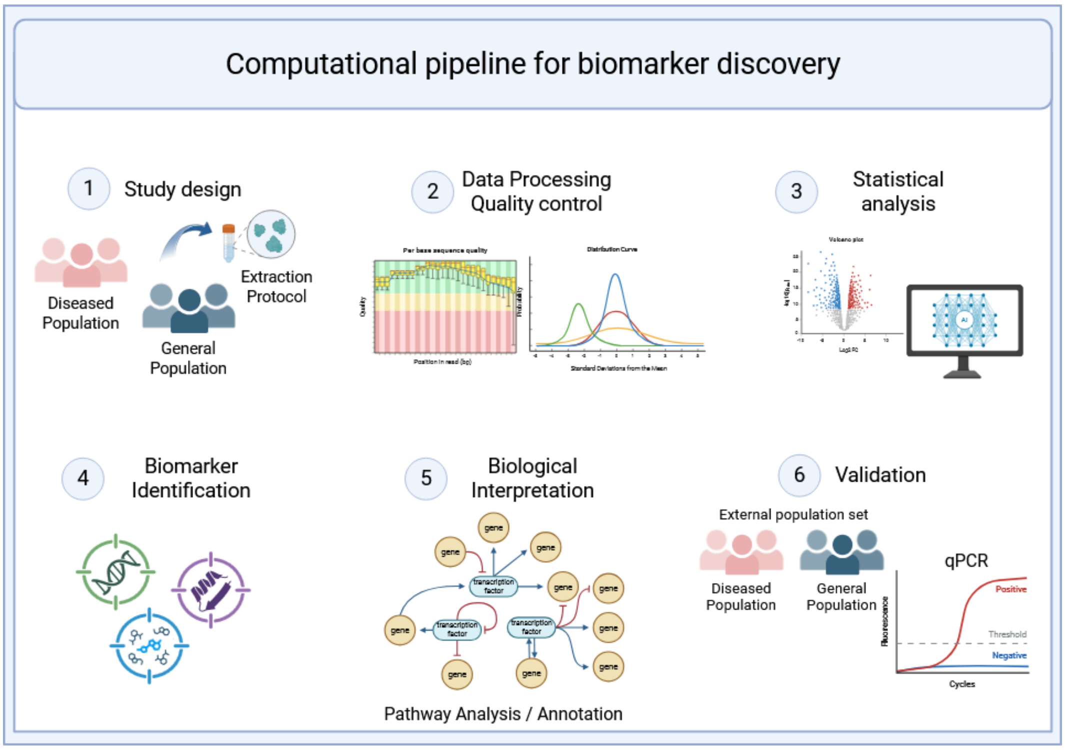


Figure 4. Generalized pipeline for the computational process of biomarker discovery.

Even when all these steps are rigorously followed, the clinical translation of biomarkers remains a multi-factor challenge. As described in the pipeline for biomarkers' discovery (Figure 3), robust statistical validation is critical to ensure that observed molecular signals are not artifacts but reflect true biological dysregulation, to avoid noise signals and ensure clinical relevance. One of the major barriers is the need for adequately powered sample cohorts, which is especially difficult to achieve for rare diseases within single-center studies. Overcoming this requires multi-center collaborations using harmonized protocols for sample collection, processing, and analysis. Such coordination across diverse sites introduces challenges related to procedural alignment, recruitment logistics, and increased costs.

Reproducibility and standardization come as a consequence of the previous challenges, making essential the need for strict and clear protocols at each step of the studies. Considering the effectiveness and the cost to apply a biomarker in everyday clinical life, it is crucial to develop minimally invasive and highly sensitive assays that can detect and quantitate the biomarker with precision. Developing such assays demands rigorous analytical validation and subsequent regulatory approval, typically from agencies such as the Food and Drug Administration (FDA) in the United States or the European Medicines Agency (EMA).

When considering ncRNAs, these challenges are further amplified. Although ncRNAs have shown great promise in areas such as oncology and immunology, their vast diversity, cell-type-specific expression, and context-dependent roles hinder straightforward clinical application. Many ncRNAs still lack known functions or validated targets, and novel species are continuously being identified. Moreover, their biochemical heterogeneity prevents the establishment of a universal detection platform; distinct subclasses (e.g., miRNAs, lncRNAs, circRNAs) often require tailored analytical approaches. These technical hurdles, coupled with the need for validation across large, independent cohorts, complicate their path to clinical translation, although highlighting their high potential in both diagnostics and therapeutics.

Chapter 3

Tools for Biomarker discovery of Genetic Markers in Neurodevelopmental Impairments

Identifying biomarkers is not trivial and depends on a lot of factors. In this chapter we will discuss some of the techniques used for identifying these markers through laboratory protocols and methods, bioinformatics statistical tools, and the use of artificial intelligence. The present chapter gently introduces parts that will be covered later in this thesis about the first challenge's (C1) objectives O2 and O3.

In more detail, we explore the acquisition methods of transcriptomics data, and how genetic information is extracted from them leveraging statistical tools to identify key components. We also examine how this information can be utilized to train machine learning (ML) algorithms. Furthermore, we investigate computational tools and databases used to identify and store information about ncRNA targets, which play a crucial role in gene expression regulation and network formation. Given the vast scope of bioinformatics, this thesis primarily focuses on transcriptomics and genomics. This choice aligns with the technical nature of the study, allowing for more in-depth exploration within other types of biomarkers and computational methods.

3.1 Classical Approaches with Statistics in Bioinformatics

We begin by discussing classical bioinformatics approaches that rely on statistical methods for analysis. This is because for a very long time and even nowadays bioinformatics relies on statistics providing a foundation for many computational approaches. Classical statistical techniques have been instrumental in extracting meaningful patterns from complex datasets, allowing researchers to quantify variability, assess relationships, make predictions and find candidate biomarkers for diseases. These methods offer a structured way to interpret experimental results while maintaining a degree of abstraction from wet-lab procedures. They take a step back from direct experimental procedures while still considering essential study details. Understanding the nature and peculiarities of the data remains crucial for a meaningful analysis.

Early bioinformatics approaches were largely built on statistical frameworks, including hypothesis testing, regression models, Bayesian inference, and clustering techniques. These methods have been widely used for tasks such as gene expression analysis, sequence alignment, and evolutionary studies. While modern ML algorithms have gained prominence, classical statistical methods remain essential due to their interpretability, robustness, and well-established theoretical foundations.

Statistical methods that are used in the classical approaches of bioinformatics revolve around comparison between groups (i.e., disease vs control) and aim to find important differences between them. A very important method for the comparison of two populations is the t-test where two groups' means are compared. It operates under the null hypothesis, which posits that there is no significant difference between the group. The violation of the null hypothesis suggests a significant difference between the means of the populations. The likelihood that this difference is due to chance is quantified by the p-value, which indicates the probability of observing such a difference if the null hypothesis were true. In bioinformatics the t-test is often applied to compare gene expression levels, treatment effects, or other biological measurements between two conditions or experimental groups.

A method for comparison of gene expression levels across multiple experimental groups is the *analysis of variance* (ANOVA), making it particularly valuable in bioinformatics studies involving different biological conditions. Unlike t-tests, which only compare two groups at a time, ANOVA determines whether there is a statistically significant difference in mean expression levels across three or more groups. For example, in NDDs, ANOVA can be used to analyze differential gene expression across varying severity stages (mild, moderate, and severe cases), helping in the identification of genes that may be progressively dysregulated. The method assumes normality and homogeneity of variances, which are important considerations in genomic datasets. A key extension is the two-way ANOVA, which allows testing for the interaction between two factors (e.g., genetic mutations and environmental exposure on gene expression). When ANOVA detects a significant effect, post-hoc tests (e.g., Tukey's HSD) are used to pinpoint specific group differences. Despite its utility, ANOVA has limitations, including its sensitivity to outliers and assumptions about normal data distribution, which often do not hold in transcriptomic data, necessitating alternative approaches like non-parametric tests.

Non-parametric tests such as the Mann-Whitney U test and Kruskal-Wallis test are critical in bioinformatics when analyzing datasets that do not meet the assumptions of normality and homogeneity of variance, which is often the case in high-throughput sequencing data. The Mann-Whitney U test is a rank-based alternative to the t-test, designed for comparing two independent groups when data distributions are skewed or contain outliers. It is frequently used in gene expression studies to assess whether a specific gene is differentially expressed between two conditions (e.g., control vs. disease) [73], [74]. On the other hand, the Kruskal-Wallis test extends this approach to multiple groups, serving as the non-parametric equivalent of ANOVA. It is particularly useful in studies comparing gene expression levels across multiple stages of a disease or different treatment conditions [75], [76]. Both tests rely on ranking data rather than assuming normal distribution, making them more robust in handling non-Gaussian distributed transcriptomic data. However, these methods do not provide information on effect size or directionality and often require post-hoc pairwise comparisons to determine specific group differences. Despite these limitations, non-parametric tests remain a powerful tool for exploratory data analysis in bioinformatics, particularly in cases where classical parametric assumptions do not hold.

Very often due to the high dimensionality of the datasets in bioinformatics where there is a very high ratio of features (i.e., genes) over samples, the results of statistical tests must be adjusted for multiple testing. This means that when a single test is conducted, the rejection of null hypothesis by chance is depicted by the p-value, but in the case of bioinformatics, thousands of tests are conducted simultaneously, which is expected to introduce many false positives in the results. For reference, for a p-value threshold of 0.05, - which is an arbitrary threshold of significance - for every 1000 tests (i.e., transcripts, genes), it is expected to have 50 false positive hits in the results. To reduce this effect the results must be adjusted. The most widely used method is the Benjamini-Hochberg (BH) procedure for False Discovery Rate (FDR) control, which is ideal for large datasets like RNA-seq and microarrays. It ranks *p*-values and adjusts them to balance false positives and power [77]. In contrast, the Bonferroni correction controls the family-wise error rate (FWER) by dividing the significance threshold by the number of tests, effectively reducing false positives but at the cost of increased false negatives [78]. There are other methods that are correcting for multiple testing such as the Holm-Bonferroni method which is less conservative alternative to Bonferroni, offering better power while maintaining FWER control and Storey’s q-value which estimates the proportion of true null hypotheses and adjusts p-values adaptively [79]. In fact, Storey's q-value is often used in reports of computational methods next to FDR. Nevertheless, due to its greater sensitivity in large datasets, BH is the preferred method in applications like RNA-seq and microarray.

Beginning with genomic analysis, we previously discussed examples of gene mutations that can directly cause diseases, such as thalassemia, which results from mutations in genes encoding the haemoglobin protein complex. In other cases, mutations increase an individual's predisposition to developing certain conditions, such as pathogenic variants in the BRCA1 and BRCA2 genes, which elevate the risk of breast and ovarian cancer.

As mentioned in an earlier section, Genome-Wide Association Studies (GWAS) identify genomic loci that statistically correlate with disease susceptibility. This is achieved by detecting genetic variations (such as single nucleotide polymorphisms, or SNPs) that occur more frequently in individuals with a particular condition compared to a control group (Figure 5) [80]. By analysing the genomic proximity of these variations, researchers can pinpoint associated regions and potential candidate genes. However, GWAS does not necessarily identify causative variants, but rather, it highlights loci associated with disease risk due to linkage disequilibrium.

In cases of monogenic diseases like thalassemia, where a single gene mutation is responsible, direct mutation screening is the preferred approach for diagnosis rather than GWAS. However, for complex diseases like cancer, cardiovascular disorders, and diabetes, GWAS helps identify genetic predispositions by associating specific SNPs with increased risk. Once validated, these genetic markers can be used in personalized medicine for risk assessment and early intervention strategies.

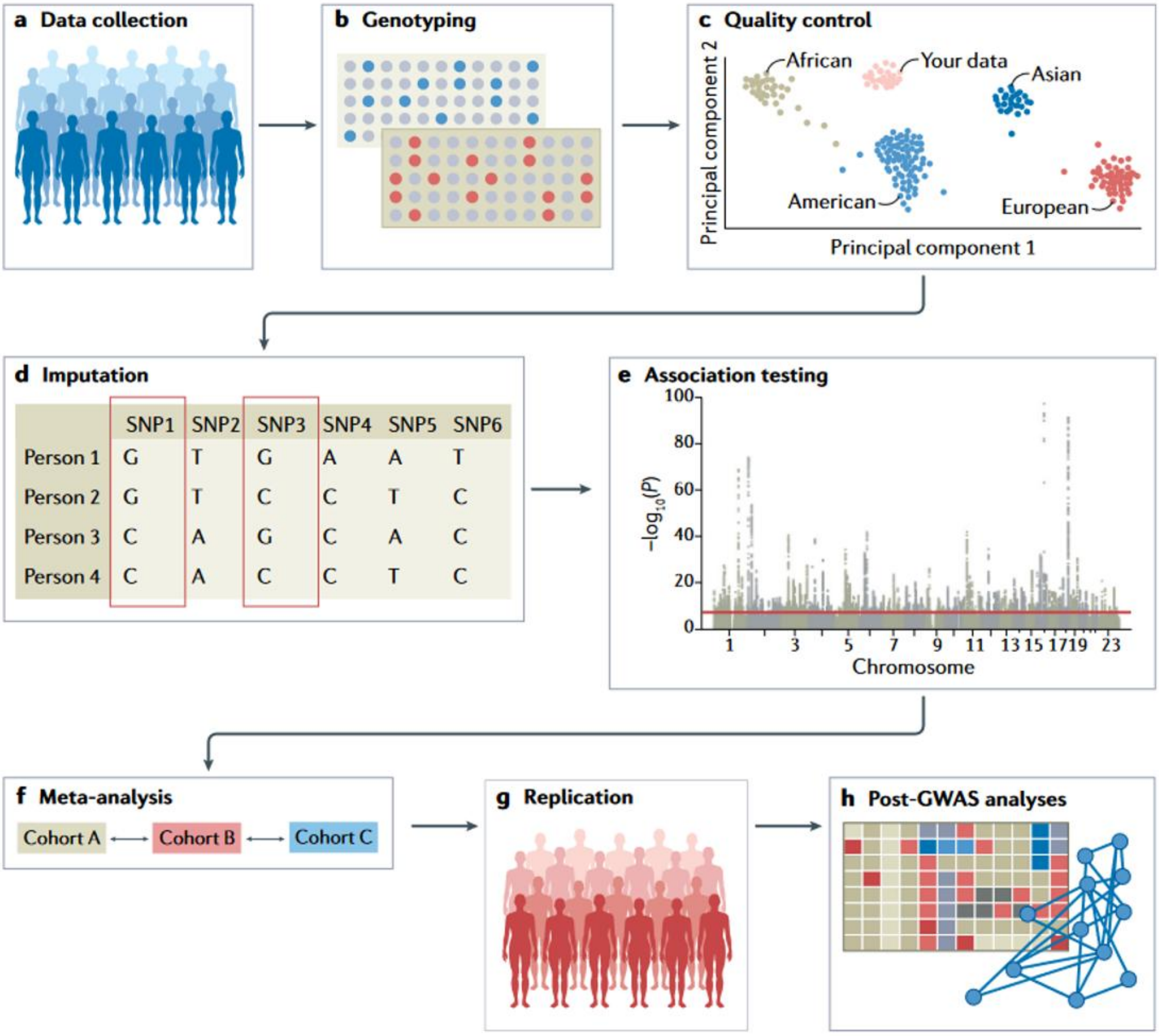


Figure 5. Illustration of an overview for the steps of genome-wide association study from the collection of data (a) to the identification of frequently occurred variations (d) and their statistical analysis (e), to the validation phase (g, h). Source: E. Uffelmann, et.al., 2021, Genome-wide association studies [80].

As with almost all approaches in any field, there are specific steps that need to be taking care of before proceeding to the analysis of data. These are the quality control tests that are applied on the data and briefly include the overall distributions examination, signal integrity and quality, and outlier identification. For solving potential problems different methods can be used such as removing outliers, normalizing the data within and among individuals, and impute missing values if there are any. All those solutions have alternative ways to be applied based on the peculiarities of each dataset and according to the end goal of the study. Additionally, there are tools developed for taking care of different parts of the quality control and the preprocessing that can be used to tackle one issue, or they might be in the form of pipelines where the results of one tool become the input of the following to automate the preprocessing procedure.

In the context of NDDs, GWAS have significantly advanced our understanding of NDDs by identifying genetic variants associated with these complex conditions, and the integration of large-scale genomic data has revealed various genetic factors, including copy number variations (CNVs) and other rare variants that contribute to the etiology of NDDs [81]. Additionally, there are studies suggesting an overlap in the risk factors between differently characterized NDDs and psychiatric disorders [82], [82], [83].

Moving to transcriptomics, which is more relevant to the present thesis since there are tools developed for transcriptomics datasets, there are different methods that were developed over the years for their analysis. But first, it is important to explain what transcriptomics is and describe the data we are working with. By definition, transcriptomics is the study of the transcriptome· the whole set of transcripts (RNA molecules) that an organism -or a cell- expresses. This includes both the coding (translatable to protein) and the non-coding molecules. The transcriptome is relatively dynamic compared to the genome and is affected by cell-cycle, tissue, age and environmental factors [84].

In transcriptomics there are different high throughput technologies that have been developed over the years for identifying transcripts starting from 1991 with only a handful of published mRNA sequences from the human brain [85]. Other technologies existed before this time from the beginning of the 70's but there were studies of single transcripts or low throughput in the 80's. Since then, the dominant methods that are still used until this day are microarrays and RNA-seq which developed in the mid 90's and early 2000's [85].

Microarrays

Microarrays' base principle is the hybridization of transcripts to predefined complementary probes. These probes are arranged in a grid of small dots on a solid surface (Figure 6). Initially, extracted RNA is reversed transcribed into complementary DNA (cDNA) and labeled with fluorescent dyes. The labeled sample is then applied to the microarray for hybridization. At the end of the process, transcripts that have not bound to their complementary probes are washed away, while those that have hybridized remain attached. Due to the fluorescent labeling, bound probes emit light on a specific wavelength dictated by the dye, which is detected using a scanner. The scanner measures fluorescence intensity at each spot, generating a table of expression values. This technology allowed the assessment of thousands of transcripts simultaneously and created the need for complex methods for the analysis of the data, which includes data cleaning (background noise removal), probe summarization, normalization, and other types of preprocessing before they can be analyzed and interpreted.

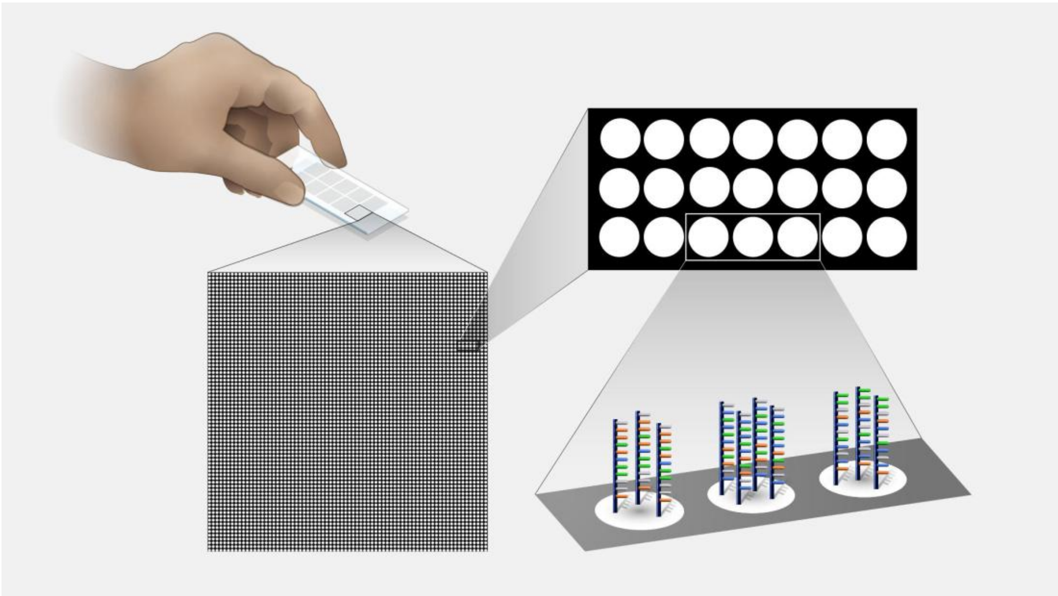


Figure 6. Illustration of a microarray structure. Each array has multiple wells where different samples are applied. In every well there is an identical grid of dots where the probes are placed.

Microarrays can be classified based on their labeling strategy into one-channel (single-color) and two-channel (dual-color) approaches. In one-channel microarrays, each RNA sample is independently labeled with a single fluorescent dye and hybridized to a separate microarray, producing absolute gene expression values. This method is well-suited for large-scale studies where data consistency across multiple samples is essential. In contrast, two-channel microarrays use two different fluorescent dyes (e.g., Cy3 and Cy5) to label two RNA samples, which are then co-hybridized onto the same microarray. The fluorescence intensity ratio between the two dyes provides a relative measure of gene expression differences between conditions. While two-channel microarrays are particularly useful for direct comparative analyses (e.g., treated vs. control), they require additional normalization due to potential dye biases. Both approaches have been widely used in transcriptomics, with the choice depending on the study's design and the type of expression data required.

Microarrays have evolved significantly over time, with various manufacturers developing distinct designs to optimize gene expression analysis. Major companies, such as Affymetrix (now part of Thermo Fisher Scientific), Agilent Technologies, and Illumina, have introduced different platforms, each with unique probe configurations and detection methods. Affymetrix arrays use short oligonucleotide probes synthesized in situ, requiring complex computational processing for analysis. Agilent microarrays, on the other hand, employ longer probes printed directly onto glass slides, offering high hybridization specificity. Illumina introduced bead-based microarrays, where probes are attached to microscopic silica beads, allowing for high-throughput transcriptomic analysis. Additionally, custom-designed microarrays have emerged, enabling researchers to tailor probe content for specific applications. However, these differences in microarray design, probe lengths, and labeling protocols introduce additional challenges for cross-study comparisons and data harmonization. Variability in preprocessing methods, probe annotations, and hybridization conditions can complicate the integration of datasets across different platforms, further adding to the inherent technical variability of microarray experiments. Standardization efforts and computational normalization techniques are essential to ensure reproducibility and comparability across studies.

RNA-seq

Next generation sequencing (NGS) is a broad class of high-throughput sequencing technologies that allow researchers to sequence DNA or RNA rapidly and in parallel. RNA-seq is a specific application of NGS used to sequence and analyze RNA in a sample and provides insights into gene expression, alternative splicing, and other RNA-related phenomena. Unlike microarrays that the probes are predefined, RNA-seq allows for novel transcript identification through sequencing of fragments of RNA, and base on the study there are different types of RNA-seq to choose from that include exome RNA sequencing, small RNA sequencing, single cell sequencing and others.

The RNA-seq method relies on the sequencing by synthesis to rapidly and efficiently sequence RNA molecules in a sample. During this process, RNA is first converted into cDNA through reverse transcription and gets fragmented into smaller pieces and adapters that are ligated to the ends of these fragments. The fragments are loaded onto a sequencing platform, where sequencing by synthesis occurs. This involves incorporating nucleotides one by one and detecting the emitted signal as each base is added, generating a sequence of base calls that represent the original RNA.

The result of RNA-seq is a series of reads - short sequences that correspond to fragments of the original RNA. These reads are typically a few dozen to a few hundred base pairs long, depending on the sequencing platform. The depth of sequencing refers to the number of reads obtained per sample. A higher sequencing depth increases the likelihood of detecting rare transcripts or lowly expressed genes, but comes at a higher cost. The number of reads mapped to a specific gene is known as the read count, which is used to quantify gene expression levels. Higher counts indicate higher expression levels of the corresponding genes in the sample. The process is visualized in Figure 7 below.

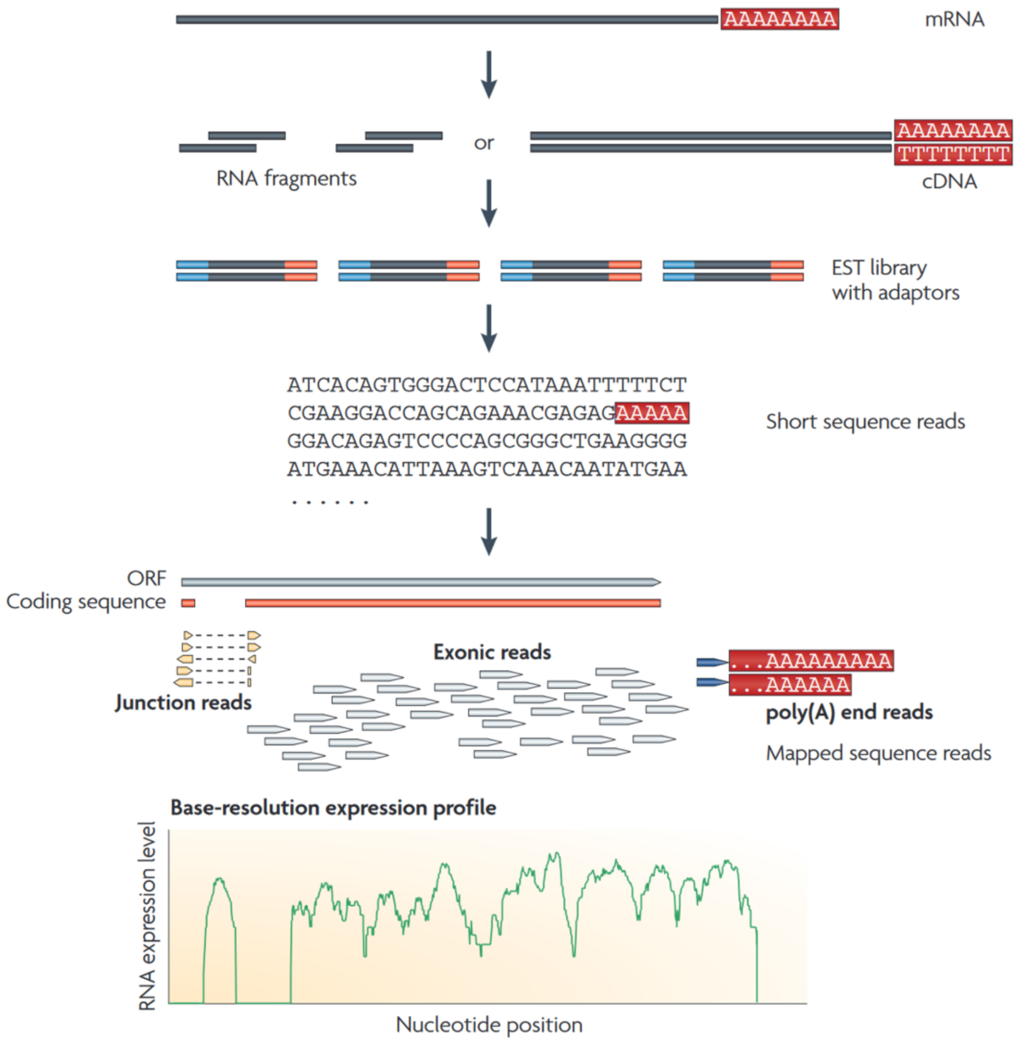


Figure 7. Illustration of a typical RNA-sequensing process. The isolated RNA is converted into cDNA and gets fragmented before entering the sequenser where they are sequensed in short reads. The reads are alligned on the reference genome and computationally are counted and normalized to produce the expression profile of the sample. Source: Zhong Wang, et.al., 2009 [86]

Once the reads are generated, they are mapped to a reference genome or transcriptome to determine their origin. This mapping step allows for the identification of which genes are expressed and how much of each gene is transcribed. Several tools, such as HISAT2, STAR, and Bowtie2, are commonly used for efficient read alignment. After mapping, normalization methods are applied to account for differences in sequencing depth, gene length, and other technical biases. Common normalization techniques include TPM (Transcripts Per Million), FPKM (Fragments Per Kilobase of transcript per Million reads), and

RPKM (Reads Per Kilobase of transcript per Million reads), which allow for meaningful comparisons between samples or conditions.

After introducing the main experimental tools—microarrays and NGS—and their principles, it is useful to explore computational methods used in bioinformatics to analyze the resulting data and gain insights into the mechanisms underlying biological processes and conditions. Before discussing the primary methods for biomarker identification, it is important to highlight techniques for dimensionality reduction and data projection, which are essential for exploratory analysis (e.g., outlier detection, global structure visualization, and clustering).

One widely used method is Principal Component Analysis (PCA), which linearly transforms data into a new coordinate system, where the principal components represent the directions capturing the greatest variance [87]. In addition to PCA, non-linear dimensionality reduction methods help preserve local relationships between data points while mapping high-dimensional data into a lower-dimensional space. One such approach is t-distributed Stochastic Neighbor Embedding (t-SNE), which optimizes the positioning of similar objects to remain close together while ensuring that dissimilar objects are mapped farther apart with high probability [88].

These dimensionality reduction and exploratory techniques provide an initial understanding of the structure and variability within the data, helping to detect patterns, remove confounding factors, and refine sample grouping. With this foundation, more targeted computational methods, such as differential expression analysis (DEA), enrichment analysis, and network-based approaches, can be applied to identify potential biomarkers and uncover biological mechanisms.

Differential Expression Analysis

Differential expression (DE) analysis is a method used to identify genes whose expression levels differ significantly between experimental conditions or populations. Different statistical methods are applied, with one of the most common being the t-test, which compares expression levels between two groups for each gene. However, the t-test may not be suitable for RNA-seq data due to its discrete and overdispersed nature. For both microarrays and RNA-seq, more advanced approaches like linear models are frequently used which can account for other cofounders in the data which as an example can be the sex, the ethnicity or the age of the populations in study. These methods report statistical metrics that help in downstream analysis, including p-values and adjusted p-values for each gene, indicating whether its expression is significantly altered between the two conditions. Additionally, a log2 fold-change (log2FC) value is often used to measure the magnitude of gene expression changes, showing whether a gene is upregulated or downregulated. The log2 fold-change is particularly useful because, in microarray analysis, raw intensity values are typically transformed to a logarithmic scale (base 2) to handle the large variation in raw data, while in RNA-seq this transformation is taking place usually after normalization to account for sequencing depth and other technical factors.

From the DE analysis, we usually end up with a list of genes which are up or down regulated between the populations of study and the level of confidence that accompanies this change in expression. The list of these differentially expressed genes can be used in pathway analysis and in network-based analysis. The pathway analysis helps in the interpretation of the biological significance of the findings by linking the genes to known biological processes. One type of pathway analysis is the Gene Set Expression Analysis (GSEA) which identify biological pathways enriched in the given list by ranking them based on their logFold change and checking if biological terms are associated with the highly upregulated or highly downregulated genes. Another method is the Over-representation analysis (ORA) that identifies whether a predefined gene set is significantly represented in different conditions. The procedure here is simpler, where the method tests if specific – predefined – gene sets (pathways) are over-represented among the DE genes identified in the dataset. Overall, both GSEA and ORA are effective approaches and help in the interpretation of the DE genes and identify key biological processes or pathways that are impacted by the conditions in the dataset.

Network analysis

Network analysis is a powerful computational approach for identifying key biological features, such as genes, transcripts, and proteins, that play significant roles in disease processes. These analyses come in multiple forms, depending on the type of biological data used and the nature of the relationships between the nodes. One of the most widely used approaches is undirected gene co-expression networks, where genes with similar expression patterns are grouped into modules based on correlation measures. Within these networks, highly interconnected genes may act as hub genes or biomarkers, as their co-expression suggests functional relevance. For instance, in cancer research, genes within oncogenic signalling pathways can be identified as potential diagnostic or prognostic biomarkers using this approach. Weighted Gene Co-expression Network Analysis (WGCNA) is a widely used method for constructing such networks.

Another key category of biological networks is the protein-protein interaction (PPI) network, where nodes represent proteins and edges represent physical or functional interactions between them. PPI networks help reveal central regulatory proteins that influence disease mechanisms. For example, network centrality measures (e.g., degree, betweenness, closeness) can be used to identify proteins that act as drug targets in diseases such as cancer or neurodegenerative disorders. PPI data is often sourced from databases like STRING, BioGRID, and IntAct.

Regulatory networks offer an additional layer of biological insight by mapping interactions between molecules of the same or different types (e.g., miRNA-mRNA, transcription factors-DNA). These networks are directed, meaning that edges indicate the regulatory effect (activation, inhibition, repression, etc.). Non-coding RNAs, such as miRNAs and lncRNAs play crucial roles in gene regulation at the post-transcriptional level. By integrating RNA-seq data with miRNA-target interaction networks, researchers can uncover disease-relevant RNA biomarkers that influence key biological pathways, aiding in the identification of potential therapeutic targets.

Network-based approaches offer significant advantages in biomarker discovery because they consider interactions between molecules rather than analyzing genes in isolation. By integrating pathway-based statistical models with network clustering techniques, researchers can move beyond traditional differential expression analysis and identify multi-gene biomarker signatures that are more robust, biologically relevant, and clinically actionable.

3.1.1 Computational tools for ncRNAs

The role of ncRNAs as key regulatory molecules in various biological processes has been proven in recent years, including gene expression, chromatin remodeling, and signal transduction. Their dysregulation has been linked to numerous diseases, such as cancer, neurodegenerative disorders, and metabolic syndromes, making them valuable candidates for biomarker discovery. Given their functional significance, understanding the roles of small non-coding RNAs (sncRNAs) such as microRNAs (miRNAs), long non-coding RNAs (lncRNAs), and piwi-interacting RNAs (piRNAs) require specialized computational tools and databases.

A wide range of bioinformatics tools and databanks have been developed to facilitate ncRNA identification, annotation, and functional analysis. These tools enable researchers to predict ncRNA targets, study their interactions with mRNAs, and integrate expression data to uncover their regulatory roles in disease pathways. Furthermore, network-based and machine learning approaches are increasingly applied to analyze ncRNA-mediated interactions, improving biomarker identification for clinical applications.

This section explores key computational tools designed for the study of sncRNAs, focusing on identification and classification, target prediction, functional annotation, performing network analysis, and visualizing. The following content has been partially extracted from the conference article published:

A review on computational tools for analytical visualisation and molecular interactions of sncRNAs: prospects in NDDs

1st International Workshop in Neurodevelopmental Impairments in Preterm Children - Computational Advancements (DETERMINED 2022).

CEUR. Volume 3363, November 2023. doi: https://doi.org/10.5281/zenodo.8009881

3.1.1.1 Introduction

The complexity of sncRNA interactions in disease manifestation necessitates the use of computational approaches, particularly interaction networks. Studying individual relationships is insufficient to unravel the dynamics of these pathologies. Instead, a holistic view is required, integrating multi-layer networks that encompass various levels of complexity and domains (e.g., RNAs, proteins, diseases, functions) [36]. Computational modelling bridges the gap between high-throughput technologies and systems-level studies, enabling the exploration of disease pathogenesis, molecular relationships, and the development of treatments, drug discoveries, and precision medicine [38].

Numerous computational tools have been developed to predict sncRNA interactions, particularly for miRNAs. These tools aim to predict targets, pathways, and mechanisms involved in various diseases and disorders [39], [40]. Additionally, several open-access databases provide experimental and computational information on ncRNAs, facilitating the development of predictive models, biomarker discovery, and therapeutic target design.

In this section, we review some of the most widely used molecular biology-related databases for the characterization and functionality of sncRNAs, and the state-of-the-art of computational tools for the analysis of these RNA molecules in various comorbidities, such as NDDs observed in some preterm infants [41], [42]. The main objective is to comprehensively collect in one article information on the effectiveness and usability of biological databases and databanks, as well as some computational tools for different types of bioinformatics analysis that are considered or could be considered in the future for research in the field of neurodevelopmental disorders, also considering preterm infants.

3.1.1.2 Features of sncRNAs

The biogenesis and interaction mechanisms of regulatory sncRNAs are not only biologically significant but also form the basis for computational tools that predict their targets. Common features used by these tools include:

- **Seed match**: The complementarity between the miRNA seed region (nucleotides 2-8) and the target sequence.
- **Conservation sequences**: Evolutionary conservation of target sites across species.
- **Free energy**: The stability of the miRNA-mRNA interaction.
- **Site accessibility**: The physical accessibility of the target site on the mRNA [48].

These features are critical for the development of computational algorithms that predict sncRNA targets and interactions, enabling researchers to explore their roles in disease mechanisms and therapeutic applications. Different tools focus on various features either empirically guided by literature, or using ML and statistics are trying to predict the sncRNA targets.

3.1.1.3 Bioinformatics Databases

With the emergence of clinical and biological databases, coupled with advancements in sequencing technologies (e.g. micro-arrays, next generation sequencing (NGS), tiling arrays, etc.), new opportunities arose for computational biology and the exploration of microscopic and macroscopic processes. Methods and tools started developing to tackle the challenges of massive amounts of data and the complication of biological systems. Results of multiple focused experiments started gathering and the findings were made easily accessible for further analysis. Additionally, databases started implementing online tools for processing information, predicting, and storing multiple-level entities because of the interoperability between different databases [89]. The ever-increasing number of databases along with the availability of data due to the new technologies of high throughput techniques led to the development of new tools, methods, and pipelines for handling the amount of available data and the extraction of new knowledge.

Biological Databases

The need for databases comes from the scattered information in literature. Having a comprehensive dataset helps researchers -especially in the clinical industry- to use the knowledge obtained from multiple and different experiments easily and find associations between instances leading to a better understanding of some conditions and processes. Biological databases can be manually or automatically curated, which means that they are constantly updated with new knowledge coming either from experiments or computational predictions. Additional to databases of linked information, there are databanks where raw data from experiments are stored. This, except for being a source of information for the databases, allows for meta-analysis of the data and merging of experiments to increase the amount of data in individual studies.

In the last decades, many efforts have been made to summarize the information about ncRNAs, as it is a game-changer in the study of cellular processes and gene regulation. Two commonly used sources of raw data are the Gene Expression Omnibus (GEO) [90] and the Sequence Read Archive (SRA) [91]. Both are from the National Institutes of Health (NIH), a part of the United States Department of Health and Human Services. In GEO, one can find collections of genomic data grouped by studies for multiple instances, and information about the protocols followed in the conducted studies. Genome browsers such as Ensembl, UCSC, and NCBI, provide interactive and comprehensive annotations of the genes on the human genome, as well as multiple tools for further bioinformatics analysis such as variant predictors and sequence comparison tools. The sequences they contain for ncRNAs are usually imported from other sources that have been created for storing information. GeneCards and HUGO Gene Nomenclature Committee (HGNC) are examples of generic databases containing information about both coding and non-coding genes. Location, aliases, description, and links to other databases can be found here, but still they only host the information contained in the sncRNA-specific databases. There are also multiple browser-based available tools for the analysis of the datasets such as gene identification tools for DE analysis on two or more groups. SRA is a repository of high throughput sequencing data, containing raw sequences and alignment information, promoting reproducibility and new discoveries through data analysis. Similar to GEO and still from the NIH is the database of Genotypes and Phenotypes (dbGaP), which provides also a controlled access space, meaning that some of the datasets stored needs authorization to get access. Finally, ArrayExpress [92] supported by the European Bioinformatics Institute (EMBL-EBI), stores high-throughput data from functional genomics experiments. The difference with the previous databanks is that ArrayExpress contains both the processed data and the raw sequences as well as links to the European Nucleotide Archive (ENA). All of these repositories contain both coding and non-coding sequences which are the building blocks for the biological databases holding information about the structure, attributes, and interactions of molecules.

General Biological Databases

A large number of biological databases for sncRNAs have been created through the years with diverse purposes such as annotation, structural information, function, interactions, location, sequence, and others. There is a large part of overlapping and redundant results contained in the databases, because of the interoperability and the information exchange between different providers. For many years now, there have been efforts to map and annotate all genes and transcripts, especially in the ever-increasing field of non-coding RNAs. The reason for non-coding-specific databases is that knowing the sequence of these molecules is the most crucial information for finding their interactions and developing computational tools for their analysis.

sncRNA sources

miRBase. The miRBase founded in 2003 [93] is among the most significant databases for miRNA sequences storage and annotation, with the latest version v22.1 (2019) containing 1917 hairpin instances and 2500 mature miRNAs for the human species alone. miRbase integrated multiple tools for sequence annotation, target prediction, and new sequence registration [94]. Additionally, it includes both experimentally verified and computationally predicted active sites and targets, and it is one of the main sources of miRNA information for other databases. Currently, there is an effort to synchronize the miRbase with Rfam; a collection of RNA families including sncRNAs with additional information about secondary structures. Both of these databases contain classifications for microRNA families but so far obtained with different methods and have a consensus of only 28% between them.

miRTarBase is a biological database that mainly provides generally validated experimentally miRNA-Target Interactions (MTI) collected in a manual way [95]. miRTarBase contains more than 4.4M interactions of about 3000 miRNAs for humans and has search filters based on specific miRNA names, their targets, and diseases. **miRCarta** [96] implements the information of precursor and mature miRNAs coming from miRBase as well as predicted ones resulted from the online pipeline **miRMaster** [97]. The import of these predicted miRNAs which are based on the sequence of the sample data, results in a huge number of miRNAs in the database which is around 25k mature miRNAs and 15k precursors for the human species alone.

An interesting and recently published comprehensive database for circulating sncRNAs is **EVAtlas** [98]. It contains information for multiple families of non-coding RNAs from disease and control datasets originating from different tissues and sources. Data collection for EVAtlas is made from 57 GEO and SRA manually reviewed registries, making it a great tool for circulating biomarker studies.

Interactions and Targets databases

miRNet [99] visualization of miRNA and other molecule interactions, can be used for multilayer network construction and ceRNA networks. It links miRNAs to coding and non-coding molecules, transcription factors, and diseases. These features make miRNet a great tool for multi-layer network reconstruction.

Pathways and enrichment analysis databases

Other than databases with structural details and interactions for ncRNAs, there are databases containing information on the involvement of ncRNAs in molecular pathways and processes, linking them in a functional role beyond their immediate first-degree interactions. The Kyoto Encyclopedia of Genes and Genome (**KEGG**) is among the most used databases for pathways, storing genomic and pathway information, and providing manually drawn maps of interactions, regulations, and signal cascading. Despite the fact that it is so well organized, KEGG has a limited amount of information about sncRNAs, and most of them are related to cancer. **Reactome**, is another generic human curated biological pathway database, which cross-references its information with NCBI, Ensebml, KEGG and others [100]. It implements online tools for analyzing and interpreting interactions and visualization of networks, but it also has relatively limited information about ncRNAs. For this reason, **miRPathDB** [101] has been created to indirectly link the regulatory information of miRNAs to the molecular pathways. Although miRPathDB does not calculate the interactions, it uses context mining techniques to gather information from different enrichment analysis and pathway generic sources (KEGG, GO), linking them to information of ncRNA databases such as miRBase or miRCarta.

RISE is a repository for RNA-RNA interactions coming mainly from transcriptome-wide studies [102]. Although RISE contains information about interactions between sncRNAs and other RNA molecules, it mostly focuses on lncRNA interactions. Thus, the use in sncRNA studies can be used in a validation step of a ceRNA network. **NPinter** [103] contains interactions between ncRNAs (except tRNAs and rRNAs) and biomolecules (proteins, RNAs, and DNAs)

with the additional feature of visualizing the network of first-degree interactions between the query and the target. The drawback of this database is the limitation to interactions.

Lastly, a broader open-source RNA interaction database is starBase or **ENCORI** [104] which integrates information for 23 species from which it has more than 4.1 million miRNA-ncRNA interactions and 2.9 million miRNA-mRNA interactions. The data for ENCORI comes from the analysis of high throughput datasets, gene co-expression analysis, and signaling pathways sources. ENCORI offers the option of searching for interactions based on the type of interaction (miRNA-Target, RNA-RNA) as well as ceRNA-Network and pathways based on KEGG terms.

3.1.1.4 Computational Tools to Investigate Molecular Mechanisms and Characteristics of sncRNAs

Bioinformatics tools are used to make the analysis of complex biological systems possible, fast and reliable. Once the sequence of sncRNAs is known through experiments and/or prediction techniques (e.g. miRMaster), and the information of interactions is available in databases, the analysis usually proceeds with the creation of networks. Networks of single or multiple-level instances such as molecules, diseases, and pathways coexist and interact in one graph. In the case of novel transcripts, where there is no experimental evidence or previous knowledge of the targets of ncRNAs, computational tools try to predict the most probable interactions of these molecules in several ways. A list of the databases and tools discussed in this work can be found in Table 1

Target prediction tools

Binding site prediction for sncRNAs is usually referred to miRNAs and siRNA targets which are calculated based on thermodynamic criteria, anti-correlation of target genes, and miRNA/siRNA expression, but most significantly by nucleotide sequences in the target's 3'UTR MREs. Many tools have developed in the last decades for this difficult task, with the most popular one being the **TargetScan**. An online computational tool for target prediction of miRNAs, based on the complementarity between the query gene transcript and the seed of the miRNA along with other multiple features related to the nucleotide sequence of the targets [105]. **DIANA** is a set of tools with the microT algorithm predicting miRNA targets in canonical (3'UTR) regions and the microT-CDS [106] algorithm for the non-canonical (coding) regions. Additionally, DIANA implements the LncBase and TarBase [107] databases for experimentally verified miRNA-target interactions with non-coding and coding transcripts respectively, and mirPath tool for identifying potential altered pathways based on miRNA expression profiles. There is a plethora of other tools and databases related to target prediction such as **miRecords** [108] or **miR2Disease** [109] which contains information about miRNAs related to specific diseases, but they are not as comprehensive or updated as the previously mentioned ones even though they hold valuable information and are sources for databases.

Network reconstruction and visualization

The use of networks in molecular interactions is crucial to depicting and tackling the complexity of biological systems. One of the uses of biological networks is the visualization of interactions, which in small networks is easy to interpret but when there are hundreds or thousands of nodes and edges it gets overwhelming for a human to handle. So, a more useful application for these systems is the analysis based on graph theory. Metrics of centrality and affinity can be used to evaluate significant nodes and pathways, leading to important conclusions such as potential therapeutic targets [110]. Moreover, instances belonging to different categories (e.g. genes, variations, and phenotypes) can be integrated into an interactive network and help to draw conclusions about difficult problems. Tools that are used in bioinformatics for visualization of networks and analysis derive from generic network-reconstruction tools that are based on maths and graph theory. Functionality related to the field of biology was added through the years, mostly in the form of add-on modules that extend the basic metrics and enrich them with biological information through the available databases.

Pajek [111] is a generic, more than 20 years old, Microsoft Windows-based network visualization tool, initially implemented for social network analysis. It is also considered an immensely powerful application for analysis and visualization of massive networks because it can easily visualize a million nodes with billions of connections in an average computer. For Pajek there are available implementations that are optimized to handle faster and with a lower need for memory larger structures (Pajek-XXL or Pajek-3XL). It also implements numerous features such as Graph layout, node merging, neighborhood detection, identification of strongly connected components, clustering, and many other network analysis metrics and tools. This feature makes it a great tool for massive networks but with lower quality visualization potential.

Gephi [112] is a free offline open-source, leading visualization and exploration software and runs on all main operating systems. It is not designed specifically for biological networks, rather it is a general-purpose tool for exploratory data analysis, social networks, and biological network analysis. In Gephi there are multiple plugin modules designed for clustering of nodes and statistical analysis. It is user-friendly, allowing for customization in the visualization and due to its flexible multi-task architecture is very fast even for large datasets.

Cytoscape [113] is probably the most popular open-source desktop application for 2D network visualization in biology and health sciences. It supports all kinds of networks (e.g. weighted, unweighted, bipartite, directed, undirected, and multi-edged) and comes with an enormous library of plugins with more than 250 modules. It was initially designed for research related to biology, as its first aim was to analyze molecular interaction networks and biological pathways, integrating them with other state data such as gene expression profiles. It can handle big networks, but it requires more memory and time for clustering and layout routines than other tools which makes it less scalable, and it is recommended to run such processes in the command line and then load the results as node/edge attributes. It is a good compromise between analysis and visualization, and it comes with a great plethora of layout, clustering, and topological network analysis algorithms, such as AutoSOME, Eisen's hierarchical and k-Means clustering (in the ClusterMaker plugin), and the basic network metrics of average connectivity betweenness centrality and others. Finally, plugins for the connection of biological databases of functional enrichment, GO annotations, data retrieval, and others have been developed making it very convenient to work with.

