Ph.D. Thesis

XXXVI cycle

"Machine learning approaches for personalized medicine"

Ph.D. candidate : Mauro Nascimben
Supervisor : Prof. Lia Rimondini

UNIVERSITÀ DEL PIEMONTE ORIENTALE

Ph.D. Program in Medical Sciences and Biotechnology
Department of Health Sciences
Academic Discipline [Area 06 - SSD med/50]

Contents

Summary (in english)

The work carried out during the Ph.D. in Medical Sciences and Biotechnology course tested the application of machine learning models in several precision medicine topics. In bioinformatic biomarker analysis, the publications focused on discretizing the gene expression levels to obtain a manageable and insightful granularity. The works demonstrated novel analysis pipelines to detect survival and tumor stages from oncologic patients' biomarkers. The same chapter presented a procedure for a public health decision support system based on machine learning, which has also been demonstrated on the same dataset. The chemoinformatics numerical experiments for drug toxicity, bioaccumulation prediction, or P450 enzyme bioactivity evaluation all exploited spiking neural networks, showing the ability of this technique to handle structural information of the compounds for predictive analysis. For clinical precision medicine, an algorithm has been tested fusing clinical variables (ordinal and binary) from nearly 300 patients to forecast the risk of developing lymphedema after breast cancer therapy. Moreover, free software has been released to measure the volumetry of the affected limb in case of edema or other pathologies requiring tracking of body parts over time. Another chapter reported the development of a free Python library to run equivalence tests in the biomedical sector, focusing on advanced visualization of the statistical outcomes. This library also fills a gap in the biostatistical tools available to Python users requiring biomedical equivalence analysis. Regarding regenerative medicine, a study has been introduced to track octacalcium phosphate synthesis through a machine-learning methodology centered on a novel algorithm exploiting an ad-hoc solution on merged XRD and FTIR peak descriptors. Octacalcium phosphate is found in biological systems, particularly in the early bone formation and mineralization stages. It is a precursor to hydroxyapatite, the main mineral component of bones and teeth. The last chapter introduced a mass spectrometry proteomic analysis sequence to detect aberrant protein expression levels. The procedure has been tested on mesenchymal stem cells' extracellular vesicle protein content cultured on biomaterials doped or not with metallic ions.

Sommario (in italiano)

Le attività svolte durante il corso di Dottorato in Scienze Mediche e Biotecnologie si sono concentrate sull'applicazione di modelli basati sul machine learning in diversi settori della medicina di precisione. Per l'analisi bioinformatica dei marcatori tumorali, le pubblicazioni si sono concentrate sulla discretizzazione dei valori di espressione genica per ottenere una granularità dei dati più gestibile e rilevante per classificare i pazienti. I lavori svolti hanno dimostrato l'uso di nuove sequenze di analisi per determinare sia la sopravvivenza che gli stadi tumorali partendo dai biomarcatori nei pazienti oncologici. Lo stesso capitolo ha presentato una procedura per creare un sistema di supporto alle decisioni di sanità pubblica basato sull'apprendimento automatico, dimostrato sullo stesso tipo di dati. Gli esperimenti numerici nell'ambito della informatica chimica si sono concentrati sulla tossicità delle molecole, la previsione del bioaccumulo delle sostanze

negli esseri viventi o la valutazione della bioattività dell'enzima P450. Tutti questi lavori hanno sfruttato le reti neurali spiking per l'analisi predittiva, dimostrando la capacità di questa tecnica di gestire le informazioni contenute nella sola struttura chimica dei composti. Per la medicina clinica di precisione, è stato testato un algoritmo che fonde le variabili cliniche (ordinali e binarie) di quasi 300 pazienti per predeterminare il rischio di sviluppare linfedema dopo la terapia del cancro al seno. Inoltre, è stato rilasciato un software gratuito per misurare la volumetria dell'arto interessato in caso di edema o altre patologie che richiedono il monitoraggio nel tempo della morfologia degli arti. Un ulteriore capitolo ha descritto lo sviluppo di una libreria Python gratuita per eseguire i test di equivalenza specifici del settore biomedico, concentrandosi sulla visualizzazione tramite grafici dei risultati statistici. Questa libreria colma anche una lacuna negli strumenti biostatistici disponibili per gli utenti Python che necessitano dei test per l'equivalenza in medicina. Per quanto riguarda la medicina rigenerativa, un algoritmo capace di tracciare la sintesi del fosfato ottacalcico attraverso una metodologia di apprendimento automatico ha sfruttato una soluzione ad–hoc per quantificare le fasi di produzione partendo da nove descrittori delle caratteristiche dei picchi nei segnali XRD e FTIR. Il fosfato ottacalcico si trova in diversi tessuti anatomici, in particolare nelle prime fasi di formazione e mineralizzazione delle ossa. È un precursore dell'idrossiapatite, il principale componente minerale di ossa e denti. L'ultimo capitolo ha introdotto una sequenza di analisi nel campo della proteomica che ha utilizzato i dati di spettrometria di massa per rilevare livelli abnormali di espressione proteica. La procedura è stata testata sul contenuto proteico delle vescicole extracellulari nelle cellule staminali mesenchimali coltivate su diversi biomateriali drogati con ioni metallici o puri.

1 Introduction

Precision medicine, also known as personalized medicine, is an approach to healthcare that considers individual variability in genes, environment, and lifestyle for each person. It seeks to customize medical treatment to each patient's unique characteristics, aiming to improve treatment effectiveness and minimize side effects. Precision medicine considers factors such as a person's genetic makeup, the molecular profile of their disease, and other specific characteristics to tailor prevention, diagnosis, and treatment strategies. By utilizing advanced technologies, such as genetic sequencing, molecular diagnostics, and big data analytics, precision medicine allows healthcare professionals to make more informed decisions when determining patients' most suitable treatment plans. This approach enables the identification of specific treatments that are more likely to be effective for certain individuals or groups, leading to better health outcomes and potentially reducing healthcare costs in the long run. Precision medicine has applications across various medical disciplines, including oncology, cardiology, neurology, and infectious diseases. It represents a shift from the traditional "one-size-fits-all" approach to medicine toward more targeted and personalized treatments that consider each patient's unique characteristics.

Machine learning (i.e., ML) is a subset of artificial intelligence that focuses on developing algorithms and statistical models that enable computers to learn from and make predictions or decisions based on data without being explicitly programmed for those tasks. Machine learning algorithms allow computer systems to improve their performance on a specific task over time as they are exposed to more data. There are several types of machine learning techniques, exemplified in Figure 1.1.

Each machine learning approach has its strengths and weaknesses, making them suitable for different types of problems and datasets.

- Supervised Learning: In supervised learning, the algorithm is trained on a labeled dataset, where the corresponding correct output accompanies the input data. The algorithm learns to map the input to the output and make predictions or decisions based on new data. Popular supervised learning algorithms include linear regression, logistic regression, decision trees, support vector machines, and neural networks.

- Unsupervised Learning: Unsupervised learning involves training the algorithm on data that is not labeled or classified. The algorithm must find patterns and relationships within the data on its own. Clustering and association are typical tasks in unsupervised learning. Popular unsupervised learning algorithms include k-means clustering, hierarchical clustering, and association rule learning.

- Semi-Supervised Learning: Semi-supervised learning combines elements of both supervised and unsupervised learning. It uses a small amount of labeled data and

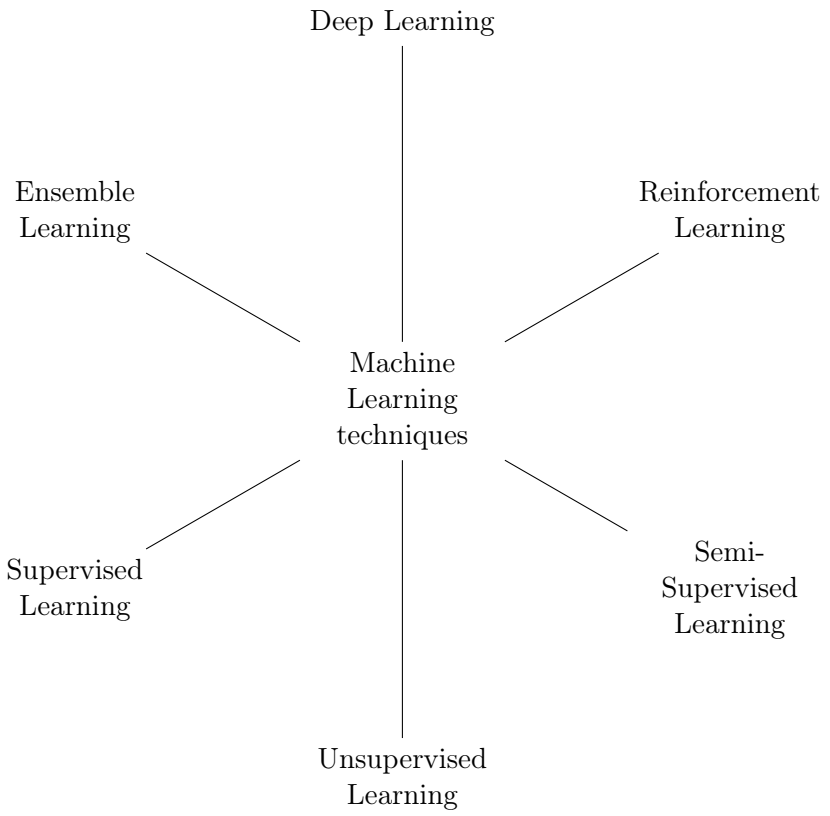

Figure 1.1: Main learning methodologies

a more significant amount of unlabeled data to improve learning accuracy. This technique is beneficial when obtaining a large amount of labeled data is difficult or expensive.

- Reinforcement Learning: Reinforcement learning involves training the algorithm to make decisions in a specific environment to achieve a goal. The algorithm learns through trial and error, receiving positive or negative feedback based on its actions. Popular reinforcement learning algorithms include Q–learning and deep Q–networks.

- Deep Learning: Deep learning is a subset of machine learning that uses artificial neural networks with multiple layers to model and understand complex patterns and relationships within data. Deep learning excels in image and speech recognition, natural language processing, and other complex pattern recognition tasks. Convolutional Neural Networks and Recurrent Neural Networks are popular deep learning architectures.

- Ensemble Learning: Ensemble learning involves combining multiple machine learning models to improve the performance and robustness of the system. Techniques such as bagging, boosting, and stacking are used to create ensemble models, such as random forests and gradient boosting machines.

In recent years, the intersection of machine learning and precision medicine has opened up new avenues for personalized and effective healthcare interventions. Algorithm–based medicine can be considered an active component of precision medicine, as it involves the application of computational algorithms and decision rules to guide clinical decision-making and patient management. On the other hand, precision medicine is an approach to healthcare that customizes medical treatment and interventions to individual characteristics, such as genetic makeup, environmental factors, and lifestyle choices. Algorithm–based medicine contributes to personalized medicine by providing standardized and efficient protocols for diagnosing, treating, and managing patients. By leveraging computational algorithms and decision-support systems, healthcare providers can make more informed and data-driven decisions that are tailored to the individual patient's needs. These algorithms can help interpret complex patient data, identify potential treatment options, and predict patient outcomes, thereby supporting the delivery of personalized care in line with the principles of precision medicine. While algorithm-based medicine is a valuable tool within the broader framework of precision medicine, it is just one of the many components that contribute to the goal of delivering targeted and personalized healthcare interventions to optimize patient outcomes. Integrating various technologies, data analytics, and clinical expertise plays a crucial role in advancing the practice of precision medicine and improving patient care. Under this view, evidence–based medicine emphasizes the integration of clinical expertise, patient values, and the best available research evidence in the decision-making process for patient care. In the context of precision medicine, evidence–based medicine is critical in ensuring that medical decisions are based on the most current and reliable scientific evidence. By incorporating the

findings of rigorous clinical research, such as randomized controlled trials, systematic reviews, and meta–analyses, evidence-based medicine helps healthcare professionals make informed decisions about the most effective treatments and interventions for individual patients. The principles of evidence-based medicine are essential in selecting appropriate diagnostic and therapeutic strategies, thereby contributing to delivering precise and effective healthcare interventions. By integrating evidence-based practices with a personalized approach to patient care, precision medicine can optimize treatment outcomes and improve patient satisfaction and quality of life. Indeed, the evidence–based paradigm of medicine could be integrated with machine learning (or algorithm–based medicine), playing important roles in improving healthcare outcomes [Subbiah, 2023].

1.1 Overview of machine learning models for precision medicine

Precision medicine, or personalized medicine, is an approach to medical treatment and healthcare that considers individual variability in patients' genes, environments, and lifestyles. Machine learning is crucial in advancing precision medicine by helping healthcare professionals make more accurate and tailored treatment decisions for individual patients [MacEachern and Forkert, 2021]. Integrating machine learning into medicine promises to improve patient care, reduce costs, and advance our understanding of diseases and treatments. Indeed, machine learning is crucial in advancing precision medicine by helping healthcare professionals make more accurate and tailored treatment decisions for individual patients. Some specific ways in which machine learning is utilized in precision medicine include:

1. ML algorithms could analyze vast amounts of genomic data; this encompasses identifying genetic mutations, understanding gene expression patterns, and predicting disease risks based on an individual's genetic makeup. Machine learning can help uncover hidden patterns and associations in genomic data. In this context, ML could be applied to discover and validate biomarkers, that are molecular or genetic indicators of disease presence, progression, or response to treatment. For clinical trials management, ML could identify suitable candidates for clinical investigations based on their genetic profiles, increasing the chances of successful trial outcomes and the development of personalized therapies.

2. ML could be employed for drug discovery and development to predict how specific drugs interact with a patient's genetic profile [Vamathevan et al., 2019]. By virtual screening and quantitative structure-activity relationship analyses, researchers can help in drug repurposing, identifying potential drug candidates, and optimizing the design of clinical trials. Additionally, in chemoinformatics, ML can predict how individual patients will respond to specific medications, allowing for more precise and effective treatment plans with fewer adverse effects.

3. Another crucial aspect of tailoring treatment and intervention strategies to specific

patient populations is identifying subgroups with similar characteristics and disease profiles. Stratifying patients at a population level can help healthcare systems and providers identify high-risk patient groups, allocate resources accordingly, and deliver patient-centered care. Accurate patient evaluation also has the final goal of enabling clinicians to adjust treatment plans and interventions in real-time. Moreover, identifying health trends and risk factors for population health management informs public health initiatives and interventions. For example, analyzing medical imaging data, such as MRI and CT scans, algorithms could classify subtle patterns and markers that may not be apparent to the human eye, aiding early disease diagnosis and treatment planning [Plant and Barton, 2021].

This summary of the ML involvement in precision medicine only reported a few application areas where predictive algorithms could be implemented [Wilkinson et al., 2020; Nayarisseri et al., 2021]. For example, more in-depth discussion could be addressed on genomics ML-based in-silico models operating on genomic sequence analysis, gene expression analysis, functional annotation, genomic variation analysis or association studies, epigenomics, metagenomics, or single-cell genomics. Indeed, genomics is the study of the complete set of an organism's genes, including their sequences and structures. However, other "omics" techniques offer exploitable biological data for ML algorithms to understand biological systems at different molecular levels. They include transcriptomics, proteomics, metabolomics, metagenomics, epigenomics, phenomics, lipidomics, glycomics, and pharmacogenomics. Omics fields are highly interdisciplinary, involving biology, genetics, bioinformatics, and various laboratory techniques to generate and analyze large datasets. The big datasets created with omics outputs have transformed biological and medical research by providing comprehensive insights into complex biological systems, enabling advances in personalized medicine, drug discovery, and understanding the molecular basis of diseases. Indeed, the term multi-omics refers to the integration and analysis of data from multiple "omics" fields; this approach should enable a more comprehensive understanding of complex biological systems by considering various molecular layers simultaneously. ML provides methods to combine and integrate different omics sources for multi-omics dataset building, aligning them to provide a holistic view of biological systems. The multi-source data could result in many featured genes, proteins, and metabolite expression indicators. Therefore, in a standard ML analysis pipeline (a summary in Figure 1.2), selecting a subset of relevant descriptors helps focus on the most informative variables for downstream analyses. Alternatively, dimensionality reduction could be employed to produce a lower set of virtual features summarizing the characteristics of the original ones. Selecting relevant features and reducing the dimensionality of data is crucial for identifying the most important factors contributing to a patient's health condition and improving the interpretability of models. The outcomes of ML-centered multi-omics analysis might have the potential to drive advancements in fields like personalized medicine, not only by predicting various biological outcomes or clinical responses, but also constructing biological pathways or networks to highlight interactions (gene-gene interactions, protein-protein interactions, and regulatory networks), show significant differences between experimental conditions which is essential for understanding

the molecular basis of diseases and variations, or discover biomarkers by identifying signatures associated with specific clinical diseases.

Also, drug development, metabolism, and toxicity are topics that ML has successfully applied in their various connotations. Machine learning is used to predict the properties of molecules and identify potential drug candidates. By analyzing large datasets of chemical and biological information, ML models can suggest novel compounds that are highly likely effective against specific diseases [Wale, 2011]. Additionally, ML algorithms can help identify potential drug targets by analyzing complex biological data [Kalinin et al., 2018]. They can sift through genetic, proteomic, and other biological data to pinpoint specific molecules or pathways crucial in disease development. For predictive toxicology, machine learning models are used to predict the toxicity of new drug candidates, helping researchers identify potential safety issues early in the development process. Understanding compound toxicology can help prioritize the most promising ones and reduce the time and costs associated with drug development. As optimization tools, ML is used to manage clinical trials by identifying the most relevant patient populations, predicting patient responses to treatments, and improving trial design. Proper management can accelerate the development process and improve the chances of success in clinical trials. With "drug repurposing", machine learning can help identify new uses for existing drugs by analyzing large datasets of biological and clinical information. By understanding the molecular mechanisms of different diseases, researchers can repurpose existing drugs for new indications, potentially reducing the time and costs associated with traditional drug development. In the view of precision medicine, ML-driven drug discovery allows for the production of targeted therapies tailored to a patient's unique genetic makeup and disease characteristics.

Patient stratification, also known as patient segmentation, is one core concept of precision medicine, and it refers to dividing patients into subgroups based on specific characteristics in terms of genome to phenotype expression [Glaab et al., 2021], disease biomarkers, or clinical presentation. Data-driven ML plays a crucial role in patient stratification in medicine, enabling more precise and personalized approaches to treatment and care by identifying distinct disease subtypes. In particular, machine learning algorithms can predict disease progression, treatment responses, and patient outcomes by analyzing patient-specific factors. One goal is to build decision-support systems assisting healthcare professionals by integrating patient data, clinical guidelines, and research evidence. Eventually, such decision systems able to work in real-time can strengthen doctors' confidence in making more accurate and personalized treatment decisions, ultimately improving patient outcomes and safety. However, the effective implementation of decision support systems for patient stratification requires robust data quality, model interpretability, and validation using real-world clinical data [Beaulieu-Jones et al., 2021].

Another aspect ML can introduce to precision medicine is the concept of causality [Sanchez et al., 2022]. Causal machine learning refers to using machine learning techniques to infer and understand causal relationships between variables in a system. Unlike traditional machine learning, which focuses on predicting outcomes based on patterns in data, causal machine learning seeks to identify cause-and-effect relationships, allowing for a deeper understanding of how changes in one variable affect another. Causal ma-

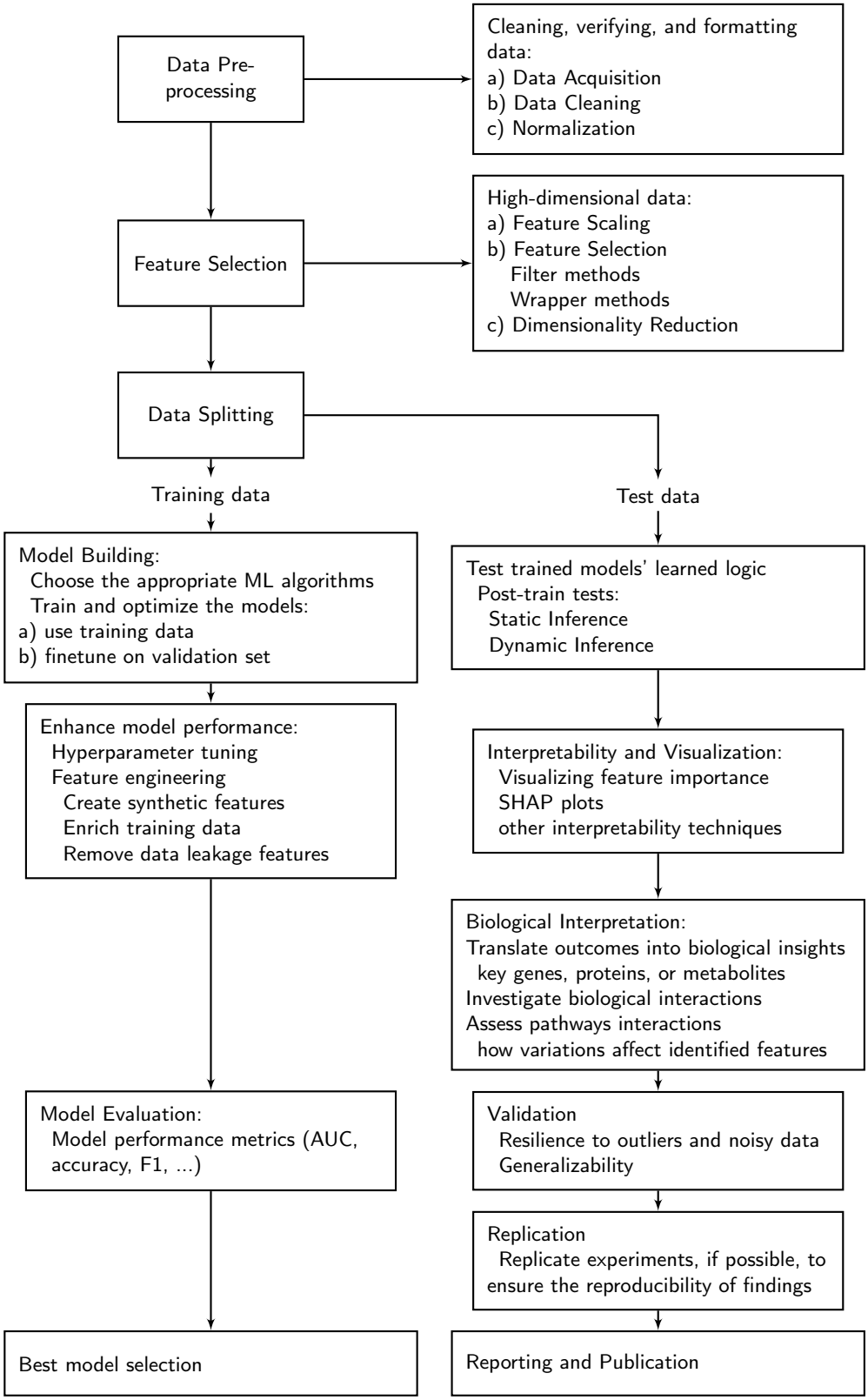

Figure 1.2: A typical machine learning workflow for omics data

chine learning could play a crucial role in precision medicine by helping researchers and healthcare practitioners uncover causal relationships between variables, identify treatment effects, and make more informed decisions about patient care. Precision medicine aims to tailor medical treatments to the individual characteristics of each patient, taking into account genetic, environmental, and lifestyle factors. Causal machine learning techniques contribute to this goal by providing a deeper understanding of the causal mechanisms underlying diseases and treatment responses. For example, causal machine learning methods, such as propensity score matching, instrumental variable analysis, and regression discontinuity design, are used to estimate the causal effect of a treatment or intervention: it can aid in determining the effectiveness of specific treatments for individual patients based on their characteristics. Moreover, for biomarker identification, causal inference methods help identify biomarkers and genetic factors that are causally linked to diseases [Lecca, 2021]. Understanding the causal relationships between genetic variations and disease outcomes can guide the development of targeted therapies. Additionally, by analyzing causal relationships in patient data, machine learning models can assist in developing personalized treatment plans. To optimize treatment outcomes, these plans consider various factors, such as genetic information, patient history, and lifestyle. Causal machine learning can enhance the design of clinical trials by identifying patient subgroups that are more likely to benefit from a particular treatment, leading to more efficient and targeted clinical trials, reducing costs, and expediting the development of effective therapies. Another interesting point is that causal models enable counterfactual reasoning, allowing researchers to simulate "what-if" scenarios [Karim et al., 2023]. It is beneficial in assessing the potential outcomes of different treatment strategies and understanding how alternative interventions might impact patient outcomes. In clinical settings, causal reinforcement learning is applied to develop dynamic treatment regimes that adapt over time based on patient responses and changing conditions. This approach is precious for chronic and complex diseases where treatment strategies may need to be adjusted throughout the disease.