Other solutions for network analysis may include market products or whole pipelines of processing. These solutions are usually less customizable but require less knowledge of the underlying methods and fewer resources of computational power from the user. One such example is InSyBio's suite, which implements multiple tools from the level of RNA-sequence analysis up to the network analysis by InSyBio BioNets [114] for identifying important nodes and potential biomarkers using machine learning approaches.

Bellow in Table 1, there are all the biological databases and the bioinformatics tools presented in this section. Additionally, there are their hyperlinks for easy access.

	Name	URL
Bio Databases	**Genome Browsers**	
	Ensembl	http://www.ensembl.org/index.html
	NCBI	https://www.ncbi.nlm.nih.gov/datasets/genome/
	UCSC	https://genome.ucsc.edu/
	Non-codingRNA related	
	miRBase	https://www.mirbase.org
	miRTarBase	https://mirtarbase.cuhk.edu.cn
	miRCarta	https://mircarta.cs.uni-saarland.de
	EVAtlas	http://bioinfo.life.hust.edu.cn/EVAtlas
	miRNet	https://www.mirnet.ca
	Molecular Pathways	
	KEGG	https://www.genome.jp/kegg
	REACTOM	https://reactome.org
	miRPathDB	https://mpd.bioinf.uni-sb.de
	RISE	http://rise.life.tsinghua.edu.cn
	NPinter	http://bigdata.ibp.ac.cn/npinter4
	ENCORI	https://starbase.sysu.edu.cn
Bioinformatics Tools	**Target prediction**	
	miRMaster	https://ccb-compute.cs.uni-saarland.de/mirmaster2
	TargetScan	https://www.targetscan.org/vert_80
	DIANA	https://diana.e-ce.uth.gr/home
	miRecords	http://c1.accurascience.com/miRecords
	miR2Disease	http://www.mir2disease.org
	Network Visualization	
	Pejek	http://mrvar.fdv.uni-lj.si/pajek
	Gephi	https://gephi.org
	Cytoscape	https://cytoscape.org

Table 1. List of databases and tools discussed in the present work and the uniform resource locator (URL) for each of these sources.

3.1.1.5 Conclusion

Computer science and bioinformatics have had a transformative impact on systems biology. The development of computational tools and well-organized, curated, and open-source databases has revolutionized the way biological studies are conducted. These tools have shifted the focus from small-scale experiments to large-scale, data-driven approaches, allowing researchers to model biological systems more accurately. Additionally, the availability of datasets in databanks, combined with the interoperability and organization of information in databases, has facilitated the use of multi-layer networks. These networks provide a holistic view of biological interactions, enabling broader and more comprehensive computational approaches.

Even though we focused on neurodevelopmental disorder, none of the tools discussed in this work are specifically built for this category of disorders. However, their general applicability makes them valuable for studying NDDs and other neurological conditions. The simplicity of sncRNA structures, lacking specific biochemical features, means that individual tools analyzing these molecules are inherently general. Nevertheless, combining these tools in a pathology-specific manner can lead to insights into the emergent properties of complex structures, such as tissues, organs, and diseases.

This review aimed to compile the most well-known and important tools and databases, providing insights into their functionality and effectiveness. While the tools presented are not exclusively designed for NDDs, they can be applied to identify common molecular pathways and comorbidities in preterm infants with NDDs.

Future Directions

The field of bioinformatics is continuously evolving, with a constant need for new tools and additional functionalities. While this review focused on high-level tools for genomic data analysis and network reconstruction, the methodologies for data analysis represent a vast and specialized area that warrants further exploration.

In conclusion, the tools and databases presented here serve as a foundation for exploring the role of sncRNAs in NDDs and other complex disorders. Their application, combined with ongoing advancements in computational biology, holds great promise for uncovering the molecular underpinnings of neurodevelopmental disorders and improving diagnostic and therapeutic strategies.

3.1.2 Machine learning in Bioinformatics

The increasing volume and the complexity of biological data have necessitated the adoption of advanced computational methods for biomarker discovery. Machine Learning (ML) has emerged as a powerful approach in bioinformatics, offering the ability to model intricate relationships, extract meaningful patterns, and improve predictive accuracy beyond traditional statistical techniques. Unlike classical statistical methods, which often rely on predefined hypotheses and assumptions, ML algorithms learn from data, enabling the discovery of novel molecular biomarkers with minimal prior knowledge. And it has been proven that the use of ML methodologies for biomarker discovery is beneficial and complementary with classical statistical methods [115].

Supervised and Unsupervised Learning for Biomarker Discovery

ML methods can be broadly categorized into two groups: supervised and unsupervised learning approaches. Supervised learning techniques, such as Support Vector Machines (SVM), Random Forest (RF), and Gradient Boosting Machines (GBM), are widely used for classification and regression tasks in biomarker discovery [116]. These models are trained on labelled datasets, where the presence or absence of a disease, or the expression level of a biomarker, serves as the target variable. For instance, RF has been used to identify potential protein markers for ASD prediction helping in improving the performance of statistical models alone [117].

On the other hand, unsupervised learning methods, such as k-means clustering, hierarchical clustering, and PCA, are employed when labelled data is unavailable. These approaches facilitate the discovery of hidden patterns and molecular subtypes in diseases by grouping similar samples based on their molecular profiles. For example hierarchical clustering can be used for differentiating between cancer subtypes [118], even though in recent years studies focus more on supervised learning.

Deep Learning for High-Dimensional Biological Data

Deep learning, a subset of ML which has gained traction in bioinformatics due to its capability to handle large-scale and high-dimensional datasets. Neural networks have been applied to genomics, transcriptomics and proteomics, enabling predictions with high accuracy and the selection of features potentially useful as biomarkers [119]. Despite their success, deep learning models require substantial computational resources and large training datasets, which remain a challenge in bioinformatics research especially for rare diseases.

Feature Selection and Dimensionality Reduction

Biological datasets, in particular omics data, frequently contain thousands of variables but fewer samples, resulting in large features to samples rations. This imbalance presents a significant challenge known as the curse of dimensionality, where distances between data points become less meaningful and distributions become sparse, often reducing machine learning performance. Methods like t-test or mutual information that rank features based on their association with the target phenotype allowing to keep a subset of the most relevant features [120]. Other methods are fitting models like Lasso, random-forests and SVM-based Recursive Feature Elimination (RFE) help to identify the relevant biomarkers, reducing noise and overfitting [121]. On the other hand, dimensionality reduction techniques like PCA and t-SNE transform the data into a lower-dimension space enabling visualization of complex multi-dimensional data, facilitating a better understanding of underlying biological processes. As dimensionality increases, models require exponentially more data to generalize effectively, making these techniques essential additional to statistical uncertainties rising from multiple testing.

Challenges and Future Directions

Several challenges persist in applying ML to bioinformatics, including small sample sizes, data heterogeneity, the risk of overfitting and the integration of multi-omics. Additionally, the standardization of ML pipelines and the development of open-access benchmark datasets will be essential for ensuring reproducibility and widespread adoption of ML-driven biomarker discovery in biomedical research.

In conclusion, ML has revolutionized biomarker discovery in bioinformatics by enabling the analysis of high-dimensional data, identifying novel disease subtypes, and improving predictive modelling. However, addressing challenges related to model interpretability, data integration, and generalizability remains crucial for translating ML-based discoveries into clinical applications.

3.2 The Need for Data Harmonization

Despite significant advancements in bioinformatics and genetics, some of the most persistent challenges remain variability between studies, small sample sizes, and technical differences. These challenges become even more pronounced when studying complex, multifactorial disorders such as NDDs, where large sample sizes are essential to detect meaningful biological differences. Although new technologies such as NGS have improved data generation, previous studies remain highly valuable. Setting up new studies is often challenging due to ethical considerations, strict data-sharing agreements, high costs, and logistical difficulties in recruiting sufficiently large and well-matched cohorts. As a result, researchers increasingly rely on existing datasets, using approaches like meta-analysis and data merging to maximize the utility of available information.

To make the most of these existing datasets, researchers often take one of two main approaches: *meta-analysis* or *data harmonization and merging*. In a meta-analysis, statistical measures (e.g., effect sizes, p-values) are calculated separately for each dataset and then aggregated to produce a summary estimate, allowing insights across studies without directly combining raw data. In contrast, data harmonization and merging involve aligning datasets to ensure compatibility and then integrating them into a single analysis after addressing potential biases.

However, merging datasets presents its own challenges, including differences in study protocols, population characteristics, batch effects, and methodological variations. Despite these difficulties, merging data increases sample sizes, thereby enhancing statistical power and improving result robustness.

It is important to recognize that variability in data can stem from both biological and technical sources. Biological variability reflects genuine differences between samples, such as genetic diversity or disease-specific changes. Technical variability, on the other hand, arises from factors such as laboratory protocols, environmental conditions (e.g., temperature, humidity), equipment differences, and variations in sample handling by different technicians. Established methods, such as batch effect correction techniques (e.g., ComBat), can mitigate technical variability by incorporating known confounding factors. For unknown sources of variability, surrogate variable analysis (SVA) can identify hidden patterns in the data that may reflect systematic biases.

A key challenge in data correction is overcorrection—where excessive adjustments remove not only technical artifacts but also biologically meaningful signals. While reducing batch effects is essential for improving data consistency, an overly aggressive correction can obscure true disease-related variations, potentially leading to misleading conclusions. Striking a balance between removing technical noise and preserving true biological differences is critical for ensuring reliable and interpretable results.

While meta-analysis provides valuable insights by summarizing findings across studies, it does not allow for direct investigation of individual-level data. In contrast, merging datasets enables a more detailed analysis by increasing statistical power, reducing variability, and allowing for the identification of subtle associations that may be missed in smaller studies. This is particularly important in the study of NDDs, where patient heterogeneity and the multifactorial nature of these disorders make large sample sizes essential for detecting meaningful patterns.

Additionally, in fields where collecting new data is difficult - such as rare diseases or studies requiring invasive procedures - pooling existing datasets is often the only way to achieve the necessary statistical strength for reliable conclusions. However, effective data merging requires careful handling of batch effects, technical variability, and study-specific biases to ensure that biological signals are preserved.

3.3 Merging of Neurodevelopmental related microarray datasets

Despite the advent of next-generation sequencing (NGS) technologies, microarray technology remains a valuable tool in genomic research. One of its key advantages is the extensive availability of historical data, leading to the creation of large repositories that enable meta-analyses across multiple studies. However, a major limitation of microarrays is the variability introduced by study protocols and different manufacturers, each employing distinct probe designs that target different gene regions, making direct comparisons between datasets challenging. Thus, harmonization and merging of the data is a way to eliminate both the frequent issue of low sample sizes and the technical differences between studies.

In this work, we developed a method for integrating datasets from different microarray platforms that originate from the same tissue type. As a case study, we focus on blood related samples and neurodevelopmental disorders in infants, toddlers, and adults. Initially, each microarray dataset undergoes independent preprocessing but maintaining the same computational methods to keep a unified pipeline. The probesets are transformed into gene-level expressions by utilizing the manufacturers' information and averaging their expression. Later, a custom normalization method based on the average expression of individuals within the array is applied to each dataset keeping the steps unified and applicable to all platforms. Subsequently, the data are merged, and a batch effect correction is applied to eliminate the non-biological effects of the study and platform used. Finally, the dataset is normalized with the quantile method bringing the expression value distributions to comparable ranges. To demonstrate the effectiveness of this approach, we utilized publicly available datasets from the Gene Expression Omnibus (GEO), selecting five independent experiments (GSE6575, GSE18123, GSE42133, GSE89594, GSE111175) derived from four distinct microarray platforms (GPL570, GPL6244, GPL10558, GPL16699).

This process enables comparative analyses across datasets that were previously incompatible due to platform-specific differences, facilitating deeper insights into neurodevelopmental disorders. As a result, we created a pool of data for NDDs with four categories. The Control group enumerates 281 samples, the group of ASD has 357, William's syndrome consists of 37 samples and Language Delay of 27 samples making a total of 695 individuals and 18.839 gene symbols containing both coding and some non-coding elements. By performing differential expression analyses in the pairs of the conditions followed by enrichment analyses, we found that meaningful features found important and known from literature Gene Ontology terms found enriched.

Overall, we were able to increase the effective sizes of the control and ASD groups, successfully harmonize the datasets in a bigger unified set and extract meaningful markers in accordance with literature. All these come after a strict process pipeline tailored for each of the individual sets and their acquisition platform, although there are multiple assumptions for the data along the procedure. This method is built on data related to NDDs but can be extended to other diseases.

3.3.1 Materials and Methods

The analysis of microarray data requires careful preprocessing to ensure accurate and biologically meaningful results. Since each platform has unique characteristics, raw intensity measurements must undergo a series of transformations to account for background noise, technical variation, and batch effects. This process converts raw fluorescence intensities into a structured expression matrix suitable for downstream analysis.

The preprocessing pipeline consists of several essential steps and is summarized visually in Figure 8:

- **Background Correction** – Adjusting raw intensity values to remove systematic noise and improve the accuracy of measured expression levels.
- **Normalization** – Correcting for technical variability across arrays using methods such as quantile normalization or robust multi-array average (RMA) to ensure comparability.

- **Filtering** – Removing low-quality or non-informative probes that do not contribute meaningful biological insights.
- **Batch Effect Correction** – Addressing technical variations between datasets using methods like ComBat to enhance data integration.

All analyses were conducted in R (version 4.3.1), utilizing the limma package (version 3.56.2) for statistical analysis and Bioconductor (version 3.19) for preprocessing and quality control. Additional R packages, including affy, oligo, and sva, were employed where necessary for handling raw data and mitigating batch effects.

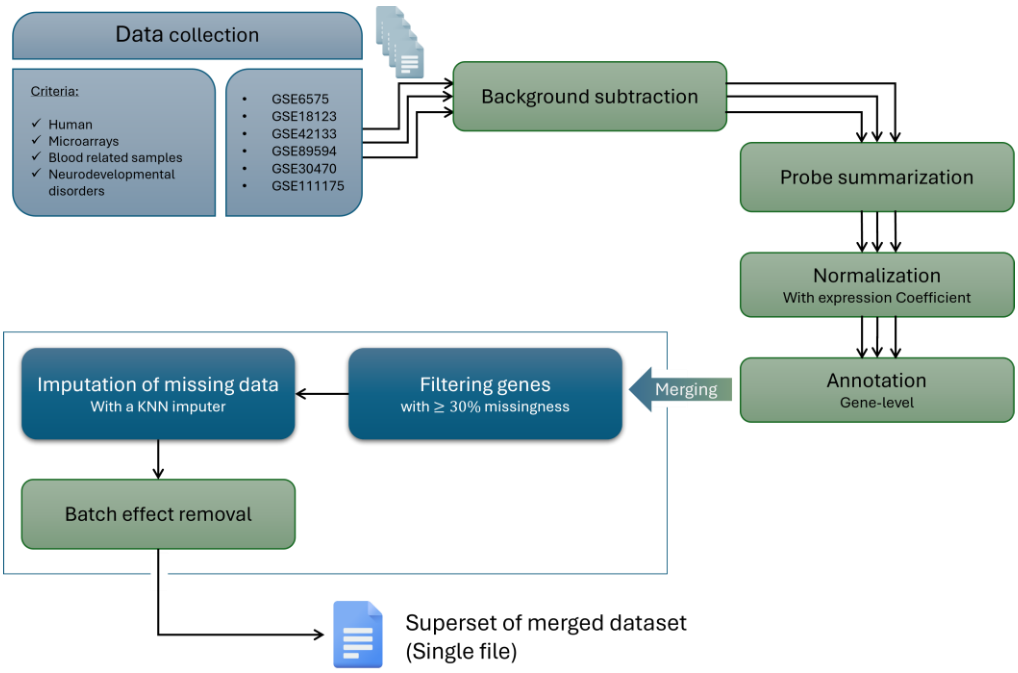


Figure 8. The Flowchart of the steps followed for merging the different datasets. From the selection of the datasets, and the criteria used until the merging of the data in a superset.

Selection and Retrieval of Data

Publicly available datasets related to neurodevelopmental disorders were retrieved from the Gene Expression Omnibus (GEO) repository. The selection focused on datasets involving preterm newborns, toddlers, and children, where circulating blood (whole blood, peripheral blood, or leukocytes) served as the sampling tissue.

Particularly, a systematic search in GEO was initially conducted using the terms “Neurodevelopmental” and “Homo sapiens”. The results were refined by including only blood-derived samples and excluding experiments with fewer than ten samples, particularly in cases where conditions could not be grouped for comparative analysis. Although the search yielded some RNA-sequencing datasets, these were not considered at this stage due to differences in processing methodologies and challenges in harmonizing RNA-seq with microarray-based data.

The final selection comprised datasets spanning from 2008 to 2019, a period marked by advancements in transcriptomic technologies and significant revisions in diagnostic criteria. During this time, the Diagnostic and Statistical Manual of Mental Disorders (DSM) transitioned from DSM-IV (1994) to DSM-5 (2013), with a subsequent revision in DSM-5-TR (2022) [122]. These temporal variations were considered to ensure consistency across the datasets. The selected datasets and their corresponding GEO accession IDs are:

- **GSE6575** [123]: Blood gene expression profiling of children with ASD compared with age and gender matched, typically developing children from the general population or IQ matched children with mental retardation or developmental delay. (2008)

- **GSE18123** [124]: ASD and matched control samples compared to investigate whether peripheral blood gene expression profiles could be used as a molecular diagnostic tool. This dataset contains data from two (GPL6244, GPL570) microarrays. (2012)

- **GSE42133** [125], [126]: Use of leukocyte gene expression levels in autistic and typically developing infants and toddlers to identify biomarkers for their distinction. (2015)

- **GSE89594** [127], [128]: Peripheral blood samples from ASD patients, William’s syndrome (WS) patients and typically development adults. (2018)

- **GSE111175** [129], [130]: Leukocyte gene expression levels in children from 1 to 4 years of age. Containing ASD patients, pervasive developmental disorder not otherwise specified (PDDNOS) patients, patients with language development, and controls. (2019)

- **GSE30470** [131]: A pilot study of the correlations of gene expression with ratings of inattention and hyperactivity/impulsivity in Tourette syndrome. This dataset contains children of age 7 to 15 years with Tourette, and no controls. (2012)

Due to the lack of control samples in the GSE30470 dataset that contains children and adolescents with Tourette syndrome, this dataset never used in the merging because we could not be sure about the normalization of the data within the array.

Data Acquisition and Preprocessing

Raw expression data were obtained using the getGEOSuppFiles function from the GEOquery package in Bioconductor using the programming language R (version 4.3.1). Processed expression matrices were retrieved using the getGEO function, which includes sample annotations, population characteristics, and relevant clinical and demographic metadata.

To enable effective data harmonization, metadata from each dataset were standardized, and preprocessing steps were applied to resolve discrepancies in sample annotations, terminologies, and missing values. The merging process involved both computational and manual curation steps to maintain data integrity and facilitate interoperability across datasets.

Raw data processing

Data loading procedures varied depending on the manufacturer and experimental design with different types of normalizations and scales. Microarray raw data contain intensity values, but variations in scanning protocols and output file formats necessitate distinct processing approaches. Due to these platform-specific differences, initial preprocessing was performed separately for each dataset before harmonizing expression data into a unified format. To keep the pipeline as uniform as possible for all datasets, we performed the recommended by the manufacturers' steps of background correction, probe summarization and control probe filtering without further normalization, and by transforming all intensities to a logarithmic base-2 scale.

Background correction and filtering

In more detail, the dataset selection process resulted in four different microarray platforms from three manufacturers. A uniform processing strategy was applied across platforms to minimize variability while accounting for platform-specific technical differences.

Starting with the Affymetrix manufacturer, the pool of data has two different platforms which are the “Human Genome U133 Plus 2.0” and the “Human Gene 1.0 ST”, with accession numbers GPL570 and GPL6244 respectively in the GEO. For both cases the Robust Multichip Average (RMA) method was used with the help of the *rma* function of the *oligo* package in R. RMA is also advised from literature to follow for Affymetrix microarrays. as recommended in the literature for Affymetrix microarrays, performs background subtraction and probe summarization, followed by an optional normalization step. To preserve the original distribution of probe set values, quantile normalization was deliberately omitted.

Second manufacturers in the list of the selected datasets is Illumina with the “HumanHT-12 V4.0 expression beadchip” that has the GPL10558 GEO platform ID. Raw data were imported with the help of limma’s package read.ilmn function which is tailored for Illumina arrays. Additionally, Illumina offers a downloadable file with the annotation for this platform through the company’s website [132] from which one can find which are the control probes and which are the actual probes measuring the expression of some transcripts. Using this knowledge, it is possible to use the *nec* function of the *limma* package in R to do a normexp background correction using the negative control probes. As with the Affymetrix platforms, quantile normalization was intentionally avoided to maintain the distribution of intensity values. Control probes were subsequently excluded based on manufacturer-provided annotations.

The final dataset category included Agilent arrays, specifically the “SurePrint G3 Human GE v2 8x60K” microarray (GPL16699). Raw scanner-generated files were imported using the read.images function in limma, which extracts mean and background intensities from the scanner’s output files. Background correction was performed using the backgroundCorrect function in limma, applying the normexp method, which accounts for both mean and background intensity measurements.

Probe-level and annotation

Following background correction, differences in preprocessing methods resulted in some datasets being transformed to a logarithmic scale, while others remained in linear space. To ensure uniformity across datasets, all expression values were $\log_2$-transformed before proceeding with further analysis.

Annotation data accompanying the processed expression matrices were initially retrieved using the getGEO and pData functions of Bioconductor. These annotations were compared against platform-specific annotation packages available in Bioconductor to resolve inconsistencies, update outdated annotations, and address multiple probe-to-gene assignments. Probes were subsequently grouped according to their corresponding target genes, and expression values were averaged to generate a gene-level expression matrix for downstream analyses.

Scaling values

To account for differences in overall gene expression levels across samples, a sample-wise scaling approach was applied. This method adjusts for variations in global expression intensity by computing a scaling coefficient for each sample based on its overall expression distribution. Specifically, samples with higher overall expression levels are scaled down, while samples with lower total expression are relatively upscaled, ensuring that all samples were brought to a comparable expression range.

Scaling coefficients were derived from a column-wise summary statistic computed across all genes in each sample, returning an array with length equal to the samples in the dataset:

$$ColumnWise\ summary_m = \sum_{n=1}^{N} x_{n,m}$$

Equation 1. Patient-wise summary: The total expression across all genes is computed for each individual sample. In the equation, $x_{n,m}$ represents the expression value of gene n in sample m, and N is the total number of genes in the dataset.

Then the average of that array is computed by adding its elements and dividing with the total number of samples:

$$Average\ summary = \frac{1}{M} \cdot ColumnWise\ summary = \frac{\sum\sum_{n}^{N} x_{n,m}}{M}$$

Equation 2. Average summary of the population: The patient-wise summaries are summed across all samples and then divided by the total number of samples (M), yielding the mean gene expression across all genes and all samples.

For every sample the corresponding coefficient is computed by taking the ratio of the average summary over the column expression summary:

$$Scale\ coefficient_m(S_m) = \frac{Average\ summary}{ColumnWise\ summary_m} = \frac{\sum\sum_{n}^{N} x_{n,m}}{M \cdot \sum_{n}^{N} x_{n,m}}$$

Equation 3. Scale Coefficient: The scaling coefficient for each sample is computed as the ratio of the average summary (mean expression across all genes and all samples) to the column-wise summary (total expression for that specific sample). This coefficient adjusts for differences in overall expression intensity between samples, ensuring comparability across the dataset.

And finally, the modified expression is the product of the original expression value and the scale coefficient.

$$Modified\ expression_m(\tilde{x}_{n,m}) = x_{n,m} \cdot S_m$$

Equation 4. Modified Expression: The adjusted expression values are obtained by multiplying the original expression of each gene in a given sample by the corresponding scaling coefficient. This transformation normalizes the overall expression levels across samples while maintaining the relative differences between genes within each sample.

This method operates under the assumption that the total gene expression across samples remains relatively constant, with only minor biological variations. Significant deviations in overall expression levels are attributed to technical differences, which are corrected using the scaling coefficient. This assumption is based on the premise that differentially expressed genes in distinct phenotypes represent a relatively small subset and do not substantially alter the total expression distribution, whereas technical artifacts can shift the entire expression profile upwards or downwards.

A toy example of how the normalization method works is shown in Figure 9 where the original expression matrix containing the raw gene expressions yields the column-wise summary. From that array the total expression summary is calculated and sequentially the mean value of that quantity. Later, the scaling coefficient for each sample is calculated by dividing the mean summary to the column-wise summary. Finally, each original expression value is multiplied by its respective scaling coefficient, producing the modified expression matrix, where global expression differences across samples are adjusted while preserving relative expression patterns.

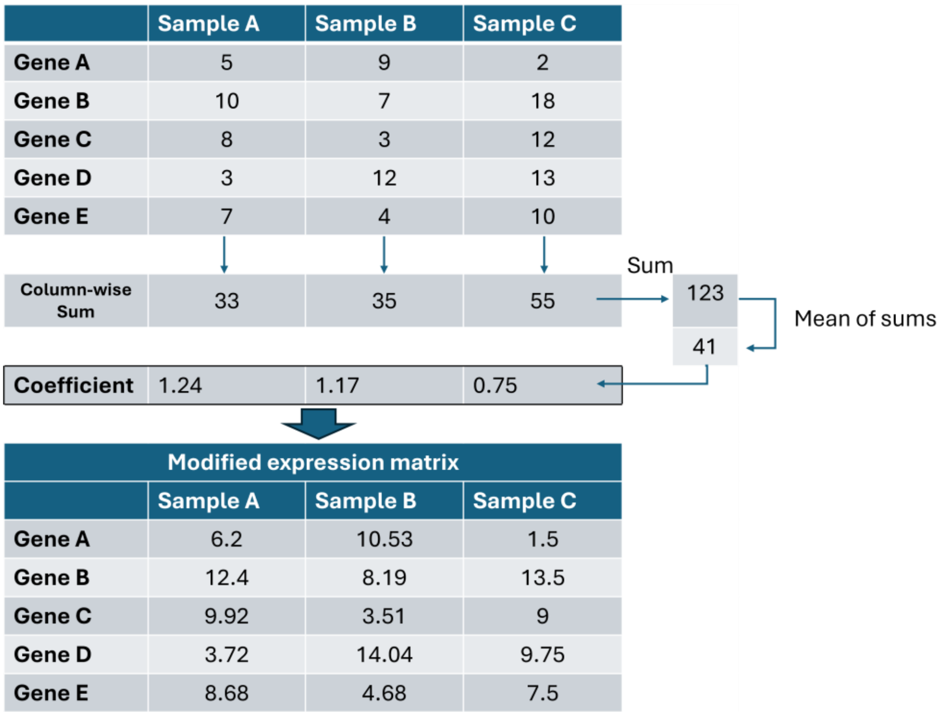


Figure 9. Illustration of the sample-wise normalization method. From the original expression matrix (top). The column-wise sum, the total expression, and average total expression across all genes are calculated. The scaling coefficient for each sample is obtained by dividing the mean of sums by the corresponding column-wise sum. Finally, each original expression value is multiplied by its respective scaling coefficient, producing the modified expression matrix (bottom).

Merging

Following preprocessing, datasets were concatenated using gene symbols as the unique identifier, creating a big expression matrix. The same method can be applied by selecting a different identifier like Ensemble gene IDs. When a gene was present in one dataset but absent from another, missing values were introduced in the corresponding sample columns of the dataset lacking that gene. Similarly, if a gene in the reference index was not present in a newly merged dataset, missing values were assigned to the corresponding samples in the new rows of the dataset.

This merging process resulted in a large gene expression matrix, where rows correspond to the union of all genes from the individual datasets, and columns represent all samples from the merged datasets. Due to differences in microarray platforms, as mentioned above, not all datasets contained probes for every gene which led to the presence of missing values in the final merged matrix. At this stage the expression values are in the original distributions of the datasets they extracted from.

Imputation of missing values

After merging, a row-wise assessment was performed to identify genes with more than 30% missing values, which were subsequently removed from the dataset. For genes with missing values below this threshold, A k-nearest neighbors (KNN) algorithm for imputation was applied using the impute.knn function from the *impute* R package with the five closest samples known as nearest neighbors (k=5). This method eliminates the missing values in the dataset either by dropping sparse rows or by introducing an approximate value for genes with a small percentage of missingness.

Adjusting for batch effects

After merging and imputation, batch effect correction was applied to mitigate technical variability introduced by differences in dataset sources and microarray platforms. Batch effects can arise due to variations in experimental conditions, platform technologies, and processing protocols, potentially masking true biological differences and introducing biases in downstream analyses.

To correct the data for these effects, batch adjustment was performed by modeling batch variables based on dataset series and platform information. Specifically, a batch factor was created using the *interaction* function in R, combining the GEO Series ID and platform ID (microarray type). The batch effect was then removed using the *batch_effect_removal* function of the *limma* library, resulting in a dataset corrected for variances depending on the study and the array used. This adjustment ensured that the expression values were primarily influenced by biological variability rather than technical artifacts, making the dataset suitable for downstream analysis.

3.3.2 Results

Following the procedure outlined above, we analysed the different datasets working on the raw data provided independently for each study to keep a consistent and coherent pipeline of data processing. The gene-level data were merged to address the problem of heterogeneity and create a pool of microarray data related to NDDs from blood samples. The results presented here are organized into three main parts: individual dataset modifications and merging result, after merging assessment, and differential expression. Collectively, these findings offer insights into how merging studies help to increase the finding of potential biomarkers and contribute to our understanding of neurodevelopmental disorders.

Beginning with the first part corresponding to the merging and harmonization of data, the individual dataset processing was followed by a merging by gene symbol step and the removal of genes with more than 30% missing values. Within-array groups that considered similar grouped together and as an example is the dataset GSE6575 where there are two groups of children diagnosed with autism (with and without regression) which grouped as one category under the term ASD (autism spectrum disorder) that aligns with the modern definition and abbreviation of autism.

Accordingly, in the GSE111175 the groups “Controls” and “Preemie No Delay” were assigned with the common group name Control and the “ASD”, “PDD-NOS” and “Autism Features” created the unified “ASD” group (see Table 2 for reference). These changes in the categories are essential for downstream analysis since the goal is to increase the sample size of the conditions and the definition -at least for autism- has changed since these studies conducted. The total number of samples after the groupings in the merged dataset is 695 samples which are the sum of 281 controls, 357 ASD, 32 William’s and 25 LD. Table 2 presents each studies conditions and sample size as well as the grouping of conditions within the study and the outer grouping that refers to the final set of conditions.

<table>
<tr><th>GEO Accession</th><th>Condition(s)</th><th>Sample size</th><th>In-array Grouping</th><th>Group size</th><th>Outer Grouping</th></tr>
<tr><td rowspan="4">GSE6575</td><td>Control</td><td>12</td><td></td><td>12</td><td>Control</td></tr>
<tr><td>Autism with regression</td><td>18</td><td rowspan="2">ASD</td><td rowspan="2">35</td><td rowspan="2">ASD</td></tr>
<tr><td>Autism without regression</td><td>17</td></tr>
<tr><td>Developmental delay</td><td>9</td><td></td><td>9</td><td>Drop group</td></tr>
<tr><td rowspan="4">GSE18123 A
(GPL570)</td><td>Control</td><td>33</td><td></td><td>33</td><td>Control</td></tr>
<tr><td>Autism</td><td>31</td><td rowspan="3">ASD</td><td rowspan="3">66</td><td rowspan="3">ASD</td></tr>
<tr><td>Asperger’s</td><td>9</td></tr>
<tr><td>PDD-NOS</td><td>26</td></tr>
<tr><td rowspan="4">GSE18123 B
(GPL6244)</td><td>Control</td><td>82</td><td></td><td>82</td><td>Control</td></tr>
<tr><td>Autism</td><td>41</td><td rowspan="3">ASD</td><td rowspan="3">104</td><td rowspan="3">ASD</td></tr>
<tr><td>Asperger’s</td><td>15</td></tr>
<tr><td>PDD-NOS</td><td>48</td></tr>
<tr><td rowspan="2">GSE42133</td><td>Control</td><td>56</td><td></td><td>56</td><td>Control</td></tr>
<tr><td>ASD</td><td>91</td><td></td><td>91</td><td>ASD</td></tr>
<tr><td rowspan="3">GSE89594</td><td>Control</td><td>30</td><td></td><td>30</td><td>Control</td></tr>
<tr><td>ASD</td><td>32</td><td></td><td>32</td><td>ASD</td></tr>
<tr><td>Williams Syndrome</td><td>32</td><td></td><td>32</td><td>Williams</td></tr>
<tr><td>GSE30470</td><td>Tourette Syndrome</td><td>21</td><td></td><td>21</td><td>Tourette</td></tr>
<tr><td rowspan="6">GSE111175</td><td>Control</td><td>70</td><td></td><td></td><td>Control</td></tr>
<tr><td>ASD</td><td>28</td><td rowspan="3">ASD</td><td rowspan="3">38</td><td rowspan="3">ASD</td></tr>
<tr><td>PDD-NOS</td><td>9</td></tr>
<tr><td>Autism Features</td><td>1</td></tr>
<tr><td>Language Delay</td><td>27</td><td></td><td>27</td><td>LD</td></tr>
<tr><td>Preemie No Delay</td><td>6</td><td></td><td>6</td><td>Control</td></tr>
</table>

Table 2. Summary of the dataset used with the list of conditions and effective sizes of every category in the data before and after merging. Conditions grouped together (“outer grouping”) to align with modern definitions and minimize the drop rate.

For the assessment of the distribution of expression values in the merged dataset, a principal component analysis with the first two components was used to find visually grouping and patterns in the data. The PCA shows clusters of datapoints in the grouped unharmonized set, and by investigating the source of the clustering it seems that the two main - if not only - factor is the platform ID and study ID, while the condition does not seem to affect the clustering as it is illustrated on the top of Figure 10 where samples are clustered based on the microarray used in each experiments and different conditions co-exist within those clusters. Additionally, there is a clear difference in the distribution of expressions by examining the boxplots of the merged dataset (Figure 10 bottom).

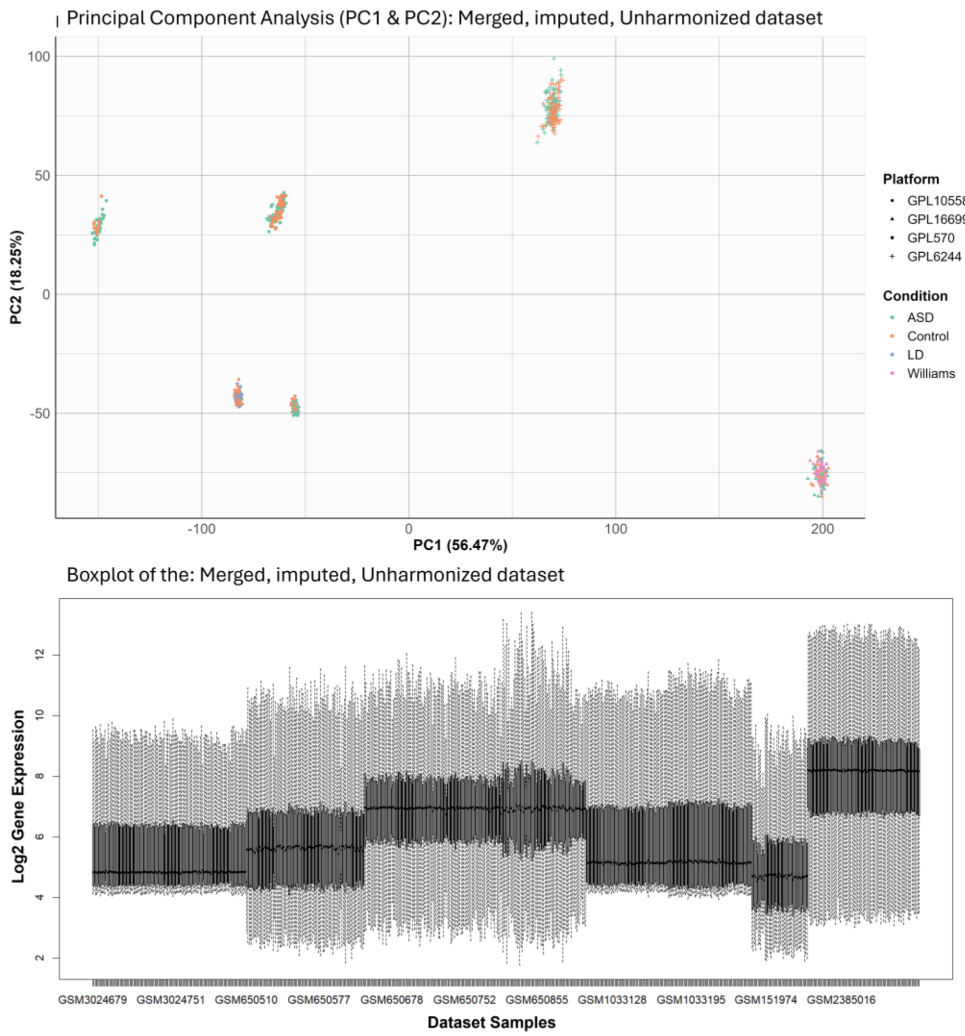


Figure 10. (Top) The PCA of the merged dataset visualizing a clear separation between samples that belong to different studies and come from different platforms, illustrating the batch effect. (Bottom) A boxplot of all samples in the merged dataset showing differences in the distributions of the data.

Normalizing the merged dataset with different methods does not seem to fix the problem of datapoints clustering alone but rather only reducing some of the variance in the lower dimension space. In contrast, by applying a batch effect correction method with the combat function using as batch the interaction term of platform ID and study ID eliminates the clustering effect as it is shown in Table 3.

	With Batch Effect		**Batch Effect Removed**	
	(% Variance Explained)			
Method	**PCA 1**	**PCA 2**	**PCA 1**	**PCA 2**
Default values	56.46	18.25	23.96	6.75
Standardization	33.3	28.99	16.03	6.46
Quantile Normalization	29.32	28.21	13.46	5.35
PNorm	31.22	27.29	24.88	6.84

Table 3. Variance explained (%) from the first two principal components of a PCA on the merged data with various normalization methods before and after the correction for the batch factors of platform ID (microarray) and GEO ID (study).

Additionally, it seems that the batch effect removal function helps in harmonizing the data by removing the effects related to the individual studies as shown also in Figure 11. The visualization of the effects the different normalization methods have before the batch effect correction can be found in the Appendix A - Chapter 3.

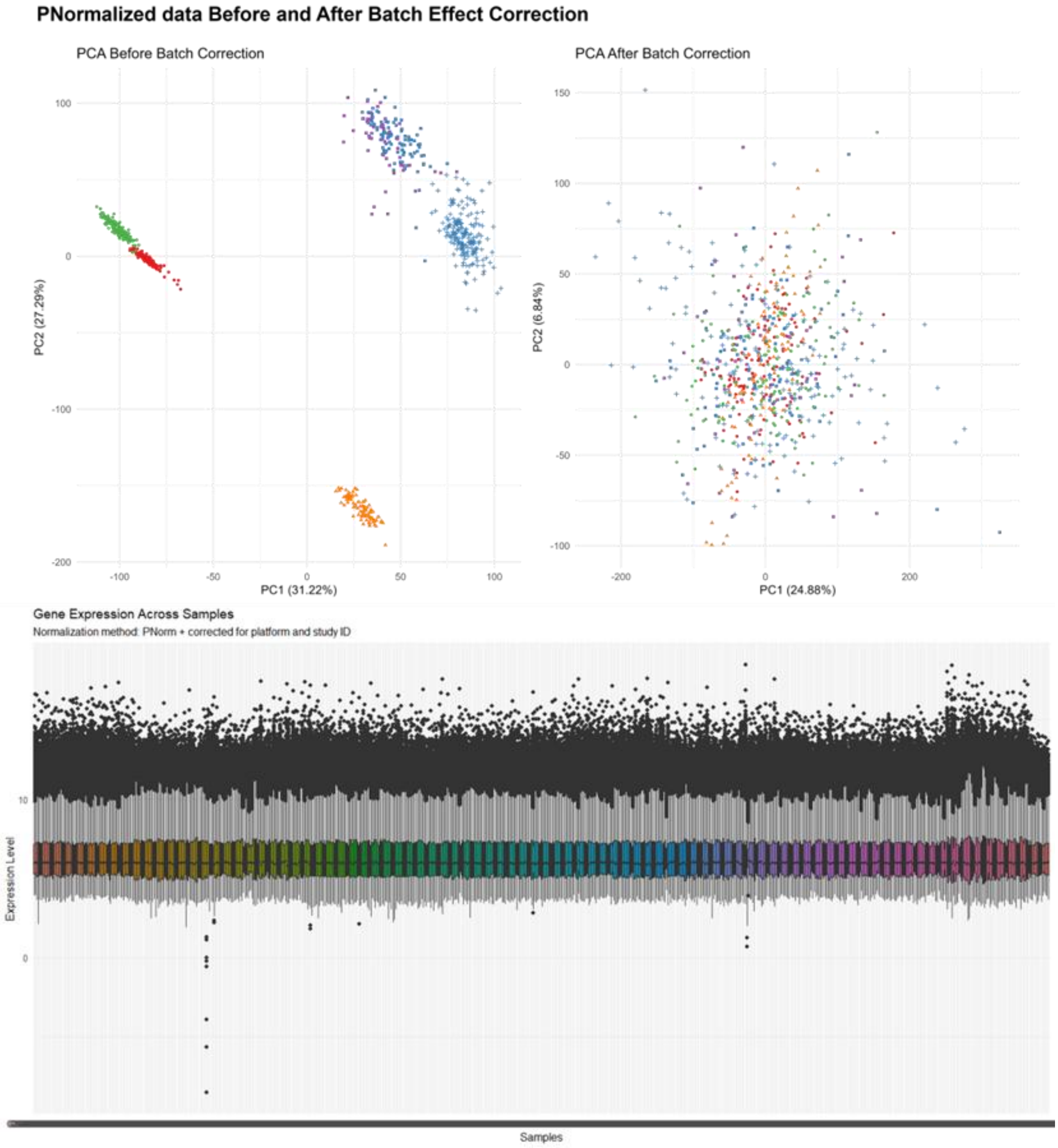


Figure 11. PCA plot comparison before (top-left) and after (top-right) the batch effect removal on the previously normalized data with the PNorm method. (Bottom) The distribution of values after merging, normalizing and correcting for batch effects.

Differential expression analysis was performed pairwise using linear models to identify genes significantly altered between conditions. For reference, the datasets were individually tested for DEGs with standard thresholds for the FDR = 0.05 and for the Log2FC = 1.5 as shown in the Table 4 below.

From this analysis, no DEG was found between the categories of the individual datasets except for the GSE89594 and the pairs of WS - Controls and ASD - WS. Between these groups only three and four genes found differentially expressed respectively.

Differentially expressed gene number with criteria: $FDR \leq 0.05, \lvert Log_2FC \rvert \geq 1.5$						
GEO ID	**ASD - Control**	**LD - Control**	**Williams - Control**	**ASD - LD**	**ASD - Williams**	**LD - Williams**
GSE6575	0	-	-	-	-	-
GSE18123 A (GPL570)	0	-	-	-	-	-
GSE18123 B (GPL6244)	0	-	-	-	-	-
GSE42133	0	-	-	-	-	-
GSE89594	0	-	3	-	4	-
GSE111175	0	0	-	0	-	-

Table 4. Results of the category pairwise Differentially Expressed Genes in each of the datasets individually with the standard thresholds often used in literature. The criteria are $FDR \leq 0.05$ and $\lvert Log_2\ FC \rvert \geq 1.5$.

For the merged dataset since it is expected to have lower log2FC differences due to the normalizations and the batch effect correction, the threshold was lowered to 0.5 with FDR remain at 0.05. Based on these thresholds the comparisons were repeated between the groups. Consequently, the differential expression analysis resulted in no significant changes between genes of ASD and Controls groups based on both criteria, but focusing only on the FDR there are 6218 genes that seem to be differentially expressed but with not an emphatic fold change. In other cases, the comparison of language delay – Control shows 4 DEGs meeting both criteria as does the pair of WS and Control (Table 5).

The number of DEGs is higher in conditions where ASD and LD have 4 genes significantly differentially expressed under both criteria opposed to the individual dataset comparison where no DEGs found. In the case of ASD vs William's Syndrome, there are 9 genes significantly different between these conditions compared to the individual dataset analysis where only 3 met the criteria of significance. Last, the comparison of Language delay and WS is enabled in the merged dataset in contrast to the individual dataset's analyses where this comparison was not possible since these groups are not studied together. In fact, there are 14 genes between these two groups that have different average expressions which are also statistically significant after accounting for multiple testing. To visualize these results, volcano plots for the pairwise comparisons were generated, as shown in Figure 12.

Significance type	ASD - Control	LD - Control	Williams - Control	ASD - LD	ASD - Williams	LD - Williams
FDR & Log2FC	0	4	4	4	9	14
FDR	6218	8	138	56	853	45
Log2FC	0	10	5	6	10	30

Table 5. Differentially Expressed Genes between the categories of the merged dataset under different criteria. The criteria are $FDR \leq 0.05$ and $|Log_2FC| \geq 0.5$.

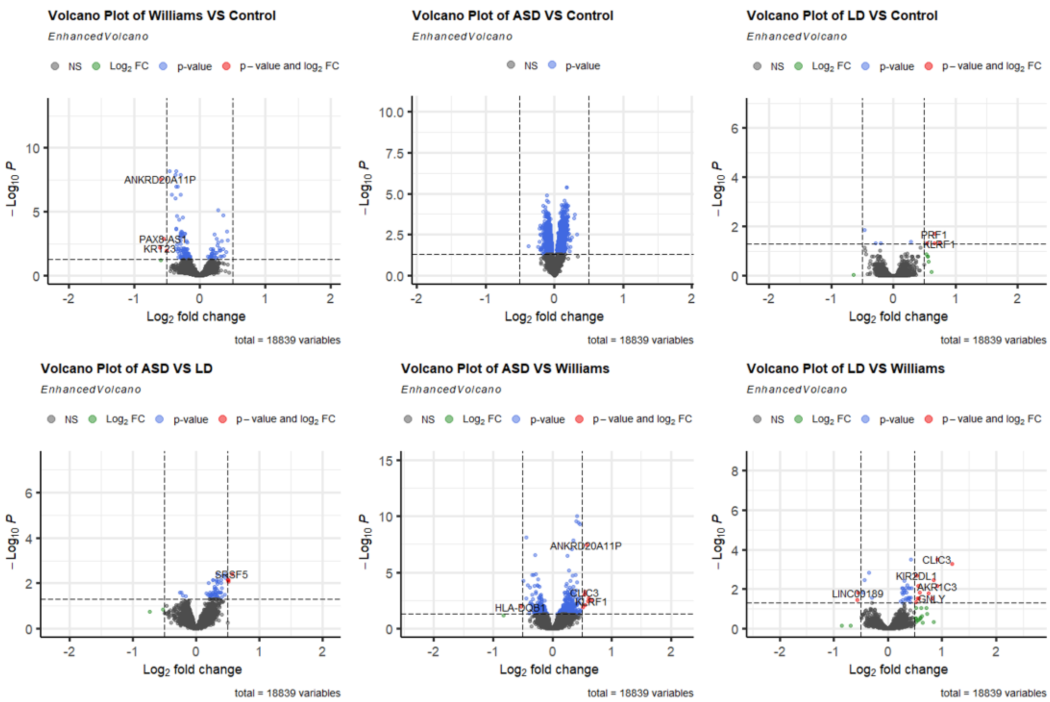


Figure 12. Volcano plots of the pairwise DEA on the merged datasets. In grey are genes that do not meet any of the significance criteria. In green those genes that have a $|Log_2FC| \geq 0.5$, in blue, genes that are statistically significantly different but with low Log_2FC and in red genes tha fulfil both criteria.

Enrichment Analysis

To elucidate the underlying biological processes associated with ASD, LD, and WS, Gene Set Enrichment Analysis (GSEA) was conducted. This analysis revealed significant enrichment of several Gene Ontology (GO) terms for each comparing pair. The following summaries highlight the distinct biological signatures of each condition, drawing upon the GSEA results.

For providing a clear and concise overview of the biological processes significantly altered in each condition, we have chosen to summarize the GO term enrichment results by condition (ASD, LD, and WS) rather than by pairwise comparison. This way we minimize redundancy and allow for a focused discussion of the unique biological signatures associated with each condition, while still referencing the relative changes observed when compared to the control group and other conditions. By highlighting the key pathways and processes enriched or depleted within each condition, we aim to offer a more integrated and interpretable representation of the findings.