1.2 In–silico ML models

In silico or computational models are mathematical or computer-based models that simulate and predict biological, chemical, or physical processes. Machine learning is increasingly applied to enhance the accuracy and predictive power of *in silico* models in various fields, including chemoinformatics [Lo et al., 2018; Fox and Kriegl, 2006; Niazi and Mariam, 2023], materials science [Morgan and Jacobs, 2020; Jablonka et al., 2020; Batra et al., 2021; Wang et al., 2020], and biology [Greener et al., 2022; Shastry and Sanjay, 2020; Cao et al., 2020; Auslander et al., 2021; Reel et al., 2021]. A visual map of the fields of application and the subbranches of machine learning in medicine and biology is displayed in Figure 1.3. However, it should be noted that in–silico models might rely on different technologies, and ML is only one of the available ones [Faulon and Faure, 2021]. The complete in–silico scenario is reported below:

- *Molecular Modeling*: This involves simulating and predicting the structure and

behavior of molecules, such as proteins, DNA, and small molecules. It includes techniques like molecular dynamics simulations and docking studies to understand molecule interactions [Noé et al., 2020; Glielmo et al., 2021].

- *Systems Biology Models*: These models integrate biological components, such as genes, proteins, and biochemical reactions, to simulate and analyze complex biological systems. They help understand the behavior of entire biological systems, including signaling pathways and regulatory networks [Kim et al., 2020].
- *Network Models*: Network models represent and analyze complex biological networks, such as gene regulatory networks, protein-protein interaction networks, and metabolic networks. They help understand the relationships and interactions between various components within a biological system [Liu et al., 2020].
- *Quantitative Structure-Activity Relationship (QSAR) Models*: These models are used in drug discovery and development to predict chemical compounds' biological activity or properties based on their structure [Carracedo-Reboredo et al., 2021]. QSAR models help in the design and optimization of new drug candidates [Chen et al., 2018].
- *Pharmacokinetic/Pharmacodynamic Models*: These models are used to study the absorption, distribution, metabolism, and excretion (pharmacokinetics) of drugs, as well as their pharmacological effects (pharmacodynamics) in the body [Ota and Yamashita, 2022]. PK/PD models help in predicting drug behavior and optimizing dosage regimens.
- *Bioinformatics Models*: Bioinformatics models involve applying computational techniques to analyze and interpret biological data, such as DNA sequences, protein structures, and gene expression data. They help understand biological processes at the molecular level and design experiments for further analysis [Koumakis, 2020; Yang et al., 2020].
- *Artificial Intelligence and Machine Learning Models*: AI and machine learning models are increasingly used in biology to analyze complex biological data, predict outcomes, and identify patterns or trends that might not be easily discernible using traditional statistical methods.

In–silico models built with machine learning are faster, cost–effective, and capable of handling large–scale data. However, validating and refining these models using experimental data is crucial to ensure their reliability and accuracy [Faulon and Faure, 2021].

Apart from the vast set of methodologies applicable to biology and medicine, all share a joint proposition called the "Goldilocks principle" (also known as the "principle of the golden mean"). The level of detail in an in-silico model should be sufficient to accurately represent the biological system or process being studied while balancing computational feasibility and practical constraints. Indeed, the Goldilocks principle in modeling underscores the importance of finding a "just right" balance for optimal outcomes regarding the size and complexity of the model [Ko and Wren, 2021].

Figure 1.3: Possible fields of application for ML in silico modeling

Phenomenological	**Mechanistic**
Phenomenological models in medicine are typically empirical and based on observed clinical data and patterns.	Mechanistic models in medicine aim describe biological mechanisms and physiological processes using mathematical equations.
They focus on describing and predicting the behavior of a disease or biological process without necessarily understanding the underlying mechanisms causing the observed phenomena.	Require a detailed understanding of the pathophysiology, molecular interactions, and physiological changes associated with the disease or treatment response.
These models are often statistical or probabilistic in nature and are useful for predicting clinical outcomes, disease progression, and treatment responses based on historical data and trends.	These models often involve the integration of biochemical, physiological, and pharmacological knowledge into mathematical or computational frameworks, such as systems biology or pharmacokinetic–pharmacodynamic models.
They are commonly used when the exact biological mechanisms contributing to a disease are not fully understood, but there is enough data to establish correlations and patterns.	Mechanistic models are parameter-dependent for simulating the effects of biological systems and demand strong mathematical and biological knowledge.
Can be handled with open source software.	Often require licensed computer simulators.

Table 1.1: Features of in–silico phenomenological vs. mechanistic models

In biology and medicine, another type of categorization of the in–silico models involves the so–called phenomenological and mechanistic models [Vert and Jacob, 2008]. They differ in their approach to understanding and predicting biological phenomena and diseases. While phenomenological medical models help make predictions based on observed correlations, mechanistic models offer an understanding of the biological processes driving disease progression and treatment responses. Both types of models play a crucial role in different aspects of medical research, including disease modeling, drug development, and personalized medicine, each providing unique insights and applications for improving patient outcomes and healthcare practices [Rodrigue and Philippe, 2010]. The Table 1.1 highlights the main differences between them.

Among mechanistic, compartmental models are a type of mathematical modeling used to represent and study the flow of quantities or substances through different compartments in a system. In these models, compartments represent distinct groups or locations where the quantities of interest are stored or transferred. The most common compartmental models in medicine are sets of mathematical equations that describe the distribution, metabolism, and elimination of drugs or other substances in the body. These models are precious in pharmacokinetics, which studies how drugs are absorbed, dis-

tributed, metabolized, and excreted by the body [Jia and Gao, 2022]. Compartments in these models represent various physiological or anatomical locations within the body, such as blood, tissues, organs, or specific physiological systems. The flow of substances between these compartments is described using mathematical equations that account for the rates of absorption, distribution, metabolism, and excretion of the substances. These models help in predicting how the concentration of a drug or substance changes over time in different parts of the body. They can also be used to optimize dosage regimens, estimate drug clearance rates, and understand how factors such as age, weight, and physiological conditions affect the pharmacokinetics of a drug [Vodovotz and An, 2019]. While compartmental models are widely used and valuable tools in various fields, including medicine, they have limitations and potential challenges. Some common issues and limitations associated with compartmental models include:

- Simplifying Assumptions: Compartmental models often rely on simplifying assumptions to describe complex physiological processes, which can lead to an oversimplification of the actual system and may only partially capture some of the dynamics and interactions.
- Limited Representation of Biological Variability: These models may need to adequately account for the individual variability in physiological parameters and responses among different patients or experimental subjects, potentially leading to discrepancies between model predictions and real-world observations.
- Complexity in Model Development: Creating a comprehensive compartmental model that accurately represents the intricate physiological processes in the body can be challenging and may require a significant amount of data and specialized expertise.
- Difficulty in Model Validation: Validating compartmental models can be complex, mainly when experimental data is limited or when the model involves numerous assumptions. Assessing the model's predictive accuracy and reliability in real-world scenarios can be challenging.
- Incorporating Time Variability: Some compartmental models may not effectively capture the dynamic changes that occur over time, especially in cases where physiological processes exhibit time-dependent behaviors.
- Assumptions of Instantaneous Mixing: Certain compartmental models assume instantaneous and uniform mixing within each compartment, which might not accurately reflect the true physiological processes, especially in cases where spatial or temporal heterogeneity exists.

Another frequent modeling strategy inside mechanistic models are agent–based models [An et al., 2009]. These computational models simulate the behavior and interactions of individual agents, such as cells, microorganisms, or individuals, within a complex biological system. Agent–based models are beneficial for studying the dynamics of disease spread, population health, and the effects of various interventions in healthcare. Compared to traditional mathematical models, these models provide a more granular and

detailed representation of the heterogeneity and interactions among agents within a system [Hadzic et al., 2009]. In the context of medicine, agent-based models are used in various areas, including:

- Epidemiology and Disease Spread: Agent–based models can simulate the spread of infectious diseases within a population by modeling interactions between individuals, incorporating factors such as contact patterns, transmission dynamics, and individual behaviors.

- Cancer Modeling: Agent–based models are used to simulate the growth and progression of tumors, the interactions between cancer cells and the immune system, and the effects of different treatment strategies on tumor development and response.

- Healthcare Delivery and Policy: Agent–based models can simulate healthcare systems to analyze the impact of various policies, resource allocation strategies, and interventions on healthcare outcomes, patient flow, and healthcare costs.

- Drug Development and Pharmacokinetics: Agent–based models can be used to simulate the pharmacokinetics and pharmacodynamics of drugs, allowing researchers to assess drug effectiveness, optimize dosing regimens, and predict potential side effects.

Agent-based models also come with specific challenges and limitations [Macal, 2020]. Some of the issues and considerations associated with the use of agent–based models include:

- Computational Complexity: Simulating large-scale systems with a high number of agents can be computationally demanding and time–consuming; this can limit the scalability of agent–based models for certain applications and may require high–performance computing resources.

- Model Validation and Calibration: Validating and calibrating agent-based models can be challenging due to the complex and stochastic nature of the simulations. It can be not easy to ensure that the model accurately represents the real–world system it is intended to simulate.

- Parameterization and Sensitivity Analysis: Agent–based models often involve numerous parameters and rules that govern agent behavior, and small changes in these parameters can lead to significantly different model outcomes. Conducting sensitivity analyses to assess the robustness of the model results is essential but can be complex.

- Data Requirements and Availability: Agent–based models may require substantial data to inform the development and calibration of the model. Obtaining detailed and accurate data on agent characteristics and interactions can be challenging, especially in the case of biological systems and complex social behaviors.

- Interpretability and Transparency: Agent–based models can be highly complex, making it challenging to interpret and understand the underlying dynamics of the simulated system. Ensuring transparency in model assumptions, rules, and outputs is crucial for facilitating stakeholders' trust and acceptance of the model.
- Model Complexity and Design: Designing an ABM that adequately captures the essential dynamics of a system while remaining computationally feasible can be a delicate balance. Simplifying the model too much can lead to losing critical details, while overly complex models may become challenging to understand and analyze. Addressing these challenges often requires a careful balance between model complexity, data availability, computational resources, and the specific goals of the simulation.

In general, the limitations of mechanistic models relate to high–quality data requirements because developing and calibrating mechanistic models often requires superior–quality data, which may only sometimes be readily available. Obtaining accurate and comprehensive data for all model parameters can be challenging, especially in cases where the biological system is complex or not well–characterized [Liang et al., 2022]. Also, mechanistic models can be highly complex, involving numerous interconnected components and processes. Managing the complexity of these models and ensuring that all relevant biological interactions are adequately represented can be challenging [Parry, 2020]. Connected with this issue is the extensive set of parameters needed to correctly estimate the process under exam and account for its uncertainty [Krivorotko et al., 2022]. Assessing model parameters accurately can be complicated, and these estimations are often associated with inherent uncertainty. It is essential to conduct sensitivity analyses and assess the robustness of model predictions to variations in parameter values. Additionally, the researcher's deep knowledge and expertise are essential to ensure correct parameter choice. Another aspect that could be relevant is the demand for computational resources. Running simulations using mechanistic models can be computationally intensive, especially for large-scale or highly detailed models. Adequate computational resources and efficient algorithms are required to handle the computational demands of these models. Also, the validation of mechanistic models can be complex, as it often involves comparing model predictions with experimental data collected under different conditions [Craver, 2006]. Ensuring that the model accurately represents the real-world system it is intended to simulate is crucial for establishing the model's reliability and predictive capability. Regarding model interpretability, communicating the intricacies of mechanistic models to non-experts can be challenging, as these models often involve complex mathematical equations and biological concepts. Ensuring the model outputs are presented clearly and understandably facilitates effective communication and decision-making [Transtrum and Qiu, 2016].

On the contrary, phenomenological models remain valuable for data–driven predictions and guiding decision–making, especially when the underlying mechanisms of the system are not well understood or when detailed data are limited [Waters et al., 2021]. Some of the common issues associated with the use of phenomenological models include the limited insight offered into biological mechanisms. Phenomenological models often

focus on correlations and patterns without fully elucidating the underlying biological processes. Another factor is the model generalizability, which should be ensured through a robust data analysis to extend the outcomes toward different contexts or populations. Sometimes, it might be difficult to distinguish between causality and correlation in phenomenological models, which may lead to misinterpretations of relationships between variables and phenomena [Attanayake et al., 2020]. Correlation is a statistical measure describing the extent to which two variables change together. In other words, it quantifies the degree to which a change in one variable is associated with a change in another variable. Correlation does not imply causation. Conversely, causality refers to a cause-and-effect relationship between two variables, where a change in one variable directly influences a change in the other. Establishing causality is more complex than establishing correlation and typically requires experimental design or sophisticated statistical methods to control for confounding variables. A final remark about phenomenological models is about data availability: they heavily rely on available observational data, and their accuracy and reliability are contingent upon the quality and quantity of the data used for model development and validation [White and Marshall, 2019].

Machine learning algorithms can be used to build phenomenological models, which are often data–driven and rely on statistical relationships between input features and output variables. Phenomenological models in machine learning can be helpful when the primary objective is to make accurate predictions or classifications based on available data without requiring an explicit understanding of the underlying causal relationships. These models are particularly effective when the underlying mechanisms of the system are complex or not fully understood.

1.3 Data mining and machine learning

Data mining is discovering patterns and knowledge from raw data, and machine learning provides the computational tools to automate and enhance this process. Machine learning plays a significant role in data mining by providing techniques and algorithms that extract valuable patterns, knowledge, and insights from large datasets. Machine learning algorithms can handle large volumes of data and automate extracting valuable information. Scalability is essential in data mining, where the datasets can be massive and complex, and machine learning algorithms excel at recognizing patterns in data [Hirschman et al., 2002]. They can identify complex relationships and trends within large datasets, a crucial aspect of data mining. Machine learning techniques such as decision trees, support vector machines, and neural networks are often used for classification and prediction tasks in data mining [Weber et al., 2009]. These algorithms can categorize data into different classes and predict future trends based on historical data. Also, clustering algorithms in machine learning, like k-means or hierarchical clustering, are utilized in data mining to group similar data points together; clustering helps discover inherent structures within the data, revealing natural divisions or patterns. Association rule mining is a technique used in data mining to discover relevant relationships, patterns, or associations among a set of variables or items in datasets. It aims to uncover

hidden patterns that indicate relationships between variables, helping to identify rules that describe the co-occurrence of certain events or items in a dataset [Tzanis, 2014]. For example, it could identify associations between symptoms and diseases in healthcare. An association rule is a statement that describes the relationship between two sets of items. It is typically written in the form "if A, then B", where A is the antecedent (the condition) and B is the consequent (the result). The rule is considered attractive if it satisfies predefined support and confidence thresholds. Another concept in association rules and data modeling is the *lift*, which is a measure that indicates how much more likely itemset B is to occur when itemset A is present compared to when itemset B is considered independent of A. Lift values greater than 1 suggest a positive association, while values less than 1 suggest a negative or unlikely association. For data mining, machine learning models, mainly unsupervised learning algorithms, can be applied to identify anomalies or outliers in datasets; in data mining, it is helpful in detecting unusual patterns that may indicate aberrant or other interesting phenomena. Another technique widely adopted is regression analysis, which supports modeling the relationship between variables. It is advantageous in data mining when trying to understand the strength and nature of relationships between different factors. With machine learning, selecting relevant features or variables from large datasets is possible, improving the efficiency and effectiveness of data mining processes: feature extraction methods help transform raw data into a more suitable representation for analysis. In summary, by integrating machine learning into data mining processes, organizations can make more informed decisions based on the patterns and insights uncovered.

2 Aim of the thesis

By leveraging machine learning capabilities, precision medicine can transform healthcare, making it more targeted, effective, and patient–centric, ultimately leading to improved health outcomes and a better quality of life for patients. In the present thesis, various works have been undertaken to address the challenges of precision medicine and illustrate the potential applications of machine learning within the field. By delving into these specific applications, the aim was to provide a comprehensive understanding of how machine learning is shaping and advancing the field of precision medicine and the potential challenges and opportunities that lie ahead. The machine learning algorithms and software tools developed during the Ph.D. tackled some specific problems related to precision medicine in different fields: biomarkers analysis in bioinformatics, virtual screening in chemoinformatics, dedicated volumetric software, and risk stratification in clinical medicine, a free-to-use programming library for biostatistics, biomaterials production tracking for regenerative medicine, and proteomics analysis in biological samples. In all these works, machine learning has been demonstrated to provide supportive and assistive technologies to advance the classic paradigm of evidence-based medicine. The common thread of all the works has been applying machine learning methods to solve various issues connected with the analysis. In chemoinformatics, a novel methodology has been proposed to the scientific community, exploiting the efficient power management provided by spiking neural networks to solve computationally intensive tasks. In bioinformatics, algorithms using biomarker expression values discretization proved adequate computer methods to predict tumor stage and outcome in bladder cancer patients. Machine learning has been studied on clinical datasets to stratify patients' risk of developing lymphedema, employing patients' factors and variables. For biomaterials production tracking, a sequence of operations has been conceived to merge heterogeneous data sources and categorize the stages of octacalcium phosphate synthesis. Working on a proteomic dataset, machine learning offered a solution to highlight proteins with aberrant expression.

Another aspect of the works presented throughout the thesis is that all machine learning models were developed on commodity hardware. This term refers to implementing machine learning algorithms and models on standard, readily available hardware, such as traditional computers, laptops, or servers, without requiring specialized or high-end computing infrastructure. This approach has gained significant popularity due to the widespread availability and affordability of commodity hardware, which makes it accessible to a broader audience, including researchers, developers, and small businesses. Leveraging commodity hardware for machine learning applications can significantly reduce costs, eliminating the need for expensive specialized hardware or cloud-based computing resources. Additionally, desktop computers offer a high degree of flexibility, allowing users

to configure and customize their machine–learning environment based on their specific requirements and computational needs. While budget hardware may have limitations in terms of processing power and memory, it can still support scalable machine-learning tasks, especially when combined with efficient algorithms and optimization techniques. Despite its advantages, machine learning on commodity hardware may face challenges related to limited processing capabilities, memory constraints, and longer processing times for complex tasks compared to specialized hardware or cloud-based solutions. One key aspect of all models presented is efficient data preprocessing and model optimization strategies to help mitigate the challenges associated with limited computational resources [Jordan and Mitchell, 2015].

2.1 Chemoinformatics

Chemoinformatics is a field that involves the application of informatics techniques to solve problems in chemistry, particularly in the analysis and interpretation of chemical data. In computational chemistry, the Ph.D. workflow has primarily focused on the development and implementation of advanced methods for the prediction of toxicity in chemical compounds from their structural configuration only. The research has made significant contributions to the field by introducing novel ML techniques, such as spiking neural networks, to verify the efficiency on toxicity and activity prediction. The advantages of applying spiking neural networks in chemoinformatics include efficient and biologically inspired processing. Additionally, improved versions of the algorithms have been successfully tested on bioaccumulation and P450 enzyme bioactivity (Chapter 3). The investigations overcame several challenges, the most important one being the representation of the chemical structures in a way that captures relevant information for machine learning models. Moreover, models in chemistry are computationally intensive, requiring substantial resources for training and inference. Additionally, datasets in chemoinformatics often suffer from class imbalance, where certain classes or outcomes are underrepresented.

2.2 Clinical precision medicine and risk stratification

Machine learning plays a crucial role in enabling the implementation of clinical precision medicine by leveraging computational algorithms to analyze complex patient data and provide personalized treatment strategies. In this field, a collaboration with medical doctors led to software production and digital methodologies to assess upper limb volume changes due to pathological conditions such as lymphedema. Upper limb volumetry, which involves measuring the volume of the upper extremities (arms), is a valuable assessment in clinical settings, particularly for conditions such as lymphedema, post-surgical evaluation, and monitoring of certain diseases. However, several challenges are associated with conducting upper limb volumetry in clinical settings or hospitals. Different healthcare professionals may use varying techniques for upper limb volumetry, leading to measurement inconsistencies. Achieving consistent and reproducible measurements

requires standardized patient positioning. For example, some clinics may only have the equipment needed for water displacement calculation, and this methodology can take time and effort. This may lead to practical challenges in incorporating volumetric assessments into routine patient care in a busy clinical setting. The volumetric software developed employed computerized reconstructions of the patient's upper limb to calculate the volume and facilitate this evaluation over time (Chapter 4).

Additionally, ML algorithms can analyze patient data, including genetic information, medical history, and lifestyle factors, to predict the risk of developing specific diseases or adverse health outcomes. In this topic, a stratification algorithm has been published to evaluate the risk of lymphedema by analyzing several patient-specific factors in breast cancer. In general, clinical datasets may suffer from missing or incomplete information, errors, and inconsistencies, which can impact the performance of machine learning models. Obtaining large, well-labeled datasets for training robust machine learning models can be challenging in the clinical domain. Additionally, clinical practitioners often require models to be interpretable and explainable to trust and understand the reasoning behind predictions. These challenges were tackled, and an innovative solution was proposed for patient risk stratification (Chapter 4).

2.3 Bioinformatics' biomarkers analysis

Genetic biomarkers play a critical role in various healthcare and medical research aspects, offering valuable insights into an individual's health, disease susceptibility, and treatment response. Genetic biomarkers can help assess an individual's susceptibility to certain diseases; by identifying genetic variations associated with particular health conditions, healthcare professionals can evaluate a person's risk and implement preventive measures or personalized screening protocols. Moreover, by analyzing an individual's genetic profile, healthcare providers can identify potential disease markers before the onset of symptoms, allowing for timely interventions and more effective treatment strategies. Also, genetic biomarkers can guide the selection of appropriate treatments for individual patients: they can help predict a patient's response to specific medications or therapies, enabling healthcare professionals to tailor treatment plans based on genetic information, ultimately improving treatment efficacy and reducing the risk of adverse reactions.

Genetic biomarkers contribute to population-level studies and research on the prevalence and distribution of certain diseases. They aid in understanding the genetic basis of diseases within specific populations, guiding public health initiatives and preventive healthcare strategies. By analyzing specific genetic factors, healthcare providers can predict the likelihood of disease advancement, allowing for better disease management and more informed decision–making regarding treatment options and patient care. In the present work, bladder cancer biomarkers have been analyzed to predict survival and identify cancer staging (Chapter 5). Biological systems are complex and exhibit inherent variability: genetic expression patterns may vary across individuals, tissues, and time. Furthermore, employing interpretable models is crucial for understanding the biological relevance of the identified biomarkers. The proposed solutions were codified in analysis

pipelines that were able to predict patient status through prognostic maps and efficient data discretization preprocessing.

2.4 Biostatistics: equivalence analysis

Equivalence testing in biostatistics is a statistical method used to determine whether the effect of a treatment or intervention is within a pre-specified equivalence margin rather than focusing solely on whether there is a statistically significant difference. It is particularly relevant in clinical trials and studies where demonstrating the absence of a meaningful difference is as important as detecting a significant effect. Equivalence testing helps researchers assess whether the new treatment is not clinically worse than the standard treatment or whether two treatments are essentially similar within a predetermined margin of equivalence.

During the Ph.D., a software library was developed and released to the general audience to run specific statistical analyses for equivalence testing; these statistical functions were not previously available to the Python community, and the developed software filled this gap. Equivalence testing is essential in fields such as pharmaceutical research and clinical trials, where demonstrating the non-inferiority or bioequivalence of a new drug compared to an established treatment is crucial for regulatory approval and clinical practice. It allows researchers to assess whether the new treatment can be considered a viable alternative without compromising efficacy or safety compared to existing standards (Chapter 6).

2.5 Regenerative medicine: biomaterials production tracking

Machine learning is increasingly being applied in materials science and regenerative medicine to accelerate the discovery and development of innovative materials and therapies. Machine learning algorithms are being used to design and optimize biomaterials for tissue engineering and regenerative medicine applications. These algorithms can analyze the structure-property relationships of various biomaterials, facilitating the identification of optimal material compositions and characteristics that promote tissue regeneration and integration. During the development of the Ph.D., a specific work tested a novel algorithm to track the production phases of bone replacement biomaterials. By leveraging the capabilities of machine learning in materials science and regenerative medicine, researchers can expedite the development of advanced biomaterials (Chapter 7). Applying machine learning to biomaterials involves various challenges due to the complexity of materials science: biomaterials datasets, especially those with specific properties or applications, may be limited in size. Biomaterials often have complex compositions, and characterizing them may involve many features and multiple experimental techniques, resulting in multi-modal data. The solution found merged different sources of information to characterize biomaterials synthesis, including feature selection methods, dimensionality reduction techniques, and domain knowledge that can aid in managing this kind of data.