ASD Summary

Autism Spectrum Disorder (ASD) exhibited significant alterations in RNA processing and protein synthesis, with notable upregulation of terms like “RNA splicing,” “mRNA processing,” “ribonucleoprotein complex biogenesis,” and “macroautophagy”. These findings suggest substantial dysregulation of RNA metabolism and cellular turnover. Conversely, ASD showed significant downregulation in sensory perception-related processes, including “detection of chemical stimulus” and “sensory perception of chemical stimulus,” indicating marked disruptions in sensory processing. Changes in cellular interactions were also observed, with downregulation of “cell-cell adhesion” related terms (reference at Sup_Fig. 10). Additionally, processes related to “Golgi vesicle transport” and “ribosome biogenesis” were upregulated, reflecting broader changes in cellular organization. When compared to language delay and WS, ASD showed an even more pronounced dysregulation in these categories.

The enrichment analysis on KEGG pathways from the ranked gene list between ASD and Control showed upregulations in the terms of “Spliceosome”, “Nucleocytoplasmic transport”, “Ribosome”, “RNA degradation”, and “ATP-dependent chromatin remodeling” with normalized enriched scores (NES) above 2.3 and FDR < 2.2e-16. From that analysis, we found the pathways of “Olfactory transduction” and “Neuroactive ligand-receptor interaction” downregulated with NES -3.2 and -2.9 respectively, and FDR <2.2e-16. Additionally, the terms “Staphylococcus aureus infection”, “Insulin secretion” and “Protein digestion and absorption” found downregulated, with NES around -2.1 and FDR near to 2.5e-4.

Language Delay Summary

Language Delay demonstrated significant perturbations in cell cycle regulation and immune response pathways, with downregulation of terms like “negative regulation of cell cycle G2/M phase transition” and “pattern recognition receptor signalling pathway.” These findings suggest potential alterations in cell division and immune signalling. Upregulation of “telencephalon cell migration” and “forebrain cell migration” indicates changes in neurodevelopmental processes. “Cytoplasmic translation” was also upregulated, reflecting changes in protein synthesis. Downregulation of terms related to “interleukin production” and “immune response-regulating signalling pathway” suggests dysregulation of inflammatory responses. Compared to ASD, language delay showed less dramatic changes in RNA processing and sensory perception. When compared to WS, language delay showed less antigen processing and presentation, and more sensory perception related down regulation.

On the GSEA for KEGG pathways, the significantly upregulated terms meeting the adjusted p-value criterion are the “Graft-versus-host disease” (hsa05332) and “Ribosome” (hsa03010) pathways (FDR = 0.0095972 and 0.023624, respectively, NES = 0.55866 and 0.42797). The respectively downregulated pathways in LD include the “Lysosome” (hsa04142) and “Phagosome” (hsa04145) (FDR = 0.0046480 and 0.048029, respectively, NES = -0.42554 and -0.36910). There are other KEGG terms both up- and down-regulated in LD which had significant p-value but after the correction for multiple testing failed to pass the threshold. Nevertheless, those findings are reported in the Appendix A - Chapter 3 and the Sup_Fig. 11.

William's Syndrome Summary

William's Syndrome (WS) was characterized by a striking upregulation of antigen processing and presentation via MHC class II, with significant enrichment of terms like “antigen processing and presentation of exogenous peptide antigen via MHC class II” and “MHC protein complex assembly.” This highlights a potentially altered immune response profile, specifically in the context of antigen presentation. Consistently, pathway analysis revealed an enrichment of the 7q11.23 copy number variation syndrome, which is the genetic hallmark of WS. In contrast to ASD and language delay, WS showed less pronounced changes in RNA processing and sensory perception. However, there were some down regulations of sensory perception when compared to control.

3.3.3 Discussion

In this study, we demonstrate the benefits of merging datasets for analyzing differences between groups of individuals with neurodevelopmental impairments. Data integration enhances our ability to detect statistically significant differentially expressed genes using linear models. Furthermore, the merging and batch effect correction of data enabled the comparison of diverse conditions that were not originally studied within a single experimental framework. For example, we successfully compared the language delay group to individuals with WS, despite these populations originating from distinct studies.

Although we were able to increase the sample sizes of individual groups, substantial variability remained within each group which can be attributed to several factors. Firstly, the differences between unique expression profiles of different people are inherited in the merged dataset and amplified after merging datasets with different experimental protocols and data collection procedures. The most important source of variation is the different platforms that were used which yield different expression distributions as is illustrated in Figure 10.

Secondly, there is the heterogeneity of the studied populations in each experiment. As an example, control groups and ASD groups may exhibit differences in ages, race, socioeconomic backgrounds, and environmental exposures, all of which can influence gene expression. These factors add extra variability in the already diverge expression profile data and highlight the need for a batch effect correction to mitigate confounding variables and improve the reliability of downstream analyses.

Data exclusion

Considering the dataset with the Tourette syndrome (GSE30470), we decided to exclude it from the merging process due to the lack of matched control samples. The inclusion of a microarray dataset with no controls presents a significant challenge to the interpretation of the findings. While we have employed a seemingly robust normalization and batch correction methods, the absence of a baseline comparison limits our ability to definitively assess the magnitude and direction of gene expression changes. Consequently, any conclusions drawn from this specific dataset must be interpreted with caution, focusing primarily on within-dataset comparisons and relative gene ranking. For this reason, the dataset was excluded, even though it could be interesting to compare the Tourette patients within that dataset to the general population and other conditions.

Imputation

KNN imputation was employed to address missing gene expression values resulting from variations in probe coverage between the various microarray platforms in the mined datasets. It is crucial to acknowledge that, unlike missingness due to technical variability or random missingness, these missing values are systematic. Consequently, imputed values may be influenced by nearest neighbors from distinct phenotypes exhibiting similar overall gene expression patterns. This potential bias could particularly impact differential expression analysis of genes with high imputation rates and gene set enrichment analysis, which relies on accurate gene ranking. For this reason, we decided to impute genes that have a relatively low missing rate of 30%.

Reducing the log2FC in the DE

The observed reduction in log2FCs following microarray dataset merging is an expected consequence of our analytical pipeline. Combining datasets, while increasing statistical power, inherently introduces variability from batch effects, platform differences, and biological heterogeneity. Subsequent batch effect correction and imputation steps, even if necessary for data harmonization, can obscure genuine biological signals and smooth extreme expression values, resulting in compressed log2FCs. Given this anticipated attenuation, we adjusted the log2FC threshold from 1.5 to 0.5 for the DEA. This adjustment acknowledges the increased noise and potential signal dilution inherent in the merged dataset, allowing us to capture biologically relevant, possibly less pronounced, differential expression patterns while mitigating the risk of overlooking subtle but meaningful changes. As a matter of fact, the presented results in Table 5 and the Venn diagrams in Sup_Fig. 9, show that genes with low corrected p-value can have diluted Log2FC that excludes them from categorized as significant.

Enrichment analysis

While from the differential expression analysis one can find genes that are significantly over or under expressed between groups, enrichment analysis further contextualizes these findings by comparing them to known sets of genes and revealing collective functions. Thus, enrichment analysis is crucial for the interpretation of the findings. In this work we decided to work with the GSEA methodology which ranks all genes in our dataset based on a specific metric before performing the comparison with the known gene sets. This decision was taken because of the low Log2FC from the DEA which is the result of the sequential normalizations and batch corrections.

In the case of the ASD-Control comparison, the biological processes of the GO terms are in accordance with bibliography where sensory function seems to be altered in the ASD population as well as processes related to splicing, transcription and translation. This is enhanced by looking at the enriched KEGG pathways in ASD, where we find terms related to RNA splicing and other regulatory pathways which are in agreement with the bibliography [133]. Also, pathways related to altered olfactory function and seem to be downregulated in our results, are reported in literature as significant for ASD and other neurodevelopmental disorders [134].

On the pair of language delay and Control, we see that the GSEA reveals both downregulated and upregulated biological processes in the LD group which are aligned with the findings of the implicating pathways related to cell cycle, inflammation, neuronal development, and metabolism. It is possible that these findings are affected by the sample tissues of the groups where the LD population comes only from leukocytes while in the Control population is the product of merging different samples of whole blood, peripheral blood and leukocytes.

Lastly, investigating the results of WS against the Control group, it was expected to find genes and processes that are related to the genes of the 7q11.23 locus since the syndrome is chromosome-specific and is associated with modification on this part of the genome. The significant upregulation of "MHC protein complex assembly" and "antigen processing and presentation" at both the GO and KEGG levels suggests an altered immune response in individuals with WS. MHC proteins are crucial for presenting antigens to T cells, and their increased assembly could indicate heightened immune activation or altered antigen presentation. This is further supported by the upregulation of pathways related to "Intestinal immune network for IgA production" and "Asthma", both of which involve immune dysregulation.

The upregulation of the GO term "response to ischemia" could point to altered cellular stress responses or vascular function, which aligns with known cardiovascular features of WS because of elastin deficiency [135]. The upregulation of "regulation of postsynaptic membrane neurotransmitter receptor levels" and "neuromuscular synaptic transmission" suggests altered neuronal signaling, which may relate to known cognitive and behavioral characteristics of WS [136].

Limitations and Future Developments

Despite the strengths of this approach, several limitations should be acknowledged and addressed with caution. One major limitation is that the data used for analyzing neurodevelopmental disorders (NDDs) are derived from peripheral blood, raising the assumption that gene expression in blood serves as a meaningful proxy for gene expression in the brain. While some studies have reported correlations between blood and brain gene expression, this relationship is gene-specific and influenced by various biological factors. As a result, findings based on blood-derived gene expression should be interpreted with caution.

Further, free extracellular RNA molecules are known to circulate in the blood [137], and the role of the blood-brain barrier has been widely studied [138], it remains uncertain whether these circulating molecules adequately represent brain-specific gene expression. Furthermore, their presence and abundance may not be sufficient to reliably indicate the existence of specific neurodevelopmental conditions since the samples are not taken directly from the affected tissue - in this case the brain.

In this work, we decided for reasons related to statistical power to neglect conditions in the datasets with less than 20 samples. This choice led to a reduced number of conditions in the merged set to three and the typically developed population. Additionally, in the merged dataset we allowed William's syndrome which is a neurodevelopmental disorder caused by known genetic factors (alternations in Chromosome 7q11.23) which makes this population a characterized group. This choice was made because William's syndrome group was first of all a sub-group in the study along with ASD and typically developed individuals except for satisfying the criterion of more than 20 samples.

This method of merging has the limitations of transforming the transcripts to gene symbols before the merging which leads to information loss since there might be multiple transcripts with the same gene symbol that could have been studied independently. The reason for this choice was that the microarray designs were diverse and the overlap in probesets and transcripts was sparse. This same reason prevented us from including conditions like Cerebral palsy in the merging process because the only available dataset was on circular RNA which is a type of non-coding molecules. As an addition to this, studies on NDDs that conducted with NGS technologies were not considered since the measurement method differs dramatically from microarrays. Thus, in the future it is important to develop methods that can work on the transcript level for merging and be able to integrate datasets both from arrays and sequencing methods.

As another future perspective, an extensive evaluation of the findings is needed in the level of differential expressed genes between groups in contrast to ML models and meta-analysis methods. This will show on one hand if ML models can work on the merged dataset and identify similar and/or different important markers, and on the other hand, the comparison with meta-analysis methods might show an enhanced biomarker discovery effect or a dilution of the differences due to multiple normalizations. Finally, in the future implementing methods for network reconstruction would help in the understanding of the interconnection of genes in the dataset and offer a different perspective of analysis based on graphs by finding key nodes/genes, clusters of genes which are co-expressed and might provide a better picture of enriched terms, and associations among genes of the dataset. Overall, these steps will be beneficial both for the interpretation of the data but also for the evaluation of the method of harmonizing indicating possible improvements.

Chapter 4

Machine Learning Algorithms for Biomarker Discovery in Cross-Sectional Clinical and Omics Data

Through this chapter, we are going to cover some of the ML methods used in the discovery of biomarkers in structured datasets such as expression metrices in bioinformatics. This chapter is also aligned with the thesis' objective O3 in which the goal is to investigate omics datasets with the use of multi-objective algorithms who can efficiently search the vast feature space of these data.

The shift from hypothesis-driven biomarker discovery to data-driven approaches marks a paradigm shift in computational biology, and while classical statistical techniques remain essential, modern ML-based and network-based methods have revolutionized the field by allowing for unbiased, large-scale biomarker identification. These advancements have an essential role in the analysis of complex diseases, where biomarkers are often subtle, polygenic, and context-dependent. The integration of multi-omics data, ML, and network-based computational methods is expected to further enhance precision medicine, enabling earlier diagnoses and more effective, personalized treatments. As computational biology continues to evolve, the synergy between big data, informatics, and biological research will be the key factor in identifying novel biomarkers and help uncover previously unknown mechanisms.

It is important to note that over the years, a vast number of methods have been developed including various artificial intelligence techniques, network-based methods, and classical or more sophisticated ML algorithms. Most of those methods are out of the scope of the present thesis which focuses on ML-related approaches with a greater focus on evolutionary algorithms (EA). To better understand EAs, it is essential to introduce the concepts of heuristic and metaheuristic methods, where the former is family of methods that EAs belong to.

Heuristic algorithms are a class of problem-solving approaches designed to deliver satisfactory solutions within reasonable computational times, especially when traditional, exhaustive methods prove too resource-intensive. By leveraging experience-based techniques and rule-of-thumb strategies, heuristics provide approximate answers to complex optimization and search problems—a necessity in fields such as biomarker discovery where datasets can be vast and multifaceted.

In many cases, heuristics facilitate efficient exploration of large or intractable search spaces by incorporating strategies such as greedy selection, simulated annealing, genetic evolution, and tabu search. These methods exploit inherent problem structures to quickly converge on acceptable solutions, even though they cannot always guarantee global optimality. However, their performance is often sensitive to specific problem parameters and initial conditions, and they may converge prematurely on local optima. This highlights the need for more versatile approaches when facing diverse and complex datasets.

While heuristics offer efficient, problem-specific shortcuts, metaheuristics provide a higher-level framework that systematically guides the overall search process across diverse optimization challenges. Unlike their heuristic counterparts, metaheuristics are designed to be generally applicable, combining various heuristic techniques to effectively explore the solution space while avoiding premature convergence on local optima.

Common metaheuristic methods include Genetic Algorithms, Simulated Annealing, and Particle Swarm Optimization [139]. These techniques balance exploration and exploitation by integrating mechanisms that allow them to adapt and fine-tune their search strategies across a wide range of applications, including the optimization challenges inherent in biomarker discovery. This robustness makes metaheuristics particularly valuable for developing machine learning models capable of navigating complex, high-dimensional spaces. Many of these metaheuristic methods simulate natural mechanisms to more efficiently explore the feature space of a given problem. Building on this idea, evolutionary algorithms have emerged as a distinct family of methods inspired by biological evolution. In the following section a broader explanation for this family is given as it is one of the interests of this thesis (O3).

4.1 Evolutionary Algorithms

Evolutionary algorithms is a family of algorithms that share the main operation principles following natural selection, where a population of potential solutions competes within a resource-limited environment. This competition drives the selection of fitter individuals, leading to an overall improvement in population quality [140]. EAs or evolutionary computation, include the genetic algorithms (GA), differential evolution, and genetic programming among others, where each of those techniques is a different variation of the main principles.

At the beginning, a set of candidate solutions are randomly generated. These individual solutions can be seen as individuals within a population with their own unique encoding. In the case of GAs, the solutions are seen as chromosomes, and every feature of the chromosome is a gene. After the initialization of the solutions, the individuals are evaluated using a predefined, or a set of predefined objective functions. These functions serve as a measure of fitness, with higher values indicating better solutions. Based on these fitness scores, the most promising candidates are selected to produce the next generation through recombination and mutation [141].

Recombination involves combining characteristics from two or more selected candidates (parents) to create new candidates (offspring). Mutation, on the other hand, introduces random changes to a single candidate, generating a modified version. The newly generated candidates are then evaluated for fitness and compete with the other individuals of their generation for survival. This iterative process continues until a satisfactory solution is found or computational limits are reached.

Evolutionary algorithms rely on two fundamental mechanisms. The first mechanism is *Variation* which includes the recombination of parental solutions and the mutation of offspring. Variation introduces diversity in the population which promotes innovation and better search of the feature space while most of the times this variation may lead to unnecessary complexity or even decrease in the fitness. The second mechanism is *Selection* where the fittest individuals are promoted to produce offspring in the next generation. Selection can improve the solution quality over generations but may also reduce the innovation. For this reason, EAs usually implement both mechanisms in balance to achieve both the improvement over the generations and the feature space exploration [141] (Figure 13).

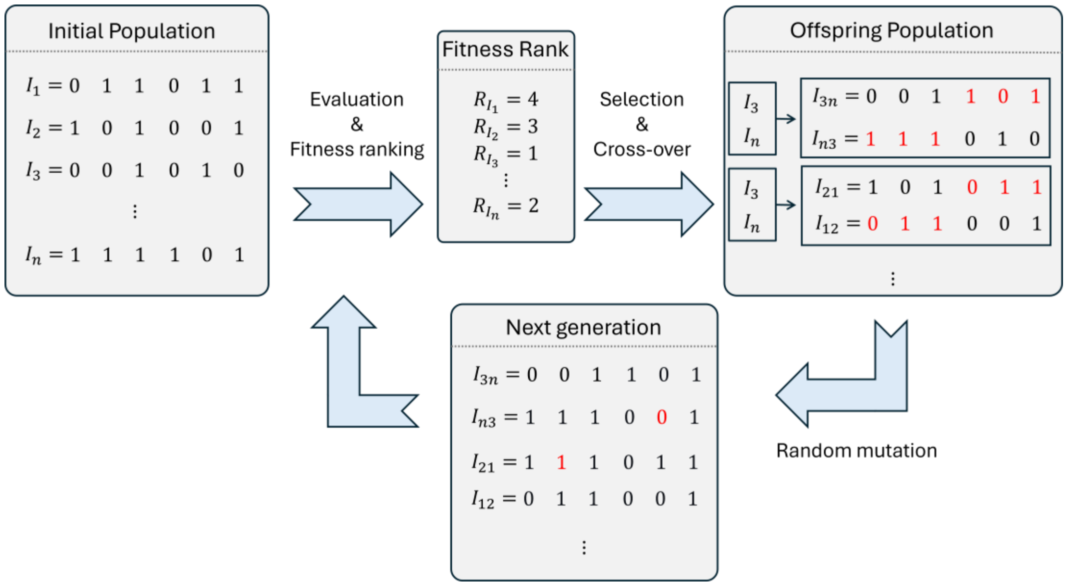


Figure 13. Fundamental processes of an evolutionary algorithm. Starting from a random initial population (left) the solutions are getting evaluated and ranked (midle). A selection methods matches the parental solutions which will produce the new generation through recombination (right). Finally, the new generation will undergo a random mutation process (bottom) and the cycle will continue untill the termination criteria met.

Although EAs are gradually increasing the performance within the population, it is worth noting that they do not guarantee the finding of the best solution. In fact, some of the applications these algorithms are used in do not have a single best solution, but rather a set of feasible solutions. This, including the stochastic character of EAs makes them extremely useful in complex problems and in vast feature spaces [142].

Overall, this process can be interpreted as an optimization strategy, progressively refining solutions toward an optimal outcome. Alternatively, it can be viewed as an adaptive mechanism, where fitness represents how well solutions meet environmental constraints. Over time, the population becomes increasingly well-suited to its given context.

4.2 A Mult-objective Evolutionary Algorithm for Biomarker Discovery Application

In the context of using ML for biomarker discovery in bioinformatics datasets, a pipeline for omics data was developed during this research line. For this purpose, a multi-objective EA created where the objectives except for higher performance metrics focus on the reduction of the model sizes. Thus, the algorithm tries to retain a compact model both in terms of the number of features (genes or transcripts) used, and in the complexity of the model itself. This method is using an XGBoost classifier in its core to train the predictive models and in parallel a Pareto frontier to find the fittest solutions. For avoiding early convergence and promoting the feature space search, similar solutions are degraded and penalized based on their similarity using proximity metrics. At the end of the evolutionary process, a set of feasible solutions which ended up in the first Pareto frontier is returned to the users.

The content of the following subsections is based on the published work:

MEvA-X: a hybrid multi-objective evolutionary tool using an XGBoost classifier for biomarkers discovery on biomedical datasets.

Bioinformatics, Volume 39, Issue 7, July 2023.
Doi: 10.1093/bioinformatics/btad384

4.3.1 Abstract

Motivation: Biomarker discovery is one of the most frequent pursuits in bioinformatics and is crucial for precision medicine, disease prognosis, and drug discovery. A common challenge of biomarker discovery applications is the low ratio of samples over features for the selection of a reliable not-redundant subset of features, but despite the development of efficient tree-based classification methods, such as the extreme gradient boosting (XGBoost), this limitation is still relevant. Moreover, existing approaches for optimizing XGBoost do not deal effectively with the class imbalance nature of the biomarker discovery problems, and the presence of multiple conflicting objectives, since they focus on the training of a single-objective model. In the current work, we introduce MEvA-X, a novel hybrid ensemble for feature selection (FS) and classification, combining a niche-based multi-objective EA with the XGBoost classifier. MEvA-X deploys a multi-objective EA to optimize the hyperparameters of the classifier and perform FS, identifying a set of Pareto-optimal solutions and optimizing multiple objectives, including classification and model simplicity metrics.

Results: The performance of the MEvA-X tool was benchmarked using one omics dataset coming from a microarray gene expression experiment, and one clinical questionnaire-based dataset combined with demographic information.

MEvA-X tool outperformed the state-of-the-art methods in the balanced categorization of classes, creating multiple low-complexity models and identifying important nonredundant biomarkers. The best-performing run of MEvA-X for the prediction of weight loss using gene expression data yields a small set of blood circulatory markers which are sufficient for this precision nutrition application but need further validation.

The code for this tool is available at: https://github.com/PanKonstantinos/MEvA-X, and an implementation can be found in the developed platform for the purpose of this thesis and is thoroughly discussed in Chapter – 6.2. The platform is accessible at: http://130.192.23.168:8501

4.3.2 Introduction

Due to the exponential increase of computational power in the last decades, life sciences and biology increasingly rely on informatics to tackle the complexity of the systems they examine. In addition, in the time of -omics, where the feature space is vast and the number of samples is usually small, finding reliable nonredundant biomarkers is one of the most frequent quests for scientists [143]. Biomarkers are useful indicators for the early detection of various pathologies such as neurodegenerative diseases, and different types of cancer and can help in the surveillance of the progression of these pathologies with low-cost and minimally invasive techniques [144], [145], [146]. Even with previous technologies such as microarrays, the number of genes detected - ranging from 2000 up to more than 20 000 - is many times greater than the number of samples [147], [148]. Bioinformatics emerged from these needs, aiming to bridge the gap between those fields by introducing new computational tools tailored to the demands of biomedical applications [149].

In the last decades, many algorithms have tried to deal with the complexity of real-life problems which possess multiple and conflicting objectives. Hybrid wrapper and ensemble classification techniques have been developed by combining EAs and machine learning methods such as k-nearest neighbors, artificial neural networks [150], and others [151], [152]. Heuristic and metaheuristic techniques have been introduced to search for acceptable solutions in a wide feature space without providing any mathematical proof for the optimality of the revealed solutions [148], [153]. Inspired by the theory of evolution in nature, EAs are among the most used methods in these cases with multiple variations being proposed over the years [154], [155], [156], [157]. Multi-objective EA can be combined with Pareto techniques that apply selective pressure to yield a set of equally good solutions, instead of trying to find the global best, since the nature of these problems does not allow for a trivial or easy way to converge to one solution only [158]. Moreover, it is important to allow the exploration of a wide set of the dominant solutions of the Pareto optimal set to avoid premature convergence in non-optimal solutions. One of the successful techniques that have been proposed to deal with these requirements is niching, where solutions of the same Pareto set get penalized according to their similarity [159]. Recently, a novel gradient-boosting technique called XGBoost has been proposed, and since then it has gained a lot of attention in computer science challenges hosted by Kaggle competitions [160].

It has been demonstrated that the combination of an XGBoost model with a genetic algorithm (GA) can substantially enhance its performance by allowing for the search for the optimal hyperparameters of the classifier [156]. Despite XGBoost classifiers performing reasonably well with their default parameters on various datasets, optimizing their parameters in an unbiased way is important for achieving higher performance. XGBoost possesses many hyperparameters and their effect on the performance of the trained models is important. Deng et al. (2022) proposed a similar method where the XGBoost algorithm was used as an ensemble-based feature selection (FS) method to get a subset of the initial number of genes, which later they used as input for the GA they use for training multi-objective models [161]. Even though this approach showed promising results, as a greedy algorithm, XGBoost may neglect genes that might have some statistical meaning. In their work, Corthésy et al. (2018) implemented a similar solution to the proposed method, but with a different classifier [154].

XGBoost has already been applied alone in many bioinformatics and biomedical applications [162], [163] while the use of metaheuristics for optimizing their hyperparameters has also been successfully performed in many other applications [164], [165], [166]. However, most of these approaches do not deal with multiple objectives and are not able to effectively handle the data missingness, the class imbalance, and the high dimensionality of the data which are used for the discovery of predictive biosignatures. Multi-objective EAs have already been applied for this purpose but without using XGBoost as a classifier. For instance, a support vector classifier was used as an estimator [154], neglecting at that time the potential discrimination power of boosting algorithms. In a more recent application, a novel framework, named AUTODC [167], has been introduced and tested for disease classification using other boosting and tree-based classification and optimizing them with a novel heuristic method based on a two-layer Multi-Armed Bandit framework. This method outperformed existing tools for biomarker discovery by effectively selecting features and hyperparameters for the classification datasets. However, it neither included XGBoost in the classification methods nor provided a solution for multiple objectives optimization, class imbalance, or multiple solutions for the same problem.

In the present work, we introduce MEvA-X, a multi-objective hybrid ensemble EA framework for optimizing the hyper-parameters of an XGBoost classifier for biomarkers discovery applications. The boosting algorithm has been selected as a binary discriminator, and its hyperparameters get optimized over the generations of the EA. In the presented method, we use a niched Pareto frontier scheme, which helps to conserve the diversity of the solutions by distributing the population over several different peaks (niches) and avoiding premature convergence of the algorithm [168]. This allows for a general-purpose biomarker discovery tool that works with numerical features not only for omics but also for clinical features. MEvA-X was benchmarked against other established state-of-the-art methods, such as the XGBoost algorithm alone, Random Forests, and XGBoost combined with other standard FS techniques to validate its performance.

Two publicly available datasets were used to test the proposed method for evaluation purposes. The first dataset is from a study that investigated the relationship between weight loss through lifestyle interventions and gene expression profiles in peripheral blood [169], [170]. This study collected data over the span of 1 year from people with high cardiac risk. Our interest was to use the baseline measurements at the beginning of the study and create models to predict if a person will lose weight with the examined intervention, as well as to find biomarkers that can help us decide if this intervention will be suitable for other subject based on their gene expression profile. Toward this goal, MEvA-X retrieved 24 solutions from the dominant Pareto frontier and exceeded 75% receiver operating characteristic (AUC) with the best performing one including only nine selected genes and significantly improved the classification of the minority class compared with the state-of-the-art methods considering that this is an imbalanced dataset with a very low samples-to-features ratio.

The second dataset is linked to the retrospective analysis of a population study on 631 chronic pain patients who were prescribed opioid painkillers, and who were treated with topical analgesics formulations. The information contained in this dataset comes from the answers of the subjects in the Brief Pain Inventory (BPI) at the beginning (baseline) and at the follow-up period, and it is encoded in numerical scaled values [171]. The study aimed to document changes in 3 months in four categories: BPI pain severity, pain interference relating to Quality-of-Life components, other medication usage, and the qualitative health complaints of these patients [172].

In this dataset, MEvA-X achieved high improvement for all four labels, since the imbalance is even greater which favors the proposed method compared with other techniques. There was an increase of 16%–32% in the weighted geometric mean (wGM) in absolute numbers and an 8%–10% increase in balanced accuracy, while 8–19 features were used for the four labels.

Several configurations and combinations of the parameters for the EA were tested to evaluate the whole pipeline and to assess the robustness of the method against the baseline and the models with a selected subset of features (for reference see in Appendix - A Chapter - 4). In all these cases, we found that MEvA-X classification outperforms the state-of-the-art methods while providing a non-redundant set of features as biomarkers.

4.3.3 Materials and methods

4.3.3.1 Datasets

The first dataset was obtained by merging two different data sources downloaded from the Gene Expression Omnibus (GEO) databank. In greater detail, the accession number GSE66175 [169], [170] contains 26 patients, plus another 63 patients who are collected from an older dataset (GSE46097) [173], and 71 control subjects. The above-mentioned 89 patient subjects in total were people having either a coronary artery disease (CAD) event or being at high risk of developing CAD, following a therapeutic strategy characterized by a combination of an active lifestyle and diet modification targeting a relevant change in their weight in 3 and 12 months. The active lifestyle changes consisted of 3 h/week of aerobic exercise and 1 h of daily stress management, while the diet modification refers to a low-fat vegetarian diet known as Ornish. Blood samples were drawn at the beginning of the study (baseline) and later during the re-examination periods. The gene expression levels in peripheral whole blood samples were measured for these patients using the Affymetrix Human Genome U133A 2.0 Array platform.

The two patient cohorts contained in GSE66175 have been unified by a batch effect correction method [174], and the duplicated gene names were grouped, taking their average values. Therefore, the merged dataset consists of 89 patients. In this dataset, weight loss information transformed into a binary label: participants with a weight loss higher than a selected threshold equal to 10% were considered to belong in the positive class (Responders), whereas participants with lower weight loss or even increased weight were considered to belong in the negative class (Non_responders). Based on the above-defined threshold, the label distribution of the dataset is 35 “Non_responders” and 54 “Responders,” with 13 239 unique gene and transcript names.

The second dataset included data from the Optimizing Patient Experience and Response to Topical Analgesics (OPERA) study. The OPERA dataset consists of 631 patients with chronic pain in the intervened group, answering the BPI-validated questionnaire along with some supplementary questions before and after a follow-up period for a total of 50 survey questions as features [172]. Therefore, this is a dataset made of 631 patients each with 50 features.

The target of the researchers releasing the OPERA dataset was to investigate the effect of replacing opioid therapies with topical analgesics on patients with chronic pain and record the changes in multiple aspects of their life such as pain and medicine reduction, interference with everyday activities, and reduction of complaints. Consequently, four different labels in this dataset refer to the changes the scientist measured during the study (for reference see Sup_Table 9), and so, all four of them were considered as different datasets and used to validate the performance of the proposed method.

4.3.3.2 Methods

A hybrid ensemble algorithm for FS and classification, based on a multi-objective EA has been developed and introduced following the procedure in Figure 14. MEvA-X optimizes the hyperparameters of an XGBoost Classifier while selecting nonredundant features (potential biomarkers) to reduce the dimensionality of the given problem. The presented method follows four main individual steps which are:

data preprocessing, training of individual models, evaluation of the trained models, and population update based on the evolutionary processes.

Specifically, MEvA-X, except the main framework of the EA with the known operators (selection, crossover, mutation), implements a niched Pareto frontier ranking scheme which makes it appropriate for searching big feature spaces without converging to one local minimum [168]. The niches are essentially different peaks (local maxima) in the objective function and this method keeps a relative balance between the distribution of the solution in these valleys to preserve pluralism and promote the searching of the feature space. After this ranking, the selection of solutions indicates which of the individuals in the population will pass their genes to the next generations through their offspring by recombining their chromosomes in pairs in the crossover operation. The mutation operator is also allowed in the offspring introducing further stochasticity in the process to allow better exploration of the search space.

The preprocessing steps of the pipeline include the automatic read of the dataset and the extraction of feature and sample names, the transformation of nominal values to numeric, imputation of missing values, normalization of the features, and merging of duplicated features.

4.3.4 The MEvA-X evolutionary framework

4.3.4.1 Feature selection

As an additional preprocessing step, features are selected using four different univariate and multivariate FS methods, namely, SelectKBest, Wilcoxon Rank Sums, Joint mutual information (JMI), and Minimum Redundancy Maximum Relevance (mRMR) [175], [176]. This way some of the solutions of the population can randomly select features from a lower-dimensional space which is based on statistical methods and can help in creating some niches (local minima) to improve the search for good feasible solutions. In the chromosome of each solution, there are specific “genes” that drive the individual to select features based on some FS methods and other “genes” responsible for the parameters of these methods.

4.3.4.2 Initialization of population

The population of the first generation is generated in a pseudo-random manner by selecting values for the parameter genes following a uniform distribution for the range of the minimum and maximum allowed values for every parameter. The selection and activation of feature genes is also a random process, but the initial total number of active genes in any individual chromosome is constrained to less than 30, which are then filtered according to the FS genes.

4.3.4.3 Stratified cross-validation data splitting

A stratified 10-fold cross-validation scheme with a different random state in every generation is used in MEvA-X, splitting the data into training and validation sets to provide concrete results of the trained models' performance dealing effectively with the imbalanced nature of most of the clinical and biological datasets. In this way, the results are less prone to biases caused by single splits, and the ratio of the positive and negative classes is preserved while keeping the same random state across the given generation allowing for comparable results between the solutions.

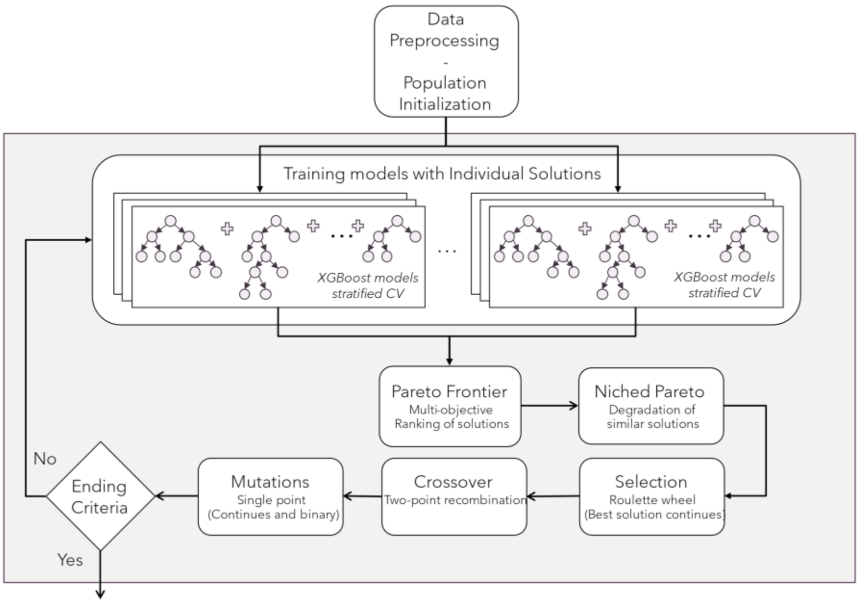


Figure 14. Flow chart of MEvA-X. The evolutionary process begins with data preprocessing and population initialization. The collection of individual solutions that encode the information in the form of chromosomes is used to train independent ensemble models using the XGBoost classifier in a 10-fold cross-validation framework. The solutions are ranked in frontiers through the Pareto Frontier method and similar solutions belonging in the same niche are degraded. The evolutionary operations (selection, crossover, and mutation) apply on the solutions and if the end criteria are not met the procedure starts over.

4.3.4.4 Classification and hyperparameter tuning

In the presented method, XGBoost classifiers are used to discriminate the samples of the dataset according to the labels. The instructions to build each classifier are encoded in every individual solution in the last seven parameter genes (see Sup_Table 10). The information contained in the chromosomes included seven hyperparameters of the classifier, namely the learning rate, the maximum depth of each tree, two pruning parameters, two generalization parameters, and a balancing parameter. The selected XGBoost hyperparameters were chosen to ensure a balance between fast convergence and robust generalization. Specifically, parameters such as *learning rate*, *max depth*, and regularization terms (*lambda*, *alpha*) are used to tune and accelerate training while minimizing the risk of overfitting. The selected value ranges were chosen based on prior literature and recommendations provided by the development team.

In every iteration of the cross-validation, the XGBoost classifier denotes the performance of both training and validation sets on the area under the AUC curve. In cases where the validation AUC does not improve for more than 50 iterations, the training is terminated to avoid overfitting, and the algorithm returns the ensemble up to the last best validation iteration.

4.3.4.5 Network creation and bioinformatics analysis

The analysis of the data was made with the programming languages Python version 3.9 and R version 4.1.2. For the Ornish diet dataset, a co-expression network constructed with the help of InSyBio BioNets [114] tool using the Spearman correlation [177] and node PageRank centrality [178] metrics for the selection of the most significant nodes. The visualization and further analysis of the network were done with Cytoscape (Shannon et al. 2003) version 3.9.1 [113] and the tissue specificity analysis was conducted with the GTExPortal [179]. Additionally, the Spearman correlation and the principal component analysis (PCA) were also calculated for the selected features of the MEvA-X models for both datasets. The results of these analyses can be seen in Figure 16.

4.3.4.6 Fitness functions

Based on the evolutionary operators, the solutions of the population “compete” to improve their fitness and survive in every iteration/generation of the algorithm, and the competition is held based on the objectives the EA ultimately tries to optimize. In MEvA-X, the multiple objectives mostly refer to the metrics used to evaluate the performance of the trained XGBoost classifier models, which are coordinated by the chromosomes of the individual solutions in the population. The coordination is done by the parameter genes and the feature genes of the solutions, which are the blueprints for the construction of the discrimination models.

The proposed method is designed to maintain a balance between the high discrimination performance and the low complexity of the models. Thus, the evolutionary pressure drives the population towards more sparse chromosomes with as few active feature genes as possible, while maintaining high classification performance. Feature_model_complexity and Split_model_complexity are used to measure the complexity of the trained models. The value of these metrics is lower for complex models and higher for simpler ones. Particularly, the more active gene features a solution has, the lower the feature model complexity metric will be. Similarly, a model with fewer splits and a simpler structure will have a greater split complexity score than a model with more trees in the ensemble and many splits.

On the other hand, metrics that evaluate discrimination of the models based on the prediction and the probabilities of the predictions of the classes are used as well. More specifically, accuracy, wGM, F1 score, F2 score, precision, recall, balanced accuracy, and the AUC were considered as evaluation metrics.

Finally, the overall score and the weighted overall score are also calculated to give an easy comparison between the solutions. The overall score considers each of the previously referred metrics as of equal importance, while the weighted overall score is a user's choice array of weights mapped to the metrics based on the importance of each one of them. This way, the user of MEvA-X has an additional degree of freedom to drive the population toward solutions with higher scores on the preferred objectives.

4.3.4.7 Niched Pareto frontier

The trained models are ranked based on their multiple evaluation metrics through a Pareto Frontier operation [158]. A niched Pareto frontier ranking approach has been selected like the proposal of Erickson et al. (2001) to avoid the premature convergence of the algorithm and for a more exhaustive search of the feature and parameter space [180]. This way similar solutions that belong to the same Pareto front get penalized even if they reach a very high score. In this manner, the pluralism of solutions is guaranteed during the passage of generations and local maxima are avoided.

In MEvA-X, there is a hybrid approach of binary and continuous encoding on the chromosomes, and so the calculation of the distance between two solutions belonging to the same Pareto frontier is a two-step process. For measuring the difference between two individual solutions belonging to the same Pareto front, we introduced a dual distance metric. One part of the metric refers to the continuous values of the parameters to be optimized by calculating the Euclidean distance between the hyper-parameters of solutions, and the second to the binary selection of features of the dataset (see in Appendix – A, Chapter – 4). According to the closeness of the solutions, a degradation/penalizing function is applied for similar solutions to reduce the possibility of the algorithm accumulating all solutions in a local minimum.

The applied evolutionary operators, termination criteria, and additional details on the evolutionary framework of MEvA-X are provided in the Appendix – A, Chapter - 4.

4.3.5 Training final ensemble classification models

When at least one of the stop criteria is fulfilled, the EA stops operating. Then, the final solutions are ranked once more with the Pareto frontier method. The models that end up in the first Pareto frontier are the dominant ones and the solution with the highest overall score is the best compromise. Except for the overall best, it is possible to take advantage of the contradictive objectives and so the dominant solutions are combined to create an ensemble of classifiers with two majority voting methods. In the first method, the final prediction of the ensemble is the one that has the highest number of votes from the individual models in the ensemble, which is also known as “hard” majority voting. The second method implied is the so-called “soft” majority voting which considers the probability the XGBoost classifier gives to the prediction. In both cases, a filtering of the solutions participating in the voting is taking place, because some of the models end up in the first Pareto frontier simply because they are very low in complexity, but they might also have no discrimination value to add to the ensemble, and so solutions that classified all instances in the same class in the cross-validation, are excluded from the majority voting.

4.3.5 Results

The performance and robustness of MEvA-X have been compared with the existing state-of-the-art XGBoost classifier by several experiments that were conducted with both datasets (Ornish diet and OPERA) using the stratified 10-fold cross-validation framework. In these experiments, MEvA-X's performance was benchmarked against single XGBoost models (baseline) and XGBoost models trained on a subset of the features with the FS techniques of Wilcoxon ranked sum, SelectKBest, JMI, and mRMR. For comparison with the state-of-the-art methods, a radar plot was used to visualize the improvement of the models created by MEvA-X in terms of simplicity and balance in the discrimination of classes (Figure 15).

A co-expression network of the selected features and their neighborhoods was reconstructed, the distribution of the features was visualized, a correlation heatmap of the selected features was created, and a tissue specificity analysis was conducted for the revealed Ornish diet weight-loss prediction biosignature to interpret and explore the patterns and associations between the gene expression biomarkers (Figure 16).

Additionally, enrichment analysis has been conducted to reveal any underlying pathways but for the given set of biomarkers, no term was found enriched.

To benchmark the MEvA-X against another established method, we used Random Forests combined with a grid search for parameter optimization. Comparative results, as shown in Sup_Table 12, demonstrated that MEvA-X was able to substantially increase classification metrics, with a balanced accuracy increase from 61% to 76%.

The proposed algorithm was tested also on the four chronic pain-related endpoints of the second (OPERA) dataset using data from a questionnaire survey. As before, radar plots were used to depict the evaluation metrics of the different frameworks against the MEvA-X tool (Figure 17).

A Spearman's correlation feature association analysis was conducted for the features selected for all four endpoints of the dataset as shown in the heatmaps in Figure 18. For the individual endpoints, MEvA-X selected a different set of features. For the Severity Change endpoint, a subset of 14 features was selected as important, while for the second endpoint (Change in Interference), 18 features were selected (Sup_Table 15). Regarding the third endpoint in the dataset (Change in

Medicines), the algorithm ended up in a subset of eight features, and finally, in the last label (Change in Complaints) 15 features remained in the subset (Sup_Table 15).

MEvA-X overall increased the performance in the classification of the minority classes and on top of that created models with features that have low intercorrelation, leading to fewer non-redundant markers and relatively simple models.

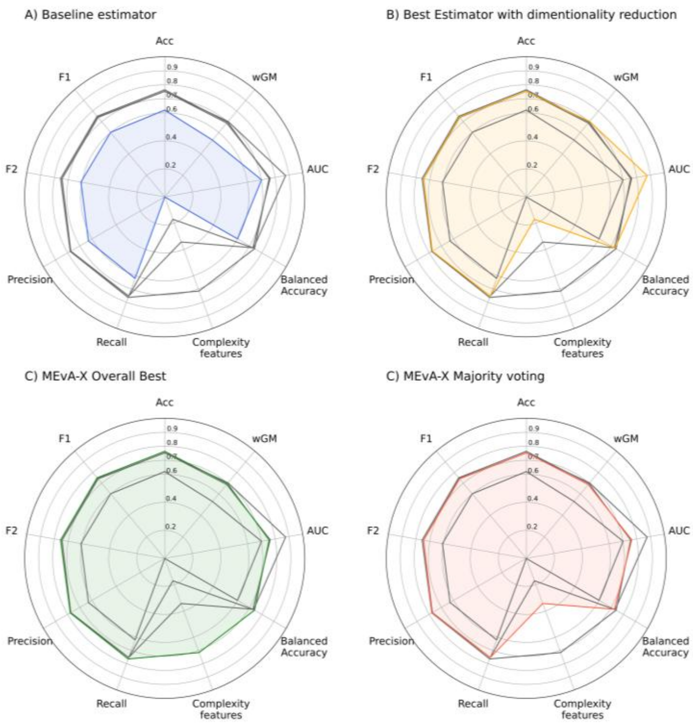


Figure 15. Comparative radar plots between the average performance of models for the Ornish diet dataset. (A) Simple XGBoost estimator, (B) best model with FS applied (JMI with k = 5), (C) MEvA-X models with the highest overall metric in the population, and (D) majority voting of the solutions in the first Pareto frontier. The metrics and their standard deviations calculated from the 10-fold cross-validation analysis are provided in the Sup_Table 12. Each panel colors the metrics of each presented method and shows in gray the rest.

4.3.6 Discussion

From the obtained results, it was shown that the MEvA-X algorithm is beneficial in two major aspects of biomedical classification problems. Optimizing the hyperparameters and features of the XGBoost classifier using the multi-objective optimization framework of MEvA-X resulted in the improvement of classification metrics, that are appropriate for imbalanced datasets (e.g. wGM) compared with XGBoost, XGboost coupled with FS methods, and grid searched optimized Random Forests. In both datasets used in the present work, the classification metrics significantly improved without a substantial corresponding decrease in the rest of the metrics. Moreover, MEvA-X increased substantially the simplicity of the final models and especially the model with the highest overall score yield by MEvA-X. The decrease in the complexity of the model is up to 60% in absolute numbers, meaning that the algorithm can identify features that improve the classification of the data without keeping redundant and low-informative ones. Especially for -omics datasets, such as the Ornish diet dataset, except for the very high features–samples ratio problem, there is high redundancy due to the co-expression of genes, making the identification of nonredundant biomarkers a challenge that our algorithm proved able to overcome.

Regarding the performance of MEvA-X on the correct classification of the minority class, it is superior to other benchmarked methods (Sup_Table 14). Without the use of the EA, most of the instances are classified as the majority class resulting in very low wGM and balanced accuracy scores. Our proposed algorithm can correct this misclassification in these difficult and imbalanced problems.

It is worth mentioning that the use of standard FS techniques alone was not proven to be efficient for both datasets. In the OPERA dataset, JMI and mRMR created very simple models of limited discrimination ability, while SelectKBest and Wilcoxon's rank sum had poor performance in both datasets. XGBoost with no FS performed much better than the FS techniques for the OPERA dataset. Rapakoulia et al. (2014) showed that EAs optimized classification models are beneficial for problems with missing values when majority voting is used [151]. A similar approach was adopted in MEvA-X, which enables the identification of multiple similarly performing prediction models. Majority voting was applied but the performance of this metaclassifier did not outperform the prediction performance of the best-performing model for each dataset, but no missing values were present in our datasets. Nevertheless, the training of multiple models with similar classification performance is extremely important for biomedical applications since different measurements are conducted in different patient cohorts with high missing values rates.

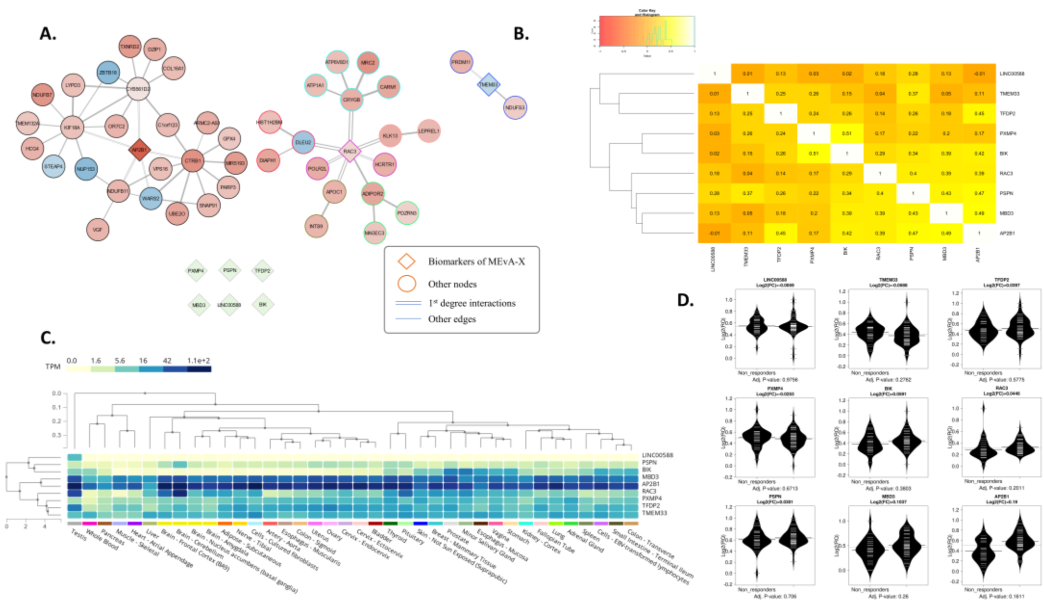


Figure 16. Bioinformatics analysis of the MEvA-X precision nutrition biosignature. (A) The co-expression network of the selected features and their immediate and two-steps-away neighbors, with the border color of the nodes representing the clusters in the network. (B) Correlation matrix of the selected features using Spearman's correlation. (C) Expression of selected genes on different tissues and organs (https://gtexportal.org). (D) Violin plots of the distribution of the selected features based on the label (responders/non-responders).