2.6 Proteomics: anomaly expression identification

In proteomics, machine learning techniques are increasingly being used for anomaly detection, which involves the identification of abnormal patterns or outliers in protein expression data. By leveraging machine learning algorithms, researchers can detect deviations from normal protein expression profiles, which may signify the presence of disease, the impact of environmental factors, or other irregular biological processes. Unsupervised machine learning algorithms, such as clustering and density-based methods, can identify unusual patterns in proteomic data without needing pre–labeled samples. These techniques help detect outliers or abnormal clusters corresponding to protein expression profiles associated with specific diseases or biological conditions. A study on a proteomics dataset has been carried out to demonstrate a novel analysis pipeline able to solve this task (Chapter 8). The nature of proteomic datasets involves highly variable and heterogeneous data, both within and across biological samples. The variability may arise from biological differences, sample preparation methods, and technical variations. Proteomic data reflect the intricate biological processes involving proteins, making it challenging to accurately capture the complexity and dynamics of cellular systems. The machine learning analysis sequence proposed on this topic can help researchers address variability and heterogeneity challenges and help evaluate biological complexity.

2.7 Data sources

The Ph.D. activities were funded through the European Union grant No. 860462 as part of the Horizon 2020 research and innovation program (i.e., "Precision Medicine for Musculoskeletal Regeneration, Prosthetics, and Active Aging", `https://premurosa.eu/`). The project involved six continental European institutions, supported by ten external partners. The main goal of this initiative was to train scientists to develop expertise and technologies addressing the various challenges associated with all aspects of musculoskeletal regeneration. The analysis carried out on Sections 7 and 8 involved data acquired during wet–lab experiments performed in the context of this European Union research project. Musculoskeletal regeneration technologies are crucial in precision medicine, aiming to provide personalized and targeted solutions for individuals with musculoskeletal disorders, injuries, or degenerative conditions [Li et al., 2021]. Notable technologies and approaches in musculoskeletal regeneration within the context of precision medicine involve genetic profiling as it can help identify optimal cell sources for regenerative therapies, such as autologous (patient's own) stem cells, ensuring compatibility and reducing the risk of rejection. Analysis through OMICS of stem cells, including mesenchymal and induced pluripotent stem cells, are being investigated for their regenerative potential in repairing damaged bone, cartilage, and muscle tissues [Lan et al., 2018]. Additionally, gene therapy may involve introducing therapeutic genes to enhance tissue regeneration, modulate inflammation, or promote the synthesis of growth factors crucial for musculoskeletal health. Another important aspect is tailoring biomaterials to match the mechanical and biological properties of the target tissue. Advanced biomaterials and scaffolds provide

a supportive cell growth and tissue regeneration environment. They can be designed to degrade over time as new tissue forms. Orthobiologics refers to medical treatments involving biological substances to promote the healing and regeneration of musculoskeletal tissues, such as bones, joints, muscles, ligaments, and tendons [Calcei and Rodeo, 2019]. These biological substances can be naturally occurring or derived from the patient's body, donors, or synthetic sources. Orthobiologics aims to enhance the body's natural healing processes and stimulate tissue repair. Understanding individual variations in growth factor responses and signaling pathways helps customize the selection and dosage of ortho-biologic agents. Orthobiologics are considered a minimally invasive and potentially safer alternative to traditional surgical interventions.

The data processing and analysis, algorithm development, and numerical experiments of Sections 3 and 5 was performed on public domain datasets. Public domain datasets are freely accessible to anyone, fostering inclusivity and providing equal opportunities for researchers and developers worldwide. They eliminate financial barriers, allowing individuals with limited resources to research and experiment in machine learning. The most important aspect is that public datasets provide a common ground for benchmarking and comparing different machine-learning algorithms and models, facilitating fair evaluations of novel approaches and techniques. Researchers can reproduce experiments and validate findings more quickly when using publicly available datasets, contributing to the transparency and reliability of research outcomes. Access to various datasets encourages creative exploration, enabling researchers to apply machine learning techniques to unconventional or interdisciplinary problems, often representing real-world scenarios, making it easier for researchers and practitioners to develop models that can be applied to practical problems. The datasets employed in the computational chemistry investigations performed in Chapter 3 were [Judson et al., 2010; Richard et al., 2020; Wu et al., 2018; Kuhn et al., 2016; Gayvert et al., 2016; Grisoni et al., 2016; Nembri et al., 2016], whereas in Chapter 5 the dataset was from [Zhang et al., 2020].

The data in Section 4 employed clinical data collected during the course of patient care and medical research in Italian hospitals. Clinical data often includes individualized and longitudinal information, and it has direct relevance on medical practice. While clinical data offers significant advantages, it's essential to note that there are challenges associated with its use, including data interoperability issues, potential biases, and variability in data quality across healthcare systems.

3 Chemoinformatics

Original contribution to knowledge

This chapter is based upon the article
Mauro Nascimben and Lia Rimondini. Molecular toxicity virtual screening applying a quantized computational SNN–based framework. *Molecules*, 28(3):1342, 2023
the book chapter
Mauro Nascimben, Silvia Spriano, Lia Rimondini, and Manolo Venturin. Molecular fingerprint based and machine learning driven QSAR for bioconcentration pathways determination. In Gabriella Bretti, Roberto Natalini, Pasquale Palumbo, and Luigi Preziosi, editors, *Mathematical Models and Computer Simulations for Biomedical Applications*, pages 193–215, Cham, 2023c. Springer Nature Switzerland. ISBN 978-3-031-35715-2
and the conference presentations
Mauro Nascimben. Molecular fingerprint based and machine learning driven QSAR for bioconcentration pathways determination. Rome, Italy, Sept 2021b. National Research Council of Italy, Virtual Workshop of Mathematical Modelling and Control for Healthcare and Biomedical Systems
Mauro Nascimben. Quantized computational QSAR framework for molecular toxicity virtual screening. Rome, Italy, July 2022. Sapienza University of Rome, 3rd Molecules Medicinal Chemistry Symposium - Shaping Medicinal Chemistry for the New Decade
Mauro Nascimben. Virtual screening by spiking neural networks: a case study on cytochrome P450. Padua, Italy, Sept 2023a. University of Padua, 18th Conference on Computational Intelligence Methods for Bioinformatics and Biostatistics
Mauro Nascimben. Low–power or resource–constrained environments for virtual screening and quantitative structure-activity relationship analysis for in silico precision medicine. Rome, Italy, Sept 2023c. Italian Association for Industrial Research, NanoInnovation Conference and Exhibition

Chemoinformatics is a specialized field that applies various informatics methods to solve chemical problems. It is a critical component of drug discovery, materials science, and other areas of chemistry. The key aspects commonly employed in chemoinformatics analysis include chemical structure representation in alphanumerical codes, chemi-

cal databases gathering chemical structures, chemical similarity and diversity analysis techniques, quantitative structure-activity relationship (QSAR), molecular docking, and specialized chemical informatics software for chemical data processing, analysis, and visualization. These methods are often integrated into a holistic workflow to address specific challenges in drug discovery, toxicity prediction, materials design, and other areas of chemistry. For instance, in chemical structure representation, SMILES and InChI (i.e., International Chemical Identifier) enable easy storage and exchange of chemical information, while chemical databases of chemical structures like PubChem, ChEMBL, and ChemSpider facilitate efficient retrieval and analysis of chemical data. For molecular docking and virtual screening, the preferred orientation of molecules can be predicted, and large compound libraries can be screened for potential drug candidates. Molecular docking predicts the preferred orientation of one molecule to a second when bound together to form a stable complex. Chemical similarity and diversity analysis could take advantage of representing molecules as binary fingerprints to quantify structural similarity, querying the molecules in a database. On the other hand, diversity analysis involves selecting diverse compounds from a chemical library to ensure a broad representation of chemical space. Cheminformatics software commonly employed for chemoinformatics purposes are RDKit, Cheminformatics Toolkit, and Open Babel for chemical data processing, analysis, and visualization.

Machine learning in chemoinformatics, which is the application of computational methods to solve chemical problems, involves chemical data analysis, including molecular structures, properties, and activities. Machine learning plays a vital role in chemoinformatics, enabling the analysis and interpretation of complex chemical data for various applications in drug discovery, molecular modeling, and chemical property prediction. The availability of high–quality, well–curated chemical datasets is crucial for training accurate and reliable machine learning models in chemoinformatics. Indeed, access to comprehensive and diverse datasets allows effective feature engineering involving selecting and extracting informative molecular descriptors, fingerprints, or molecular representations that effectively capture the relevant chemical information needed for the specific modeling task [Idakwo et al., 2018]. Thoughtful feature engineering can significantly enhance the predictive performance of ML models in chemoinformatics.

Machine learning algorithms can be applied to predict the properties or activities of chemical compounds, with clustering and classification techniques used to group compounds based on similarities or classify them into predefined categories. Also feature selection techniques help identifying relevant molecular descriptors for use in predictive models.

The primary fields of application of machine learning in chemoinformatics are:

1. Quantitative Structure-Activity Relationship (QSAR) Modeling: QSAR models use machine learning algorithms to predict chemical compounds' biological activity or properties based on their structure. These models can be used in drug discovery to prioritize compounds for further experimental testing. Usually, quantitative descriptors (molecular properties) are used as features, and statistical methods (linear regression, support vector machines, etc.) are employed to build predictive

models.

2. Virtual Screening and Compound Prioritization: Machine learning can be used to screen large chemical libraries to identify potential drug candidates or compounds with desired properties, reducing the need for costly and time-consuming experimental research. ML algorithms can be used to prioritize compounds based on their likelihood of being active against a specific target.

3. Molecular Property Prediction: Machine learning algorithms can be applied to predict various molecular properties, such as solubility, lipophilicity, and bioavailability, which are crucial in drug design and development based on molecular descriptors (physicochemical, topological, electronic, etc.). Regression models or classification algorithms are trained using descriptor values as input features.

4. De Novo Molecule Design: Machine learning can aid in generating novel chemical structures with desired properties by learning from existing data and generating new molecules that are likely to exhibit specific characteristics. Deep learning techniques, such as artificial neural networks, can be applied to learn complex patterns from chemical data for molecular generation. In particular, deep generative models like variational autoencoders (VAEs) or generative adversarial networks (GANs) can be used to create new chemical entities.

5. Chemical Reaction Prediction: Machine learning models can be used to predict the outcomes of chemical reactions, helping chemists identify the most efficient synthetic routes and optimize reaction conditions. This task is often achieved by representing each molecule (reactants and products) using molecular descriptors (numerical representations of chemical properties) and then developing a representation for the entire reaction, considering the relationships and interactions between reactants and products. Recurrent neural networks are suitable for sequence-based data, making them helpful in modeling the sequential nature of reactions. Otherwise, graph neural networks are designed for graph-structured data, making them effective for capturing the relationships between atoms and molecules in reactions, or transformers, known for their success in natural language processing, can also be applied to reactions by treating them as sequences.

Spiking neural networks have the potential to offer unique advantages in the field of chemoinformatics due to their ability to model complex temporal dynamics and process spatiotemporal information, which are essential for understanding molecular interactions and chemical processes [Xiaoxue et al., 2023; Tavanaei et al., 2019]. Indeed, SNNs could capture and model the temporal dynamics of molecular interactions and reactions. This capability allows for the analysis of complex chemical processes, such as protein-ligand binding kinetics, enzymatic reactions, and molecular signaling pathways, providing insights into the time-dependent behavior of chemical systems. Moreover, SNNs process information in an event-driven manner, making them efficient in handling sparse and asynchronous data, which is common in chemoinformatics; it enables SNNs to simulate molecular events, such as chemical reactions, diffusion processes, and signaling cascades,

more accurately than traditional neural networks. SNNs can be used for pattern recognition and feature extraction in molecular structures, aiding in identifying molecular fingerprints, functional groups, and structural motifs essential for characterizing chemical compounds and predicting their properties. Another aspect is that SNNs enable the simulation of neural networks and chemical systems, allowing researchers to explore the analogies between neural processing and chemical reactions. This interdisciplinary approach facilitates understanding complex biochemical processes and developing innovative computational models for drug discovery and design. For bioactivity and toxicity Prediction, SNNs can be utilized to predict the bioactivity and toxicity of chemical compounds by modeling their interactions with biological targets and cellular systems. By integrating molecular dynamics simulations and SNN-based predictive models, researchers can more accurately assess the pharmacological and toxicological properties of potential drug candidates. In this sector, SNNs can aid in optimizing drug design by simulating the interactions between drugs and their target receptors, predicting drug binding affinities, and facilitating the virtual screening of large chemical libraries to identify novel drug candidates with improved efficacy and specificity. Like other ML techniques, SNNs can integrate multimodal data sources, including chemical, biological, and structural information, to provide a comprehensive and dynamic view of molecular systems. This integrative approach supports the analysis of complex relationships between chemical structures, biological activities, and physiological responses, enhancing the understanding of drug action and toxicity mechanisms. By leveraging the capabilities of SNNs and considering all these aspects, researchers in chemoinformatics can advance their understanding of molecular interactions, chemical processes, and drug-target interactions, leading to more accurate predictions and insights for drug discovery, toxicity assessment, and personalized medicine applications.