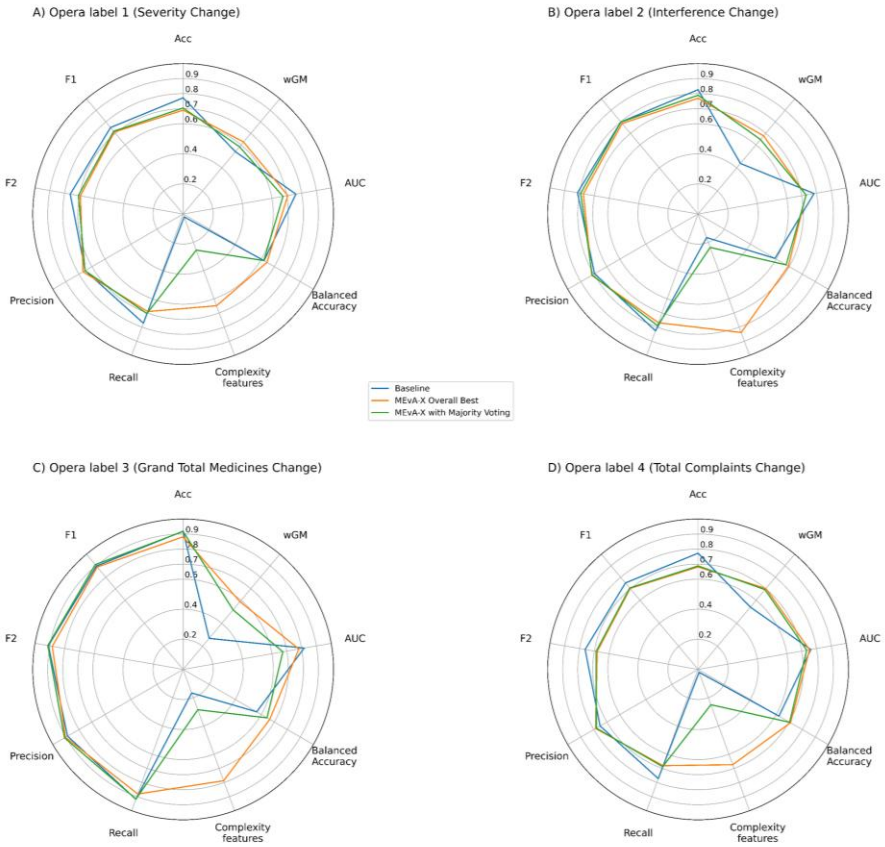


Figure 17. Comparative radar plots between the baseline and the MEvA-X solutions for the four endpoints of the OPERA dataset. MEvA-X outperforms the baseline models, both in simplicity (number of features used) and in the discrimination power of the minority class. (A) Endpoint related with the change in pain severity over the follow-up period. (B) Endpoint related to the difference in the interference of pain in everyday tasks over the follow-up period. (C) Endpoint showing the change of the total prescribed medicine to patients over the follow-up period. (D) Endpoint depicting the change of complaints of the patients over the course of the follow-up period. The metrics and their standard deviations calculated from the 10-fold cross-validation analysis are provided in the Sup_Table 14.

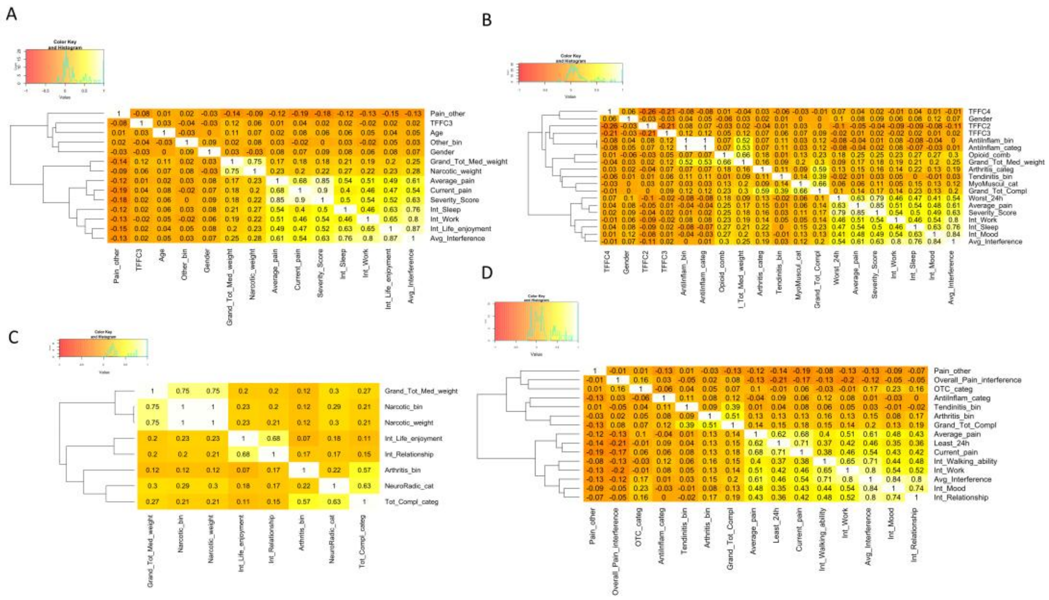


Figure 18. Correlation of the selected features by MEvA-X, for the OPERA dataset for the four labels. (A) Label 1 “Total Severity Change.” (B) Label 2 “Interference Change.” (C) Label 3 “Grant Total Medicine Change.” (D) Label 4 Total Complaints Change.” Spearman’s correlation method was used, and plots depict Spearman’s Rho coefficient.

In the precision diet dataset, further bioinformatics analysis was performed on the selected biosignature set to interpret the results (Figure 16). Based on the correlation analysis of the biosignatures, it appears that there is little correlation between them, indicating the effective selection of nonredundant features by the tool. This was confirmed when attempting to perform pathway and functional enrichment analysis with no term being significantly enriched in the revealed biosignature. Additionally, the co-expression network of the selected features with their first- and second-degree neighbors was reconstructed. Three of the selected features of MEvA-X (AP2B1, RAC3, and TMEM33) were identified as hubs according to a PageRank centrality-based analysis for their subnetworks, suggesting that these features could potentially be markers to indicate if this specific diet would be beneficial for a patient. Adapter-related protein complex 2 subunit beta 1 (AP2B1) encodes one of the large chain components of the assembly protein complex 2 whose functionality is to protein transport via transport vesicles in different membrane traffic pathways. Rac Family Small GTPase 3 (RAC3) encodes a protein that according to Gene Ontology is involved in GTP binding and calcium-dependent protein binding. Transmembrane Protein 33 (TMEM33) encodes a protein that is involved in the structural constituent of the nuclear pore and was found to maintain intracellular calcium homeostasis [181]. Furthermore, from the tissue-specificity bioinformatics analysis, it is observed that these genes are not specific in metabolism-related tissues and organs. The non-coding LINC00588 and the BIK gene are gender-specific since they are mostly expressed in the testis and prostate, but the dataset is balanced between male and female subjects that eliminate the sex bias, so their changes are most likely explained from DE between responders and not responders within the male group. Even though most of the selected genes have a notable expression in the pancreas, thyroid, liver, and stomach, no previous association has been made between these markers and weight loss except one study identifying AP2B1 as a potential marker for eosinophilic gastroenteritis [182], linking it thus indirectly to weight loss. These genes are associated with completely different molecular functions and pathways while their independent predictive potential is small with univariate statistical analysis showing marginal significance or no significant changes between responders and non-responders (Figure 16D). This suggests that the MEvA-x method was able to generate a highly accurate biosignature combining weak independent features, without being limited by the inherent assumptions of parametric tests for DE and removing redundancy from the selected features.

MEvA-X has the potential to become part of the Artificial Intelligence (AI) tools arsenal existing for solving difficult medical and biological problems. Advancements in AI have already allowed algorithms to be used in translational research applications having prospects in many diseases such as cancer, diabetes, and others [183], [184]. Another very promising field for biomarker discovery is neurodegenerative diseases with substantial progress being made lately in conditions such as Alzheimer's disease [185]. This tool can help in the identification of potential biomarkers and can become part of a pipeline for the exploration of neuro diseases.

Data availability

The Ornish diet dataset was obtained from National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) and is accessible through GEO with Series accession number GSE66175. The OPERA dataset was provided by Clarity Science LLC and is available upon request to the corresponding author with the permission of Clarity Science LLC.

Supplementary data are available at the end of this thesis in the Appendix - A section under the Chapter – 4 subsection or through Bioinformatics online at https://academic.oup.com/bioinformatics/article/39/7/btad384/7199580.

Chapter 5

Machine Learning driven risk predictor by longitudinal patient follow up: the case of cardiovascular risk in Thalassemic patients

This chapter addresses the problem of risk prediction in healthcare and the early identification of individuals at high-risk and finding clinical markers that can help in the early categorization of those patients through machine learning models.

In chronic diseases requiring continuous monitoring - such as cardiovascular complications in Thalassemia patients - the ability to track disease progression over time is essential for predicting adverse outcomes. Longitudinal patient data, which captures multiple observations across different time points, provides an opportunity for more robust predictive modeling. By taking advantage of those observations – hospital visit follow-ups – we can use ML techniques to predict risk trajectories early and provide valuable prognostics to doctors for interventions. In this case the biomarkers are not coming from a vast feature space, but as an effort to find when combinations of clinical characteristics can indicate a future health state. In other words, we try to find the pivot point from one condition state to another by investigating sequential records.

In this chapter we examine the risk prediction through an application related to a question frequently posed by hematologists managing patients with thalassemia which is: “what the risk and mechanisms underlying the development of CVD as a complication of the condition?” While thalassemia is primarily recognized as a genetic disorder affecting hemoglobin production, its systemic effects, including iron overload and chronic inflammation, contribute significantly to the elevated risk of CVD observed in these patients. A brief overview of the biological and clinical context of thalassemia will be provided to establish the medical relevance of this challenge and to frame the scope of the discussion.

At its core, this chapter demonstrates the application of ML methodologies to predict mid-term CVD risk in thalassemia patients. A classification model is developed, showcasing how ML can deliver accurate and actionable predictions.

The discussion will guide readers through the process of model design, feature selection, training, and evaluation, highlighting the interplay between data science and clinical practice. By addressing the limitations of traditional risk assessment methods, this chapter underscores the potential of ML to enable more precise and personalized risk stratification for thalassemia patients by identifying important features for predictions.

5.1 Thalassemia as a disease

Thalassemia is one of the most common blood-related disorders in populations from the Mediterranean, South Asia, West Asia, and Africa. It is a group of inherited hematologic disorders broadly classified into two main types: α-thalassemia and β-thalassemia. Symptoms vary depending on the severity of the disorder but commonly include anemia, characterized by low hemoglobin levels in red blood cells, which causes fatigue, and weakness, as well as iron overload in the body. These disorders are caused by genetic mutations that impair the genes encoding the α- or β-globin chains of hemoglobin, resulting in reduced expression and/or malfunction of these proteins [186], [187].

In more detail, the protein complex of hemoglobin (Hb) is a tetramer and consists of 2 pairs of amino acid chains: two α-chain globins and two β-chain globins (Figure 19). Each of these chains contain a heme group that carries an iron ion which binds with oxygen, and thus, one Hb can carry up to four oxygen molecules with the mechanism of cooperative binding. This mechanism essentially makes the Hb efficient in oxygen binding [188].

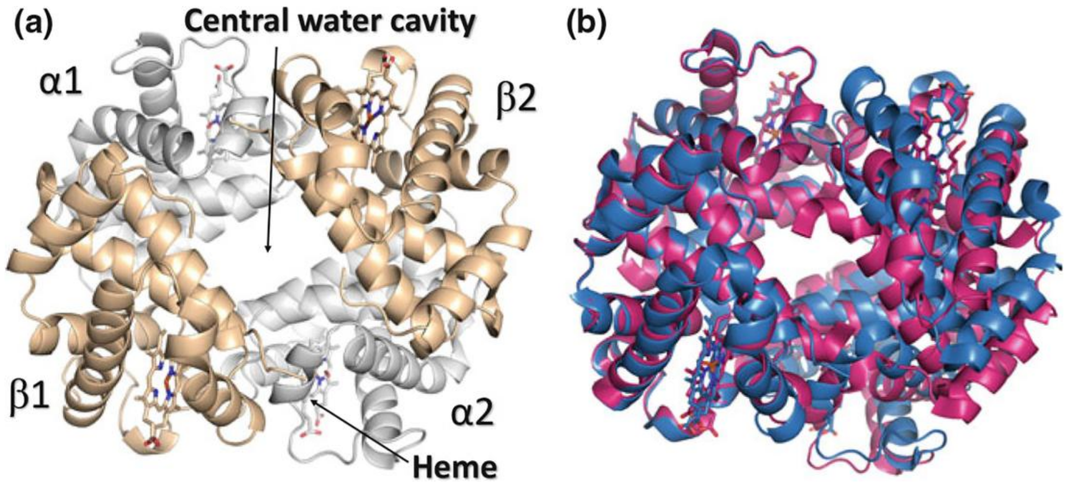


Figure 19. The crystal structure of the Hemoglobin complex (Hb). (a) The structure of the complex on which the pairs of α- and β- chains are colored. (b) the conformations of the oxygenated Hb in the Relaxed state (magenta) and the deoxygenated, tense state (blue) conformations. Source: Mostafa H. Ahmed et. al. (2020) [188].

Thalassemia has different severity levels from simple carriers that bear the trait of the disorder but do not show any symptoms and need no therapy, to severe cases that need blood transfusion and/or targeted therapies. Thalassemia can cause complications in other organs because of iron overload on the body and can be lethal for the patient if left untreated properly [189].

In more detail, iron accumulation in thalassemia results from two main processes: increased dietary iron absorption driven by ineffective erythropoiesis and iron loading from transfusions. In non-transfused patients with severe thalassemia, abnormal iron absorption from the diet adds 2 to 5 grams of iron to the body annually. In severe forms of thalassemia, patients depend on regular blood transfusions to sustain life and require chelation therapy to manage iron overload. Meanwhile, those with β-thalassemia intermedia do not need chronic transfusions but eventually develop iron overload due to ineffective erythropoiesis and hypoxia-induced hepcidin downregulation, which enhances gastrointestinal iron absorption [190], [191].

5.2 Cardiovascular Risk in Thalassemia: A Persistent Clinical Challenge

As hemoglobin is the primer carrier of oxygen in the body and is expressed in red blood cells Iron accumulation due to lack of hemoglobin in Thalassemia patients can have devastating outcomes and one of the

Although thalassemia affects approximately 18.3 out of 100,000 people [192], and there is a plethora of studies leveraging machine learning (ML), most previous research has focused on diagnosing the disease [193] or distinguishing between different conditions such as β-thalassemia minor and iron deficiency anemia [194]. Regarding how thalassemia affects human organs, in their work, Positano et. al. used deep learning models to link thalassemia to liver iron content from magnetic resonance and multiecho imaging with high accuracy [195]. Other studies have explored myocardial dysfunction in thalassemia patients using echocardiographic imaging with ML models [196] or magnetic resonance imaging [197], often overlooking blood markers. A notable recent study by Sabir et al. employed artificial intelligence to diagnose anemia by estimating hemoglobin concentration through photoplethysmography and video imaging [198]. However, their approach captures the patient's current state rather than predicting the risk of developing a disease in the future.

While these works are valuable, they do not address the critical issue of predicting CVD risk in the mid-term future using routinely acquired markers from follow-up visits. Such prognostic tools could aid clinicians in tailoring therapeutic strategies to prevent or manage cardiac events or dysfunction effectively.

5.3 Materials and Methods

5.3.1 Data source

Data is a longitudinal cohort from Italy collected by the Day Hospital of Thalassemia and Hepatopathy of the University hospital Sant' Anna. The dataset is constructed with follow-ups of patients' records from 1982 to 2022. In total the cohort consists of 287 thalassemia patients, of whom 12 have follow-up data labeled only as cardiac event, and those patients were excluded from the analysis since they could not participate in the prediction. With this, the cohort reduced to 275 individuals, with a total of 34,970 follow-up visits recorded. The dataset is balanced with respect to the sex of the patients with almost equal number of males and females (140 males, 135 females), but with a broad range of ages spanning from two to fifty-one years of age. Consequently, the available follow-up data is centered around twenty-four years of age because thalassemia patients visit the hospitals regularly after diagnosis which happens early in life, and there is a thorough record kept of those follow-up visits. Additionally, the dataset exhibits a significant imbalance in the number of patients who experienced a cardiovascular event during their lifetime compared to those who did not, with a ratio of 1:5.5 (42 patients with CVD and 233 without).

As an additional feature that was not considered while building the models is the “all cause death”. This is a feature showing if the patient has died from any clinical reason and we excluded it since it can provide information to the models from future events. Although not utilized, all cause death shows that over 90% of thalassemia patients in this dataset died, emphasizing the importance of prevention and early detection of pathologies.

Characteristic	**Range / Distribution of values**	
Sex (Male/Female)	M: 147 (51%), F: 140 (49%)	
Age at diagnosis (0-1/1-4/4-9 years old)	(0-1): 82, (1-2): 134 (2-4): 41, (4-9):19	
Age at the last follow-up in years	2 - 51	
Start of chelation age in years	0 - 29	
	Yes	**No**
Heart disease – Patient-wise	42 (15.3%)	233 (84.7%)
Heart disease – Total follow-ups	25,906 (8.4%)	284,208 (91.6%)
Heart disease – follow-up-wise (1-year)	2,155 (17.9%)	9,854 (82.1%)
Heart disease – follow-up-wise (3-years)	5,833 (16.7%)	29,137 (83.3%)
All cause death	263 (91.6%)	24 (8.4%)

Table 6. Summary of the basic population information of the dataset.

5.3.2 Data analysis and preprocessing

After filtering patients that are classified as CVD in all their follow-ups, and prior to the data analysis, the dataset was reviewed to categorize the features by datatype. Continuous variables stored as floating-point numbers and included features such as Age, Transplantation Age, Age of diagnosis, Age of first chelation therapy, weight, height, HB, HBF, HBA2, MCV, MHC, Blood unit hematocrit, Leucocytes, Eosinophils, and others, most of them with high missingness. Categorical variables on the other hand were encoded as integers, depending on their cardinality or their category. Among others, the available categorical features were Sex, Splenectomy status, HCV Ab status, Chelation therapy status, Hydroxyurea therapy status, Anti HCV therapy status, and the CVD target label.

The programming language Python 3.9 was used for the manipulation and cleaning of the data, with the library Pandas 2.2.3 and Numpy 1.26.4 for data handling and pre-processing. For the development of machine learning models, the dataset, while rich in follow-up records, contains a relatively small number of individual patients, and despite the abundance of follow-ups per patient, the dataset is heavily imbalanced with only 15.3% of individuals developing CVD. Furthermore, post cardiovascular event follow-up records constitute just 8.4% of all available data points.

To ensure consistency in temporal reference across cases and controls, we focused on specific follow-up periods. For individuals who experienced a cardiovascular event, we retained follow-up data from one-year and three-years prior to the event. For individuals without any cardiovascular problems, we included the most recent one year and three years of follow-up data, serving as the control group. After applying these criteria, the filtered dataset contained:

	1-year worth of follow-ups (individuals)	3-years' worth of follow-ups (individuals)
CVD	2,155 (233)	5,833 (233)
No CVD	9,854 (42)	29,137 (42)

Table 7. Count of follow-ups and individual patiets in the time-windows of one year and three years for CVD and non-CVD thalassema patients.

Data cleaning

The clinical information collected during each follow-up visit was utilized to generate time-related features that enable the ML models to capture temporal patterns within the data. Though, static clinical features, representing snapshots of a patient's condition at the time of the visit, were the main input features included. This approach integrates both the current clinical state and temporal dynamics of the patients. Outlier values were identified and removed during preprocessing, and features with extensive missing data were excluded. Additionally, new temporal features were engineered, such as the duration (in years) since the initial diagnosis, the number of years a patient has had diabetes, and other calculated variables, to enhance the dataset's temporal dimension.

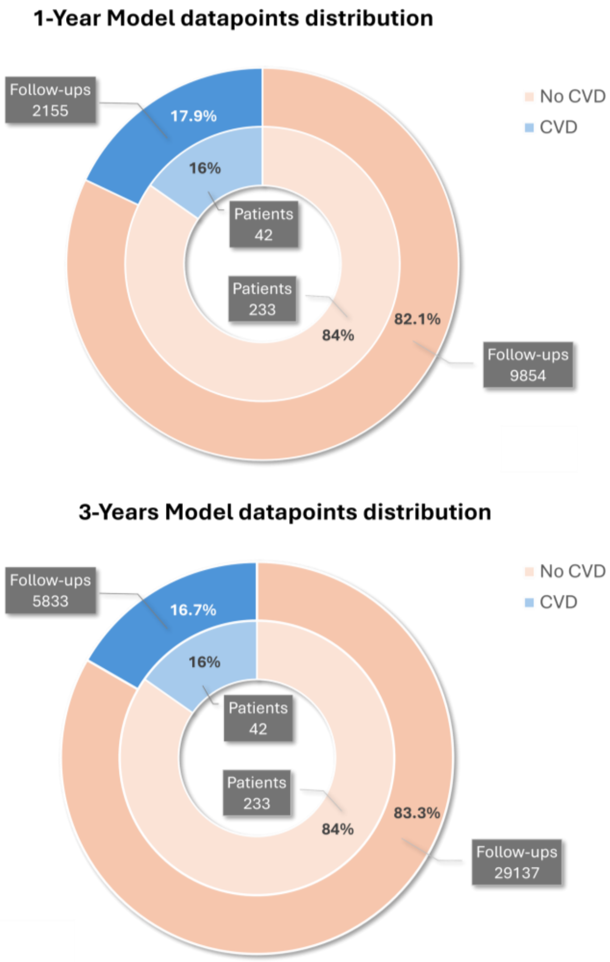


Figure 20. Pie charts of the follow-ups and individuals in the datasets for the cases of 1-year (top) and 3-year (bottom) subsets.

The follow-up visits were grouped by patient ID, and a subset of their records was selected for use in training the ML models. Patients in the cohort were categorized into three groups (Figure 21):

Event: Patients who experienced a cardiovascular event during the follow-up period were identified. For these patients, the time of the event (t_e) was used as the reference point. Follow-up records occurring after the event were excluded, as they no longer contribute to the prediction task. To train models for predicting the development of CVD in the mid-term future, we retained follow-up records from 1 and 3 years prior to the event, corresponding to the prediction windows (Δt) under investigation. Records within the prediction window (Δt) and preceding the event time (t_e) were labeled as **"**Event**"**, while earlier records were discarded.

No Event: Patients who had not developed CVD up to the current time. For these patients, the hospital follow-up records from the last 1 or 3 years were retained and labeled as "No Event."

Excluded Patients: Patients who were classified as having CVD throughout the entire follow-up period. These cases were excluded, as no prediction could be made for them.

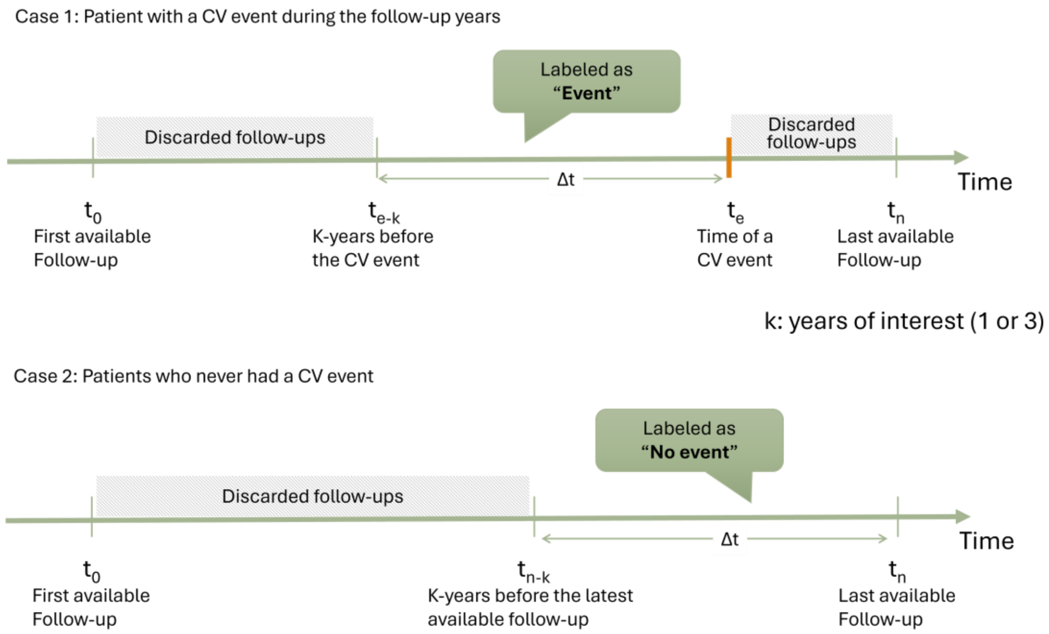


Figure 21. Schematic of the way the follow-ups are kept for the training of the ML models. On top is the case where the patients had a cardiovascular event at some time (t_e), and in those we kept all follow-ups K-years prior to the time of the event. On the bottom, there is the case where thalassemia patients never categorized as CVD patients, and for those we kept the last K-years' worth of follow-ups.

This stratification ensured that the dataset captured relevant temporal information, reduced the bias of patients with a very high number of hospital visits and at the same time is adhering to the specific requirements of the prediction task.

5.3.3 Models' development

To develop a predictive model for assessing mid-term heart disease risk in thalassemia patients, we employed a supervised machine learning approach using XGBoost [160], a widely used tree-based ensemble algorithm known for its strong performance and for providing feature importance scores for enhanced interpretability. XGBoost provides feature importance scores, which offer insight into the model's predictions, and is well-suited for clinical datasets due to its ability to handle complex, nonlinear relationships, missing values, and features with varying scales. This makes it particularly effective for our clinical dataset, which includes variables with occasional missing values and measurements on different scales.

The initial dataset comprised of 145 clinical features collected from blood tests performed at each follow-up visit and two demographics (Age and Sex). However, to prevent information leakage, we excluded any columns that directly referenced future cardiovascular outcomes or causes of death. Additionally, we removed features with more than 60% missing values, measurements related to the diagnosis of CVD, and features related to dates of measurement. This filtering reduced the total number of characteristics to 17 (see Sup_Table 17 for reference). These remaining features in the dataset correspond to clinical measurements acquired frequently in follow-up visits, patients' body-mass index (BMI), sex and age, the status of specific therapies, and time derived features linked to the duration of conditions. Namely, the features used for training the models are: Age, blood unit hematocrit, BMI, Chelation therapy status, number of Eosinophils, Hemoglobin levels, Hepatitis type C antibodies detection (HCV Ab) status, Hematocrit, Hydroxyurea therapy status, number of Monocytes, Sex, Splenectomy status, chelation therapy initiation age, years after splenectomy (if applies), years since thalassemia diagnosis, years since the start of chelation therapy (if applies) and years diagnosed with diabetes (if applies).

Given the frequent follow-up visits of thalassemia patients - sometimes multiple visits within a week - many sequential entries are nearly identical, creating a bootstrap effect that could bias the model. To address this, we engineered new temporal features to differentiate these entries based on time passed from reference points, capturing progression and trends across visits. Those temporal features are related to the time that has passed in years since: the diagnosis of thalassemia; splenectomy; the start of chelation treatment; the diagnosis of diabetes.

After the filtering of the features with high percentage of missing values and the creation of new temporal characteristics, the dataset consists of 275 individual patients, with 13.877 and 39.506 follow-ups for the 1-year and 3-years models respectively, and 18 features in total (See in the Appendix – A, Chapter - 5 for reference).

Stepping upon this foundation, we built two separate models to predict the risk of a patient developing CVD within a period of one year and three years, based on their current hospital visit. Both models follow the same framework and the same data cleaning. Additionally, no imputation and normalization applied on the data before splitting to avoid introducing any bias.

The dataset was split into training (70%) and holdout (30%) sets retaining the ratio of positive and negative classes and respecting the ID of the patients, ensuring class balance between CVD patients and controls while also preventing information leakage from one set to the other. The training set was used to train the models, and the hold-out set was used as an outside group of patients that the models never encountered before.

For model optimization, we used a nested 10-fold stratified cross-validation framework within the training set respecting again the IDs of the individuals to avoid information leakage. The inner loop was used to fine-tune hyperparameters including the learning rate, maximum tree depth, a generalization parameter (alpha), and the number of classifiers in the XGBoost ensemble classifiers. The outer loop of the cross-validation evaluates the models' generalization performance across the different splits and was used to ensure robustness and the selection of a good set of hyperparameters.

Ultimately, the whole training set was used again to train the final models. To this end, the training set was once more split in half respecting the patients' IDs and the class ratio. The first part was used to train the models on the best hyperparameters found from cross-validation and the second half was used to train a calibration model. This tool is used to ensure that the predictions made by the ML model closely match the actual observed outcomes, and we use calibration models to improve the reliability of predictions, making them more accurate and clinically useful. Calibrated models act as an additional layer on predictive models to improve the mapping of an ML model's output scores to the actual likelihood of observed outcomes. This process ensures that the predicted probabilities better reflect the true likelihood, enhancing the reliability and interpretability of the model. Calibration not only makes the model's predictions more trustworthy but also provides users with an intuitive understanding of the probability associated with a given risk, as opposed to relying on arbitrary or uncalibrated scores. Given the dataset's imbalance, we evaluated model performance using multiple metrics: balanced accuracy, F1-score, false positive rate (FPR), false negative rate (FNR), and area under the receiver operating characteristic curve (AUC). Balanced accuracy accounts for the uneven distribution of classes, while the F1-score emphasizes the trade-off between precision and recall, which is critical in clinical settings where both false positives and false negatives provide insights for the model's misclassification rate on each class.

5.4 Results

By following the procedure described in the methods, two different models are built, and both predict the calibrated risk of a patient developing CVD within the mid-term future of one and three years respectively.

5.4.1 Model performance

For the evaluation of the models' effectiveness, both the training and testing set were fed to them repeatedly to have robust measurements with a confidence interval around the mean. To achieve this in the training set a cross-validation scheme was used, while in the testing set a 1000 repetition bootstrap approach (random sampling with replacement) was chosen. In both cases, the mean and 95% confidence interval calculated and reported for the two models as shown in Table 8.

The model predicting CVD within one year scored over 93% in the Area Under the Curve (AUC) in the repeating validation with confidence interval of 0.95 spanning from 92.1% to 94.6%. The Balanced accuracy that depicts the ability of a model to classify both cases correctly has a mean value of 90.9%, an F1 score of 89.7% and a FNR of 17.7% with a relatively high variance among the validation iterations (see Table 8). Similarly, the three-years predictive model scored a mean AUC of 93.5% with very little variance and a Balanced Accuracy mean value of 92.4% ranging from 91.6% up to 93.3% by examining the confidence interval, and an F1 score of 91.6%.

Overall, both models demonstrated comparable levels of predictive performance, with the one-year model achieving slightly higher scores across most metrics of the training phase but the three-years model scoring slightly higher in the hold-out sets.

	1-year Performance [95% CI]		3-years Performance (95% CI)	
Metrics	**Training**	**Hold-out**	**Training**	**Hold-out**
Accuracy	95.9 [91.9, 98.6]	95.5 [94.6, 96.0]	95.5 [92.3, 98.5]	96.6 [96.1, 97.0]
Balanced accuracy	95.0 [90.6, 97.8]	90.9 [89.6, 91.9]	93.3 [82.2, 98.1]	92.4 [91.6, 93.3]
Precision	85.6 [74.1, 98.9]	98.6 [97.6, 99.1]	82.3 [65.6, 98.1]	99.1 [98.4, 99.6]
Sensitivity	93.7 [86.4, 99.9]	82.3 [79.6, 84.2]	90.4 [65.8, 99.9]	85.1 [83.5, 86.9]
Specificity	96.3 [92.5, 99.8]	99.6 [99.4, 99.8]	96.3 [92.3, 99.7]	99.8 [99.6, 99.9]
Negative Prediction value	98.7 [96.8, 100]	94.7 [93.6, 95.3]	98.4 [95.1, 100]	96.0 [95.5, 96.5]
False Positive Rate (FPR)	3.7 [0.2, 7.5]	0.4 [0.2, 0.6]	3.7 [0.3, 7.7]	0.2 [0.1, 0.4]
False Negative Rate (FNR)	6.3 [0.1, 13.6]	17.7 [15.8, 20.4]	9.6 [0.1, 34.2]	14.9 [13.1, 16.5]
AUC	98.9 [97.5, 99.7]	93.5 [92.1, 94.6]	98.5 [97.0, 99.7]	93.2 [92.4, 94.0]
F1 score	89.1 [81.1, 95.1]	89.7 [88.1, 90.9]	85.6 [70.8, 95.8]	91.6 [90.5, 92.5]

Table 8. Performance metrics of the training and testing set for the 1-year and 3-years prediction models.

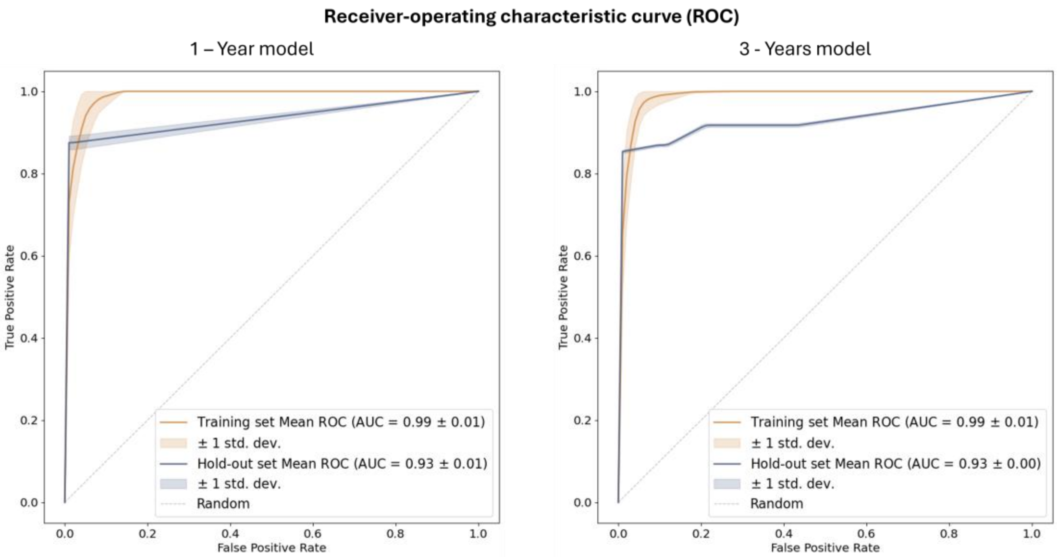


Figure 22. Receiver-operating characteristic curve (ROC) and the calculated area under the curve (AUC) for both models. In orange, the training set curves with the calculated standard deviation, and in blue the hold-out set curves with their standard deviation.

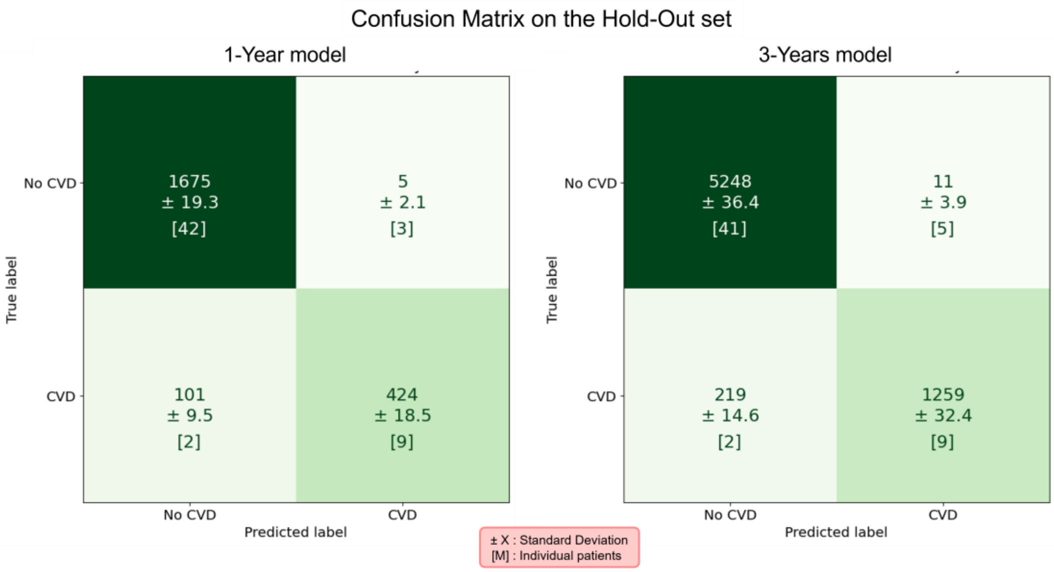


Figure 23. Confusion metrices of the bootstrapped performance of the 1-year model (left) and 3-years model (right) on the Hold-out set. The main values in the quartiles represent the number of follow-ups followed by the ± standard deviation value and in the square brackets is the number of unique patients.

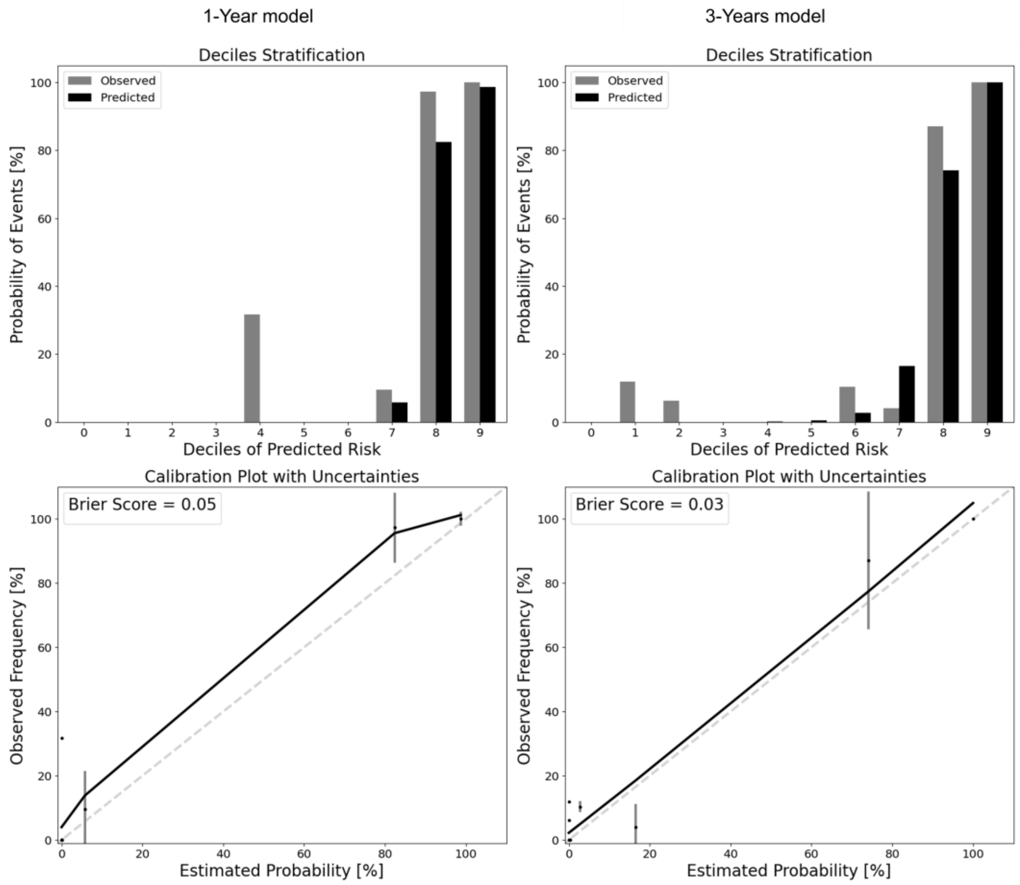


Figure 24. Decile analysis (top row) of the models' predictions over the ground truth labels, and the calibrated model plots (bottom row) of the models showing how close the models' outputs reflect to the real world probabilities.

5.4.2 Feature importances

From the feature importance analysis of the models, the top eight features with the highest scores are shared between both models, albeit with slight variations in their ranking (Figure 25). Notably, the two most important features by far are the time-dependent variables: *Years with Diabetes* and *Years from First Follow-Up*. These two features alone account for approximately 40% and 50% of the information gain metric in the one-year and three-year models, respectively. Completing the list of shared top features, but with substantially lower scores, are *Monocyte* count, *Years of Chelation*, *Blood Unit Hematocrit*, *Hematocrit*, *Eosinophil* count, and *Years after Splenectomy*.

Beyond the shared top features, the ranking of the remaining characteristics varies between the two models. The least influential features include *Sex*, *Splenectomy* *status* (yes/no), Hemoglobin B (*Hb*) levels, and Body Mass Index (*BMI*). These features contribute minimally to the models, with gain scores ranging from 0.5% to 2.4% indicating that they are not the main risk factors for cardiac events in thalassemia patients.

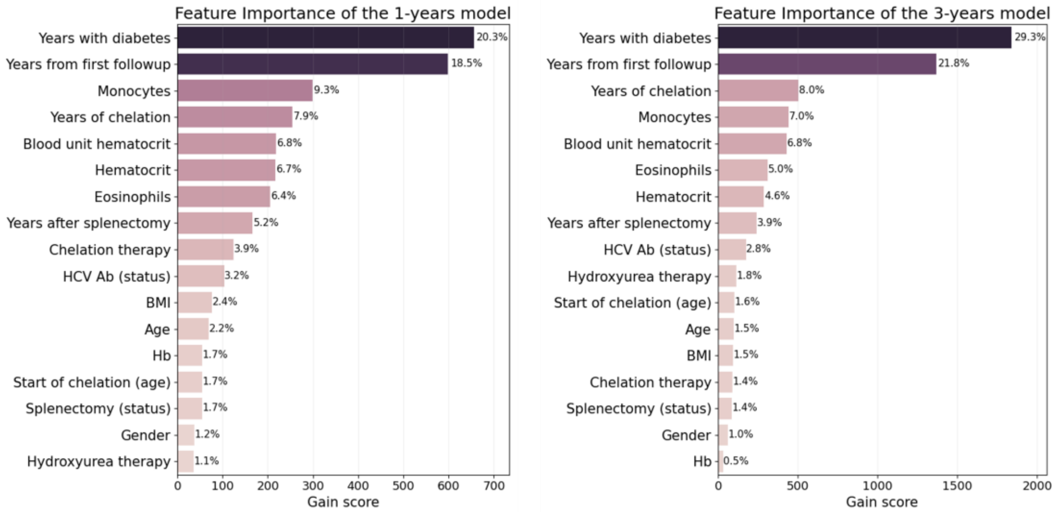


Figure 25. Feature importance of the models based on the XGBoost classifier internal ranking on the information gain on the (x-axis). Higher gain means that a feature is overall more important for generating a prediction.

5.4.3 Misclassifications

A closer examination of the results allows for the analysis of misclassifications in both the level of follow-ups and the individual patient level for the two models. This analysis reveals that they share common patients who were consistently misclassified across all follow-ups. The main difference was that the three-year prediction model misclassified one additional patient compared to the one-year model. Regarding partial misclassification across follow-ups, a similar pattern was observed. In this case, one patient was partially correctly classified by the one-year model, whereas the three-year model partially misclassified this patient and an additional one.

By examining the individuals where the models misclassified a fraction of the total follow-ups, it seems that for the CVD patient which is common in both models, the 1-year model classifies only one third of the follow-ups correctly as CVD while on the other hand the 3-year model has higher than 80% correct classifications.

Additionally, the 3-year model misclassified another individual but only for 2% of its follow-ups (Sup_Fig. 13).

Interestingly, an analysis of the models' misclassifications, focusing on patients whose follow-ups were consistently misclassified, revealed that both models incorrectly classified the same patients in this category, with the three-year model misclassifying one additional patient. Among the four shared misclassified patients, three have very few follow-up records and are all non-CVD thalassemia patients. In contrast, the remaining shared patient is labeled as CVD and has a substantial number of follow-up records: 69 in the one-year subset and 193 in the three-year subset. Finally, the additional patient misclassified exclusively by the three-year model is a non-CVD subject with only one follow-up record as it is depicted in Figure 26.

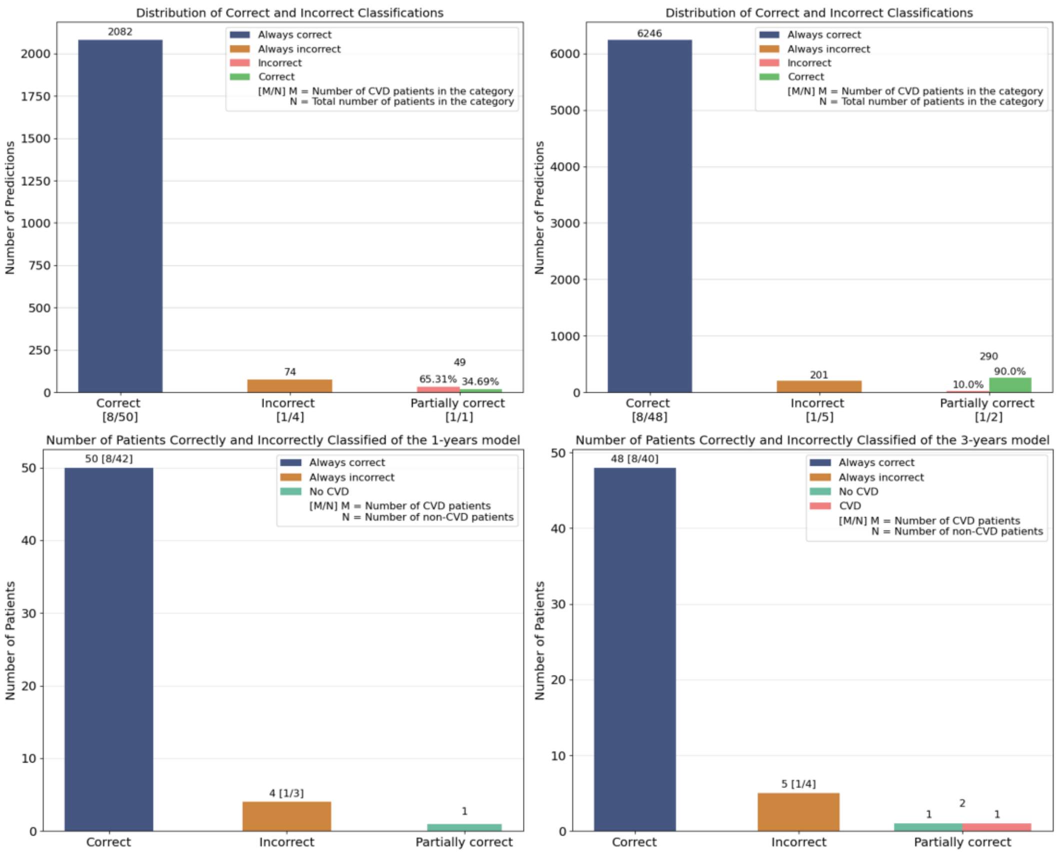


Figure 26. Barplots of the classifications. On the left side are the plots for the 1-year model, and on the right side are the 3-years model results. On the top is the analysis on the level of Follow-ups, and in the bottom row the results per patient.

5.5 Discussion

Using data from Thalassemia patients and by feature engineering some new characteristics integrating the time passed from a specific reference point, we built 2 models to predict based on the current follow up of a patient, if they will develop CVD within the mid-term future of one in three years.

By comparing the performance of both models in the evaluation and testing phase, it is obvious that the models are both stable and have the potential to predict if a patient will develop a CVD in the mid-term future by accessing only one set of examinations. The models even though trained on a plethora of datapoints, had access to a limited number of individual patients. Thus, the decision of treating each follow-up as independent input worked as a form of bootstrapping. Although the follow-ups considered independent, the splitting of the training, testing and evaluation sets made respecting the ID of the patients so that no leakage of information could take place between the training and hold-out sets.

We report high performance scores for both models with mean AUC on the hold-out set above 93% and over 90.9% mean Balanced Accuracy, demonstrating that the model is both accurate and generalizes well on unseen data. From the performance report is also clear that the model was presented with a relatively low number of positive examples because of the data imbalance, and this become obvious by focusing on the False Negative Rate and the Sensitivity where in both models the hold-out performance drops. This is a result of low number of independent patients since by analysing the misclassifications the number of patients wrongly classified is low, and by viewing the number of positive cases, one cans see that there are only two CVD cases wrongly classified. From those cases, one was wrongly classified by both models and one was partially correctly classified (Sup_Table 18 & Sup_Table 19). From the analysis of the misclassified follow-ups belonging to patients partially correctly classified, it does not seem to exist a specific pattern of theses mistakes. For example, we did not observe all the misclassification being only in the initial or in the latest visits (Sup_Fig. 17).

The results indicate that the models are robust, with overall performance metrics being similar between the training and hold-out sets. However, some metrics show greater discrepancies, primarily due to the lack of independence between follow-ups within each set. This dependency makes certain samples easier for the algorithm to classify consistently, while others are more challenging. As a result, multiple follow-ups for the same patient are either correctly or incorrectly classified, and the effective contribution of each patient to the performance metrics is uneven.

For the one-year model, it appears that the earliest follow-ups for certain patients are the ones most frequently misclassified. This suggests that only the most recent hospital visits may contain the necessary information in their features to enable correct classification for these patients. In contrast, the three-year model, which considers a broader time window, does not exhibit the same issue. Although its misclassifications are clustered around specific time periods, they do not follow a specific pattern.

Given the sparsity of the features in the dataset before the preprocessing, it was necessary to drop features that might be more relevant to the target variable. Features that are related to blood such as red or white blood cells concentration. Based on clinical knowledge, features like these can be indicators of inflammation or disease and they could be used as markers for the algorithm, but since the dataset had a high rate of missing values in those features, we decided to drop them and create features that integrate the element of time. Thus, we created the continues characteristics that measure in years the time passed from a reference point such as the diagnosis of diabetes or splenectomy in the cases where it was applicable.

As it appears from the feature importance analysis, the time interval-based features appear in the top ten of the ranked lists. This indicates a good correlation between them and cardiovascular risk. This strengthens the hypothesis that, in thalassemia patients, the medical history and trajectory of the patient may provide information about their future outcomes. Future developments of this study may also include considering other time interval-based clinical features to study more deeply their connection to cardiovascular risk.

A larger dataset with fewer missing values would be ideal for developing machine learning models for this task, particularly if it includes features known to clinicians as strong indicators of CVD or other related conditions. Further, an evaluation set of patients from different sources, such as different hospitals, could help in the understanding of the generalization power of the model. That, in combination with a larger training set could be beneficial for creating robust models. Additionally, although the element of time is integrated in these models by introducing time-from-event features, for a future development it is necessary to implement models that can handle follow-up visits as dependent, and as a sequence of time frames. This could explain better the pivot points from the healthy condition to the diseased one and provide a better understanding of the features responsible for this change.

A high-performing predictive model, such as the one developed in this study, holds significant clinical potential by enabling the early identification of thalassemia patients at elevated risk of cardiovascular disease. Specifically, the system can alert the attending physician up to three years prior to a potential cardiovascular event, providing a valuable window for implementing timely and personalized preventive strategies. These may include lifestyle modifications and adjustments in pharmacological therapy aimed at mitigating risk progression. By facilitating earlier intervention, such a tool could contribute to improved long-term outcomes, including prolonged cardiovascular event-free survival and enhanced overall life expectancy in patients with thalassemia***Future developments***

Despite these limitations, our study demonstrates that machine learning can effectively predict the risk of a thalassemia patient experiencing a cardiovascular event in the mid-term future, based on routine examinations conducted during hospital visits. While these models are not perfect, they provide valuable risk assessments that can guide doctors and patients in taking proactive measures. This may include lifestyle modifications, additional diagnostic tests for heart-related conditions, and other preventive interventions.

In conclusion, our study demonstrates the potential of machine learning models in predicting the mid-term risk of cardiovascular disease (CVD) in thalassemia patients using routinely collected clinical and demographic data. By utilizing a dataset of 275 patients with multiple follow-ups, we developed two predictive models, one for 1-year risk and another for 3-year risk which achieved high performance. The 1-year model attained an AUC of 93.5%, balanced accuracy of 90.9%, and an F1-score of 89.7%, while the 3-year model demonstrated an AUC of 93.2%, balanced accuracy of 92.4%, and an F1-score of 91.6%. These results highlight the robustness and reliability of our models in identifying high-risk individuals. By incorporating time-engineered features, such as years with diabetes and years from first follow-up, our approach captures temporal disease progression, enabling more effective risk stratification. These models offer a promising tool for clinicians to enhance patient monitoring, personalize treatment strategies, and potentially mitigate the progression of CVD in thalassemia patients.

Future research should integrate other inputs such as genetic markers, imaging, and lifestyle data to enhance model performance and interpretability. This will provide a more general and wholistic view of the patient to the models in different levels which we believe it will be beneficial for their generalization and performance. As an addition, a tailored time-series-like method or features that summarize other clinical characteristics should be developed to account for sequential measurements in the follow-ups. This requires a higher number of individual patients in the dataset which was one of the main challenges of this project as well. A prospective validation in diverse clinical settings is needed to ensure real-world applicability and evaluation of the models, while the use of explainable AI can improve clinician trust and adoption.

Chapter 6

A Unified Platform for Biomarker Discovery Tools

This chapter has a dual objective of introducing the development and integration of a Python-based platform and a Structured Query Language [SQL] database, which constitute key components of the research infrastructure. The SQL database was designed to store, organize, and manage the extensive datasets generated and utilized throughout this research. It serves as the backbone for data retrieval and analysis, ensuring efficient access and manipulation of structured information. Complementing this, the Python-based platform provides an intuitive and interactive interface, allowing users to navigate through multiple pages, each offering specialized tools developed during the research. These tools enable users to perform complex analyses with the implemented tools, visualize data, and extract meaningful insights, all while seamlessly interacting with the underlying database. Together, the platform and database form a cohesive ecosystem that not only supports the methodological advancements of the research but also enhances accessibility and usability for future applications.

The integration of these two systems represents a significant contribution to the field, bridging the gap between data management and practical utility. By combining the robustness of SQL for data handling with the flexibility and power of Python for computational tasks, this framework offers a scalable and user-friendly solution for researchers and practitioners alike. In the following sections, the architecture, functionality, and development process of both the database and the platform are detailed, highlighting their interconnected roles in advancing the objectives of the research.

6.1 Database

Purpose and Usability:

The database serves two key purposes in this work, each contributing to its overall utility. First, it acts as a centralized repository designed to gather, store, and harmonize all publicly available datasets related to NDDs from online sources and is directly linked with the objective O3 that was presented in detail in Chapter-3. It not only houses raw data but also stores processed and analyzed results, systematically linking them to their corresponding raw datasets. Over time, the database evolved into a comprehensive pool of harmonized microarray datasets from various studies, enabling efficient access and cross-study comparisons. The second crucial role of the database is supporting the unified platform (refer to section The Platform) by storing user-related information and serving as the backbone for the application's functionalities. This dual purpose ensures that the database not only facilitates data management and analysis but also enhances the platform's usability and interactivity, making it an indispensable component of the research infrastructure.

The integration of the database with the Python-based platform enables efficient data querying and retrieval tailored to specific research needs. This capability is particularly valuable for analyzing and comparing different disorders. Since the data are harmonized across studies, the database provides a larger number of samples for each class, which can enhance statistical power and improve the robustness of the findings.

The database was developed using an SQL-based framework, chosen for its robustness, scalability, and widespread adoption in managing structured data. Additionally, SQL libraries allow seamless interconnectivity with other popular programming languages which makes them wildly used in research. SQL provides a powerful and flexible way to interact with relational databases, making it an ideal choice for handling complex datasets. The relational database model was employed, organizing data into tables with predefined relationships to ensure efficient storage and retrieval. This model is particularly advantageous for maintaining data integrity and supporting complex queries. For implementation, the MySQL library was utilized, with administration handled through the phpMyAdmin software, enabling web-based management of the database.

At its core, the database schema defines the organization of data, including tables, fields, primary keys, and relationships between entities. The schema was designed to be modular and hierarchical, allowing it to accommodate diverse data types and adapt to evolving research needs. This structured approach not only simplifies data management but also enhances the database's ability to support complex analytical tasks, making it ideal for integration with the unified platform developed for this work.

6.1.1 Database for NDDs

In more detail, the part of the dataset that serves as the storage of the datasets from studies focused on NDDs was designed so that it contains information to the original data in the form of Uniform Resource Locators [URLs], Identifiers to GEO and links for downloading the data directly from the online source.

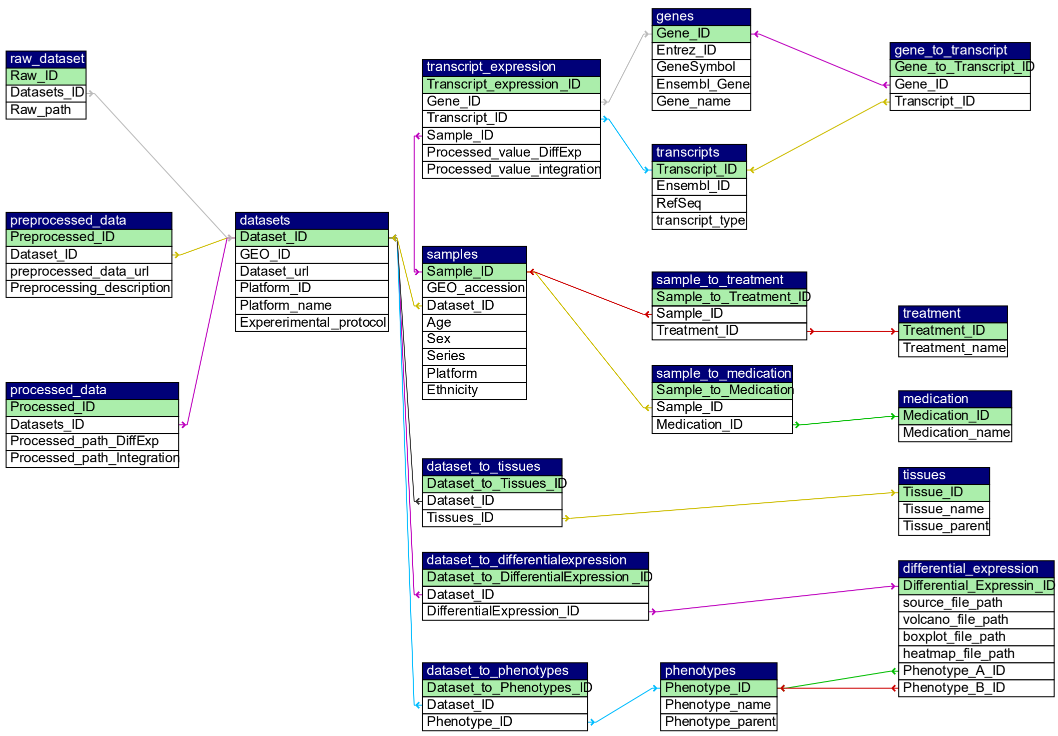


Figure 27. Structure of the SQL database created for this work, showing only the tables related to transcriptomics datasets and the results of their analysis.

The goal of this database section is to link together the different parts of individual studies as collected in Chapter – 3. 3. The merged harmonized dataset is stored in a table that is connected through foreign keys with the initial individual studies and it is available for querying from the users through the unified platform presented in this chapter in section 6.2.4 Database queries.

6.1.2 Database as authenticator and job management

The database includes a dedicated table for managing user information, which is essential for supporting the platform's functionality and ensuring secure access. This table contains columns for *email* (serving as the primary key), *Name*, *Surname,* *Password* (stored in a hashed and encrypted format for security), *role,* and *organization*. The user table plays a critical role in authenticating and authorizing platform users, enabling personalized access to tools and datasets based on their roles. By storing user details in a structured and secure manner, the table facilitates efficient user management and ensures that users can interact with the platform in a controlled and role-appropriate way. This design not only supports the platform's usability but also provides a secure way for users to access their personal result repositories.

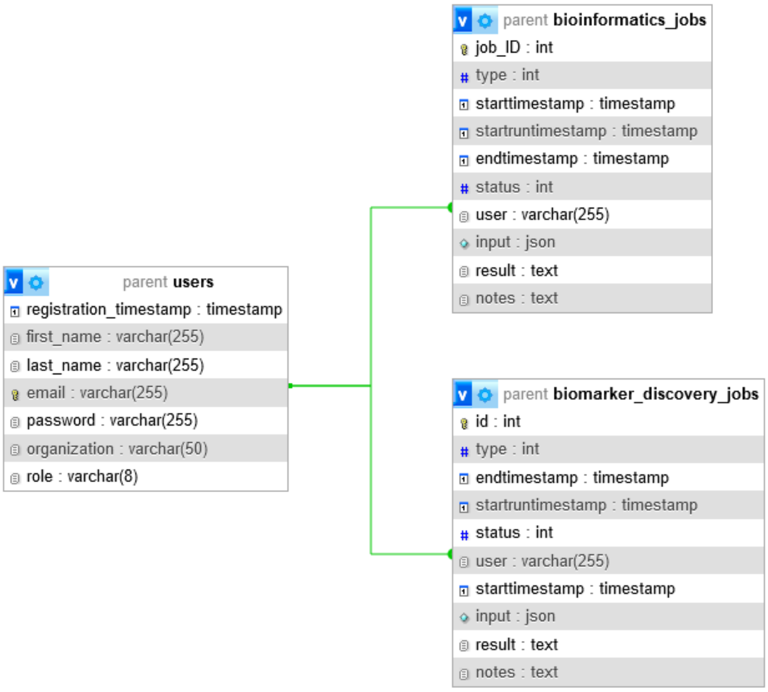


Figure 28. The structure and relationship of the tables for users and tables

The **users** table is directly linked to the **biomarker_discovery_jobs** and **bioinformatics_jobs** tables through a foreign key relationship, enabling seamless tracking and management of job requests submitted by users via the platform (Figure 28). Both job tables store critical information about pending, running, and executed jobs, including:

- **Job_ID**: A unique identifier serving as the primary key for each job.
- **type**: Specifies the type of job, which determines the back-end script to be executed.

- **Priority**: an integer that controls the importance of the job. Jobs with higher priority number will run before those with lower priority.
- **starttimestamp**: Records the timestamp when the job was requested by the user.
- **endtimestamp**: Captures the timestamp when the job completed (if it ran successfully).
- **startruntimestamp**: Indicates the timestamp when the job began execution.
- **status**: Tracks the current status of the job (e.g., pending, running, finished, error).
- **users**: A foreign key linking the job to the **Users** table, identifying the user who submitted the request.
- **input**: Stores the input parameters for the job in JSON format, ensuring flexibility and structured data handling.
- **results**: Provides the path to the directory where the job's results are stored.
- **notes**: A free-text field allowing users to add comments or notes related to the job.

This relational design ensures that every job is associated with the user who initiated it, enabling personalized tracking and management of job requests. By linking jobs to users, the platform can provide users with a clear overview of their submitted jobs, including their status, execution timelines, and results. Additionally, this structure supports efficient back-end processing, as the "*type"* column directs the platform to the appropriate script for execution, while the "*status*" column ensures transparent monitoring of job progress. Together, these tables create a robust system for managing user-driven computational tasks, enhancing both usability and accountability within the platform.

6.2 The Platform

For wrapping up the research and putting all the developed technologies and tools in a confined and easily accessible space, a Python-based full-stack platform was created. For this purpose, Python version 3.11.11 was used as the programming language with the main libraries for the development being *Streamlit* version 1.38.0, *pymysql* version 1.1.1, and *Sqlalchemy* version 2.0.35. Streamlit was used for the graphical interface and the management of the applications with pymysql and sqlalchemy being used for communication with the database. To access the platform, a public user's URL is available here: http://130.192.23.168:8501/.

The platform allows three different types of user roles with different permission levels as shown in Figure 29. First is the simple guest user type where access navigation to the pages is allowed, but not the use of the tools. The reason is that for using a tool and producing some results the platform must store the results in a user-dedicated directory which is accessible only by the specific user to ensure privacy and security. Nonetheless, simple guests can submit queries to the database to retrieve data related to the transcriptomics data for NDDs just for viewing, and without the option of downloading them.

The second level of user is the regular user who has been registered with the proper credentials and has been automatically assigned with a personalized directory. Thus, these users can submit jobs through different applications and get notifications when the jobs are finished. Additionally, the registered users can download the responses of the database to their queries into .csv or .txt formats.

Last, there is the administrator role that can be granted to users upon request. This allows users not only to fetch data from the database and download them, but also to insert or modify information of the database through the platform by submitting proper MySQL commands.

Platform's feature	Guest	Regular User	Admin
Biomarker Discovery	✗	✓	✓
CVD risk prediction	✓	✓	✓
Microarray process	✗	✓	✓
Job requests	✗	✓	✓
Upload files	✗	✓	✓
Database queries	✓	✓	✓
Database download	✗	✓	✓
Personal vault	✗	✓	✓
Database manipulation	✗	✗	✓

Figure 29. Users' roles and permissions for navigating and utilization of tools in the platform.

All users of the platform regardless of their role are allowed to access the pages of the projects GALATEA and PARENT which are the two founding projects of the research. These pages contain all the necessary information of the projects and active hyperlinks to the official websites of the projects and the European Unions funding websites.

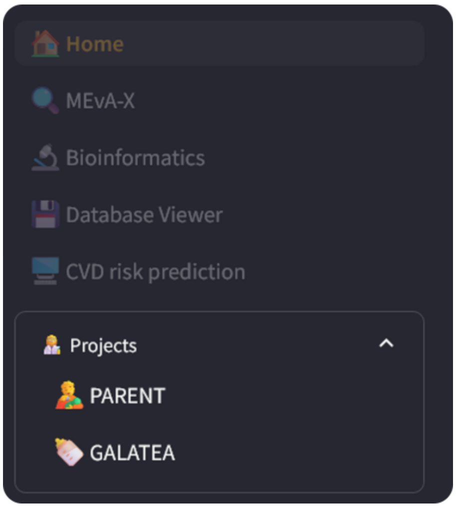


Figure 30. The navigation bar of the platform with the PARENT and GALATEA projects options highlighted.

6.2.1 The interface

The platform is built using the Streamlit Python library, designed to provide an intuitive and interactive environment for users to explore the developed tools and analyze their own data. The interface follows a clean and structured layout, ensuring ease of navigation and accessibility to various functionalities.

Upon launching the platform, users are presented with a side navigation panel, which allows seamless switching between different sections. Each page serves a distinct purpose, ranging from general platform information to specific bioinformatics applications, machine learning models and user login and registration pages. The interface is designed to handle dynamic user inputs, display graphs for visualization, and provide real-time data analysis results in a user-friendly manner.

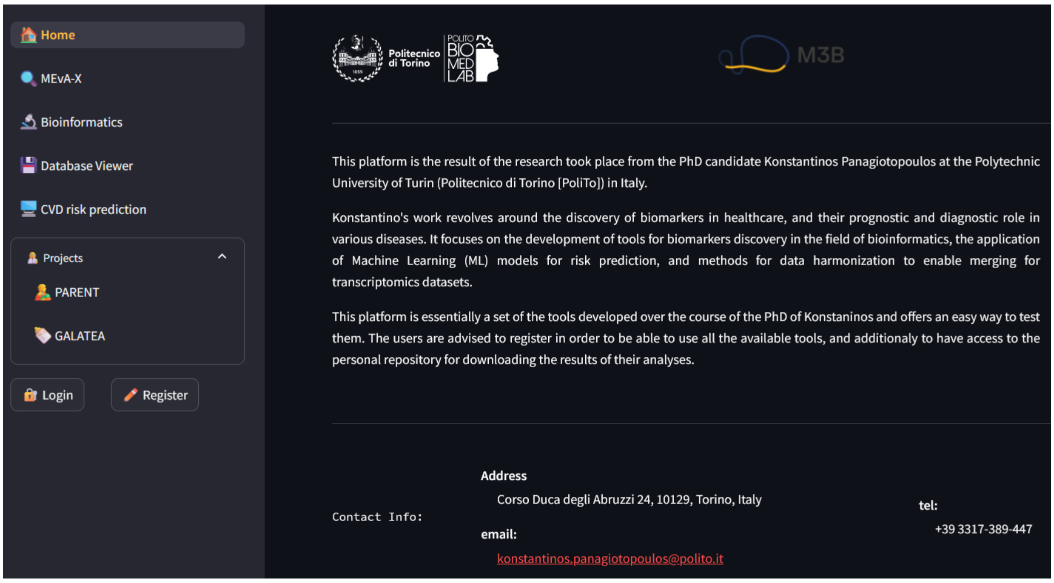


Figure 31. The Home page of the platform with the navigation bar on the right to allow users to explore the developed applications.

The platform is organized into multiple pages, each dedicated to a specific functionality or tool which are explained in more detail in the following sections.

6.2.2 User's pages

The platform includes several user-oriented pages that enable users to register and gain elevated privileges compared to guest users. The registration page (Figure 32B) presents users with a form to input their personal information, including their name, surname, email address, and an optional field for their organization. Users are also required to create a password, which must be entered twice to minimize typographical errors. For security reasons, the password fields mask input by displaying it as "●". The password must be at least five characters long, ensuring a basic level of security. Additionally, users can request administrator privileges by selecting a checkbox; however, such requests require approval from the platform's creator.

Upon submission of the registration form, the platform performs a database check to ensure that the provided email address is unique, as it serves as the primary identifier for each user. If no duplicate email is found, an automated confirmation email is sent to the user's provided address, notifying them that their account has been successfully activated.

During the creation of a new user account, a dedicated directory is automatically generated in the back-end. This directory serves as a personal storage space for the user's input files, job requests, and analysis results. By providing each user with their own directory, the platform ensures data organization and security while allowing users to easily access and manage their files.

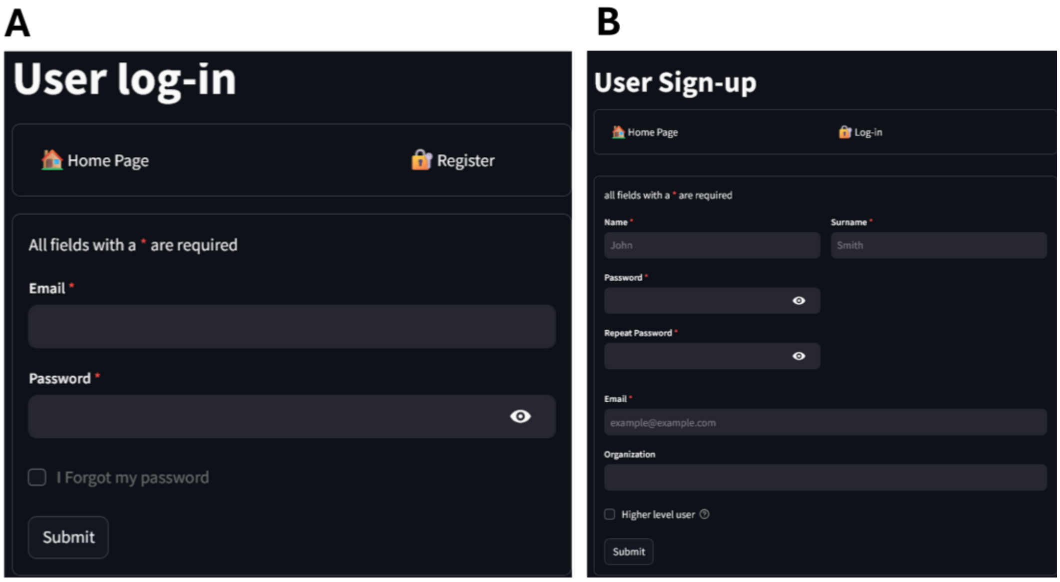


Figure 32. The *log-in* (left) and *registration* (right) pages of the platform.

Regarding the log-in process, the *log-in* page (Figure 32A) requires users to enter their registered email address and corresponding password. If a user forgets their password, they can initiate a password recovery process, which triggers an automated email containing a new, randomly generated password. This temporary password allows the user to regain access to their account.

Once logged in, users gain access to a suite of tools and features that are unavailable to guest users. These include:

- The MEvA-X tool for biomarker discovery and ML model building.
- Bioinformatics tools for processing microarray data.
- A CVD risk predictor tool specifically designed for thalassemia patients.
- Access to the platform's database, with the ability to download query results in .csv or .txt formats.

Additionally, logged-in users can view an overview of their submitted jobs, including their current status and a graphical representation of the jobs' states as it is depicted in Figure 34. Apart from that, users can download the results of finished jobs from the *Download* tab of this page by selecting the tool and the corresponding Job ID or Job Title as show in Figure 33.

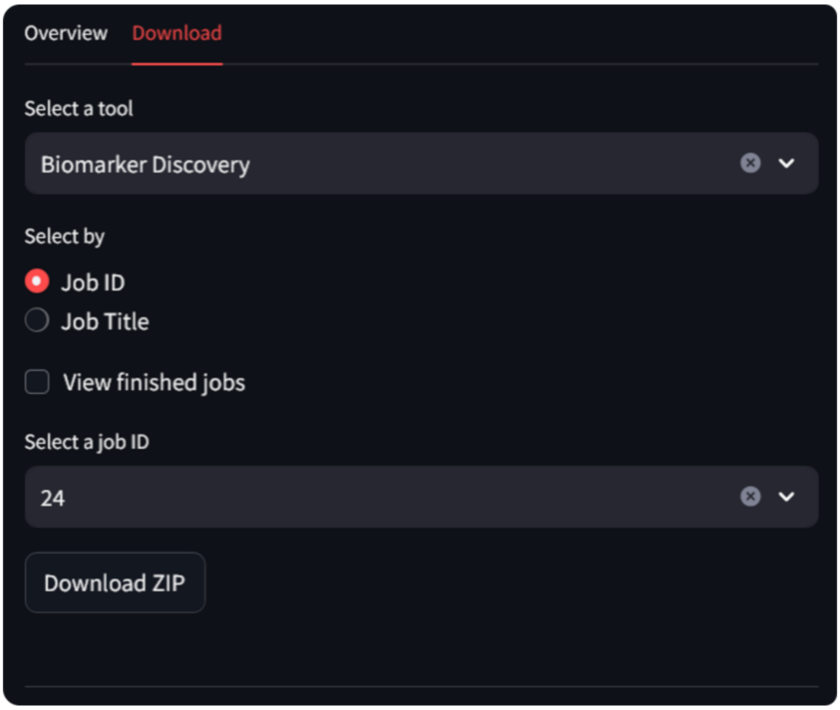


Figure 33. User's personal page, download tab view for the job with ID 24 from the Biomarker Discovery tool.

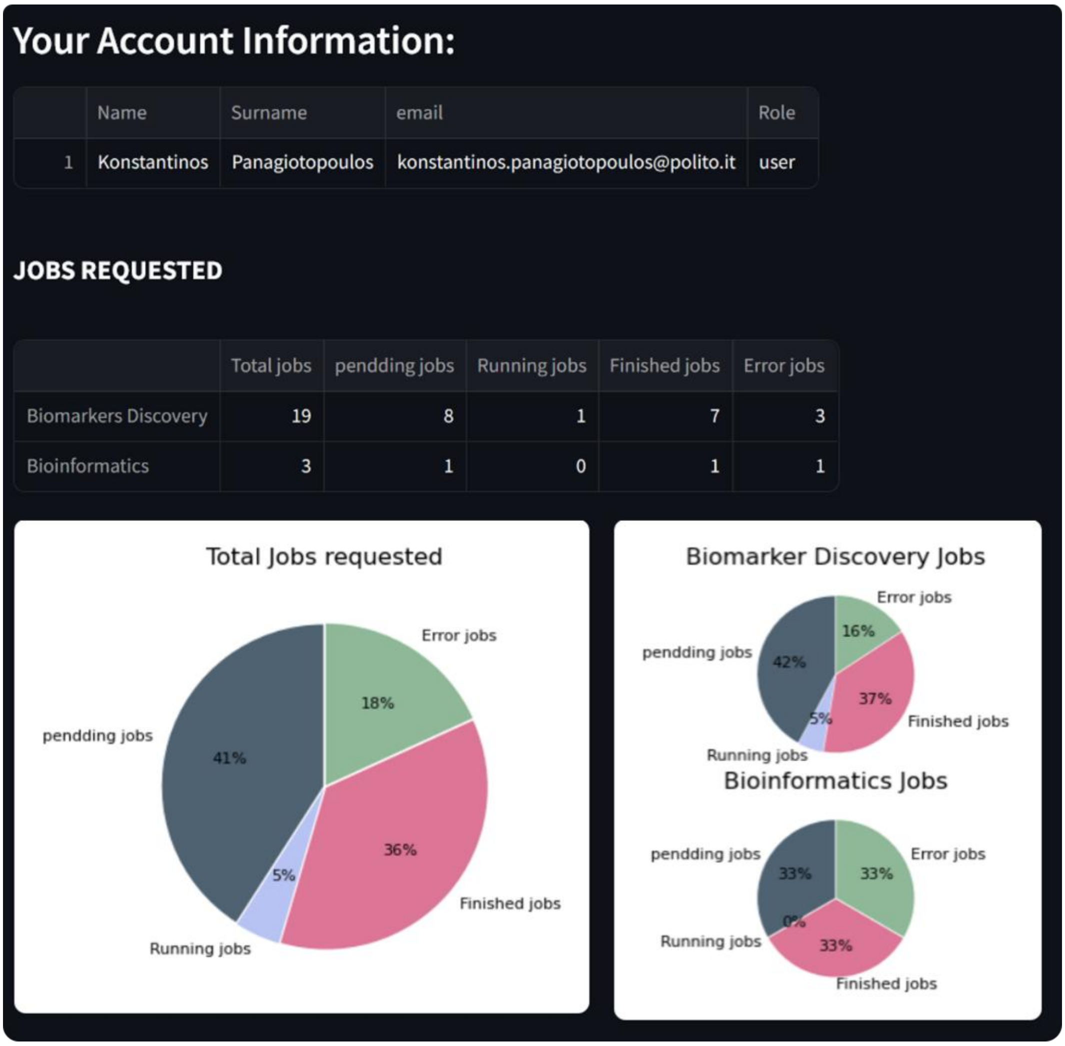


Figure 34. View of the personalized account page for the users. On top, the user information is displayed followed by the number of job requests in a table form and a visual representation in pie charts.

6.2.3 Biomarker Discovery tool

The page containing the MEvA-X tool described in Chapter – 4 is implemented in the platform under the title *Biomarker Discovery* to make more intuitive its purpose to users. In fact, a brief explanation is given at the top of the page and a reference to the publication with an active hyperlink is available for further information about the research behind the tool. As it is described in the publication, this tool is the implementation of a multi-objective evolutionary algorithm with the XGBoost classifier as predictor for - mainly - omics data [199].

In order to use the tool, users must be logged-in with their credentials so that they can upload their own data which will be stored in their dedicated directory under the job ID assigned to the submitted request. The results of the analysis are available for download to the users when the job is completed successfully, through their personal vault.

In the main frame of the page the user is presented with the necessary input fields that need to be used and these are: the dataset file in the form of a .csv, .tsv or .txt file, the file with the target labels of the samples in one of the formats mentioned before, the number of generations for the evolutionary to run, the number of solutions in every generation, and finally the number of folds for the internal cross validation for the models.

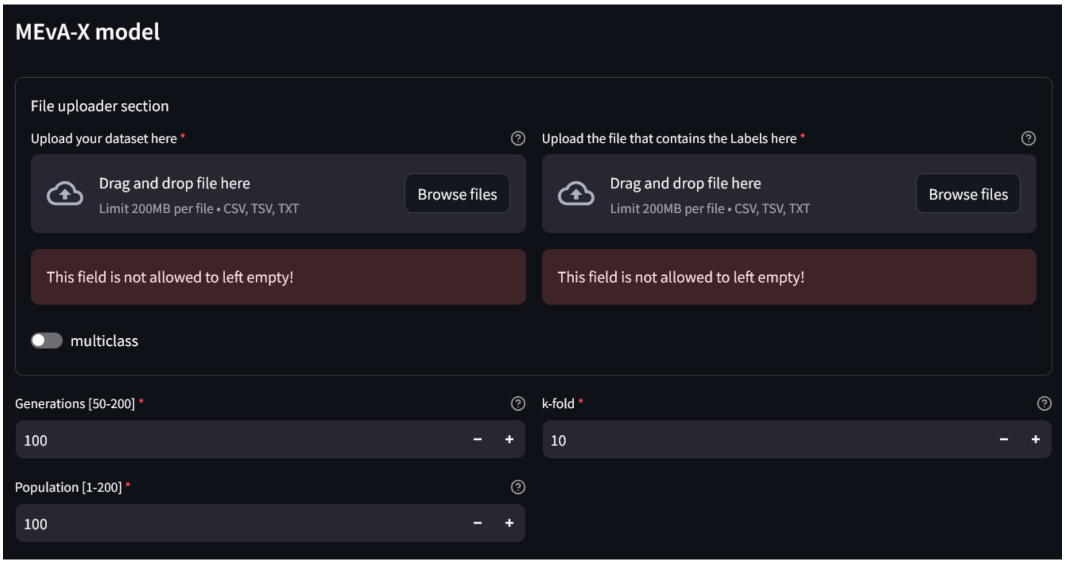


Figure 35. A view of the necessary fields for the MEvA-X tool. The file uploaders allow the user to browse their computer and upload the data from there. The numerical fields of generations, population and k-fold are assigned with their default values.

Following the necessary fields, the platform allows for changing other parameters of the algorithm starting with the preprocessing of the dataset, and there are: the imputation of missing values (True/False), and the normalization of the dataset (True/False). Additionally, users can upload a file with a precalculated feature selection subset based on the JMI, mRMR, SelectKBest and Wilcoxon rank sum algorithms that is used to initiate some of the solutions. Another set of parameters is linked to the evolutionary algorithm process that controls the mutation probability (default 5%) and the cross-over probability (default 5%) on the offspring in every generation. Last, the platform allows for custom weighting of the evaluation metrics

-also called objectives- that MEvA-X tries to optimize or the upload of a file that contains these weights (Figure 36).

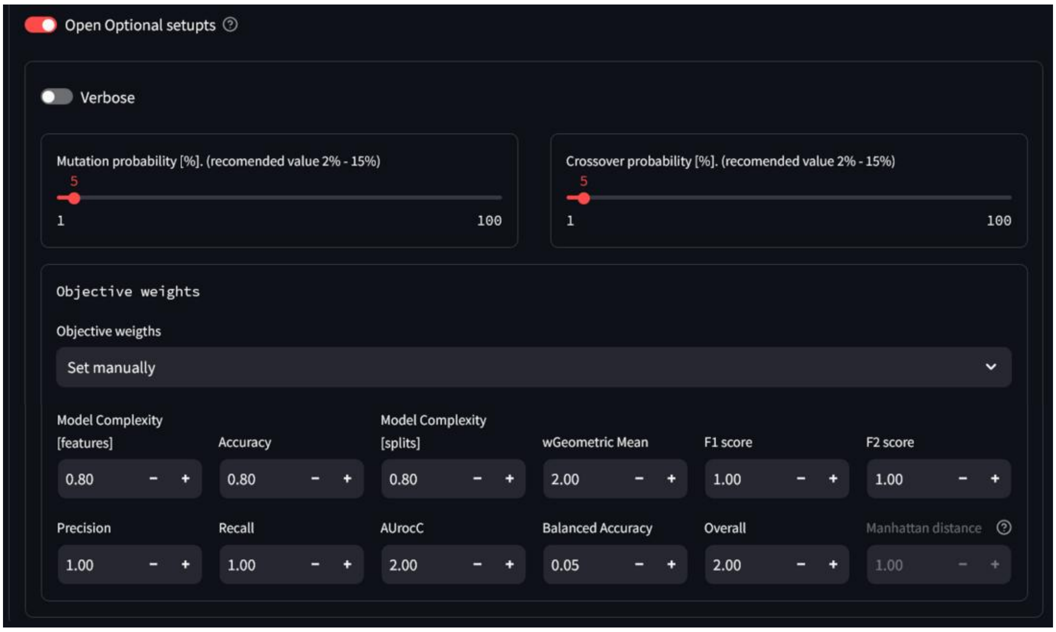
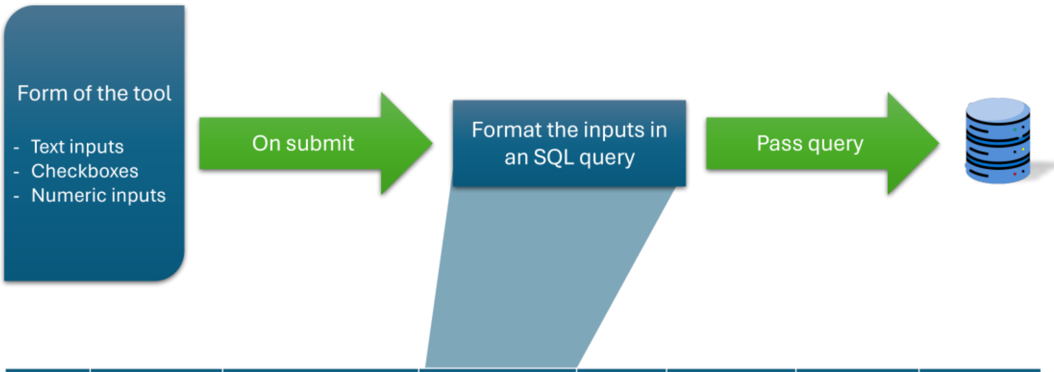


Figure 36. The view of the non-necessary inputs of the MEvA-X tool in the *Biomarker Discovery* page.

To make use of the tool the user is not necessary to have any kind of expertise in machine learning or programming since the parameters are internally transformed in a .JSON file which is stored to the database under the biomarker_discovery_jobs with a unique identifier code. Subsequently, if the scheduler (The scheduler) prioritize the job, the input file with the parameters will be translated into a command line prompt that will be executed asynchronously. Upon execution, the job status changes from “0” to “1” indicating that the job is not pending anymore but it is now running. When the job successfully finishes, the job’s status will become equal to “2” and an e-mail will be notifying the user that this specific job has been completed. In a similar way, in the case of an error, the job’s status will change into “-1” in the database and an e-mail will be sent to the user notifying them about the error.



Job ID	Time_start	User ID [email]	Inputs (Json)	Status	Type	Time_end	Time_run
[Auto]	[Auto]	St.session	text	Init = 0	From input	Init = Na	Init = Na
1	2025/01/01	admin@admin.com	{'A' : 42, 'B' : 'example'}	2	1	2024/12/13	2025/01/01

Figure 37. (Top) The transformation of the inputs from the GUI to the database by an appropriate SQL query. (Bottom) An example table of the database where the input is stored.

6.2.4 Microarray processing tools

This tool is related to the pipelines developed during the PhD research as a method to process microarray datasets retrieved from GEO of the National Center for Biotechnology Information (NCBI). As a matter of fact, this is a collection of methods developed for the different microarray platforms that used to acquire the data in Chapter – 3 of this thesis. After the thorough investigation of the data and their harmonization into a super-collection, the scripts written for this purpose were generalized so that they can be used for other datasets of the same microarray platform, following the same preprocessing and normalization steps. For this, five different retrieving and preprocessing scripts created - one for every microarray design encountered in this study (see Sup_Table 1 for reference) – that allow for seamless processing of data. Further, scripts that normalize the data and convert probes to genes were built, offering an easy converter with standardized steps as the ones followed for the merging of the neurodevelopmental disorder datasets we used in this work.

Finally, a script that handles the merging of different datasets is also available and integrated into the platform. This last module allows for an effortless merging that is also customizable from the user's interface and requires only a good level of understanding of the data the user handles.

As with the page of the biomarker discoverer tool, this page transforms the inputs of the graphical interface into a JSON file which is stored into the database along with other information for the user, the type of the requested job, and the timing (see for reference).

6.2.5 CVD risk prediction tool

A page dedicated to the tool that predicts CVD risk in thalassemia patients is implemented on the platform where users can choose between the two trained models of CVD occurrence within one- and three-years (as described in section 5.2). The two models are loaded directly on the platform and the prediction is done on-the-fly by providing a risk score to the user as a probability of a Thalassemia patient developing CVD in the mid-term future. A view of the page is shown in Figure 38 where there are the necessary input fields that are used as features from the models. The models allow for missing values and thus a user can leave some of those inputs empty.

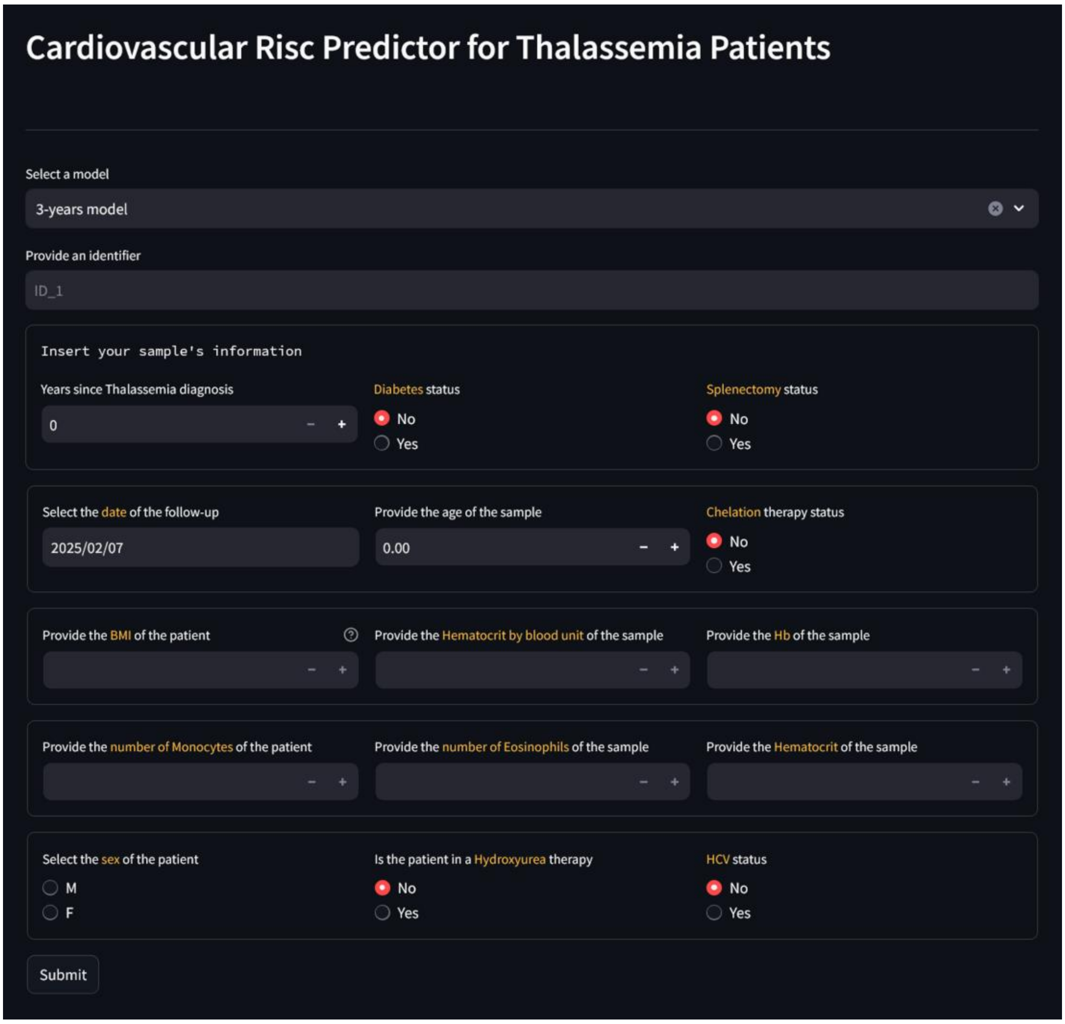


Figure 38. The view of the *CVD risk prediction* tool on the platform.

6.2.6 Database queries

7 A crucial component of the platform is the dedicated database, which as mentioned above (Database) serves as the central repository for storing data. To facilitate user interaction with this database, the platform includes the *Database Viewer* page, providing an intuitive interface for querying and retrieving information. This feature enables users to execute structured queries, explore stored data, and obtain relevant results in a streamlined manner.

On this page, users can submit queries directly to the database through the platform to retrieve information. To assist users in navigating the database structure, dedicated sub-views in the form of drop-down menus and tabs provide an overview of tables and their respective columns. Additionally, a default sample query is provided to retrieve information about the stored phenotypes, along with a multi-selection menu that automatically constructs phenotype-related queries.

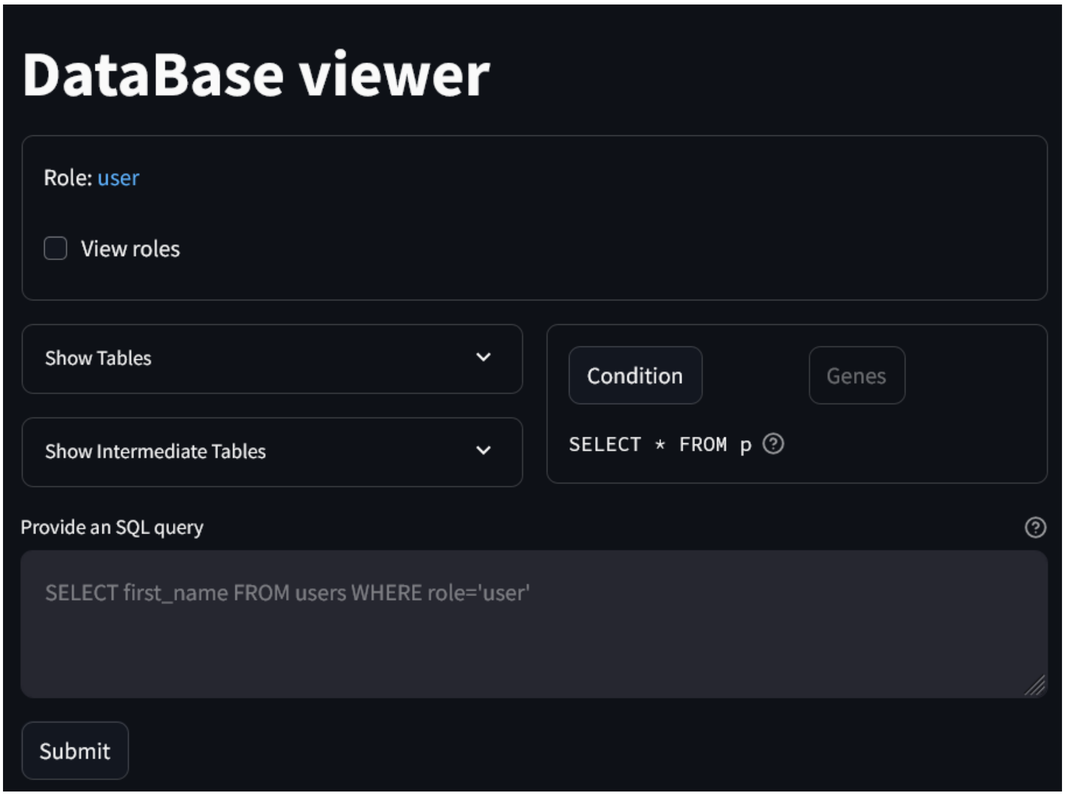


Figure 39. The view of the *Database viewer* page of the platform. The users can submit their own SQL queries to fetch data, or use the condition drop-down menu to select predefined queries.