The leaky integrate-and-fire (LIF, [Rast et al., 2010; Tal and Schwartz, 1997]) neuron model is a mathematical model used in computational neuroscience to describe the behavior of individual neurons in spiking neural networks. It is a popular and widely used model due to its simplicity and computational efficiency, making it well-suited for large-scale simulations of neural systems. The LIF neuron model is based on integrating incoming signals and generating an output spike when a certain threshold is reached. The neuron is modeled as a "leaky integrator" of its input current $I(t)$.

$$\tau_m \frac{dv}{dt} = -v(t) + RI(t)$$

with $v(t)$ representing the membrane potential at time t, τ_m the time constant and R the resistance of the membrane neuron respectively. This equation could be interpreted as a resistor–capacitor circuit where the leakage term is due to the resistor and the integration of $I(t)$ is obtained from the capacitor placed in parallel to the resistor [Orhan, 2012; Nascimben et al., 2023c]. The firing or spiking events in the LIF model are modeled when the membrane potential $v(t)$ reaches a certain threshold: in that occasion, it is instantaneously reset to a lower value called reset potential, and the process described by the above equation re-starts with the initial value equal to the reset potential. This behavior is depicted in Figure 3.1 from [Nascimben and Rimondini, 2023]. It could

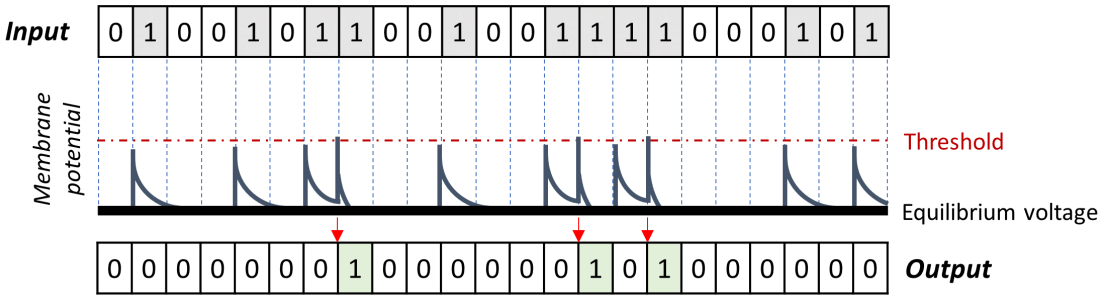

Figure 3.1: The spiking and resetting mechanisms in the LIF (from [Nascimben and Rimondini, 2023]).

be possible to add an absolute refractory period to improve the biological plausibility immediately after $v(t)$ reached the voltage threshold.

In [Nascimben et al., 2023c; Nascimben and Rimondini, 2023] and related works [Nascimben, 2021b, 2022, 2023a,c], we exploited binary molecular fingerprints as input sequences for the SNN. The choice of fingerprints as input data for the SNN is derived from the observation, according to which binary vectors are the natural input of SNNs. Indeed, SNNs receive incoming information encoded as bit sequences that simulate neuronal spike trains. Molecular fingerprints (MF) are binary (or numerical) representations of chemical compounds that capture their structural and physicochemical characteristics [Wigh et al., 2022]. These fingerprints are commonly used in chemoinformatics, computational chemistry, and drug discovery for various purposes, including compound classification, similarity analysis, and the prediction of biological activities. Molecular fingerprints represent the structural features of a chemical compound, such as its atom types, bond types, and connectivity. Each feature is typically assigned a binary (1 or 0) value or a numerical value. In a binary fingerprint, each feature is represented as a binary digit (1 or 0), indicating the presence or absence of a specific structural element or substructure in the compound. Numerical fingerprints assign numeric values to each feature, representing the frequency or occurrence of the structural elements in the compound. In chemoinformatics, molecular fingerprints are often used to calculate the similarity or dissimilarity between chemical compounds. Similarity measures, such as the Tanimoto coefficient or the Jaccard index, are applied to binary fingerprints, while numerical fingerprints can be used with various distance or similarity metrics. Applications of MF include virtual screening, used to compare potential drug candidates with known active compounds to identify molecules with similar structural features and potential biological activity, or QSAR to build predictive models that relate chemical structure to biological activity or other properties. Additionally, MF analysis helps assess the diversity of chemical libraries to ensure a broad range of compounds for drug discovery, or they are employed to find compounds containing specific chemical substructures (aka Substructure and Fragment Searching). There are various types of molecular fingerprints, including

Daylight, MACCS (Molecular ACCess System), ECFP (Extended Connectivity Fingerprint), and many others. These fingerprints vary regarding the structural information they capture and the algorithms used to generate them: the length of molecular fingerprints is determined by the number of features or substructures that are being encoded for each compound. Indeed, shorter fingerprints may capture basic structural information, such as the presence or absence of specific substructures or molecular fragments, while longer fingerprints can provide more detailed and comprehensive representations of the molecular structure, including information about bond types, atom types, and topological features.

All the experiments followed the workflow exemplified in the Figure 3.2. Regarding the ML assessment, repeated nested cross-validation was selected as a medium to evaluate both hyperparameters and model performance. Nested cross-validation is a technique used in the appraisal of machine learning models, particularly for assessing the performance and generalization ability of a model on a limited dataset. It is an extension of the standard k-fold cross-validation technique and is commonly employed when there is a need to perform both model selection and model evaluation simultaneously. The primary purpose of nested cross-validation is to provide a more accurate estimate of the model's performance by addressing the issue of overfitting during model selection. The dataset is divided into multiple folds, as in k-fold cross-validation. The outer loop of nested cross-validation splits the data into training and testing sets. Each iteration of the outer loop involves training the model on a subset of the data (training set) and evaluating its performance on the remaining data (testing set). Within each iteration of the outer loop, a separate inner loop is used for model selection. This involves further dividing the training set into multiple folds. The inner loop is used to select the best hyperparameters or features for the model. Various combinations of hyperparameters or features are tested, and the best combination is selected based on the performance metric, such as accuracy or ROC–AUC. The model's performance is then assessed using the testing set in the outer loop; it provides an unbiased estimate of the model's performance on unseen data, as the model has not been directly exposed to the testing data during the training phase. Nested cross-validation helps to address the issue of overfitting that can occur during standard cross–validation, as it separates the process of model selection from model evaluation. Performing model selection within each iteration of the outer cross-validation loop it ensures that the selected model is not biased towards the specific dataset splits used in the cross-validation process. This technique is beneficial when working with limited datasets or when the model has multiple hyperparameters that must be tuned. It helps provide a more reliable estimate of the model's performance and generalization ability, enabling more robust model selection and evaluation in machine learning tasks.

3.1 Predictive toxicity

Predictive toxicity refers to predicting the potential toxicity of a chemical, drug, or substance before it is extensively tested in animals or humans. It involves using various

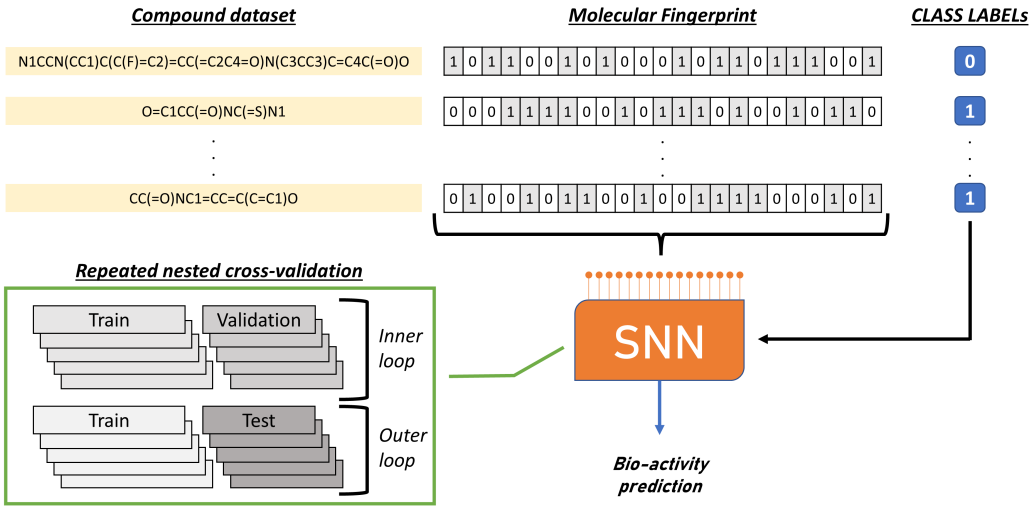

Figure 3.2: Overview of the numerical experiments involving SNNs (modified from [Nascimben and Rimondini, 2023]).

computational models, in–vitro tests, and in–silico simulations to assess the potential adverse effects of a substance on human health or the environment. This approach is crucial in toxicology and drug development, as it allows researchers and regulators to identify potentially harmful substances early in the development process, thereby reducing the need for extensive animal testing and minimizing risks to human health and the environment. Predictive toxicity is essential for several reasons: early identification of toxic substances, reduction of animal testing, regulatory compliance and drug development, and cost reduction. By predicting the toxicity of a compound at an early stage, researchers can avoid investing time and resources in the development of potentially harmful substances, thereby reducing the risk of adverse effects on human health and the environment. Predictive toxicity methods can help reduce the need for animal testing, which is often expensive, time-consuming, and ethically controversial. Using computational models and in–vitro tests, researchers can obtain valuable toxicity data without relying solely on animal experiments. Regulatory agencies often require comprehensive toxicity data before approving new chemicals or drugs. Predictive toxicity helps generate the necessary data to meet regulatory requirements, ensuring that only safe and effective substances are approved for use in various applications. Additionally, predictive toxicity methods can significantly reduce the time and costs associated with traditional toxicity testing. By employing computational models and in–vitro assays, researchers can quickly screen large numbers of compounds for potential toxicity, allowing them to focus their resources on developing safer and more effective substances. Predicting com-

pounds' toxicity is critical in the pharmaceutical industry to identify potential adverse effects of drug candidates early in the development process; this enables researchers to make informed decisions and optimize the safety profile of new drugs before they enter clinical trials, ultimately improving the success rate of drug development and reducing the risk of unexpected toxicities in human subjects.

In [Nascimben and Rimondini, 2023], we presented an innovative SNN–based framework as a virtual screening tool aiming at demonstrating SNN for toxicity prediction using datasets of compounds converted into MAACS fingerprints. Using spiking neural networks in quantitative structure-activity analysis represents a groundbreaking advancement in chemoinformatics and computational toxicology. Spiking neural networks, inspired by the functioning of the human brain, offer a promising alternative to traditional machine learning algorithms, mainly due to their energy efficiency and applicability to specialized hardware. The successful application of spiking neural networks in the evaluation of public-domain databases of compounds for toxicity prediction underscores their potential in addressing complex tasks that require significant computational resources. By achieving accuracies comparable to those of established high-quality frameworks, these networks demonstrate their capacity to handle challenging chemoinformatics tasks effectively. Furthermore, the analysis of hyperparameters and the testing of spiking neural networks on molecular fingerprints of varying lengths highlight the versatility and adaptability of these networks in accommodating different data representations and model configurations. This adaptability is crucial in handling the diverse and complex molecular structures often encountered in chemoinformatics. The potential of spiking neural networks to offer alternatives to conventional software and hardware in computationally demanding tasks, such as toxicity prediction, can pave the way for significant advancements and innovations in the field. This development not only opens up new avenues for research but also holds the promise of improving the efficiency and accuracy of predictive models in chemoinformatics, ultimately contributing to the identification and development of safer and more effective chemical compounds.

Our methodology has been tested on five public–domain toxicological datasets, each evaluated in separate numerical experiments through specific SNNs. All SNNs had in common the neuronal model, the leaky integrate-and-fire and the architecture shown in Figure 3.3.

The compounds in SMILES format (simplified molecular-input line-entry system) converted to MF were obtained from the following benchmark datasets:

- TOXCAST [Judson et al., 2010], containing results of in vitro toxicological experiments. In particular, the outcomes for "Tox21–TR–LUC–GH3–Antagonist" were considered due to the best sample ratio between labels;
- Tox21 [Richard et al., 2020], predicting the toxicity on biological targets, including nuclear receptors or stress response pathways. Activities selected were "SR–ATAD5", "NR–EL–LBD", and "NR–AR" for the relatively low number of missing entries compared to the others inside the dataset;
- BBBP [Wu et al., 2018] assessing drug's blood–brain barrier penetration;

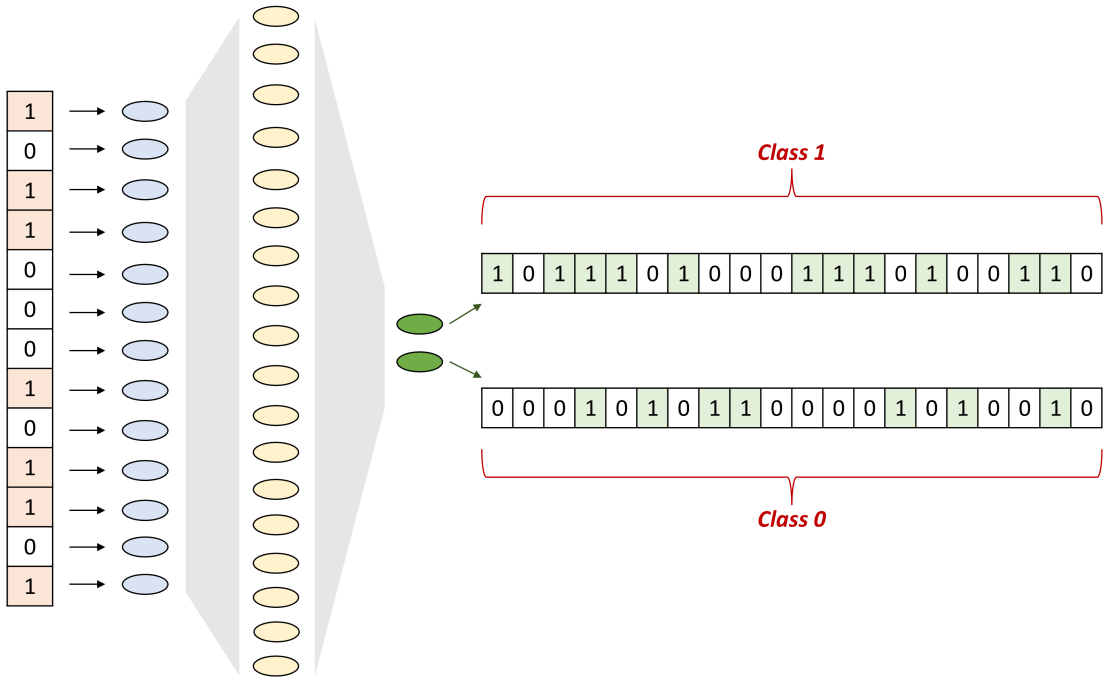

Figure 3.3: Overview of the SNN employed for predictive toxicity in [Nascimben and Rimondini, 2023]

Dataset	AUC mean	AUC st. dev.
BBBP	0.946	0.008
Clintox	0.974	0.01
ISD	0.795	0.008
NR–AR	0.988	0.002
NR–ER–LBD	0.986	0.003
NSD	0.97	0.008
SR–ATAD5	0.991	0.002
TOXCAST	0.912	0.007

Table 3.1: Summary of the best SNNs' models in [Nascimben and Rimondini, 2023].

- SIDER [Kuhn et al., 2016], employed for predicting drug's side effects on the immune and nervous systems;
- Clintox [Gayvert et al., 2016], containing drugs that failed or passed clinical trials for toxicity.

In the datasets, the dissimilarity between compounds measured by the Tanimoto index ranged from 79.2% to 66.5%, ensuring a heterogeneous aggregation of molecules to be tested. This aspect is crucial for the generalizability of the model because using mixed samples can help to solve assorted problems; these SNN models trained on diversified molecules describing composite compounds can adapt appropriately to new, previously unseen data. Under this view, a proper balance of the input data was ensured by oversampling the minority class.

The successful application of spiking neural networks using structural information derived from molecular fingerprints represents a significant breakthrough in toxicity prediction and chemoinformatics. By leveraging the structural details encoded in molecular fingerprints, SNNs have demonstrated remarkable performance (Table 3.1), as highlighted by the meta–analysis in [Nascimben and Rimondini, 2023]. The consistent performance of SNNs compared to other high-quality methods previously utilized for toxicity prediction underscores the potential of SNNs in advancing the field. These findings suggest that SNNs can effectively harness the structural information embedded in molecular fingerprints to make accurate toxicity predictions, thereby facilitating the identification and evaluation of potentially harmful chemical compounds. By leveraging the power of SNNs, researchers can enhance the efficiency and accuracy of toxicity prediction, leading to improved decision-making processes in various domains, including drug development, environmental risk assessment, and chemical safety regulation. The promising results obtained from the application of SNNs underscore their potential to revolutionize the field of chemoinformatics and computational toxicology, offering a valuable tool for researchers and practitioners to assess the safety and potential risks associated with various chemical compounds. As SNNs continue to demonstrate their efficacy and reliability in toxicity prediction, they are poised to play a pivotal role in shaping the future of predictive toxicology and facilitating the development of safer and more sustainable chemicals

and drugs.

3.2 Bioaccumulation pathways prediction

Bioaccumulation pathway prediction involves using computational models to assess and predict the potential pathways through which chemical substances may accumulate in living organisms, particularly in environmental exposure. Understanding bioaccumulation pathways is crucial for evaluating the potential risks associated with the long-term exposure of organisms to various chemicals, including pollutants and environmental contaminants. By predicting bioaccumulation pathways, researchers and regulatory agencies can assess the potential risks of chemical substances to human health and the environment. This information is crucial for designing effective risk management strategies, setting regulatory guidelines, and promoting the development of environmentally sustainable practices [Nichols et al., 2009].

In [Nascimben et al., 2023c], the use of quantitative structure-activity relationship (QSAR) models in combination with machine learning techniques represents a powerful approach for evaluating the potential risks associated with chemical compounds, particularly in terms of their bio-activity and potential for accumulation in living organisms. This integrated approach allows for predicting the bioaccumulation pathways of candidate molecules, aiding in identifying potentially risky chemicals. The previously developed SNN framework for toxicity prediction has been extended testing the a more complex SNN based on synaptic neurons that include the synaptic current modulation to simulate the information flow between pre– and post–synaptic neurons. The analysis conducted in this study compared various machine learning algorithms, including extreme gradient boosting, support vector machines, neural networks, and, notably, spiking neural networks. While the former algorithms have been previously employed in similar studies, applying spiking neural networks with molecular fingerprints as direct inputs is a novel and potentially pioneering development in computational toxicology. The analysis on the dataset from [Grisoni et al., 2016] featuring a three-class predictive problem, revealed that the support vector machines outperformed other models, demonstrating high balanced accuracies of 86.9% and 87.85% in forecasting the bioaccumulation pathways. Furthermore, the spiking neural network architectures also achieved satisfactory results, showcasing correctness levels of 83.77% (employing LIF) and 81.96% (SNN with synaptic neurons). These findings highlight the potential of advanced machine learning techniques, including spiking neural networks, in accurately predicting the bioaccumulation pathways of chemical compounds. The incorporation of molecular fingerprints as direct inputs for spiking neural networks opens up new possibilities for leveraging biologically inspired neural networks in the field of computational toxicology. The results obtained from this analysis have significant implications for the identification and assessment of potentially risky chemicals, thereby contributing to the advancement of chemical safety evaluation and environmental risk assessment.

3.3 P450 enzyme bioactivity prediction

The bioactivity of P450 enzymes, also known as cytochrome P450 enzymes, is essential in drug metabolism, toxicology, and various biochemical processes in living organisms [Ioannides and V Lewis, 2004; Reilly and Yost, 2006]. Some key reasons why P450 enzyme bioactivity is essential include:

- Drug Metabolism: P450 enzymes play a crucial role in metabolizing a wide range of drugs and xenobiotics in the body. They are involved in the oxidation and biotransformation of many foreign compounds, including medications, environmental toxins, and industrial chemicals, making them more water-soluble and facilitating their excretion from the body.

- Toxicology and Detoxification: P450 enzymes are essential in detoxifying various environmental pollutants and harmful substances. They catalyze the conversion of lipophilic toxins into more hydrophilic forms that the body can easily eliminate, thereby reducing the potential toxic effects of these compounds.

- Endogenous Metabolism: P450 enzymes metabolize endogenous compounds like steroids, fatty acids, and cholesterol. They participate in the synthesizing and degradation of various biomolecules, including hormones, signaling molecules, and lipid derivatives, essential for maintaining normal physiological functions.

- Pharmacokinetics and Drug Interactions: Understanding the bioactivity of P450 enzymes is critical for predicting drug interactions and potential adverse effects that may result from altered drug metabolism. Certain drugs can induce or inhibit specific P450 enzymes, leading to changes in the metabolism of co-administered medications and affecting their efficacy and toxicity.

- Personalized Medicine: Variations in the activity and expression of P450 enzymes can influence individual responses to medications and may contribute to inter-individual variability in drug efficacy and adverse reactions; this has implications for developing personalized medicine approaches tailored to an individual's specific metabolic profile.

- Environmental and Occupational Health: P450 enzymes metabolize environmental pollutants, such as polycyclic aromatic hydrocarbons and pesticides. Understanding their bioactivity is essential for assessing the potential health risks associated with exposure to these contaminants and implementing appropriate regulatory measures to minimize environmental and occupational hazards.

The investigation presented in [Nascimben, 2023a], reported the application of SNN to predict the complex behavior of P450 in response to the interaction with several molecules using the dataset from [Nembri et al., 2016]. The application of spiking neural networks to molecular fingerprints for predicting the bioactivity of the P450 enzyme, specifically

MF length	Val. BA MEAN	Test BA MEAN	Test BA STD
256 bits	80.46%	79.23%	±0.80%
512 bits	78.24%	80.19%	±0.94%
1024 bits	78.43%	80.05%	±0.95%
2048 bits	82.81%	81.52%	±0.88%

Table 3.2: Summary of the SNN's outcomes when tested on longer MF from [Nascimben, 2023a].

its 3A4 and 2C9 isoforms, represents a significant advancement in quantitative structure-activity analysis. The numerical experiments conducted in the study focused on evaluating different network configurations and determining the optimal fingerprint length for the prediction task. The results obtained from the experiments were consistent with those of other machine learning techniques previously utilized in related studies, indicating the effectiveness of spiking neural networks in this domain. By demonstrating the capability of spiking neural networks in effectively predicting the bioactivity of specific enzyme isoforms, the current work contributes to the growing body of evidence supporting the utility of these networks in quantitative structure-activity analysis. The findings suggest that spiking neural networks hold promise for facilitating the identification and assessment of compounds with potential interactions with the P450 enzyme, thereby aiding in drug development and the discovery of new therapeutic agents. Furthermore, the potential integration of spiking neural networks with neuromorphic hardware offers a pathway toward developing energy–efficient and accelerated virtual screening methods. Leveraging neuromorphic hardware can significantly enhance the computational efficiency of spiking neural networks, enabling rapid and cost-effective analysis of large chemical datasets. This approach can potentially revolutionize virtual screening processes, facilitating the identification of promising drug candidates and accelerating the drug discovery pipeline. The findings from this study underscore the importance of exploring the application of spiking neural networks in the field of quantitative structure-activity analysis and highlight the potential for future advancements in both methodology and hardware integration, ultimately contributing to the development of more efficient and effective drug discovery processes.

The length of molecular fingerprints is not fixed and can vary depending on the specific fingerprint generation method and the desired level of molecular detail and complexity. The numerical experiments tested the applicability of SNN to MF of different bit length, showing that longer MF improve the predictive ability of SNN on P450 bioactivity (Table 3.2, BA means Balanced Accuracy). The best accuracies pertain to the MF with 2048 bits in length, and in general, all values are higher than those obtained with MAACS MF.