All users, regardless of their role or login status, have access to the Database Viewer. However, downloading query results is restricted to logged-in users (both users and administrators), excluding guest users (see reference in Figure 29), and the results can be downloaded in .csv or .txt formats.

10.1 The scheduler

The platform is an interactive application that allows the users to navigate in the different pages and access the developed tools, but to submit a job and run it in an efficient manner without making the platform very heavy and without the need of leaving it run for a long time, an asynchronous way was chosen to run the requested jobs in the background of a server and in a different application. Thus, a scheduler that checks if there are any jobs currently running and if there are pending jobs to prioritize them had to be created. Essentially, the scheduler periodically queries the database in the bioinformatics_jobs and biomarker_discovery_jobs to check the status of all the records. If there are no running jobs and there are pending jobs in the queue, it calls the proper tool through the terminal to run and changes the job's status from "0" to "1" ("pending" to "running").

In more detail, the scheduler is a bash-based program that executes simple commands periodically every 5 minutes. These commands are:

```
WHILE True:
    // Query the database to check for rows with status = 1
    status_1_exists = DATABASE QUERY("SELECT COUNT(ID) FROM jobs
                      WHERE status = 1 LIMIT 1")
    IF status_1_exists: // A job is running, thus do nothing
        CONTINUE
    ELSE:                                   // No job is running, check if
there is a pending job to run
        status_0_exists = DATABASE QUERY("SELECT COUNT(ID) FROM
                          jobs WHERE status = 0 LIMIT 1")
        IF status_0_exists:
                tool_inputs = DATASET QUERY("SELECT input FROM jobs
                              WHERE status = 0
                              ORDER BY priority
                              LIMIT 1")
                CALL tool(tool_inputs)
    // Sleep for 5 minutes before the next iteration
    SLEEP(300)
```

10.2 Discussion

The platform is hosted by Politecnico di Torino and the M3B lab's servers. It is open access and can be reached through the following link: http://130.192.23.168:8501. It is advised to create an account for using the platform to be allowed to use all of its functionality as a signed user.

A high-level view of the interconnection of the modules synthesizing the unified platform is presented in Figure 40 showing the structure and the relationship between them. As a matter of fact, the database stores all the essential information for the users and the jobs submitted, and it is the most central part of this structure accessed by the user's interface, the scheduler and individual scripts running in the background.

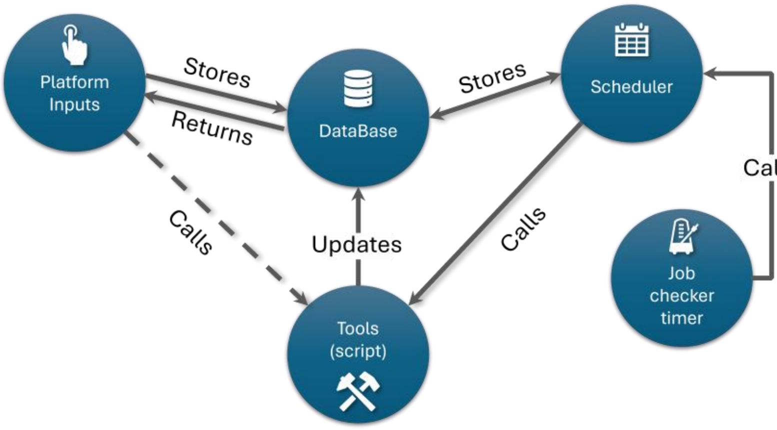


Figure 40. High level view of the interconnections of modules synthesizing and coordinating the function of the platform in the back-end. The front-end receives the users' queries and inputs which are either stored or call a tool directly. In case of asynchronous tools, a job checker scripts requests through the scheduler to run any pending job. The database store information on the status of the jobs and updates its records from the tools responses.

The current platform serves as a wrapper of the developed technologies and tools developed during this research and is a high-level graphical interface referring mostly to users that do not have a deep knowledge of programming but also is an easy way of using the produced tools. The selection of streamlit library to build it was made because of the steep learning curve for this tool, the community support and the fact that it is a python-based library. While initially other frameworks were considered for the platform such as YII-2, PHP, and Django, since this platform serves mostly as a place for users to run the developed tools and is for non-commercial use, the ease of development was prioritized compared to performance.

Considering the database, it was originally built only for the purpose of storing data related to the NDD sets and for serving as a pool of homogenized information across the different studies, prompting the cross-study analysis over a meta-analysis of the data. Over the development of the different AI and statistical tools for biomarker analysis and data processing, the purpose of the database modified to store information for the platform and to serve as its main core of data and information source. The platform sends requests to the database very frequently for updates and for retrieving necessary information for the users. Other modules such as the scheduler and the job checker are independent from the platform itself but work in synergy to find unfinished or pending jobs. This synergy allows for asynchronous execution of the jobs which allows the platform to be more light-weight and run smoother. In fact, jobs such as the biomarker discovery with the MEvA-X tool require a lot of resources from the system in terms of computational power because they also run in parallel processes to speed-up the execution. Thus, the duet of these modules works as the traffic police that allows new jobs to begin only when there is no other process running in that moment. Since the platform is not optimized or designed for large numbers of visitors, the job checker queries the database once every five minutes to see if there is no job running and no job pending at the same time.

The objective of this present platform is to serve as a unifier suit of tools developed for this thesis and the integration of the database in a user-friendly environment. Since the platform was built with a python-based library which is ideal for prototyping, the scalability is low, and it cannot serve many users simultaneously. Thus, it was decided to implement a job scheduler that prioritizes the requests to avoid overloading of the server.

Chapter 7

Conclusions and Future Perspectives

Concluding Remarks

Biomarker discovery is essential in modern medicine in the pursuit of accurate prognosis, early diagnosis, tailored therapies, drug design, and the understanding of complex diseases mechanisms. In the field of bioinformatics, classical statistical approaches have been extremely useful in the identification of molecules for these purposes and nowadays there is an ongoing integration of AI technologies to enhance biomarker discovery in all aspects of the field. Great examples of this integration are the prediction of ncRNA targets as described in Chapter–3 of the present thesis, and the non-redundant biomarker identification with tools like MEvA-X which proposed in Chapter–4. On structural clinical data, biomarkers are extremely important in supporting clinical decisions and helping doctors with the right therapy and intervention at the right time. In this thesis, we additionally presented a tool that predicts a thalassemia patient's risk of developing cardiovascular disease in the mid-term future by giving medics - and the patient - an early notice for prevention and intervention. The biomarkers here are not represented by novel molecules or unknown pathways, but essentially the ML models indicate when a change in known clinical data can lead to certain pathologies. Thus, characterizing states of a disease as of high or low risk for the emergence of a pathology.

Except for algorithms for predicting outcomes and identifying biomarkers, we discussed the need for comprehensive databanks and databases in biology in Chapter-2 and Chapter-3. This is an essential need in medicine which is a field where we constantly try to improve our understanding of disease and the models to describe them. On this matter, we covered some of the biological databases in use today and tools to summarize them focused on ncRNAs but this expands to all other aspects of the field. Additionally, we examined the need for increased sample sizes in conditions that are poorly understood. As example we covered neurodevelopmental disorders where we see the phenotypes which we are able to describe, but not fully able to explain their mechanisms and provide early diagnosis or better therapies. Considering this, we proposed a method of merging and harmonizing datasets from microarray studies on neurodevelopmental disorders where we successfully increased the effect sizes of ASD samples and the typically developed population. As a result, we were able to investigate the conditions in the unified set with an increased statistical certainty on the result, although noticing potential flaws in the overall design which are inherited from the nature of the data and sample diversity, and the methods used.

Additionally, this work explored - in a non-exhaustive manner - the concept of biomarkers across various biological levels and data types. As discussed in Chapter-2, molecular biomarkers cover a wide type of data in different scales and complexities. For this reason and as an effort to unify tools in a single wide-purpose application the platform presented in Chapter–6 was developed. It supports both synchronous and asynchronous execution modes, enabling flexible and user-specific operation. The integration of these features represents a step toward a unified suite of tools for analyzing complex biomedical data, with the potential for future expansion to accommodate emerging biomarker modalities.

Limitations and prospects

Although this thesis presents a series of promising tools and methods for the identification and investigation of biomarkers, it is essential to acknowledge its limitations. Like any scientific work, this thesis is subject to constraints arising from methodological, computational, and biological complexity. In this section, we critically reflect on the main limitations of the study and outline directions for future improvements and research.

From a methodological perspective, one key limitation of this work lies in the evaluation of the developed models and computational methods. Due to the limited availability of high-quality annotated datasets, the performance assessments presented here may not fully capture potential weaknesses or edge cases. This scarcity of data can obscure issues related to model generalizability, especially in complex biological systems where relationships between features and outputs are often nonlinear and context-dependent.

While evaluation metrics were applied, more extensive validation - ideally with larger and more diverse datasets - is necessary to reveal potential shortcomings and improve model robustness. For example, predictive models might benefit substantially from increased training data that better represent biological variability, thereby enhancing their ability to learn subtle patterns.

Similarly, methodological frameworks such as the harmonization pipeline proposed in Chapter - 3 should be rigorously benchmarked against state-of-the-art approaches to assess their practical utility and scalability. Future research should prioritize two directions: first, expanding the availability of harmonized, high-quality datasets; and second, integrating large-scale, multi-omics data to train and validate more comprehensive and biologically informed models.

Overall, the scope of this thesis is broad, and a notable inherent limitation arises from this since it spans multiple dimensions of biomarker research, including molecular and clinical biomarkers, through a combination of bioinformatics pipelines, machine learning models, data harmonization strategies, and software tools. While this interdisciplinary approach provides a comprehensive framework for biomarker discovery and investigation, it also presents substantial integration challenges. The convergence of diverse data types, analytical methodologies, and domain-specific tools often requires trade-offs between depth and breadth, as well as careful coordination across computational, biological, and clinical contexts.

The effort to unify these elements into a single prototype platform represents an important step toward a more holistic biomarker analysis ecosystem. However, this initial integration remains limited in functionality, scalability, and efficiency. Future work should aim to refine this platform, improve modularity, and provide standardized interfaces that facilitate broader adoption and seamless interaction across disciplines.

References

[1] T. Kim and C. M. Croce, “MicroRNA: trends in clinical trials of cancer diagnosis and therapy strategies,” *Exp Mol Med*, vol. 55, no. 7, pp. 1314–1321, Jul. 2023, doi: 10.1038/s12276-023-01050-9.

[2] M. H. Janeiro *et al.*, “Biomarkers in Alzheimer’s disease,” *Advances in Laboratory Medicine / Avances en Medicina de Laboratorio*, vol. 2, no. 1, pp. 27–37, Mar. 2021, doi: 10.1515/almed-2020-0090.

[3] M. Chapleau, L. Iaccarino, D. Soleimani-Meigooni, and G. D. Rabinovici, “The Role of Amyloid PET in Imaging Neurodegenerative Disorders: A Review,” *J Nucl Med*, vol. 63, no. Supplement 1, pp. 13S-19S, Jun. 2022, doi: 10.2967/jnumed.121.263195.

[4] T. E. Inder *et al.*, “Neuroimaging of the Preterm Brain: Review and Recommendations,” *The Journal of Pediatrics*, vol. 237, pp. 276-287.e4, Oct. 2021, doi: 10.1016/j.jpeds.2021.06.014.

[5] M. Bax, C. Tydeman, and O. Flodmark, “Clinical and MRI Correlates of Cerebral Palsy: The European Cerebral Palsy Study,” *JAMA*, vol. 296, no. 13, p. 1602, Oct. 2006, doi: 10.1001/jama.296.13.1602.

[6] M. Crotti, S. Genoe, N. Ben Itzhak, L. Mailleux, and E. Ortibus, “The relation between neuroimaging and visual impairment in children and adolescents with cerebral palsy: A systematic review,” *Brain and Development*, vol. 46, no. 2, pp. 75–92, Feb. 2024, doi: 10.1016/j.braindev.2023.11.002.

[7] S. Arulkumaran *et al.*, “MRI Findings at Term-Corrected Age and Neurodevelopmental Outcomes in a Large Cohort of Very Preterm Infants,” *AJNR Am J Neuroradiol*, vol. 41, no. 8, pp. 1509–1516, Aug. 2020, doi: 10.3174/ajnr.A6666.

[8] C. Peña-Bautista, T. Durand, C. Vigor, C. Oger, J.-M. Galano, and C. Cháfer-Pericás, “Non-invasive assessment of oxidative stress in preterm infants,” *Free Radical Biology and Medicine*, vol. 142, pp. 73–81, Oct. 2019, doi: 10.1016/j.freeradbiomed.2019.02.019.

[9] K. Beardsall *et al.*, “Real-time continuous glucose monitoring in preterm infants (REACT): an international, open-label, randomised controlled trial,” *The Lancet Child & Adolescent Health*, vol. 5, no. 4, pp. 265–273, Apr. 2021, doi: 10.1016/S2352-4642(20)30367-9.

[10] A. Galderisi *et al.*, “Continuous Glucose Monitoring in Very Preterm Infants: A Randomized Controlled Trial,” *Pediatrics*, vol. 140, no. 4, p. e20171162, Oct. 2017, doi: 10.1542/peds.2017-1162.

[11] R. L. Goldenberg, J. F. Culhane, J. D. Iams, and R. Romero, “Epidemiology and causes of preterm birth,” *The Lancet*, vol. 371, no. 9606, pp. 75–84, Jan. 2008, doi: 10.1016/S0140-6736(08)60074-4.

[12] H. Blencowe *et al.*, “National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications,” *The Lancet*, vol. 379, no. 9832, pp. 2162–2172, Jun. 2012, doi: 10.1016/S0140-6736(12)60820-4.

[13] E. O. Ohuma *et al.*, “National, regional, and global estimates of preterm birth in 2020, with trends from 2010: a systematic analysis,” *The Lancet*, vol. 402, no. 10409, pp. 1261–1271, Oct. 2023, doi: 10.1016/S0140-6736(23)00878-4.

[14] D. Dewey, “What Is Comorbidity and Why Does It Matter in Neurodevelopmental Disorders?,” *Curr Dev Disord Rep*, vol. 5, no. 4, pp. 235–242, Dec. 2018, doi: 10.1007/s40474-018-0152-3.

[15] D. J. Zgaljardic and R. O. Temple, “Neuropsychological Assessment Battery (NAB): Performance in a Sample of Patients with Moderate-to-Severe Traumatic Brain Injury,” *Applied Neuropsychology*, vol. 17, no. 4, pp. 283–288, Nov. 2010, doi: 10.1080/09084282.2010.525118.

[16] L. E. L. M. Vissers, C. Gilissen, and J. A. Veltman, “Genetic studies in intellectual disability and related disorders,” *Nat Rev Genet*, vol. 17, no. 1, pp. 9–18, Jan. 2016, doi: 10.1038/nrg3999.

[17] A. H. MacLennan, S. C. Thompson, and J. Gecz, “Cerebral palsy: causes, pathways, and the role of genetic variants,” *American Journal of Obstetrics and Gynecology*, vol. 213, no. 6, pp. 779–788, Dec. 2015, doi: 10.1016/j.ajog.2015.05.034.

[18] J. H. Thygesen *et al.*, “Neurodevelopmental risk copy number variants in adults with intellectual disabilities and comorbid psychiatric disorders,” *Br J Psychiatry*, vol. 212, no. 5, pp. 287–294, May 2018, doi: 10.1192/bjp.2017.65.

[19] M. K. Licari *et al.*, “The Brain Basis of Comorbidity in Neurodevelopmental Disorders,” *Curr Dev Disord Rep*, vol. 6, no. 1, pp. 9–18, Mar. 2019, doi: 10.1007/s40474-019-0156-7.

[20] P. Zhang, W. Wu, Q. Chen, and M. Chen, “Non-Coding RNAs and their Integrated Networks,” *Journal of Integrative Bioinformatics*, vol. 16, no. 3, p. 20190027, Sep. 2019, doi: 10.1515/jib-2019-0027.

[21] F. Crick, “Central Dogma of Molecular Biology,” *Nature*, vol. 227, no. 5258, pp. 561–563, Aug. 1970, doi: 10.1038/227561a0.

[22] The ENCODE Project Consortium, “An integrated encyclopedia of DNA elements in the human genome,” *Nature*, vol. 489, no. 7414, pp. 57–74, Sep. 2012, doi: 10.1038/nature11247.

[23] A. S. Cristino *et al.*, "Neurodevelopmental and neuropsychiatric disorders represent an interconnected molecular system," *Mol Psychiatry*, vol. 19, no. 3, pp. 294–301, Mar. 2014, doi: 10.1038/mp.2013.16.

[24] R. A. Clarke, S. Lee, and V. Eapen, "Pathogenetic model for Tourette syndrome delineates overlap with related neurodevelopmental disorders including Autism," *Transl Psychiatry*, vol. 2, no. 9, pp. e158–e158, Sep. 2012, doi: 10.1038/tp.2012.75.

[25] M. Sahin and M. Sur, "Genes, circuits, and precision therapies for autism and related neurodevelopmental disorders," *Science*, vol. 350, no. 6263, p. aab3897, Nov. 2015, doi: 10.1126/science.aab3897.

[26] I. V. Novikova, S. P. Hennelly, and K. Y. Sanbonmatsu, "Sizing up long non-coding RNAs: Do lncRNAs have secondary and tertiary structure?," *BioArchitecture*, vol. 2, no. 6, pp. 189–199, Nov. 2012, doi: 10.4161/bioa.22592.

[27] A. Mahfouz, M. N. Ziats, O. M. Rennert, B. P. F. Lelieveldt, and M. J. T. Reinders, "Shared Pathways Among Autism Candidate Genes Determined by Co-expression Network Analysis of the Developing Human Brain Transcriptome," *J Mol Neurosci*, vol. 57, no. 4, pp. 580–594, Dec. 2015, doi: 10.1007/s12031-015-0641-3.

[28] S.-F. Zhang, J. Gao, and C.-M. Liu, "The Role of Non-Coding RNAs in Neurodevelopmental Disorders," *Front. Genet.*, vol. 10, p. 1033, Nov. 2019, doi: 10.3389/fgene.2019.01033.

[29] S. K. Fineberg, K. S. Kosik, and B. L. Davidson, "MicroRNAs Potentiate Neural Development," *Neuron*, vol. 64, no. 3, pp. 303–309, Nov. 2009, doi: 10.1016/j.neuron.2009.10.020.

[30] C. N. Watson, A. Belli, and V. Di Pietro, "Small Non-coding RNAs: New Class of Biomarkers and Potential Therapeutic Targets in Neurodegenerative Disease," *Front. Genet.*, vol. 10, p. 364, Apr. 2019, doi: 10.3389/fgene.2019.00364.

[31] J. O'Brien, H. Hayder, Y. Zayed, and C. Peng, "Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation," *Front. Endocrinol.*, vol. 9, p. 402, Aug. 2018, doi: 10.3389/fendo.2018.00402.

[32] M. Muñoz-Culla *et al.*, "SncRNA (microRNA & snoRNA) opposite expression pattern found in multiple sclerosis relapse and remission is sex dependent," *Sci Rep*, vol. 6, no. 1, p. 20126, Feb. 2016, doi: 10.1038/srep20126.

[33] O. Noronha, L. Mesarosovo, J. J. Anink, A. Iyer, E. Aronica, and J. D. Mills, "Differentially Expressed miRNAs in Age-Related Neurodegenerative

Diseases: A Meta-Analysis," *Genes*, vol. 13, no. 6, p. 1034, Jun. 2022, doi: 10.3390/genes13061034.

[34] C. P. D. C. Gomes *et al.*, "Regulatory RNAs in Heart Failure," *Circulation*, vol. 141, no. 4, pp. 313–328, Jan. 2020, doi: 10.1161/CIRCULATIONAHA.119.042474.

[35] E. S. Smith, E. Whitty, B. Yoo, A. Moore, L. F. Sempere, and Z. Medarova, "Clinical Applications of Short Non-Coding RNA-Based Therapies in the Era of Precision Medicine," *Cancers*, vol. 14, no. 6, p. 1588, Mar. 2022, doi: 10.3390/cancers14061588.

[36] Y. Peng and C. M. Croce, "The role of MicroRNAs in human cancer," *Sig Transduct Target Ther*, vol. 1, no. 1, p. 15004, Jan. 2016, doi: 10.1038/sigtrans.2015.4.

[37] R. Bardini, G. Politano, A. Benso, and S. Di Carlo, "Multi-level and hybrid modelling approaches for systems biology," *Computational and Structural Biotechnology Journal*, vol. 15, pp. 396–402, 2017, doi: 10.1016/j.csbj.2017.07.005.

[38] Z. Ji, K. Yan, W. Li, H. Hu, and X. Zhu, "Mathematical and Computational Modeling in Complex Biological Systems," *BioMed Research International*, vol. 2017, pp. 1–16, 2017, doi: 10.1155/2017/5958321.

[39] A. Oulas *et al.*, "Prediction of miRNA Targets," in *RNA Bioinformatics*, vol. 1269, E. Picardi, Ed., in Methods in Molecular Biology, vol. 1269. , New York, NY: Springer New York, 2015, pp. 207–229. doi: 10.1007/978-1-4939-2291-8_13.

[40] A. Korfiati, K. Theofilatos, D. Kleftogiannis, C. Alexakos, S. Likothanassis, and S. Mavroudi, "Predicting human miRNA target genes using a novel computational intelligent framework," *Information Sciences*, vol. 294, pp. 576–585, Feb. 2015, doi: 10.1016/j.ins.2014.09.016.

[41] M. N. Ziats and O. M. Rennert, "Identification of differentially expressed microRNAs across the developing human brain," *Mol Psychiatry*, vol. 19, no. 7, pp. 848–852, Jul. 2014, doi: 10.1038/mp.2013.93.

[42] H. J. Kang *et al.*, "Spatio-temporal transcriptome of the human brain," *Nature*, vol. 478, no. 7370, pp. 483–489, Oct. 2011, doi: 10.1038/nature10523.

[43] P. Paul *et al.*, "Interplay between miRNAs and human diseases," *Journal Cellular Physiology*, vol. 233, no. 3, pp. 2007–2018, Mar. 2018, doi: 10.1002/jcp.25854.

[44] H. Lv *et al.*, "Neonatal hypoxic ischemic encephalopathy-related biomarkers in serum and cerebrospinal fluid," *Clinica Chimica Acta*, vol. 450, pp. 282–297, Oct. 2015, doi: 10.1016/j.cca.2015.08.021.

[45] M. Ha and V. N. Kim, “Regulation of microRNA biogenesis,” *Nat Rev Mol Cell Biol*, vol. 15, no. 8, pp. 509–524, Aug. 2014, doi: 10.1038/nrm3838.

[46] J. K. W. Lam, M. Y. T. Chow, Y. Zhang, and S. W. S. Leung, “siRNA Versus miRNA as Therapeutics for Gene Silencing,” *Molecular Therapy - Nucleic Acids*, vol. 4, p. e252, 2015, doi: 10.1038/mtna.2015.23.

[47] C. B. Assumpção *et al.*, “The Role of piRNA and its Potential Clinical Implications in Cancer,” *Epigenomics*, vol. 7, no. 6, pp. 975–984, Sep. 2015, doi: 10.2217/epi.15.37.

[48] S. M. Peterson, J. A. Thompson, M. L. Ufkin, P. Sathyanarayana, L. Liaw, and C. B. Congdon, “Common features of microRNA target prediction tools,” *Front. Genet.*, vol. 5, 2014, doi: 10.3389/fgene.2014.00023.

[49] K. T. Thomas and S. S. Zakharenko, “MicroRNAs in the Onset of Schizophrenia,” *Cells*, vol. 10, no. 10, p. 2679, Oct. 2021, doi: 10.3390/cells10102679.

[50] S. Chang, S. Wen, D. Chen, and P. Jin, “Small regulatory RNAs in neurodevelopmental disorders,” *Human Molecular Genetics*, vol. 18, no. R1, pp. R18–R26, Apr. 2009, doi: 10.1093/hmg/ddp072.

[51] M. Costa-Mattioli and L. M. Monteggia, “mTOR complexes in neurodevelopmental and neuropsychiatric disorders,” *Nat Neurosci*, vol. 16, no. 11, pp. 1537–1543, Nov. 2013, doi: 10.1038/nn.3546.

[52] I. Parenti, L. G. Rabaneda, H. Schoen, and G. Novarino, “Neurodevelopmental Disorders: From Genetics to Functional Pathways,” *Trends in Neurosciences*, vol. 43, no. 8, pp. 608–621, Aug. 2020, doi: 10.1016/j.tins.2020.05.004.

[53] G. Son *et al.*, “miR-124 coordinates metabolic regulators acting at early stages of human neurogenesis,” *Commun Biol*, vol. 7, no. 1, p. 1393, Oct. 2024, doi: 10.1038/s42003-024-07089-2.

[54] G. R. Aaluri, Y. Choudhary, and S. Kumar, “Mitochondria-Associated MicroRNAs and Parkinson’s Disease,” *J Exp Neurosci*, vol. 19, p. 26331055241254846, Jan. 2024, doi: 10.1177/26331055241254846.

[55] Z. Alkhazaali-Ali, S. Sahab-Negah, A. R. Boroumand, and J. Tavakol-Afshari, “MicroRNA (miRNA) as a biomarker for diagnosis, prognosis, and therapeutics molecules in neurodegenerative disease,” *Biomedicine & Pharmacotherapy*, vol. 177, p. 116899, Aug. 2024, doi: 10.1016/j.biopha.2024.116899.

[56] M. A. Obeid *et al.*, “Targeting siRNAs in cancer drug delivery,” in *Advanced Drug Delivery Systems in the Management of Cancer*, Elsevier, 2021, pp. 447–460. doi: 10.1016/B978-0-323-85503-7.00027-4.

[57] L. H. Madkour, "Therapeutic applications of siRNA gene delivery systems," in *Nucleic Acids as Gene Anticancer Drug Delivery Therapy*, Elsevier, 2019, pp. 65–74. doi: 10.1016/B978-0-12-819777-6.00005-6.

[58] G. Jain *et al.*, "A combined miRNA–piRNA signature to detect Alzheimer's disease," *Transl Psychiatry*, vol. 9, no. 1, p. 250, Oct. 2019, doi: 10.1038/s41398-019-0579-2.

[59] T. N. Turner and E. E. Eichler, "The Role of De Novo Noncoding Regulatory Mutations in Neurodevelopmental Disorders," *Trends in Neurosciences*, vol. 42, no. 2, pp. 115–127, Feb. 2019, doi: 10.1016/j.tins.2018.11.002.

[60] S. Subramaniyan, B. B. Kuriakose, S. Mushfiq, N. M. Prabhu, and K. Muthusamy, "Gene Signals and SNPs Associated with Parkinson's Disease: A Nutrigenomics and Computational Prospective Insights," *Neuroscience*, vol. 533, pp. 77–95, Nov. 2023, doi: 10.1016/j.neuroscience.2023.10.007.

[61] J. Z. Liu and C. A. Anderson, "Genetic studies of Crohn's disease: Past, present and future," *Best Practice & Research Clinical Gastroenterology*, vol. 28, no. 3, pp. 373–386, Jun. 2014, doi: 10.1016/j.bpg.2014.04.009.

[62] T. Munkongdee *et al.*, "Predictive SNPs for β0-thalassemia/HbE disease severity," *Sci Rep*, vol. 11, no. 1, p. 10352, May 2021, doi: 10.1038/s41598-021-89641-2.

[63] M. J. Salcedo-Arellano, B. Dufour, Y. McLennan, V. Martinez-Cerdeno, and R. Hagerman, "Fragile X syndrome and associated disorders: Clinical aspects and pathology," *Neurobiology of Disease*, vol. 136, p. 104740, Mar. 2020, doi: 10.1016/j.nbd.2020.104740.

[64] S. R. Shuken, "An Introduction to Mass Spectrometry-Based Proteomics," *J. Proteome Res.*, vol. 22, no. 7, pp. 2151–2171, Jul. 2023, doi: 10.1021/acs.jproteome.2c00838.

[65] J. Hugosson *et al.*, "A 16-yr Follow-up of the European Randomized study of Screening for Prostate Cancer," *European Urology*, vol. 76, no. 1, pp. 43–51, Jul. 2019, doi: 10.1016/j.eururo.2019.02.009.

[66] G. Kustatscher *et al.*, "Understudied proteins: opportunities and challenges for functional proteomics," *Nat Methods*, vol. 19, no. 7, pp. 774–779, Jul. 2022, doi: 10.1038/s41592-022-01454-x.

[67] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, Aug. 2021, doi: 10.1038/s41586-021-03819-2.

[68] P. Alhopuro, R. Vainionpää, A.-K. Anttonen, K. Aittomäki, H. Nevanlinna, and M. Pöyhönen, "Constitutional mosaicism for a BRCA2 mutation as a cause of early-onset breast cancer," *Familial Cancer*, vol. 19, no. 4, pp. 307–310, Oct. 2020, doi: 10.1007/s10689-020-00186-1.

[69] M. Chahrour and H. Y. Zoghbi, “The Story of Rett Syndrome: From Clinic to Neurobiology,” *Neuron*, vol. 56, no. 3, pp. 422–437, Nov. 2007, doi: 10.1016/j.neuron.2007.10.001.

[70] W. A. Gold *et al.*, “Rett syndrome,” *Nat Rev Dis Primers*, vol. 10, no. 1, p. 84, Nov. 2024, doi: 10.1038/s41572-024-00568-0.

[71] A. Pasqui *et al.*, “A proteomic approach to investigate the role of the MECP2 gene mutation in Rett syndrome redox regulatory pathways,” *Archives of Biochemistry and Biophysics*, vol. 752, p. 109860, Feb. 2024, doi: 10.1016/j.abb.2023.109860.

[72] G. Middleton, H. Robbins, F. Andre, and C. Swanton, “A state-of-the-art review of stratified medicine in cancer: towards a future precision medicine strategy in cancer,” *Annals of Oncology*, vol. 33, no. 2, pp. 143–157, Feb. 2022, doi: 10.1016/j.annonc.2021.11.004.

[73] T. W. MacFarland and J. M. Yates, “Mann–Whitney U Test,” in *Introduction to Nonparametric Statistics for the Biological Sciences Using R*, Cham: Springer International Publishing, 2016, pp. 103–132. doi: 10.1007/978-3-319-30634-6_4.

[74] P. E. McKnight and J. Najab, “Mann-Whitney U Test,” in *The Corsini Encyclopedia of Psychology*, 1st ed., I. B. Weiner and W. E. Craighead, Eds., Wiley, 2010, pp. 1–1. doi: 10.1002/9780470479216.corpsy0524.

[75] W. H. Kruskal and W. A. Wallis, “Use of Ranks in One-Criterion Variance Analysis,” *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, Dec. 1952, doi: 10.1080/01621459.1952.10483441.

[76] P. E. McKight and J. Najab, “Kruskal-Wallis Test,” in *The Corsini Encyclopedia of Psychology*, 1st ed., I. B. Weiner and W. E. Craighead, Eds., Wiley, 2010, pp. 1–1. doi: 10.1002/9780470479216.corpsy0491.

[77] Y. Benjamini and Y. Hochberg, “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 57, no. 1, pp. 289–300, Jan. 1995, doi: 10.1111/j.2517-6161.1995.tb02031.x.

[78] R. A. Armstrong, “When to use the Bonferroni correction,” *Ophthalmic Physiologic Optic*, vol. 34, no. 5, pp. 502–508, Sep. 2014, doi: 10.1111/opo.12131.

[79] J. D. Storey, “The positive false discovery rate: a Bayesian interpretation and the q-value,” *Ann. Statist.*, vol. 31, no. 6, Dec. 2003, doi: 10.1214/aos/1074290335.

[80] E. Uffelmann *et al.*, “Genome-wide association studies,” *Nat Rev Methods Primers*, vol. 1, no. 1, p. 59, Aug. 2021, doi: 10.1038/s43586-021-00056-9.

[81] S. Azidane, X. Gallego, L. Durham, M. Cáceres, E. Guney, and L. Pérez-Cano, “Identification of novel driver risk genes in CNV loci associated with neurodevelopmental disorders,” *Human Genetics and Genomics Advances*, vol. 5, no. 3, p. 100316, Jul. 2024, doi: 10.1016/j.xhgg.2024.100316.

[82] A. J. Schork *et al.*, “A genome-wide association study of shared risk across psychiatric disorders implicates gene regulation during fetal neurodevelopment,” *Nat Neurosci*, vol. 22, no. 3, pp. 353–361, Mar. 2019, doi: 10.1038/s41593-018-0320-0.

[83] S. Gao, C. Shan, R. Zhang, and T. Wang, “Genetic advances in neurodevelopmental disorders,” *Medical Review*, Sep. 2024, doi: 10.1515/mr-2024-0040.

[84] E. C. Plunk, W. S. Chambers, and S. M. Richards, “System biology,” in *Metabolomics Perspectives*, Elsevier, 2022, pp. 3–25. doi: 10.1016/B978-0-323-85062-9.00001-5.

[85] R. Lowe, N. Shirley, M. Bleackley, S. Dolan, and T. Shafee, “Transcriptomics technologies,” *PLoS Comput Biol*, vol. 13, no. 5, p. e1005457, May 2017, doi: 10.1371/journal.pcbi.1005457.

[86] Z. Wang, M. Gerstein, and M. Snyder, “RNA-Seq: a revolutionary tool for transcriptomics,” *Nat Rev Genet*, vol. 10, no. 1, pp. 57–63, Jan. 2009, doi: 10.1038/nrg2484.

[87] D. Granato, J. S. Santos, G. B. Escher, B. L. Ferreira, and R. M. Maggio, “Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective,” *Trends in Food Science & Technology*, vol. 72, pp. 83–90, Feb. 2018, doi: 10.1016/j.tifs.2017.12.006.

[88] L. van der Maaten and G. Hinton, “Visualizing Data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, Nov. 2008, [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html

[89] A. D. Baxevanis, “The Importance of Biological Databases in Biological Discovery,” *CP in Bioinformatics*, vol. 34, no. 1, Jun. 2011, doi: 10.1002/0471250953.bi0101s34.

[90] T. Barrett *et al.*, “NCBI GEO: archive for functional genomics data sets—update,” *Nucleic Acids Research*, vol. 41, no. D1, pp. D991–D995, Nov. 2012, doi: 10.1093/nar/gks1193.

[91] “Database resources of the National Center for Biotechnology Information,” *Nucleic Acids Res*, vol. 44, no. D1, pp. D7–D19, Jan. 2016, doi: 10.1093/nar/gkv1290.

[92] A. Athar *et al.*, "ArrayExpress update – from bulk to single-cell expression data," *Nucleic Acids Research*, vol. 47, no. D1, pp. D711–D715, Jan. 2019, doi: 10.1093/nar/gky964.

[93] S. Griffiths-Jones, "The microRNA Registry," *Nucleic Acids Research*, vol. 32, no. 90001, pp. 109D – 111, Jan. 2004, doi: 10.1093/nar/gkh023.

[94] S. Griffiths-Jones, H. K. Saini, S. Van Dongen, and A. J. Enright, "miRBase: tools for microRNA genomics," *Nucleic Acids Research*, vol. 36, no. Database, pp. D154–D158, Dec. 2007, doi: 10.1093/nar/gkm952.

[95] H.-Y. Huang *et al.*, "miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database," *Nucleic Acids Research*, p. gkz896, Oct. 2019, doi: 10.1093/nar/gkz896.

[96] C. Backes *et al.*, "miRCarta: a central repository for collecting miRNA candidates," *Nucleic Acids Research*, vol. 46, no. D1, pp. D160–D167, Jan. 2018, doi: 10.1093/nar/gkx851.

[97] T. Fehlmann *et al.*, "miRMaster 2.0: multi-species non-coding RNA sequencing analyses at scale," *Nucleic Acids Research*, vol. 49, no. W1, pp. W397–W408, Jul. 2021, doi: 10.1093/nar/gkab268.

[98] C.-J. Liu *et al.*, "EVAtlas: a comprehensive database for ncRNA expression in human extracellular vesicles," *Nucleic Acids Research*, vol. 50, no. D1, pp. D111–D117, Jan. 2022, doi: 10.1093/nar/gkab668.

[99] L. Chang, G. Zhou, O. Soufan, and J. Xia, "miRNet 2.0: network-based visual analytics for miRNA functional analysis and systems biology," *Nucleic Acids Research*, vol. 48, no. W1, pp. W244–W251, Jul. 2020, doi: 10.1093/nar/gkaa467.

[100] A. Fabregat *et al.*, "The Reactome pathway Knowledgebase," *Nucleic Acids Research*, vol. 44, no. D1, pp. D481–D487, Jan. 2016, doi: 10.1093/nar/gkv1351.

[101] T. Kehl *et al.*, "miRPathDB 2.0: a novel release of the miRNA Pathway Dictionary Database," *Nucleic Acids Research*, vol. 48, no. D1, pp. D142–D147, Jan. 2020, doi: 10.1093/nar/gkz1022.

[102] J. Gong *et al.*, "RISE: a database of RNA interactome from sequencing experiments," *Nucleic Acids Research*, vol. 46, no. D1, pp. D194–D201, Jan. 2018, doi: 10.1093/nar/gkx864.

[103] X. Teng *et al.*, "NPInter v4.0: an integrated database of ncRNA interactions," *Nucleic Acids Research*, p. gkz969, Oct. 2019, doi: 10.1093/nar/gkz969.

[104] J.-H. Li, S. Liu, H. Zhou, L.-H. Qu, and J.-H. Yang, "starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks

from large-scale CLIP-Seq data," *Nucl. Acids Res.*, vol. 42, no. D1, pp. D92–D97, Jan. 2014, doi: 10.1093/nar/gkt1248.

[105] V. Agarwal, G. W. Bell, J.-W. Nam, and D. P. Bartel, "Predicting effective microRNA target sites in mammalian mRNAs," *eLife*, vol. 4, p. e05005, Aug. 2015, doi: 10.7554/eLife.05005.

[106] M. D. Paraskevopoulou *et al.*, "DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows," *Nucleic Acids Research*, vol. 41, no. W1, pp. W169–W173, Jul. 2013, doi: 10.1093/nar/gkt393.

[107] D. Karagkouni *et al.*, "DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA–gene interactions," *Nucleic Acids Research*, vol. 46, no. D1, pp. D239–D245, Jan. 2018, doi: 10.1093/nar/gkx1141.

[108] F. Xiao, Z. Zuo, G. Cai, S. Kang, X. Gao, and T. Li, "miRecords: an integrated resource for microRNA-target interactions," *Nucleic Acids Research*, vol. 37, no. Database, pp. D105–D110, Jan. 2009, doi: 10.1093/nar/gkn851.

[109] Q. Jiang *et al.*, "miR2Disease: a manually curated database for microRNA deregulation in human disease," *Nucleic Acids Research*, vol. 37, no. Database, pp. D98–D104, Jan. 2009, doi: 10.1093/nar/gkn714.

[110] P. N. Deepthi and R. Anitha, "Applications of Network Analysis in Bioinformatics," in *Advances in Computational and Bio-Engineering*, vol. 16, S. Jyothi, D. M. Mamatha, S. C. Satapathy, K. S. Raju, and M. N. Favorskaya, Eds., in Learning and Analytics in Intelligent Systems, vol. 16. , Cham: Springer International Publishing, 2020, pp. 79–84. doi: 10.1007/978-3-030-46943-6_9.

[111] A. Mrvar and V. Batagelj, "Analysis and visualization of large networks with program package Pajek," *Complex Adapt Syst Model*, vol. 4, no. 1, p. 6, Dec. 2016, doi: 10.1186/s40294-016-0017-8.

[112] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An Open Source Software for Exploring and Manipulating Networks," *ICWSM*, vol. 3, no. 1, pp. 361–362, Mar. 2009, doi: 10.1609/icwsm.v3i1.13937.

[113] P. Shannon *et al.*, "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks," *Genome Res.*, vol. 13, no. 11, pp. 2498–2504, Nov. 2003, doi: 10.1101/gr.1239303.

[114] K. Theofilatos, C. Dimitrakopoulos, C. Alexakos, A. Korfiati, S. Likothanassis, and S. Mavroudi, "InSyBio BioNets: an efficient tool for network-based biomarker discovery," *EMBnet j.*, vol. 22, no. 0, p. 871, Dec. 2016, doi: 10.14806/ej.22.0.871.

[115] D. Ledesma, S. Symes, and S. Richards, "Advancements within Modern Machine Learning Methodology: Impacts and Prospects in Biomarker Discovery," *CMC*, vol. 28, no. 32, pp. 6512–6531, Oct. 2021, doi: 10.2174/0929867328666210208111821.

[116] Y. Xie *et al.*, "Early lung cancer diagnostic biomarker discovery by machine learning methods," *Translational Oncology*, vol. 14, no. 1, p. 100907, Jan. 2021, doi: 10.1016/j.tranon.2020.100907.

[117] L. Hewitson, J. A. Mathews, M. Devlin, C. Schutte, J. Lee, and D. C. German, "Blood biomarker discovery for autism spectrum disorder: A proteomic analysis," *PLoS ONE*, vol. 19, no. 12, p. e0302951, Dec. 2024, doi: 10.1371/journal.pone.0302951.

[118] S. Kim, A. Kim, J.-Y. Shin, and J.-S. Seo, "The tumor immune microenvironmental analysis of 2,033 transcriptomes across 7 cancer types," *Sci Rep*, vol. 10, no. 1, p. 9536, Jun. 2020, doi: 10.1038/s41598-020-66449-0.

[119] K. Hamanaka *et al.*, "Large-scale discovery of novel neurodevelopmental disorder-related genes through a unified analysis of single-nucleotide and copy number variants," *Genome Med*, vol. 14, no. 1, p. 40, Dec. 2022, doi: 10.1186/s13073-022-01042-w.

[120] R. Fernandez Rojas, X. Huang, and K.-L. Ou, "A Machine Learning Approach for the Identification of a Biomarker of Human Pain using fNIRS," *Sci Rep*, vol. 9, no. 1, p. 5645, Apr. 2019, doi: 10.1038/s41598-019-42098-w.

[121] M. Turewicz, M. Ahrens, C. May, K. Marcus, and M. Eisenacher, "PAA: an R/bioconductor package for biomarker discovery with protein microarrays," *Bioinformatics*, vol. 32, no. 10, pp. 1577–1579, May 2016, doi: 10.1093/bioinformatics/btw037.

[122] American Psychiatric Association, Ed., *Diagnostic and statistical manual of mental disorders: DSM-5-TR™*, Fifth edition, text Revision. Washington, DC: American Psychiatric Association Publishing, 2022.

[123] J. P. Gregg *et al.*, "Gene expression changes in children with autism," *Genomics*, vol. 91, no. 1, pp. 22–29, Jan. 2008, doi: 10.1016/j.ygeno.2007.09.003.

[124] S. W. Kong *et al.*, "Characteristics and Predictive Value of Blood Transcriptome Signature in Males with Autism Spectrum Disorders," *PLoS ONE*, vol. 7, no. 12, p. e49475, Dec. 2012, doi: 10.1371/journal.pone.0049475.

[125] T. Pramparo *et al.*, "Prediction of Autism by Translation and Immune/Inflammation Coexpressed Genes in Toddlers From Pediatric

Community Practices," *JAMA Psychiatry*, vol. 72, no. 4, pp. 386–394, Apr. 2015, doi: 10.1001/jamapsychiatry.2014.3008.

[126] T. Pramparo *et al.*, "Cell cycle networks link gene expression dysregulation, mutation, and brain maldevelopment in autistic toddlers," *Molecular Systems Biology*, vol. 11, no. 12, p. 841, Dec. 2015, doi: 10.15252/msb.20156108.

[127] R. Kimura *et al.*, "Integrative network analysis reveals biological pathways associated with Williams syndrome," *Child Psychology Psychiatry*, vol. 60, no. 5, pp. 585–598, May 2019, doi: 10.1111/jcpp.12999.

[128] R. Kimura *et al.*, "An epigenetic biomarker for adult high-functioning autism spectrum disorder," *Sci Rep*, vol. 9, no. 1, p. 13662, Sep. 2019, doi: 10.1038/s41598-019-50250-9.

[129] M. V. Lombardo *et al.*, "Atypical genomic cortical patterning in autism with poor early language outcome," *Sci. Adv.*, vol. 7, no. 36, p. eabh1663, Sep. 2021, doi: 10.1126/sciadv.abh1663.

[130] V. H. Gazestani *et al.*, "A perturbed gene network containing PI3K–AKT, RAS–ERK and WNT–β-catenin pathways in leukocytes is linked to ASD genetics and symptom severity," *Nat Neurosci*, vol. 22, no. 10, pp. 1624–1634, Oct. 2019, doi: 10.1038/s41593-019-0489-x.

[131] Y. Tian *et al.*, "Correlations of gene expression with ratings of inattention and hyperactivity/impulsivity in tourette syndrome: a pilot study," *BMC Med Genomics*, vol. 5, no. 1, p. 49, Dec. 2012, doi: 10.1186/1755-8794-5-49.

[132] "HumanHT-12 v4.0 Gene Expression BeadChip Product Files." Illumina, Inc., 2013. [Online]. Available: https://emea.support.illumina.com/downloads/humanht-12_v4_product_file.html

[133] F. Ayhan and G. Konopka, "Regulatory genes and pathways disrupted in autism spectrum disorders," *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 89, pp. 57–64, Mar. 2019, doi: 10.1016/j.pnpbp.2018.08.017.

[134] J. R. Sweigert *et al.*, "Characterizing Olfactory Function in Children with Autism Spectrum Disorder and Children with Sensory Processing Dysfunction," *Brain Sciences*, vol. 10, no. 6, p. 362, Jun. 2020, doi: 10.3390/brainsci10060362.

[135] R. T. Collins, "Cardiovascular disease in Williams syndrome," *Current Opinion in Pediatrics*, vol. 30, no. 5, pp. 609–615, Oct. 2018, doi: 10.1097/MOP.0000000000000664.

[136] C. B. Mervis and A. E. John, "Cognitive and behavioral characteristics of children with Williams syndrome: Implications for intervention approaches,"

American J of Med Genetics Pt C, vol. 154C, no. 2, pp. 229–248, May 2010, doi: 10.1002/ajmg.c.30263.

[137] O. Pös, O. Biró, T. Szemes, and B. Nagy, “Circulating cell-free nucleic acids: characteristics and applications,” *Eur J Hum Genet*, vol. 26, no. 7, pp. 937–945, Jul. 2018, doi: 10.1038/s41431-018-0132-4.

[138] H. Kadry, B. Noorani, and L. Cucullo, “A blood–brain barrier overview on structure, function, impairment, and biomarkers of integrity,” *Fluids Barriers CNS*, vol. 17, no. 1, p. 69, Dec. 2020, doi: 10.1186/s12987-020-00230-3.

[139] W. Wong and C. I. Ming, “A Review on Metaheuristic Algorithms: Recent Trends, Benchmarking and Applications,” in *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, Sarawak, Malaysia, Malaysia: IEEE, Jun. 2019, pp. 1–5. doi: 10.1109/ICSCC.2019.8843624.

[140] A. E. Eiben and J. E. Smith, “What Is an Evolutionary Algorithm?,” in *Introduction to Evolutionary Computing*, in Natural Computing Series. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2015, pp. 25–48. doi: 10.1007/978-3-662-44874-8_3.

[141] P. A. Vikhar, “Evolutionary algorithms: A critical review and its future prospects,” in *2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC)*, Jalgaon, India: IEEE, Dec. 2016, pp. 261–265. doi: 10.1109/ICGTSPICC.2016.7955308.

[142] A. Slowik and H. Kwasnicka, “Evolutionary algorithms and their applications to engineering problems,” *Neural Comput & Applic*, vol. 32, no. 16, pp. 12363–12379, Aug. 2020, doi: 10.1007/s00521-020-04832-8.

[143] J. K. Aronson and R. E. Ferner, “Biomarkers—A General Review,” *CP Pharmacology*, vol. 76, no. 1, Mar. 2017, doi: 10.1002/cpph.19.

[144] L. M. Chahine, M. B. Stern, and A. Chen-Plotkin, “Blood-based biomarkers for Parkinson’s disease,” *Parkinsonism & Related Disorders*, vol. 20, pp. S99–S103, Jan. 2014, doi: 10.1016/S1353-8020(13)70025-7.

[145] G. A. Ganepola, “Use of blood-based biomarkers for early diagnosis and surveillance of colorectal cancer,” *WJGO*, vol. 6, no. 4, p. 83, 2014, doi: 10.4251/wjgo.v6.i4.83.

[146] A. Leuzy, N. Mattsson-Carlgren, S. Palmqvist, S. Janelidze, J. L. Dage, and O. Hansson, “Blood-based biomarkers for Alzheimer’s disease,” *EMBO Mol Med*, vol. 14, no. 1, p. e14408, Jan. 2022, doi: 10.15252/emmm.202114408.

[147] A. Mukhopadhyay and M. Mandal, “Identifying Non-Redundant Gene Markers from Microarray Data: A Multiobjective Variable Length PSO-

Based Approach," *IEEE/ACM Trans. Comput. Biol. and Bioinf.*, vol. 11, no. 6, pp. 1170–1183, Nov. 2014, doi: 10.1109/TCBB.2014.2323065.

[148] A. Boucheham, M. Batouche, and S. Meshoul, "An Ensemble of Cooperative Parallel Metaheuristics for Gene Selection in Cancer Classification," in *Bioinformatics and Biomedical Engineering*, vol. 9044, F. Ortuño and I. Rojas, Eds., in Lecture Notes in Computer Science, vol. 9044. , Cham: Springer International Publishing, 2015, pp. 301–312. doi: 10.1007/978-3-319-16480-9_30.

[149] S. R. Manisekhar, G. M. Siddesh, and S. S. Manvi, "Introduction to Bioinformatics," in *Statistical Modelling and Machine Learning Principles for Bioinformatics Techniques, Tools, and Applications*, K. G. Srinivasa, G. M. Siddesh, and S. R. Manisekhar, Eds., in Algorithms for Intelligent Systems. , Singapore: Springer Singapore, 2020, pp. 3–9. doi: 10.1007/978-981-15-2445-5_1.

[150] N. Salari, S. Shohaimi, F. Najafi, M. Nallappan, and I. Karishnarajah, "A Novel Hybrid Classification Model of Genetic Algorithms, Modified k-Nearest Neighbor and Developed Backpropagation Neural Network," *PLoS ONE*, vol. 9, no. 11, p. e112987, Nov. 2014, doi: 10.1371/journal.pone.0112987.

[151] T. Rapakoulia, K. Theofilatos, D. Kleftogiannis, S. Likothanasis, A. Tsakalidis, and S. Mavroudi, "EnsembleGASVR: a novel ensemble method for classifying missense single nucleotide polymorphisms," *Bioinformatics*, vol. 30, no. 16, pp. 2324–2333, Aug. 2014, doi: 10.1093/bioinformatics/btu297.

[152] D. Kleftogiannis, K. Theofilatos, S. Likothanassis, and S. Mavroudi, "YamiPred: A Novel Evolutionary Method for Predicting Pre-miRNAs and Selecting Relevant Features," *IEEE/ACM Trans. Comput. Biol. and Bioinf.*, vol. 12, no. 5, pp. 1183–1192, Sep. 2015, doi: 10.1109/TCBB.2014.2388227.

[153] A. Swiercz *et al.*, "Unified encoding for hyper-heuristics with application to bioinformatics," *Cent Eur J Oper Res*, vol. 22, no. 3, pp. 567–589, Sep. 2014, doi: 10.1007/s10100-013-0321-8.

[154] J. Corthésy *et al.*, "An Adaptive Pipeline To Maximize Isobaric Tagging Data in Large-Scale MS-Based Proteomics," *J. Proteome Res.*, vol. 17, no. 6, pp. 2165–2173, Jun. 2018, doi: 10.1021/acs.jproteome.8b00110.

[155] A. Lambora, K. Gupta, and K. Chopra, "Genetic Algorithm- A Literature Review," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, Faridabad, India: IEEE, Feb. 2019, pp. 380–384. doi: 10.1109/COMITCon.2019.8862255.

[156] J. Chen, F. Zhao, Y. Sun, and Y. Yin, "Improved XGBoost model based on genetic algorithm," *IJCAT*, vol. 62, no. 3, p. 240, 2020, doi: 10.1504/IJCAT.2020.106571.

[157] S. Katoch, S. S. Chauhan, and V. Kumar, "A review on genetic algorithm: past, present, and future," *Multimed Tools Appl*, vol. 80, no. 5, pp. 8091–8126, Feb. 2021, doi: 10.1007/s11042-020-10139-6.

[158] A. Abraham, L. Jain, and R. Goldberg, Eds., *Evolutionary Multiobjective Optimization*. in Advanced Information and Knowledge Processing. London: Springer-Verlag, 2005. doi: 10.1007/1-84628-137-7.

[159] J. Zou, Q. Deng, Y. Liu, X. Yang, S. Yang, and J. Zheng, "A Dynamic-Niching-Based Pareto Domination for Multimodal Multiobjective Optimization," *IEEE Trans. Evol. Computat.*, vol. 28, no. 5, pp. 1529–1543, Oct. 2024, doi: 10.1109/TEVC.2023.3316723.

[160] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[161] X. Deng, M. Li, S. Deng, and L. Wang, "Hybrid gene selection approach using XGBoost and multi-objective genetic algorithm for cancer classification," *Med Biol Eng Comput*, vol. 60, no. 3, pp. 663–681, Mar. 2022, doi: 10.1007/s11517-021-02476-x.

[162] K. Li *et al.*, "Efficient gradient boosting for prognostic biomarker discovery," *Bioinformatics*, vol. 38, no. 6, pp. 1631–1638, Mar. 2022, doi: 10.1093/bioinformatics/btab869.

[163] B. Ma, G. Yan, B. Chai, and X. Hou, "XGBLC: an improved survival prediction model based on XGBoost," *Bioinformatics*, vol. 38, no. 2, pp. 410–418, Jan. 2022, doi: 10.1093/bioinformatics/btab675.

[164] V. S. Desdhanty and Z. Rustam, "Liver Cancer Classification Using Random Forest and Extreme Gradient Boosting (XGBoost) with Genetic Algorithm as Feature Selection," in *2021 International Conference on Decision Aid Sciences and Application (DASA)*, Sakheer, Bahrain: IEEE, Dec. 2021, pp. 716–719. doi: 10.1109/DASA53625.2021.9682311.

[165] N. Ghatasheh, I. Altaharwa, and K. Aldebei, "Modified Genetic Algorithm for Feature Selection and Hyper Parameter Optimization: Case of XGBoost in Spam Prediction," *IEEE Access*, vol. 10, pp. 84365–84383, 2022, doi: 10.1109/ACCESS.2022.3196905.

[166] A. H. Syed, T. Khan, and N. Alromema, "A Hybrid Feature Selection Approach to Screen a Novel Set of Blood Biomarkers for Early COVID-19

Mortality Prediction," *Diagnostics*, vol. 12, no. 7, p. 1604, Jun. 2022, doi: 10.3390/diagnostics12071604.

[167] Y. Bai, Y. Li, Y. Shen, M. Yang, W. Zhang, and B. Cui, "AutoDC: an automatic machine learning framework for disease classification," *Bioinformatics*, vol. 38, no. 13, pp. 3415–3421, Jun. 2022, doi: 10.1093/bioinformatics/btac334.

[168] J. Horn, N. Nafpliotis, and D. E. Goldberg, "A niched Pareto genetic algorithm for multiobjective optimization," in *Proceedings of the First IEEE Conference on Evolutionary Computation. IEEE World Congress on Computational Intelligence*, Orlando, FL, USA: IEEE, 1994, pp. 82–87. doi: 10.1109/ICEC.1994.350037.

[169] H. L. Blackburn, S. McErlean, G. L. Jellema, R. Van Laar, M. N. Vernalis, and D. L. Ellsworth, "Gene expression profiling during intensive cardiovascular lifestyle modification: Relationships with vascular function and weight loss," *Genomics Data*, vol. 4, pp. 50–53, Jun. 2015, doi: 10.1016/j.gdata.2015.03.001.

[170] D. L. Ellsworth *et al.*, "Importance of substantial weight loss for altering gene expression during cardiovascular lifestyle modification," *Obesity*, vol. 23, no. 6, pp. 1312–1319, Jun. 2015, doi: 10.1002/oby.21079.

[171] C. S. Cleeland and K. M. Ryan, "Pain assessment: global use of the Brief Pain Inventory," *Ann Acad Med Singap*, vol. 23, no. 2, pp. 129–138, Mar. 1994.

[172] J. A. Gudin, M. Brennan, E. Harris, P. Hurwitz, D. Dietze, and J. Strader, "Changes in pain and concurrent pain medication use following compounded topical analgesic treatment for chronic pain: 3- and 6-month follow-up results from the prospective, observational Optimizing Patient Experience and Response to Topical Analgesics study," *JPR*, vol. Volume 10, pp. 2341–2354, Oct. 2017, doi: 10.2147/JPR.S143513.

[173] D. L. Ellsworth *et al.*, "Intensive Cardiovascular Risk Reduction Induces Sustainable Changes in Expression of Genes and Pathways Important to Vascular Function," *Circ Cardiovasc Genet*, vol. 7, no. 2, pp. 151–160, Apr. 2014, doi: 10.1161/CIRCGENETICS.113.000121.

[174] Y. Zhang, G. Parmigiani, and W. E. Johnson, "ComBat-seq: batch effect adjustment for RNA-seq count data," *NAR Genomics and Bioinformatics*, vol. 2, no. 3, p. lqaa078, Sep. 2020, doi: 10.1093/nargab/lqaa078.

[175] B. Rosner, R. J. Glynn, and M. Ting Lee, "Incorporation of Clustering Effects for the Wilcoxon Rank Sum Test: A Large-Sample Approach," *Biometrics*, vol. 59, no. 4, pp. 1089–1098, Dec. 2003, doi: 10.1111/j.0006-341X.2003.00125.x.

[176] J. R. Vergara and P. A. Estévez, “A review of feature selection methods based on mutual information,” *Neural Comput & Applic*, vol. 24, no. 1, pp. 175–186, Jan. 2014, doi: 10.1007/s00521-013-1368-0.

[177] R. Artusi, P. Verderio, and E. Marubini, “Bravais-Pearson and Spearman Correlation Coefficients: Meaning, Test of Hypothesis and Confidence Interval,” *Int J Biol Markers*, vol. 17, no. 2, pp. 148–151, Apr. 2002, doi: 10.1177/172460080201700213.

[178] M. Coppola, J. Guo, E. Gill, and G. C. H. E. De Croon, “The PageRank algorithm as a method to optimize swarm behavior through local analysis,” *Swarm Intell*, vol. 13, no. 3–4, pp. 277–319, Dec. 2019, doi: 10.1007/s11721-019-00172-z.

[179] J. Lonsdale *et al.*, “The Genotype-Tissue Expression (GTEx) project,” *Nat Genet*, vol. 45, no. 6, pp. 580–585, Jun. 2013, doi: 10.1038/ng.2653.

[180] M. Erickson, A. Mayer, and J. Horn, “The Niched Pareto Genetic Algorithm 2 Applied to the Design of Groundwater Remediation Systems,” in *Evolutionary Multi-Criterion Optimization*, vol. 1993, E. Zitzler, L. Thiele, K. Deb, C. A. Coello Coello, and D. Corne, Eds., in Lecture Notes in Computer Science, vol. 1993. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 681–695. doi: 10.1007/3-540-44719-9_48.

[181] M. Arhatte *et al.*, “TMEM33 regulates intracellular calcium homeostasis in renal tubular epithelial cells,” *Nat Commun*, vol. 10, no. 1, p. 2024, May 2019, doi: 10.1038/s41467-019-10045-y.

[182] B. Zhang, W. Huang, M. Yi, and C. Xing, “Gene Differential Expression and Interaction Networks Illustrate the Biomarkers and Molecular Mechanisms of Atherosclerotic Cerebral Infarction,” *Journal of Healthcare Engineering*, vol. 2022, pp. 1–8, Jan. 2022, doi: 10.1155/2022/3912697.

[183] A. Bohr and K. Memarzadeh, “The rise of artificial intelligence in healthcare applications,” in *Artificial Intelligence in Healthcare*, Elsevier, 2020, pp. 25–60. doi: 10.1016/B978-0-12-818438-7.00002-2.

[184] G. Rompianesi, F. Pegoraro, C. D. Ceresa, R. Montalti, and R. I. Troisi, “Artificial intelligence in the diagnosis and management of colorectal cancer liver metastases,” *WJG*, vol. 28, no. 1, pp. 108–122, Jan. 2022, doi: 10.3748/wjg.v28.i1.108.

[185] O. Hansson, “Biomarkers for neurodegenerative diseases,” *Nat Med*, vol. 27, no. 6, pp. 954–963, Jun. 2021, doi: 10.1038/s41591-021-01382-x.

[186] A. Aessopos, M. Tsironi, A. Andreopoulos, and D. Farmakis, “Heart Disease in Thalassemia Intermedia,” *Hemoglobin*, vol. 33, no. sup1, pp. S170–S176, Jan. 2009, doi: 10.3109/03630260903351676.

[187] H. Bajwa and H. Basit, “Thalassemia,” in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2024. Accessed: Oct. 13, 2024. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK545151/

[188] M. H. Ahmed, M. S. Ghatge, and M. K. Safo, “Hemoglobin: Structure, Function and Allostery,” in *Vertebrate and Invertebrate Respiratory Proteins, Lipoproteins and other Body Fluid Proteins*, vol. 94, U. Hoeger and J. R. Harris, Eds., in Subcellular Biochemistry, vol. 94. , Cham: Springer International Publishing, 2020, pp. 345–382. doi: 10.1007/978-3-030-41769-7_14.

[189] K. M. Musallam *et al.*, “Risk of mortality from anemia and iron overload in nontransfusion-dependent β-thalassemia,” *American J Hematol*, vol. 97, no. 2, Feb. 2022, doi: 10.1002/ajh.26428.

[190] R. Mariani, P. Trombini, M. Pozzi, and A. Piperno, “Iron metabolism in thalassemia and sickle cell disease,” *Mediterr J Hematol Infect Dis*, vol. 1, no. 1, p. e2009006, Oct. 2009, doi: 10.4084/MJHID.2009.006.

[191] S. Basu, M. Rahaman, T. K. Dolai, P. C. Shukla, and N. Chakravorty, “Understanding the Intricacies of Iron Overload Associated with β-Thalassemia: A Comprehensive Review,” *Thalassemia Reports*, vol. 13, no. 3, pp. 179–194, Jul. 2023, doi: 10.3390/thalassrep13030017.

[192] Y. Tuo *et al.*, “Global, regional, and national burden of thalassemia, 1990–2021: a systematic analysis for the global burden of disease study 2021,” *eClinicalMedicine*, vol. 72, p. 102619, Jun. 2024, doi: 10.1016/j.eclinm.2024.102619.

[193] Y. Song, S. Chang, J. Tian, W. Pan, L. Feng, and H. Ji, “A Comprehensive Comparative Analysis of Deep Learning Based Feature Representations for Molecular Taste Prediction,” *Foods*, vol. 12, no. 18, p. 3386, Sep. 2023, doi: 10.3390/foods12183386.

[194] N. Akiki, M. H. Hodroj, R. Bou-Fakhredin, K. Matli, and A. T. Taher, “Cardiovascular Complications in β-Thalassemia: Getting to the Heart of It,” *Thalassemia Reports*, vol. 13, no. 1, Art. no. 1, Mar. 2023, doi: 10.3390/thalassrep13010005.

[195] V. Positano *et al.*, “Deep Learning Staging of Liver Iron Content From Multiecho MR Images,” *Magnetic Resonance Imaging*, vol. 57, no. 2, pp. 472–484, Feb. 2023, doi: 10.1002/jmri.28300.

[196] H. Taleie *et al.*, “Left Ventricular Myocardial Dysfunction Evaluation in Thalassemia Patients Using Echocardiographic Radiomic Features and Machine Learning Algorithms,” *J Digit Imaging*, vol. 36, no. 6, pp. 2494–2506, Dec. 2023, doi: 10.1007/s10278-023-00891-0.

[197] P. Kirk *et al.*, "Cardiac T2* Magnetic Resonance for Prediction of Cardiac Complications in Thalassemia Major," *Circulation*, vol. 120, no. 20, pp. 1961–1968, Nov. 2009, doi: 10.1161/CIRCULATIONAHA.109.874487.

[198] H. Sabir *et al.*, "Fingertip Video Dataset for Non-Invasive Diagnosis of Anemia Using ResNet-18 Classifier," *IEEE Access*, vol. 12, pp. 68880–68892, 2024, doi: 10.1109/ACCESS.2024.3398353.

[199] K. Panagiotopoulos, A. Korfiati, K. Theofilatos, P. Hurwitz, M. A. Deriu, and S. Mavroudi, "MEvA-X: a hybrid multiobjective evolutionary tool using an XGBoost classifier for biomarkers discovery on biomedical datasets," *Bioinformatics*, vol. 39, no. 7, p. btad384, Jul. 2023, doi: 10.1093/bioinformatics/btad384.

[200] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011, [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html

[201] P. Keerin, W. Kurutach, and T. Boongoen, "Cluster-based KNN missing value imputation for DNA microarray data," in *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Seoul, Korea (South): IEEE, Oct. 2012, pp. 445–450. doi: 10.1109/ICSMC.2012.6377764.

[202] N. M. A. AL-Salami, "Evolutionary Algorithm Definition," *American J. of Engineering and Applied Sciences*, vol. 2, no. 4, pp. 789–795, Apr. 2009, doi: 10.3844/ajeassp.2009.789.795.

[203] M. J. Reddy and D. N. Kumar, "Optimal Reservoir Operation Using Multi-Objective Evolutionary Algorithm," *Water Resour Manage*, vol. 20, no. 6, pp. 861–878, Oct. 2006, doi: 10.1007/s11269-005-9011-1.

Appendix A

List of Appendix tables

List of Appendix figures

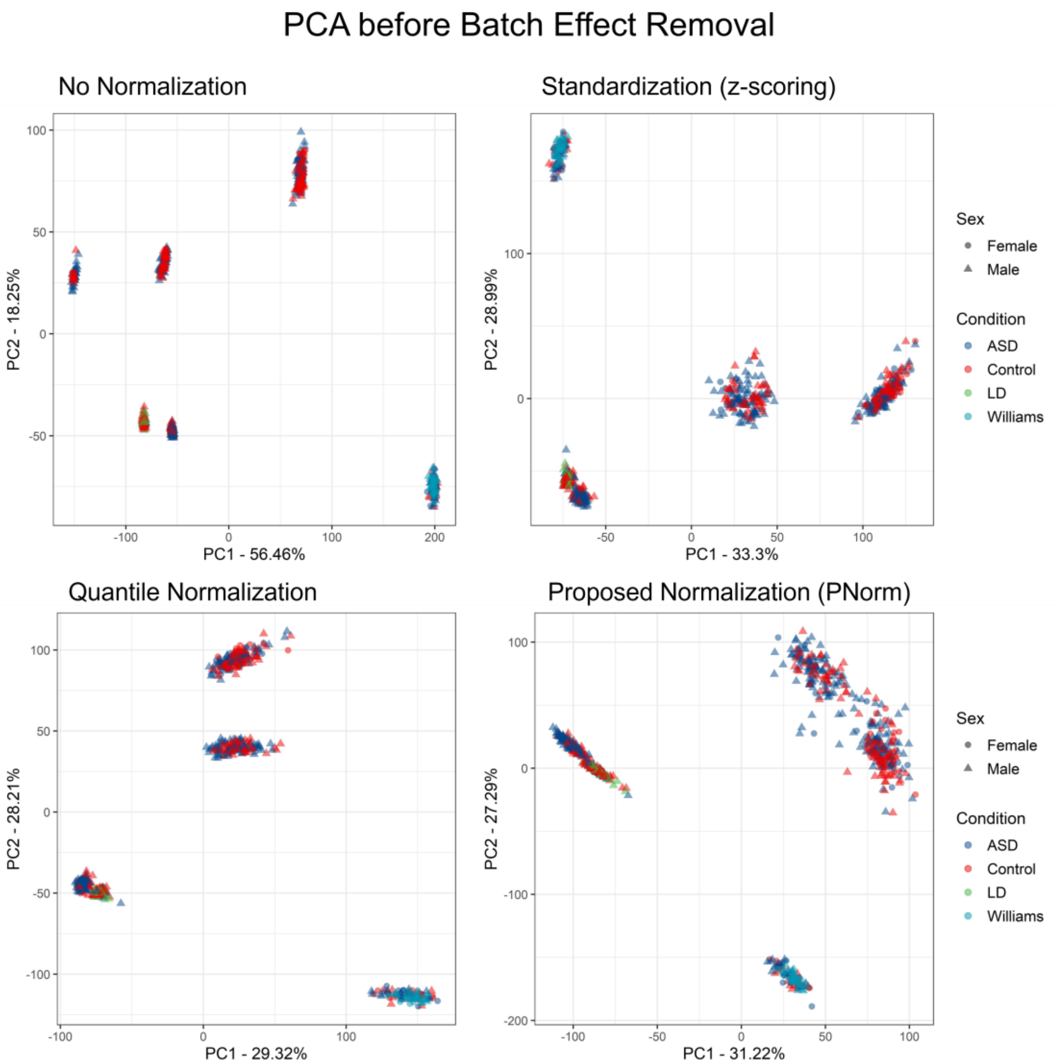
Chapter – 3

In this section further explanations, tables and figures related to chapter 3 are presented to support the claims and results of the work. In Sup_Table 1 below, is a summary of information of the different platforms used in the studies used for the merging and harmonization of NDD datasets. All "tissue" elements are related to blood either as whole blood, peripheral blood or leukocyte cells.

GEO Accession	Tissue	Platform ID	Microarray name	Manufacturer
GSE6575	Whole blood	GPL570	Human Genome U133 Plus 2.0	Affymetrix
GSE18123 A	Peripheral blood	GPL570	Human Genome U133 Plus 2.0	Affymetrix
GSE18123 B	Peripheral blood	GPL6244	Human Gene 1.0 ST	Affymetrix
GSE42133	Leukocyte	GPL10558	HumanHT-12 V4.0 expression beadchip	Ilumina
GSE89594	Peripheral blood	GPL16699	SperPrint G3 Human GE v2 8x60K	Agilent
GSE30470	Peripheral blood	GPL570	Human Genome U133 Plus 2.0	Affymetrix
GSE111175	Leukocyte	GPL10558	HumanHT-12 V4.0 expression beadchip	Ilumina

Sup_Table 1. Information table related to the studies of this analysis concerning the biological tissue used and the different platforms.

Sup_Fig. 1 depicts the projection of the datasets with the different normalization methods applied to them, after merging but before the correction for the batch effects of studies and platforms. From this, it is clear that no normalization method alone can eliminate the non-biological differences and technical variabilities of the different datasets. Nevertheless, the custom normalization method that was used in this work seems to be as efficient as other established methods such as standardization and quantile normalization.



Sup_Fig. 1. PCA of the first two components for the data without normalization after merging (top-left) and with the various normalization methods used before the Batch effect removal step. Normalization methods alone is not able to account for systematic errors like the GEO ID and the Platform ID.

Differential expression analysis

Contrast	Gene Symbol	logFC	Aver. Expr	P-val	FDR
ASD – LD	SRSF5	0.572447	10.39922	3.91E-07	0.003821
	ARPC1B	0.471634	9.477668	5.02E-07	0.003821
	TNPO1	0.335223	7.38046	6.09E-07	0.003821
	STK38	0.419603	9.085803	9.61E-07	0.004526
LD – Control	ARPC1B	-0.46807	9.477668	7.58E-07	0.014271
	PRF1	0.663183	9.353933	2.07E-06	0.019521
	REEP3	0.281486	5.906302	6.61E-06	0.041492
	SNRNP35	-0.19672	6.564368	1.24E-05	0.047676
William's – Control	TBL2	-0.36306	6.523251	5.75E-13	6.32E-09
	EIF4H	-0.45814	8.909363	6.71E-13	6.32E-09
	ABHD11	-0.29034	5.846867	2.11E-12	1.25E-08
	BAZ1B	-0.37203	6.863545	2.64E-12	1.25E-08
ASD – William's	BAZ1B	0.412877	6.863545	5.54E-15	1.04E-10
	RFC2	0.405984	6.396063	2.76E-14	2.60E-10
	DNAJC30	0.437909	6.69084	6.09E-14	3.82E-10
	EIF4H	0.469874	8.909363	1.01E-13	4.75E-10
	SHC2	-0.4371	5.325696	1.90E-12	7.17E-09
	BUD23	0.350607	7.382656	4.23E-12	1.33E-08
	ANKRD20A11P	0.58043	5.538706	1.35E-11	3.63E-08
	TBL2	0.32944	6.523251	3.36E-11	7.92E-08
	TMEM38A	-0.28923	5.591573	1.36E-10	2.84E-07
LD – William's	BAZ1B	0.420685	6.863545	2.62E-08	0.000305
	CLIC3	0.910751	7.802743	3.23E-08	0.000305
	KLRF1	1.183479	7.98448	8.51E-08	0.000535
	COL8A2	-0.35355	5.793827	3.11E-07	0.001466
	KIR2DL1	0.538844	6.020627	5.43E-07	0.002044
	SHC2	-0.43129	5.325696	1.24E-06	0.00344
	SH2D1B	0.848936	7.182838	1.28E-06	0.00344
	LIMK1	0.307143	6.024266	1.54E-06	0.003634
	RFC2	0.355843	6.396063	2.98E-06	0.006247
	AKR1C3	0.912484	7.023975	4.01E-06	0.006927

	ANKRD20A11P	0.567042	5.538706	4.04E-06	0.006927
	BCL7B	0.394993	7.406025	6.35E-06	0.008607
	RFPL2	0.328577	4.861659	6.64E-06	0.008607
	CFH	0.377771	4.433916	7.09E-06	0.008607

Sup_Table 2. Results of the top genes from the Differential Expression Analysis between pairs of conditions

In Chapter 3, and section 3.3.2, the effect of normalization followed by a correction method for batches on the merged microarray data is illustrated. In that case, only the proposed method of normalization (PNorm) is shown, and thus, here we list the illustrations for the other two methods of Quantile normalization (QNorm) and z-score standardization prior to the batch effect removal. As with the proposed method, the normalization on the merged data before the batch effect can not eliminate the non-biological differences and has a small effect on the result after the correction. This occurs if one inspects the variance explained by the principal components of the PCA plots in Sup_Fig. 2 and focus on the top and bottom plots after merging where the explained variance is comparable between the two.

Sup_Fig. 2. PCA plot of the first 2 components before (left) and after (right) the batch effect correction for the study ID and the microarray ID for the methods of standardization (top) and Quantile normalization (bottom).

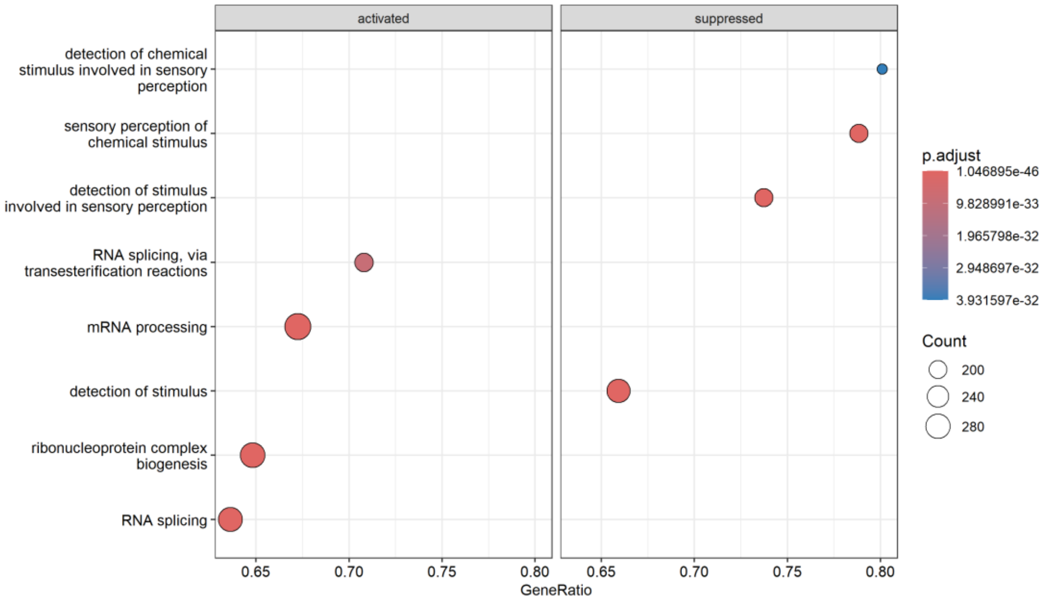
Enrichment analysis

From the GSEA to find up- or down-regulated enriched GO terms between the pairs of conditions evaluated, in the tables below we present the result of the top 20 GO terms based on the absolute normalized enrichment score (NES) and the adjusted p-value of each term (Sup_Table 3 - Sup_Table 8). Additionally, we present the GSEA plots for each comparison for the first 4 upregulated and 4 down regulated terms (Sup_Fig. 3 - Sup_Fig. 8).

<u>ASD – Control:</u>

GO ID	Description	Set Size	NES	p.adjust	Regulation Direction
GO:0022613	ribonucleoprotein complex biogenesis	435	3.426	1.05E-46	Up
GO:0006397	mRNA processing	458	3.387	1.35E-45	Up
GO:0008380	RNA splicing	426	3.357	1.29E-43	Up
GO:0050906	detection of stimulus involved in sensory perception	278	-3.358	1.47E-38	Down
GO:0051606	detection of stimulus	399	-3.026	8.53E-35	Down
GO:0007606	sensory perception of chemical stimulus	260	-3.284	1.75E-34	Down
GO:0000375	RNA splicing, via transesterification reactions	291	3.301	8.65E-33	Up
GO:0050907	detection of chemical stimulus involved in sensory perception	211	-3.371	3.93E-32	Down
GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	287	3.309	4.84E-32	Up
GO:0000398	mRNA splicing, via spliceosome	287	3.309	4.84E-32	Up
GO:0009593	detection of chemical stimulus	244	-3.233	1.43E-31	Down
GO:0007608	sensory perception of smell	193	-3.263	3.22E-29	Down
GO:0050911	detection of chemical stimulus involved in sensory perception of smell	171	-3.358	4.93E-29	Down
GO:0034470	ncRNA processing	424	2.847	5.68E-27	Up
GO:0048193	Golgi vesicle transport	286	3.027	4.71E-25	Up
GO:0042254	ribosome biogenesis	290	3.014	4.82E-25	Up
GO:0022618	ribonucleoprotein complex assembly	197	3.219	9.36E-23	Up
GO:0071826	ribonucleoprotein complex subunit organization	205	3.199	1.23E-22	Up
GO:0072594	establishment of protein localization to organelle	432	2.665	1.23E-22	Up
GO:0006417	regulation of translation	404	2.739	3.44E-22	Up

Sup_Table 3. Enrichment analysis results on the top regulated GO terms between ASD and Control groups.



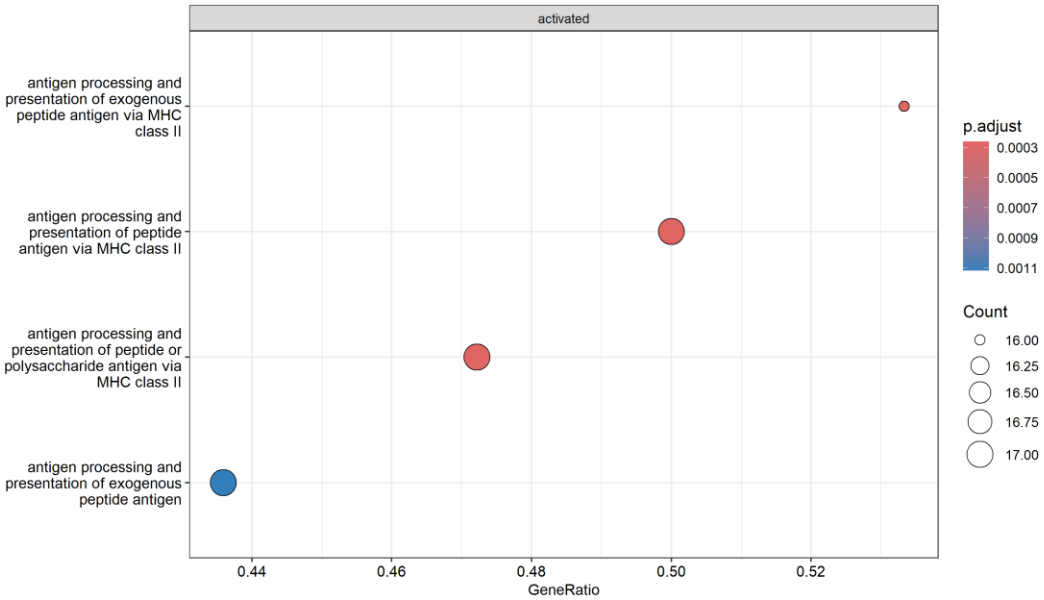
Sup_Fig. 3. ASD-Control GSEA plot of the first 4 up-regulated terms and the first 4 down-regulated GO terms based on the normalized enrichment score (NES) and the adjusted p-value.

William's – Control:

GO ID	Description	Set Size	NES	p.adjust	Regulation Direction
GO:0019886	antigen processing and presentation of exogenous peptide antigen via MHC class II	30	2.679	2.61E-04	Up
GO:0002495	antigen processing and presentation of peptide antigen via MHC class II	34	2.618	2.61E-04	Up
GO:0002504	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	36	2.609	2.61E-04	Up
GO:0002478	antigen processing and presentation of exogenous peptide antigen	39	2.476	1.12E-03	Up
GO:0002396	MHC protein complex assembly	19	2.479	5.53E-03	Up
GO:0002501	peptide antigen assembly with MHC protein complex	19	2.479	5.53E-03	Up
GO:0019884	antigen processing and presentation of exogenous antigen	48	2.300	9.21E-03	Up
GO:0048002	antigen processing and presentation of peptide antigen	67	2.097	1.28E-02	Up

GO:0002399	MHC class II protein complex assembly	15	2.454	1.88E-02	Up
GO:0002503	peptide antigen assembly with MHC class II protein complex	15	2.454	1.88E-02	Up

Sup_Table 4. Enrichment analysis results on the top regulated GO terms between William's Syndrome and Control groups.



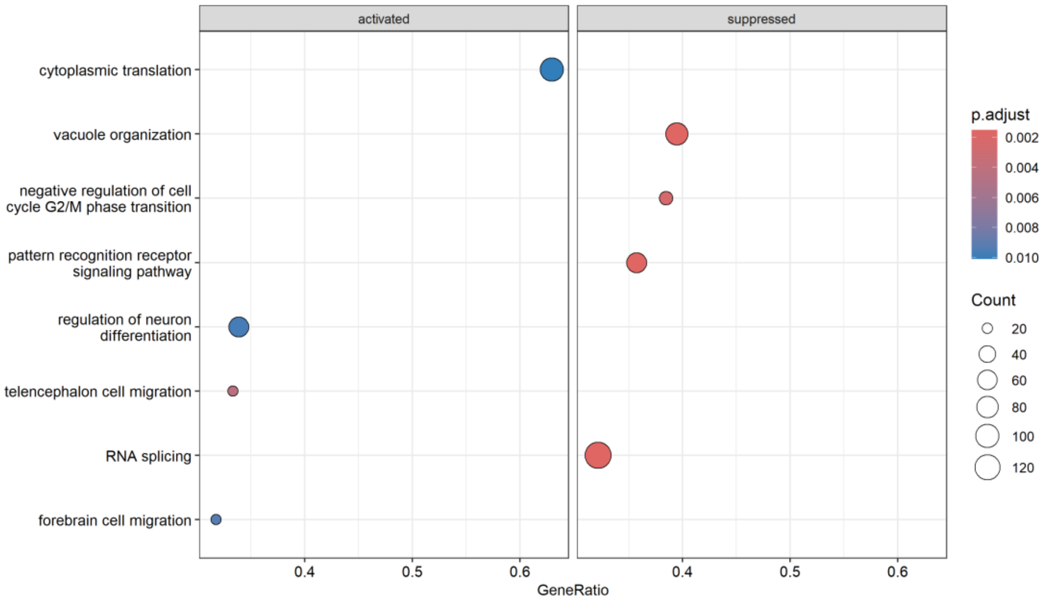
Sup_Fig. 4. William's-Control GSEA plot of the first 4 up-regulated terms and the first 4 down-regulated GO terms based on the normalized enrichment score (NES) and the adjusted p-value.

Language delay – Control:

GO ID	Description	Set Size	NES	p.adjust	Regulation Direction
GO:0022029	telencephalon cell migration	60	2.026	4.25E-03	Up
GO:0002717	positive regulation of natural killer cell mediated immunity	28	2.011	2.19E-02	Up
GO:0021795	cerebral cortex cell migration	47	2.004	1.88E-02	Up
GO:0021885	forebrain cell migration	63	1.966	9.32E-03	Up
GO:0014854	response to inactivity	12	1.957	3.89E-02	Up
GO:0021895	cerebral cortex neuron differentiation	27	1.955	2.73E-02	Up
GO:0017156	calcium-ion regulated exocytosis	64	1.780	4.93E-02	Up
GO:0002181	cytoplasmic translation	154	1.776	1.01E-02	Up
GO:0045664	regulation of neuron differentiation	186	1.715	9.68E-03	Up

GO:1900227	positive regulation of NLRP3 inflammasome complex assembly	18	-2.323	5.18E-03	Down
GO:1902750	negative regulation of cell cycle G2/M phase transition	65	-2.206	2.36E-03	Down
GO:0044818	mitotic G2/M transition checkpoint	51	-2.196	3.98E-03	Down
GO:0010972	negative regulation of G2/M transition of mitotic cell cycle	63	-2.181	4.22E-03	Down
GO:0007095	mitotic G2 DNA damage checkpoint signaling	35	-2.179	8.65E-03	Down
GO:0002224	toll-like receptor signaling pathway	66	-2.132	4.21E-03	Down
GO:0033194	response to hydroperoxide	17	-2.086	3.38E-02	Down
GO:0032731	positive regulation of interleukin-1 beta production	59	-1.996	1.67E-02	Down
GO:0007040	lysosome organization	101	-1.985	4.21E-03	Down
GO:0080171	lytic vacuole organization	101	-1.985	4.21E-03	Down
GO:0032732	positive regulation of interleukin-1 production	68	-1.967	2.19E-02	Down

Sup_Table 5. Enrichment analysis results on the top regulated GO terms between Language delay and Control groups.

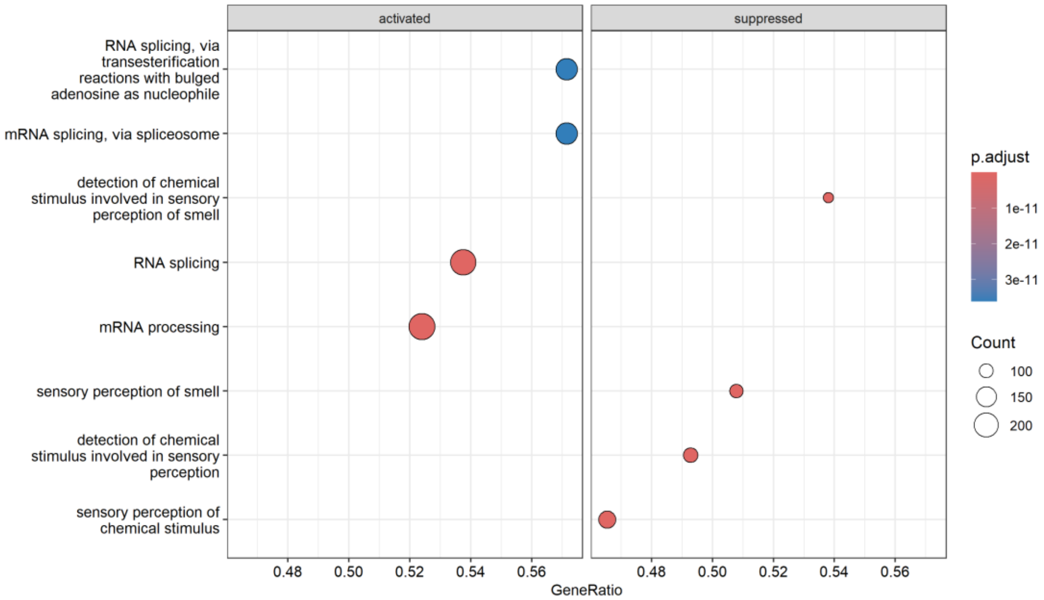


Sup_Fig. 5. LD-Control GSEA plot of the first 4 up-regulated terms and the first 4 down-regulated GO terms based on the normalized enrichment score (NES) and the adjusted p-value.

ASD – William's:

GO ID	Description	Set Size	NES	p.adjust	Regulation Direction
GO:0050907	detection of chemical stimulus involved in sensory perception	211	-3.085	6.45E-31	Down
GO:0007606	sensory perception of chemical stimulus	260	-2.891	1.31E-29	Down
GO:0050911	detection of chemical stimulus involved in sensory perception of smell	171	-3.144	2.24E-28	Down
GO:0007608	sensory perception of smell	193	-2.912	9.83E-25	Down
GO:0009593	detection of chemical stimulus	244	-2.689	1.17E-22	Down
GO:0050906	detection of stimulus involved in sensory perception	278	-2.629	2.04E-22	Down
GO:0051606	detection of stimulus	399	-2.345	9.12E-20	Down
GO:0008380	RNA splicing	426	2.145	3.86E-15	Up
GO:0006397	mRNA processing	458	2.089	3.58E-14	Up
GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	287	2.129	3.62E-11	Up
GO:0000398	mRNA splicing, via spliceosome	287	2.129	3.62E-11	Up
GO:0000375	RNA splicing, via transesterification reactions	291	2.147	4.84E-11	Up
GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules	144	-2.369	2.18E-09	Down
GO:0043062	extracellular structure organization	309	-1.952	6.92E-09	Down
GO:0045229	external encapsulating structure organization	310	-1.948	6.92E-09	Down
GO:0002181	cytoplasmic translation	154	2.230	7.92E-09	Up
GO:0010498	proteasomal protein catabolic process	497	1.840	2.18E-08	Up
GO:0016236	macroautophagy	315	1.977	2.82E-08	Up
GO:0031424	keratinization	72	-2.532	3.32E-08	Down
GO:0030198	extracellular matrix organization	308	-1.948	4.44E-08	Down

Sup_Table 6. Enrichment analysis results on the top regulated GO terms between ASD and William's Syndrome groups.



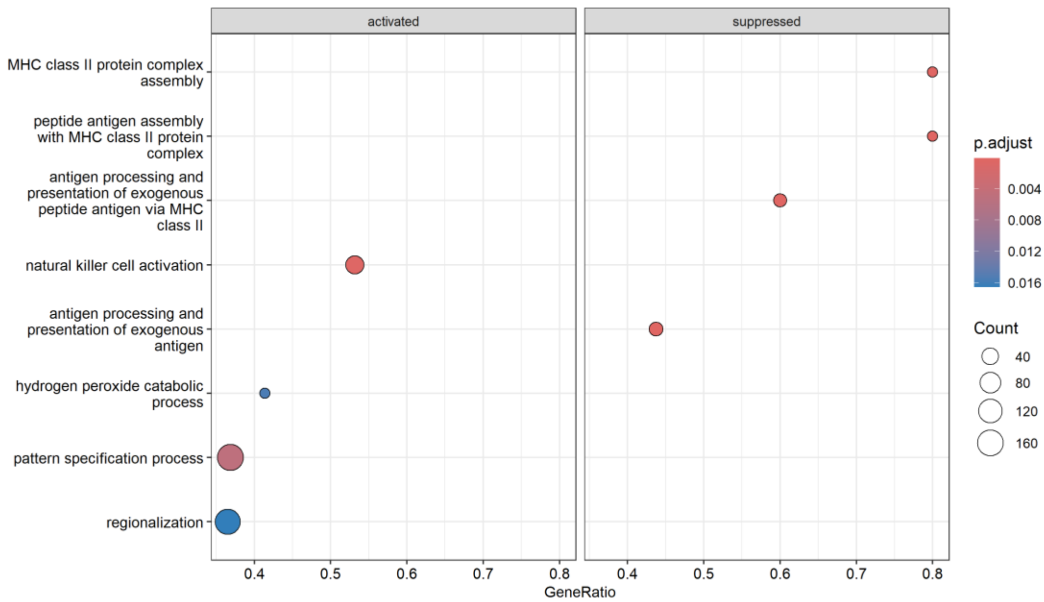
Sup_Fig. 6. ASD-William's GSEA plot of the first 4 up-regulated terms and the first 4 down-regulated GO terms based on the normalized enrichment score (NES) and the adjusted p-value.

Language delay – William's:

GO ID	Description	Set Size	NES	p.adjust	Regulation Direction
GO:0002399	MHC class II protein complex assembly	15	-2.725	1.06E-04	Down
GO:0002503	peptide antigen assembly with MHC class II protein complex	15	-2.725	1.06E-04	Down
GO:0019886	antigen processing and presentation of exogenous peptide antigen via MHC class II	30	-2.674	1.06E-04	Down
GO:0019884	antigen processing and presentation of exogenous antigen	48	-2.643	1.06E-04	Down
GO:0002478	antigen processing and presentation of exogenous peptide antigen	39	-2.639	1.06E-04	Down
GO:0002504	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	36	-2.611	1.06E-04	Down
GO:0019882	antigen processing and presentation	110	-2.099	1.85E-04	Down
GO:0002396	MHC protein complex assembly	19	-2.610	2.76E-04	Down
GO:0002501	peptide antigen assembly with MHC protein complex	19	-2.610	2.76E-04	Down

GO:0002495	antigen processing and presentation of peptide antigen via MHC class II	34	-2.575	2.76E-04	Down
GO:0030101	natural killer cell activation	94	1.971	4.56E-04	Up
GO:0048002	antigen processing and presentation of peptide antigen	67	-2.159	2.11E-03	Down
GO:0016064	immunoglobulin mediated immune response	141	-1.836	4.91E-03	Down
GO:0007389	pattern specification process	456	1.507	5.02E-03	Down
GO:0019724	B cell mediated immunity	144	-1.791	5.30E-03	Down
GO:0002381	immunoglobulin production involved in immunoglobulin-mediated immune response	73	-2.063	6.77E-03	Down
GO:0042744	hydrogen peroxide catabolic process	29	2.053	1.55E-02	Up
GO:0032755	positive regulation of interleukin-6 production	97	-1.830	1.55E-02	Down
GO:0003002	regionalization	411	1.477	1.66E-02	Up
GO:0002440	production of molecular mediator of immune response	237	-1.587	1.66E-02	Down

Sup_Table 7. Enrichment analysis results on the top regulated GO terms between Language delay and William's Syndrome groups.

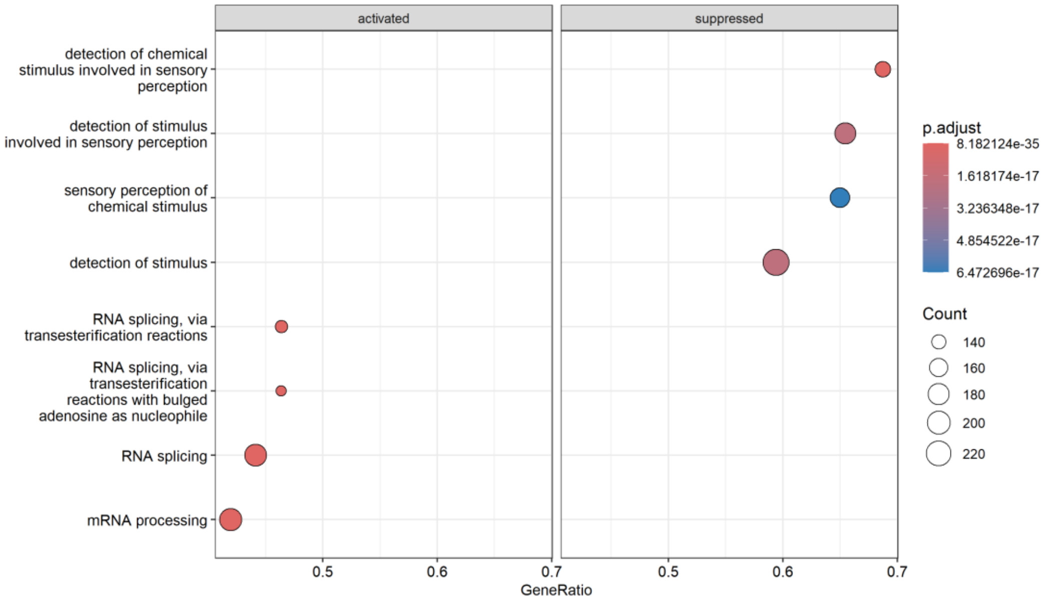


Sup_Fig. 7. LD-William's GSEA plot of the first 4 up-regulated terms and the first 4 down-regulated GO terms based on the normalized enrichment score (NES) and the adjusted p-value.

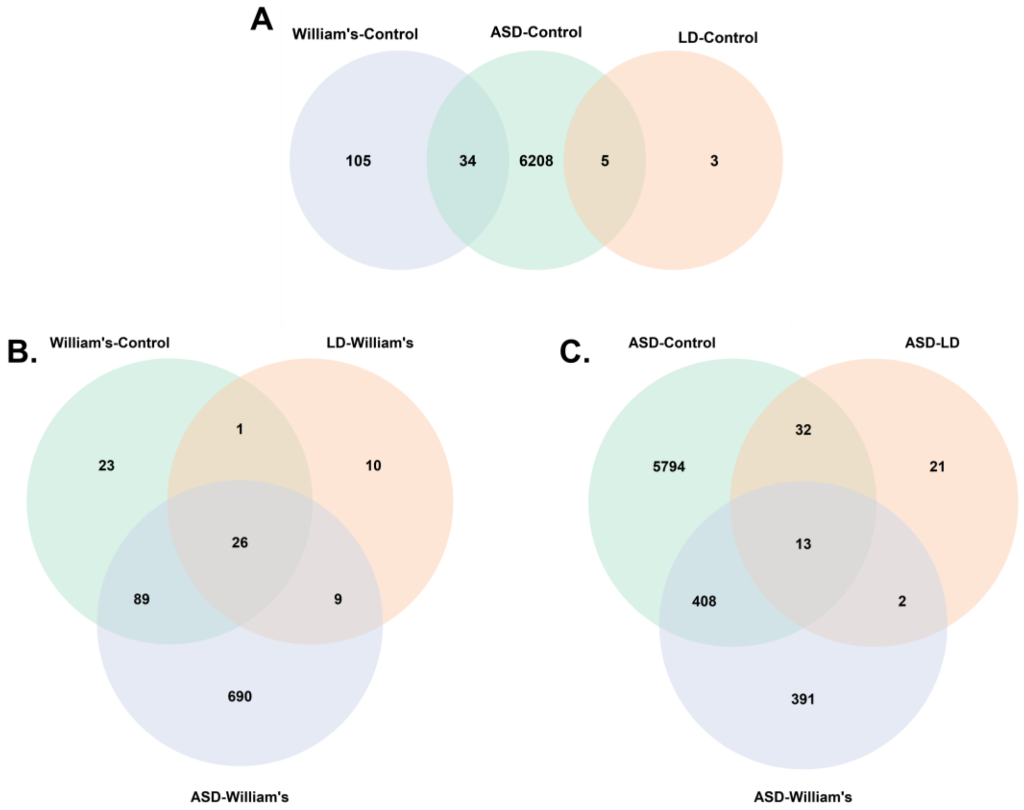
ASD – Language delay:

GO ID	Description	Set Size	NES	p.adjust	Regulation Direction
GO:0008380	RNA splicing	426	2.929	8.18E-35	Up
GO:0006397	mRNA processing	458	2.808	7.62E-32	Up
GO:0000375	RNA splicing, via transesterification reactions	291	2.793	9.55E-22	Up
GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	287	2.767	9.55E-22	Up
GO:0000398	mRNA splicing, via spliceosome	287	2.767	9.55E-22	Up
GO:0050907	detection of chemical stimulus involved in sensory perception	211	-2.664	2.02E-19	Down
GO:0050906	detection of stimulus involved in sensory perception	278	-2.504	1.91E-17	Down
GO:0051606	detection of stimulus	399	-2.321	1.91E-17	Down
GO:0016236	macroautophagy	315	2.498	6.33E-17	Up
GO:0007606	sensory perception of chemical stimulus	260	-2.485	6.47E-17	Down
GO:0050911	detection of chemical stimulus involved in sensory perception of smell	171	-2.684	9.32E-17	Down
GO:0022613	ribonucleoprotein complex biogenesis	435	2.296	1.39E-16	Down
GO:0007608	sensory perception of smell	193	-2.582	6.60E-16	Down
GO:0007033	vacuole organization	213	2.650	7.20E-16	Up
GO:0009593	detection of chemical stimulus	244	-2.442	1.05E-15	Down
GO:0010498	proteasomal protein catabolic process	497	2.128	1.25E-14	Up
GO:0043161	proteasome-mediated ubiquitin-dependent protein catabolic process	432	2.206	1.31E-14	Up
GO:0072594	establishment of protein localization to organelle	432	2.104	1.46E-12	Up
GO:0048193	Golgi vesicle transport	286	2.306	1.05E-11	Up
GO:0016050	vesicle organization	363	2.102	7.06E-11	Up

Sup_Table 8. Enrichment analysis results on the top regulated GO terms between ASD and Language delay groups.



Sup_Fig. 8. ASD-LD GSEA plot of the first 4 up-regulated terms and the first 4 down-regulated GO terms based on the normalized enrichment score (NES) and the adjusted p-value.

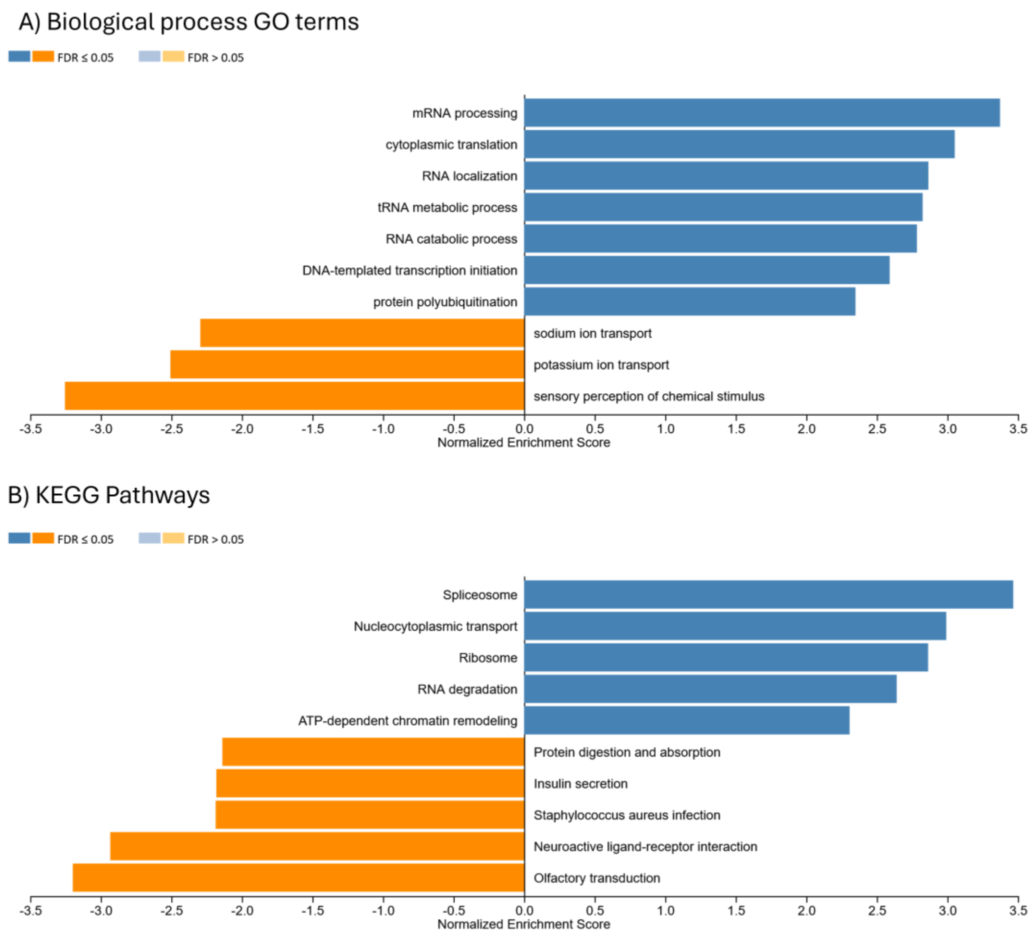


Sup_Fig. 9 Venn-diagrams of the genes found significantly different (FDR<0.05, independent of the |Log2FC|), between the compared groups. (A) Control was used as the common group. (B) William's syndrome was the reference group for this diagram and on (C) the reference was ASD.

KEGG - Pathway analysis

In this part, we present the findings of the pathway analysis conducted on the ranked genes after the pair-wise differential analysis of the harmonized merged data. The criteria used for the filtering of the results are the minimum number of genes in each category to be at least five, and the FDR < 0.05. Additionally, we present only the top 10 upregulated and top 10 downregulated pathways returned from the analysis.

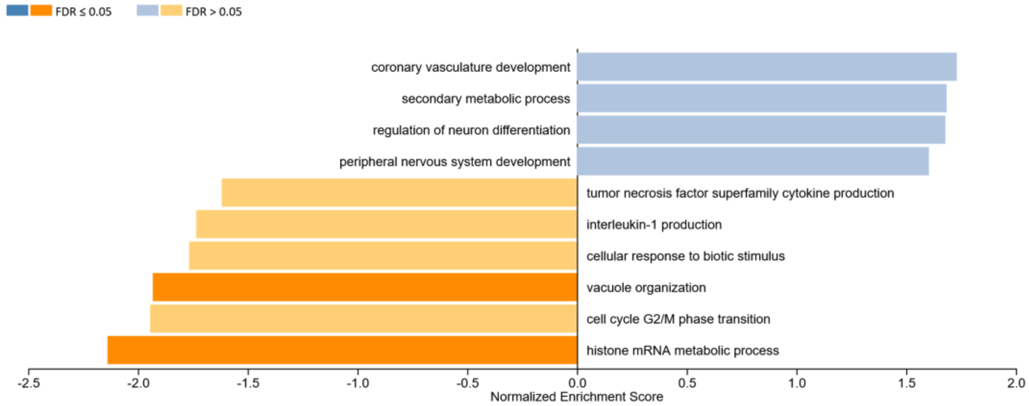
<u>ASD – Control</u>



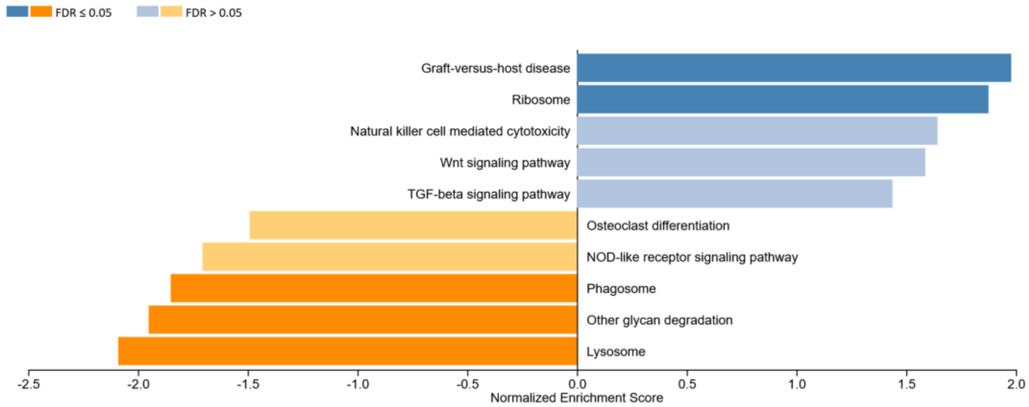
Sup_Fig. 10. Bar plots of the enriched GO terms for Biological processes (A) and the enriched KEGG pathways (B) for the ASD - Control comparison.

LD - Control

A) Biological process GO terms



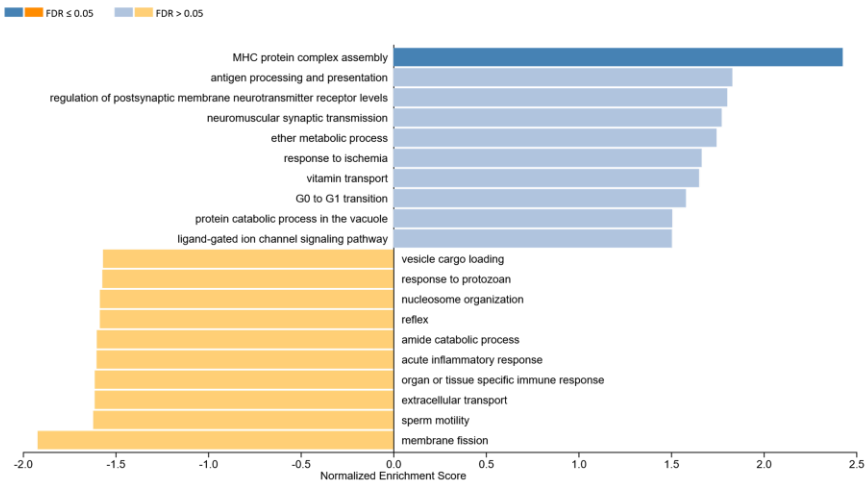
B) KEGG Pathways



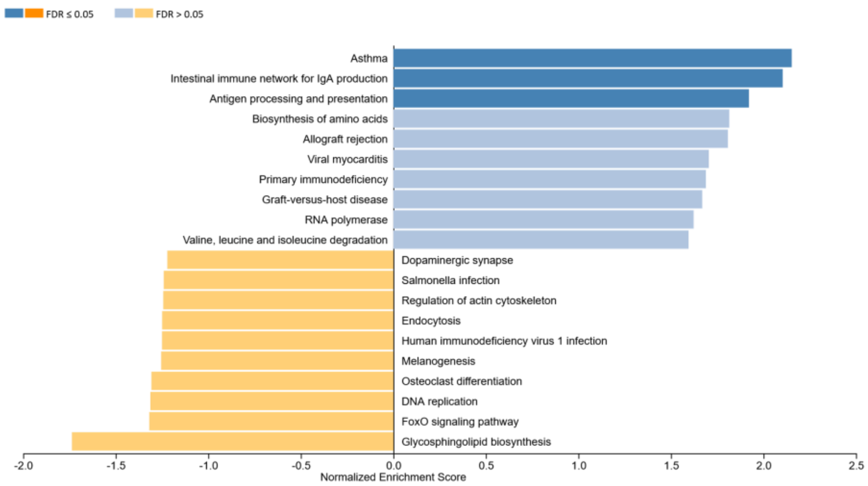
Sup_Fig. 11. Bar plots of the enriched GO terms for Biological processes (A) and the enriched KEGG pathways (B) for the LD - Control comparison.

William's Syndrome - Control

A) Biological process GO terms



B) KEGG Pathways



Sup_Fig. 12. Bar plots of the enriched GO terms for Biological processes (A) and the enriched KEGG pathways (B) for William's Syndrome - Control comparison.

Chapter – 4

MEvA-X supplementary

Preprocessing

The input data and labels for MEvA-X are transformed into numerical values, in case their original form is nominal or categorical, with the method of Label Encoding [200]. The dataset is searched for missing and abnormal values which then are imputed with a K-nearest neighbors [KNN] scheme [201]. Duplicated feature names are merged by getting their mean to have unique features in the dataset. Even though the normalization of the data is not necessary for the tree-based XGBoost classifier, for keeping a consistent pre-processing process with the other methods, the data are normalized feature-wise within the range [0,1] by the MinMaxScaler method [200]. The final product of the pre-processing method is a two-dimensional data frame object where the rows correspond to the features and the columns to the samples of the dataset, while the labels are kept in a different one-dimensional array corresponding to each example and have only numerical values.

Initialization of population

The population of the first generation is created in a pseudo-random manner by selecting values for the parameter-genes following a uniform distribution for the range of the minimum and maximum allowed values for every parameter. The selection and activation of feature genes is also a random process, but the initial number of active genes in any individual chromosome is constrained to be less than 30 genes, and the selected feature-genes are then filtered according to the feature selection genes.

Calculation of distance between solutions

Distance of parameter-genes calculation:

The distance of parameter genes between two solutions belonging to the same Pareto:

$$D_{pg}(a,b) = \sqrt{\frac{\sum_{i=1}^{\#param}\left(\frac{gene_{a,i} - gene_{b,i}}{(gene_i)\ - min\ (gene_i)}\right)^2}{\#param}}$$

Distance of feature genes calculation:

Logical XOR for the rounded gene values divided by the total number of unique active genes:

$$D_{fg}(a,b) = \frac{\sum_{j=1}^{\#Features} g}{genes_a \cup genes_b}$$

where $g = \begin{cases} 1, if\ gene_{a,j} \neq gene_{b,j} \\ 0, if\ gene_{a,j} = gene_{b,j} \end{cases}$

Pairwise Total distance:

To calculate the total distance of two solutions that belong in the same Pareto frontier, the parameter distance (D_{pg}) and the feature distance (D_{fg}) are used to calculate their average.

$$D_{tot}(a,b) = \frac{D_{pg} + D_{fg}}{2}$$

Degradation of evaluation metrics: (niche (peak))

In the proposed method, a Niched Pareto method that penalizes similar solutions is acting on the individuals that belong in the same Pareto frontier, is deployed to keep a balance between good solutions and pluralism.

$$\sigma_{share} = \frac{0.5}{\sqrt[10]{features}}$$

$$m_{a_{shared}} = \begin{cases} \sum_{b \in Pareto} \left(1 - \left(\frac{D_{tot}(a,b)}{\sigma_{share}}\right)^2\right), if\ D_{tot} \leq \sigma_{share} \\ 0, \quad otherwise \end{cases}$$

where σ_{share} is the radius of the niche (neighborhood) around each solution.

Finally, the degradation of the solutions is calculated as:

$$f_{a_{shared}} = \frac{f_{pareto\,max}}{m_{a_{shared}}}$$

where $f_{pareto\,max}$ is the highest values of the whole Pareto frontier the solution a belongs to, and $m_{a_{shared}}$ is the previously calculated shared factor based on the adjacency of solution a with the other solutions in the same Pareto frontier.

Evolutionary Operators

Every EA has some basic operations that are applied to differentiate the initial population of solutions and allows these algorithms to converge to better solutions. Different approaches and variations have been proposed through the years, but the main operators are conserved as the building blocks of the algorithm, including Selection, Crossover and Mutation operators [202], [203].

In our approach, the solution with the highest weighted overall score is considered the best solution of the generation and passes with its original chromosome unchanged to the next generation. Later, all solutions including the one with the highest overall score get a probability of being selected for pairing proportional to their weighted overall score over the total weighted overall score for all individual solutions.

Despite the numerous alternatives that have been proposed through the years, in MEvA-X the two-point cross-over was chosen as the recombination of genes for the two offspring chromosomes. The two-point approach translates into the exchange of a specific region of the parental genomes between two randomly defined points (P1, P2) with a fixed length (L). Crossover is an operation that is not necessarily happening in every mating of the parental genomes, and the probability of this operation is predefined at 90%; while the rest of the times (1 out of 10) there is no crossover whatsoever, and the parental chromosomes pass to the next generation unchanged.

For the mutation of the offspring, a probability of 5% is selected to allow for the exploration of the feature space without introducing dramatic changes in the population. Single point mutations -similar to SNPs- are used by MEvA-X in the feature genes, toggling the state of these genes from active to inactive and vice versa. The number of point mutations is determined randomly and is not allowed to be higher than six genes to change the new chromosome without bringing huge changes in it. For the parameter-genes on the other hand, since they are coded in continuous values, the mutations on these genes follow a Gaussian distribution around the mean of every parameter.

Termination criteria

There are two conditions that whenever either one of them is met, the evolutionary algorithm terminates. The evolutionary algorithm in MEvA-X is terminated under two conditions and if either one of them is fulfilled. The first termination criterion is if the maximum number of generations selected by the user have reached, while the number of generations that have passed compared to the maximum generations number the user has selected. If the algorithm reaches the number of generations given by the user, it exits the evolutionary process and stores the final solutions. The second ending criterion is the convergence of the solutions to a single niche. This allows the algorithm to stop the evolutionary process if the population has reached very similar solutions and the evolutionary operations will not allow for further exploration of the feature and parameter space.

Metrics

For the evaluation of solutions, multiple classical machine learning metrics for classification are used. Among these metrics, two custom model complexity measures were also implemented to measure the number of features in each model and the number of splits in each learner of the ensemble of the XGBoost model.

Model’s complexity (features):

$$Complexity_{features} = \frac{a}{a + \#selected\ features}$$

Model’s complexity (splits):

$$Complexity_{splits} = 1 - \frac{\text{number of splits}}{\max(allowed\ depth) * 2^{(\max(depth)-1)}}$$

Additionally, metrics that consider the imbalance of the dataset were also used to provide a better understanding of the overall performance of the models. These metrics include the weighted geometric mean, the balanced accuracy, the F1 and F2 scores.

Weighted geometric mean (wGM):

$$wGM = \frac{1}{N}\sum_{i\in I}\left(\sqrt{Sensitivity \cdot Specificity} \cdot support_i\right) = \frac{1}{N}\sum_{i\in I}\left(\sqrt{\frac{TP}{TP+FN}\cdot\frac{TN}{TN+FP}} \cdot support_i\right)$$

Balanced Accuracy (bAcc):

$$bAcc = \frac{Sensitivity + Specificity}{2} = \frac{\frac{TP}{TP + FN} + \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}}{2}$$

F1 Score:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

F2 Score:

$$F2 = 5 \cdot \frac{Precision \cdot Recall}{(4 \cdot Precision) + Recall}$$

Algorithm

```
data preprocessing
pseudo-random initialization of solutions
While max_num_generations not reached AND Similarity_ending_criterion not reached:
        for k ← 0,10:
                train of N models
                Evaluation of model on the selected metrics
        'Best' individual pass to the next generation
        Pareto frontier ranking of solutions.
        Nitch degradation of similar solutions (genome closeness) in the same Pareto front
        Selection of N-1 solutions (randomly, non-uniformly) to pass to the next generation
        crossover on pairs of solutions (90% chance)
        Mutation of single solutions on feature- and/or parameter-genes (5% chance)
        Calculate the similarity of solutions
for k ← 0,10:
        train of N models
        Evaluation of the model on the selected metrics
        Pareto frontier ranking of solutions
Save all models of the 1st Pareto frontier
```

Supplementary Tables

Label Name	Class 0	Class 1	Total	Ratio [0:1]
Severity Change	506	125	631	4.05
Interference Change	523	108	631	4.84
Change Grand Total Medicines	581	50	631	11.62
Change In Total Complaints	481	150	631	3.21

Sup_Table 9. Distribution of the binary classes, and the ratio between them in the four different outcomes of the OPERA study datasets.

	Name	Gene index	Description	Value Range (min – max)
Feature Selection	FS	0	Feature selection method	0 - 4
	FS type	1	Type of feature filtering	0 - 3
	K-NN	2	K nearest neighbors in JMI & mRMR	4 - 10
	K-Best (SKB)	3	Number of features	1 - 100
XGBoost Classifier	Eta	4	Learning rate	0.01 - 0.35
	Max depth	5	Learner maximum depth	1 - 7
	Gamma	6	Pruning parameter	0 - 10
	Lambda	7	Generalization (L2) parameter	0 - 10
	Alpha	8	Generalization (L1) parameter	0 - 8
	Min child weight	9	Pruning parameter	0 - 15
	Scale pos weight	10	Imbalance correction	0 - 5

Sup_Table 10. Parameter-genes of MEvA-X. Genes [0-3] control the feature selection methods to be applied on the dataset, while the rest [4-10] are the hyper-parameters of the XGBoost algorithm.

	AUroC	**Weighted GM**	**F2 score**	**Overall**	**Balanced Accuracy**
Baseline model (No feature selection)	0.702 ± 0.153	0.530 ± 0.217	0.608 ± 0.174	0.583 ± 0.121	0.6 ± 0.147
Wilcoxon	0.677 ± 0.165	0.512 ± 0.207	0.595 ± 0.139	0.571 ± 0.116	0.584 ± 0.143
Select K best (K=100)	0.599 ± 0.167	0.422 ± 0.229	0.544 ± 0.131	0.517 ± 0.112	0.519 ± 0.133
JMI k = 4	0.877 ± 0.121	0.661 ± 0.219	0.732 ± 0.132	0.689 ± 0.109	0.717 ± 0.144
JMI k = 5	**0.880 ± 0.115**	0.669 ± 0.2	0.730 ± 0.127	0.689 ± 0.104	0.718 ± 0.136
JMI k = 6	0.870 ± 0.124	0.669 ± 0.193	0.723 ± 0.131	0.683 ± 0.106	0.712 ± 0.137
JMI k = 8	0.873 ± 0.120	**0.687 ± 0.181**	**0.740 ± 0.126**	**0. 696 ± 0.102**	**0.728 ± 0.132**
JMI k = 10	0.85 ± 0.132	0.683 ± 0.168	0.726 ± 0.125	0.685 ± 0.102	0.718 ± 0.134
mRMR k=4	0.82 ± 0.130	0.653 ± 0.193	0.720 ± 0.128	0.674 ± 0.105	0.704 ± 0.136
mRMR k = 5	0.835 ± 0.131	0.675 ± 0.153	0.719 ± 0.125	0.678 ± 0.098	0.708 ± 0.131
mRMR k = 6	0.849 ± 0.123	0.670 ± 0.183	0.730 ± 0.122	0.685 ± 0.1	0.716 ± 0.131
mRMR k = 8	0.864 ± 0.123	0.666 ± 0.2	0.729 ± 0.126	0.687 ± 0.103	0.72 ± 0.134
mRMR k = 10	0.864 ± 0.119	0.652 ± 0.218	0.722 ± 0.134	0.68 ± 0.111	0.71 ± 0.146

Sup_Table 11. Comparison of the XGBoost models with prior feature selection (with different parameters) and the baseline XGBoost classifier for the Ornish diet dataset. All the experiments were conducted with a stratified 10-fold Cross-validation.

	AUC	Weighted GM	F2 score	Balanced Accuracy	Feature complexity	Overall
Baseline model (No feature selection)	0.7 ± 0.15	0.53 ± 0.22	0.6 ± 0.15	0.6 ± 0.15	7E-4 ± 0.00	0.58 ± 0.12
No EA with JMI (k=5)	**0.88 ± 0.12**	**0. 7 ± 0.16**	0.75 ± 0.13	0.73 ± 0.14	0.17 ± 0.01	0. 7 ± 0.1
MEvA-X Overall	0.76 ± 0.14	0.7 ± 0.19	**0.76 ± 0.14**	**0.74 ± 0.14**	**0.76 ± 0.11**	**0.76 ± 0.15**
MEvA-X Majority Voting	0.76 ± 0.14	0.69 ± 0.17	0.74 ± 0.13	0.73 ± 0.14	0.34 ± 0.19	0.74 ± 0.13

Sup_Table 12. Comparative results of MEvA-X with the baseline (simple XGBoost) and the best feature selection technique (JMI k=5) with XGBoost for the Ornish diet dataset.

Hyper-parameters				**Evaluation metrics**		
				AUC	**F1**	**B.Accuracy**
# Estimators	**Max Depth**	**Min sample split**	**Min sample leaf**	**mean**	**mean**	**mean**
145	2	3	4	0.59 ± 0.26	0.64 ± 0.17	0.61 ± 0.18
133	4	4	7	0.59 ± 0.28	0.63 ± 0.15	0.61 ± 0.17
84	8	8	8	0.60 ± 0.22	0.63 ± 0.15	0.60 ± 0.17
424	2	7	3	0.59 ± 0.25	0.63 ± 0.14	0.60 ± 0.16
740	3	2	1	0.55 ± 0.26	0.63 ± 0.14	0.60 ± 0.16

Sup_Table 13. Table of the results of the trained models with the Random Forest classification algorithm on the Ornish Dataset in a stratified 10-fold cross-validation framework using grid search for the optimization of the parameters. 5 runs were performed and the best-performing solution in each run is described in each row of the table.

Label	Model	AUC	F2 score	Balanced Accuracy	Feature complexity	Overall
Label 1: “Total Severity Change”	Baseline	**0.72 ± 0.06**	**0.76 ± 0.03**	0.55 ± 0.05	0.17 ± 0.0	0.62 ± 0.04
	MEvA-X	0.7 ± 0.07	0.69 ± 0.05	**0.64 ± 0.07**	**0.65 ± 0.0**	**0.7 ± 0.04**
Label 2: “Interference Change”	Baseline	**0.78 ± 0.06**	**0.81 ± 0.04**	0.59 ± 0.06	0.16 ± 0.0	0.67 ± 0.05
	MEvA-X	0.77 ± 0.05	0.77 ± 0.05	**0.69 ± 0.07**	**0.84 ± 0.0**	**0.78 ± 0.04**
Label 3: “Grant Total Medicine Change”	Baseline	**0.82 ± 0.09**	**0.9 ± 0.02**	0.56 ± 0.07	0.16 ± 0.0	0.71 ± 0.05
	MEvA-X	0.78 ± 0.08	0.89 ± 0.03	**0.67 ± 0.09**	**0.79 ± 0.0**	**0.81 ± 0.04**
Label 4: “Total Complaints Change”	Baseline	**0.76 ± 0.06**	**0.76 ± 0.04**	0.62 ± 0.06	0.02 ± 0.0	0.64 ± 0.04
	MEvA-X	0.75 ± 0.07	0.68 ± 0.06	**0.71 ± 0.06**	**0.67 ± 0.0**	**0.72 ± 0.04**

Sup_Table 14. Comparative results between MEvA-X and the baseline (simple XGBoost) for the OPERA dataset. The lines are grouped by the four different labels of the dataset.

	Labels of the OPERA dataset			
	Label 1 Severity Change	**Label 2 Interference Change**	**Label 3 Total Drug Change**	**Label 4 Complaints change**
1	TFFC3	TFFC2	NeuroRadic_cat	Arthritis_bin
2	Age	TFFC3	Arthritis_bin	Tendinitis_bin
3	Gender	TFFC4	Int_Relationship	Grand_Tot_Compl
4	Other_bin	Gender	Int_Life_enjoyment	Pain_other
5	Pain other	Arthritis_categ	Narcotic_weight	Least_24h
6	Grand_Tot_Med_weight	Grand_Tot_Med_weight	Grand_Tot_Med_weight	Overall_Pain_interference
7	Current pain	Tendinitis_bin	Narcotic_bin	Current_pain
8	Severity Score	Grand_Tot_Compl	Tot_Compl_categ	Average_pain
9	Int Work	Worst_24h		Int_Mood
10	Int Sleep	Average_pain		Int_Walking_ability
11	Life enjoyment	Severity_Score		Int_Work
12	Avg Interference	Int_Mood		Int_Relationship
13	Narcotic weight	Int_Work		Avg_Interference
14	Average pain	MyoMuscul_cat		OTC_categ
15		Int_Sleep		AntiInflam_categ
16		Avg_Interference		
17		AntiInflam_categ		
18		Opioid_comb		

Sup_Table 15. Lists of the selected features by the MEvA-X algorithm for the labels of the OPERA dataset. The features are not ranked based on their importance but only listed as important.

	Original name of the variable in the dataset	Alias used
1	AgeatSurvey1	Age
2	GenderRECODE	Gender
3	Q10S1Least	Least_24h
4	Q11S1Average	Average_pain
5	Q12S1RightNow	Current_pain
6	Q13S1	Overall_Pain_interference
7	Q14S1GeneralActivity	Int_Gen_Activity
8	Q15S1Mood	Int_Mood
9	Q16S1WalkingAbility	Int_Walking_ability
10	Q17S1NormalWork	Int_Work
11	Q18S1RelationshipsWithOtherPeople	Int_Relationship
12	Q19S1Sleep	Int_Sleep
13	Q1S1ArthritisRecode	Arthritis_bin
14	Q1S1ArthritisTotal	Arthritis_categ
15	Q20S1EnjoymentofLife	Int_Life_enjoyment
16	Q26S1	Meds_3days
17	Q272829S1HowManyCategoriesOfMeds	Categ_of_meds
18	Q27S1OTCRecode	OTC_bin
19	Q27S1OTCTotal	OTC_categ
20	Q28S1AntiInflamREcode	AntiInflam_bin
21	Q28S1AntiInflamTotal	AntiInflam_categ
22	Q28S1AntiInflamWeight5	AntiInflam_weight
23	Q29S11or2OpioidsNoAnticonvulsants	Opioids_No_Anticonvulsants
24	Q29S1NarcoticRecode	Narcotic_bin
25	Q29S1NarcoticTotal	Narcotic_categ
26	Q29S1NarcoticWeight10	Narcotic_weight
27	Q2S1NeuroRadicRecode	NeuroRadic_bin
28	Q2S1NeuroRadicTotal	NeuroRadic_cat
29	Q3S1MyoMusculPainORSpasmTotal	MyoMuscul_cat
30	Q3S1MyoMuscuRecode	MyoMuscul_bin
31	Q4S1TendinitisRecode	Tendinitis_bin
32	Q4S1TendinitisTotal	Tendinitis_cat
33	Q5S1OtherRecode	Other_bin
34	Q5S1OtherTotal	Other_cat
35	Q7S1	Pain_other
36	Q9S1Worst	Worst_24h
37	S1GrandTotalMedicines	Grand_Tot_Med
38	S1GrandTotalMedicinesWEIGHTED	Grand_Tot_Med_weight

39	S1GrandTotalofAllComplaintsExceptOverallOther	Grand_Tot_Compl
40	S1InferenceScoreOutof10	Avg_Interference
41	S1NotOnAntiConvulsant	Not_on_Anticonvulsant
42	S1NumberofPainCategoriesIncludingOther	Pain_categ
43	S1OnOpioidAloneorOpioidPlus	Opioid_comb
44	S1Q29OpioidAloneorOthers	Opioid_others
45	S1SeverityScoreOutof10	Severity_Score
49	S1TotalComplaintCategoriesNotOther	Tot_Compl_categ
47	TFourFormulationCategories_1	TFFC1
48	TFourFormulationCategories_2	TFFC2
49	TFourFormulationCategories_3	TFFC3
50	TFourFormulationCategories_4	TFFC4

Sup_Table 16. List of the original names in the OPERA dataset and the aliases used in the ML model. Features starting with Q come from the questionnaire, those starting with S are treatment data and features starting with T are the outcomes that were used as target labels.

Chapter – 5

Supplementary tables

	Feature name	Description	Value range	Mean value	Median value
1	Age	Age at each follow-up	2 - 51	25.3	25
2	Start of Chelation (age)	Age of Chelation (common for all follow-ups of the same patient)	0 – 29, NaN	4.84	4
3	BMI	Body Mass Index	3 – 42, Nan		22.5
4	Blood unit hematocrit	Percentage of red blood cells compared to the total blood volume	0.58 – 1, NaN	0.7	0.65
5	Splenectomy (Status)	Has the patient had splenectomy?	Yes/No/NaN	Yes	Yes
6	HCV Ab (Status)	has the patient been infected with the hepatitis C virus	Yes/No/NaN	Yes	Yes
7	Years from first follow-up	Time since diagnosis	0 - 66	24.7	30.3
8	Hb	Hemoglobin levels measured in g/dL	5.4 - 16.5, Nan	10.8	10.6
9	# Monocytes	Percentage of white blood cells being Monocytes	0.07 - 7.17, NaN	1.06	0.94
10	# Eosinophils	Count of eosinophils per µL of blood	0.0 .57, NaN	0.3	0.24
11	Hematocrit	Ratio of red blood cells volume to the total blood volume	21.4 - 63.2, NaN	33.7	33.3
12	Years of chelation	Time-increasing variable since the start of chelation therapy	0.0 - 50.2, NaN	32.5	36
13	Years with diabetes	Time-increasing variable since the diagnosis of diabetes	0 – 40, NaN	9.7	8.5
14	Years after splenectomy	Time-increasing variable since the procedure	0.86 - 54.95, NaN	30.15	32.6
15	Gender	Binary feature	F (0), M (1)	M	M
16	Chelation Therapy	Binary feature	Yes/No/NaN	Yes	Yes
17	Hydroxyurea Therapy	Is the patient under Hydroxyurea therapy	Yes/No/NaN	No	No

Sup_Table 17. Descriptive table of the 17 features used to build the predictive models for predicting CVD risk in thalassemia patirnts.

From the analysis of the mistakes of the two models in the cases where all the follow-ups of patients were classified incorrectly, it occurs that the patient with the ID “LPRTCLRAERFLHYE” who is also a CVD patient is missed completely and has a sufficient number of follow-ups to back-up the claim that the model fails. The rest of the patients even if they have all their follow-ups incorrectly classified, they only have from one to three visits which is not a statistically significant number to extract any conclusions.

		1-year model	**3-Years model**	
ID	**Class**	**total**	**total**	**Category**
BNDINTNNLEMXMNC	0	1	2	Always incorrect
DCVEMSSM__MNKSV	0	1	2	Always incorrect
LPRTCLRAERFLHYE	1	69	193	Always incorrect
VNTRLTZC__FNMQZ	0	3	3	Always incorrect
CNTPLNREBOFBJBV	0	-	1	Always incorrect

Sup_Table 18. List and information of patients where the models failed to classify any of their follow-ups correctly.

A similar analysis conducted for patients that the models failed to classify correctly but only in part of their follow-ups as shown in the table below and the Sup_Fig. 16.

		1-year model			3-Years model		
ID	**Class**	**Correct**	**Incorrect**	**total**	**Correct**	**Incorrect**	**total**
MRTSSTFN_FWCWY	1	17	32	49	112	26	138
	0	-	-	-	149	3	152

Sup_Table 19. List and information of partially correctly classified patients and their follow-ups.

Training Features list

From the Nested cross validation scheme used to tune the hyperparameters of the models, below are the best set according to the grid search performed in the training set.

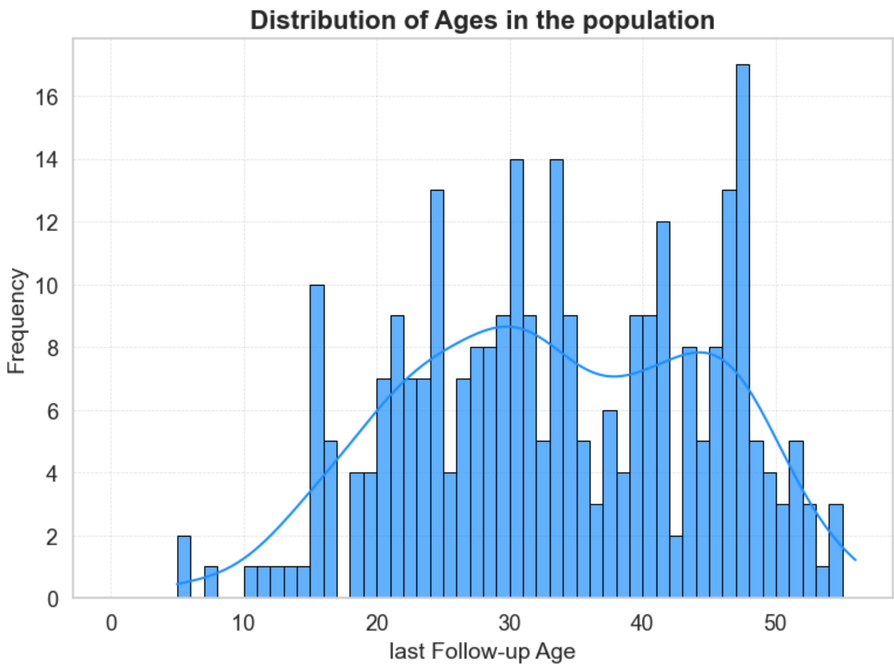
Model's hyperparameters

	1-Year Model	3-Years Model	Cross validation
Parameter Name	**Parameter Value**	**Parameter Value**	**Parameters**
learning_rate	0.01	0.01	0.01, 0.03, 0.05, 0.1
max_depth	5	3	3, 5, 7
n_estimators	700	500	500, 700, 1000
alpha (L1regression)	0.5	0.01	0.01, 0.05, 0.1, 0.5
colsample_bytree	0.5	0.5	0.5
scale_pos_weight	5.875	5.875	negative_class/positive_class
Early stopping rounds	200	200	
objective	Binary:Logistic	Binary:Logistic	
eval_metric	logloss	logloss	
tree_method	hist	hist	

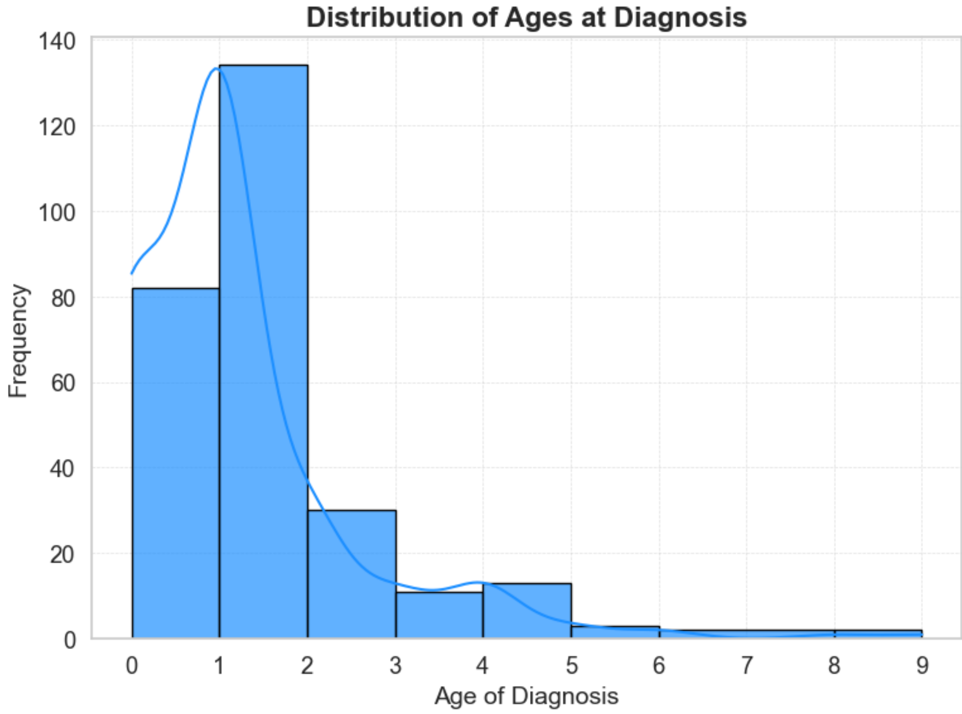
Sup_Table 20. Hyper-parameters and model parameters used to train the XGBoost models in the nested cross-validation scheeme for the risk prediction of cardiovascular disease in Thalassemia patients for the 1 & 3 years models.

Supplementary figures

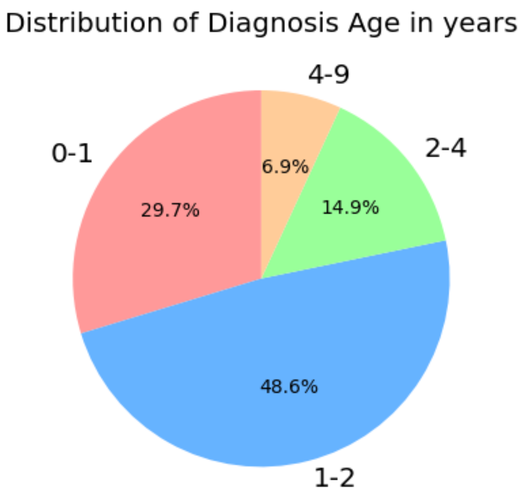
After data cleaning and selection of the last one and three years of follow-ups for the thalassemia patients, a descriptive analysis was made to visualize basic characteristics of the data. As shown in the figure below, we visualized the distribution of age of the patients in their last visit at the hospital (Sup_Fig. 13) showing that most of the patients are adults ranging from 18-56 years old. Additionally, a plot for the distribution of the age of diagnosis was drawn (Sup_Fig. 14) and the difference of those two features shows the years that a patient has been diagnosed with thalassemia. As a supportive graph, we generated a pie chart with age groups of diagnosis as shown in Sup_Fig. 15.



Sup_Fig. 13. Distribution of the samples in their last recorded follow-up visit.

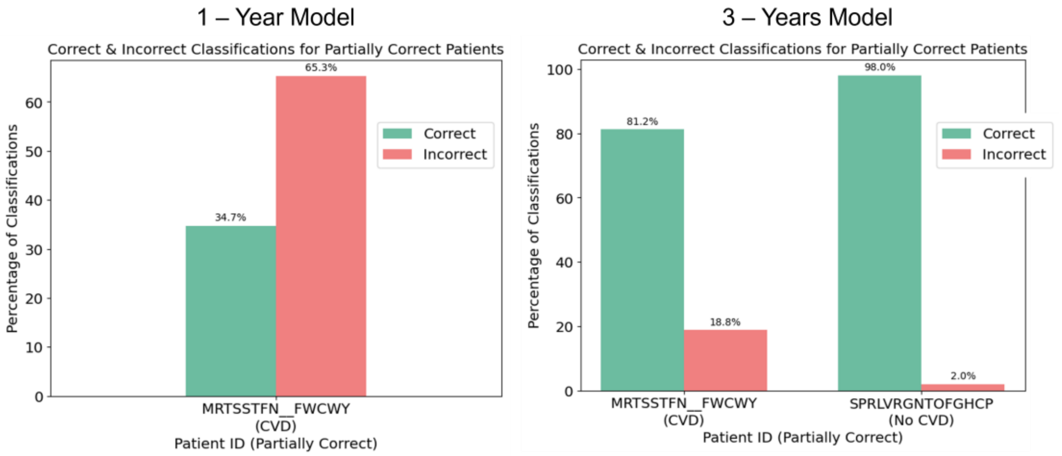


Sup_Fig. 14. Distribution of the age of the patients when they have been diagnosis with thalassemia



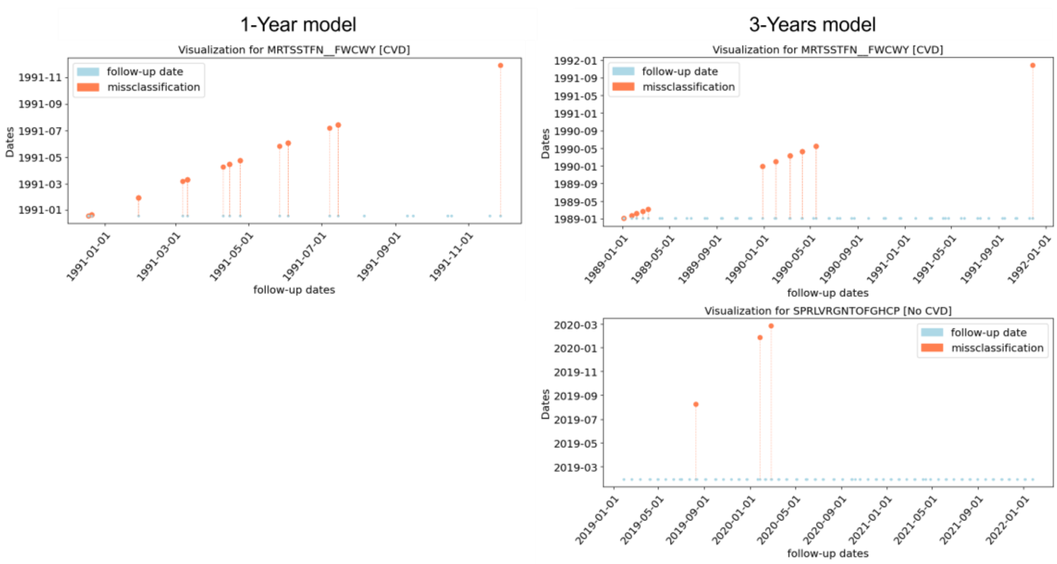
Sup_Fig. 15. Age group distribution of patients at the time of diagnosis.

From the analysis of the mistakes of the models described also in Sup_Table 19, it occurs that one patient is shared in both models as incorrectly classified in some of the follow-ups, and in the three years model there is another patient where 3 out of 152 hospital visits were falsely classified as "CVD".



Sup_Fig. 16. Analysis of the patients that had a mixture of correct and incorrect classifications of their follow-ups.

From the time analysis of the mistakes, no clear pattern is shown for these cases even though they are only two patients, and so, no specific explanation can be given visually for why the models fail in these cases.



Sup_Fig. 17. Time-based visualization of the misclassifications for the patients whose follow-ups were partially correctly classified.