Moreover, it resulted that also in this investigation the LIF neuron achieved slightly better performance compared to the SNN employing the synaptic neuron model. The Figure 3.4 depicts a part of the hyperparameter space of one numerical experiment with LIF–based SNN; it shows the Torch optimizers and the associated learning rate over a

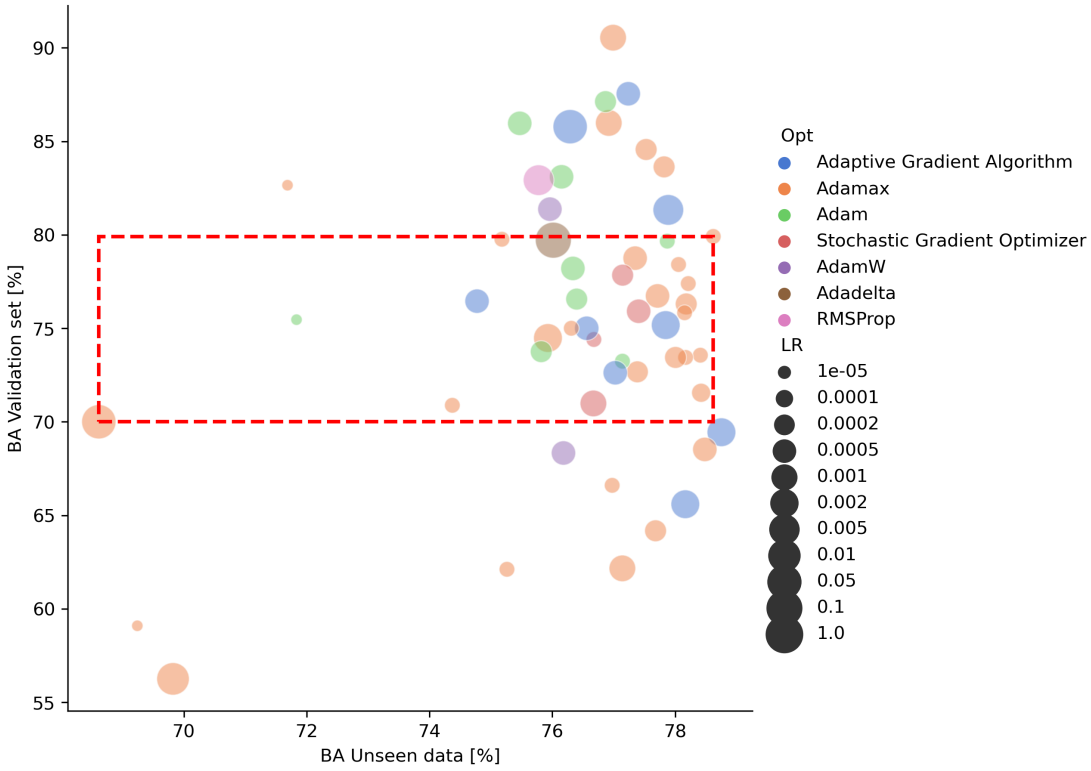

Figure 3.4: The figure illustrates the Torch optimizer and the associated learning rate for each validation and test set BA pair of LIF–based MAACS MF analysis. The red rectangle might reference the non–overfitting area where both validation and test BA lay below 2% (from [Nascimben, 2023a]).

scatterplot with test versus validation sets BA. The red rectangle shows the area where the difference between BAs was below 2%, meaning reduced overfitting. The top–right corner of the rectangle contains several values from the Adamax optimizer with small learning rates. This kind of hyperparameter space visual analysis highlights patterns and characterize the SNN behavior for this P450 dataset.

The prediction of P450 enzyme activity has garnered significant interest among researchers, primarily owing to the crucial role of this enzyme in the metabolism of xenobiotics, which are foreign chemical substances that enter the body. Understanding the interaction of small molecules with the P450 enzyme is vital in various fields, including pharmacology, toxicology, and environmental sciences, as it influences the efficacy and safety of drugs, as well as the potential toxicity of environmental chemicals. The P450 enzyme system is responsible for the metabolism and elimination of numerous exogenous compounds, including drugs, environmental pollutants, and other xenobiotics, making it a critical determinant of their bioavailability and potential effects on living organisms.

3.4 Final remarks

The works presented in this chapter introduced to the chemoinformatics community the possibility of using SNN as a virtual screening or QSAR tool employing chemical structures. Several aspects of the application of SNN on structural descriptors were explored and investigated. Future work might merge the functional information of the compounds with the structural data to produce more powerful insights. So far, SNNs were tested on traditional hardware; however, applying SNN to neuromorphic hardware could produce an energy–efficient platform for chemoinformatics as introduced in [Nascimben, 2023c]. Neuromorphic hardware represents a promising avenue for the future of computing, particularly in the realm of artificial intelligence and cognitive computing. These hardware systems are inspired by the architecture of the human brain and aim to emulate the parallel processing and energy efficiency observed in biological nervous systems. Neuromorphic hardware is particularly well-suited for implementing and running spiking neural networks, which are computational models based on the biological principles of how neurons communicate through discrete, asynchronous spikes. Several research initiatives and projects around the world are focused on advancing neuromorphic hardware: companies (i.e., Intel, IBM, Samsung among others) and academic institutions are investing in the development of neuromorphic chips and architectures to unlock their potential for a variety of applications. Researchers are actively working on testing systems able to make neuromorphic computing practical and widely applicable to offer alternatives to traditional computing structures, which are based on the von Neumann architecture [Chen et al., 2017; Jeong and Hwang, 2018].

4 Clinical precision medicine and risk stratification

Original contribution to knowledge

This chapter is based upon the articles
Mauro Nascimben, Lorenzo Lippi, Nicola Fusco, Marco Invernizzi, and Lia Rimondini. A software suite for limb volume analysis applicable in clinical settings: upper limb quantification. *Frontiers in Bioengineering and Biotechnology*, 10:863689, 2022a
Mauro Nascimben, Lorenzo Lippi, Nicola Fusco, Alessandro de Sire, Marco Invernizzi, and Lia Rimondini. Technical aspects and validation of custom digital algorithms for hand volumetry. *Technology and Health Care*, 31(5):1835–1854, 2023b
Mauro Nascimben, Lorenzo Lippi, Alessandro De Sire, Marco Invernizzi, and Lia Rimondini. Algorithm–based risk identification in patients with breast cancer–related lymphedema: A cross-sectional study. *Cancers*, 15(2):336, 2023a
Lorenzo Lippi, Alessio Turco, Stefano Moalli, Mauro Nascimben, Claudio Curci, Alessandro de Sire, Lia Rimondini, and Marco Invernizzi. Quantitative assessment of upper-limb volume: Implications for lymphedema rehabilitation? *Applied Sciences*, 13(17):9810, 2023b
and the conference presentation
Lorenzo Lippi, Mauro Nascimben, Alessandro de Sire, Arianna Folli, Nicola Fusco, Lia Rimondini, and Marco Invernizzi. A novel free–to–use software for upper limb volume quantification in breast cancer related lymphedema: implementing cutting–edge technology in the individualized therapeutic approaches of breast cancer survivors. *Cancer Research*, 83(Supplement 5):P5–08–18–P5–08–18, 03 2023a. ISSN 0008–5472. doi: 10.1158/1538-7445.SABCS22-P5-08-18. URL `https://doi.org/10.1158/1538-7445.SABCS22-P5-08-18`

Clinical precision medicine customizes medical care and treatment strategies based on a patient's specific characteristics and needs. It involves the integration of various data sources, including genetic information, molecular profiling, clinical data, and lifestyle factors, to inform the development of personalized healthcare interventions. Clinical precision medicine aims to improve patient outcomes, enhance treatment efficacy, and minimize adverse effects by tailoring medical decisions and therapies to each patient's unique profile. Risk stratification, conversely, is the process of categorizing patients into

different risk groups based on specific criteria or predictive factors. In healthcare, risk stratification is often used to identify individuals at a higher risk of developing certain diseases or experiencing adverse health outcomes. By stratifying patients based on risk profiles, healthcare providers can prioritize interventions, allocate resources more efficiently, and implement targeted preventive measures to reduce the likelihood of adverse health events. In clinical practice, the integration of precision medicine and risk stratification allows healthcare professionals to:

- Tailor Treatment Approaches: By understanding the unique genetic, molecular, and clinical profiles of individual patients, clinicians can select the most appropriate treatment strategies and medications that are more likely to be effective and well-tolerated based on the patient's specific characteristics.
- Predict Disease Progression: Risk stratification models can help predict the likelihood of disease progression and identify patients who may benefit from early intervention or aggressive treatment approaches, thereby improving disease management and patient outcomes.
- Optimize Preventive Measures: By identifying individuals at higher risk for certain diseases, healthcare providers can implement targeted preventive measures, such as lifestyle interventions, regular screenings, and vaccination programs, to mitigate the risk of disease development and promote overall wellness.
- Facilitate Patient-Centered Care: Integrating clinical precision medicine and risk stratification promotes patient-centered care by empowering patients to participate in their treatment decisions actively, understand their risk profiles, and make informed choices regarding their healthcare management and preventive measures.

By leveraging the principles of clinical precision medicine and risk stratification, healthcare providers can deliver more personalized, proactive, and effective care tailored to individual patients' specific needs and risk profiles.

Breast cancer is a type of cancer that begins in the cells of the breast. It can occur in men and women but is far more common in women. Breast cancer usually starts in the inner lining of milk ducts or the lobules that supply them with milk. The diagnostic path for breast cancer often begins in various ways, depending on the context and the individual's circumstances: screening, follow-up with clinical examination, or emergencies related to a new palpable nodule. Regardless of the initial pathway, the diagnostic process for breast cancer often involves a series of steps, including imaging studies (like mammography, ultrasound, or MRI), biopsy for tissue sampling, and pathological analysis to confirm the presence of cancer and characterize its type and characteristics. Once a diagnosis is confirmed, a treatment plan is developed based on the case's details [Ginsburg et al., 2020].

According to statistics from 2023, breast cancer affected 55900 individuals in Italy, leaving 834200 women living with the diagnosis. Notably, survivorship at the five-year mark after diagnosis was 88%, while the probability of living an extra four years after the first year post-diagnosis was 91% (from [Associazione italiana oncologia medica, 2023]).

These numbers provide insight into the state of breast cancer in Italy and highlight the importance of early detection and effective treatment.

Breast cancer treatment, significantly more intensive and prolonged regimens, can have significant impacts on employment and contribute to a greater healthcare burden for individuals undergoing treatment. Intensive treatment protocols, such as aggressive chemotherapy or extended surgeries, may require extended periods away from work; this can result in lost income and potentially impact job security. Prolonged treatment and recovery periods may interrupt an individual's career trajectory. It could affect opportunities for advancement, job promotions, or the ability to pursue specific career goals. Moreover, the side effects of cancer treatments, such as fatigue, nausea, pain, and cognitive issues, can affect an individual's ability to work during and after treatment. Recovery time varies from person to person, and some may need an extended period to regain full strength and functionality. Also, the costs associated with cancer treatment, including medical bills, medications, and additional expenses related to managing side effects, can lead to financial strain; it may further contribute to the stress associated with the disease [Greenup et al., 2019]. In public health, intensive breast cancer treatment often involves numerous medical appointments, tests, and follow-up care. Managing this healthcare burden can be challenging, requiring coordination of various treatment and follow–up care aspects. Regarding the emotional and psychological spheres, the emotional and psychological toll of a breast cancer diagnosis and treatment can affect an individual's ability to cope with work-related stressors. Emotional involvement may lead to changes in priorities, perspectives, and career choices.

Research has shown that surgical procedures, such as lymph node or axillary node dissection, radiation therapy, and systemic treatments like chemotherapy can lead to an upper limb impairment or considerable decline in upper extremity disability among women [Chrischilles et al., 2019]. Furthermore, women with lower income, health literacy, and previous diabetes, arthritis, or shoulder pain diagnoses exhibited a more significant decline in upper extremity movements. Patients with worse upper extremity disability also report a poorer quality of life.

4.1 Upper arm volumetry software

Post–breast cancer lymphedema is the condition where lymphedema develops as a result of treatments for breast cancer, particularly surgeries and radiation therapy involving the lymph nodes. Lymphedema is a chronic condition characterized by the accumulation of lymphatic fluid, resulting in swelling, usually in the arms or legs. In the context of breast cancer, lymphedema typically affects the arm on the same side as the treated breast. Breast cancer treatments that can lead to lymphedema include Lymph Node Removal or Radiation Therapy. Surgical procedures such as axillary lymph node dissection or sentinel lymph node biopsy can disrupt the normal flow of lymphatic fluid, leading to lymphedema. Also, radiation treatment for breast cancer can cause scarring and damage to the lymphatic system, impairing its ability to drain fluid effectively and resulting in lymphedema. Post-breast cancer lymphedema can manifest in various ways, including

persistent swelling in the arm, hand, fingers, or chest on the side where the breast cancer treatment was administered, sensation of heaviness, tightness, or discomfort in the affected arm, reduction in the arm's range of motion due to the swelling and stiffness, increased risk of developing infections due to the compromised lymphatic system. Early detection and proactive management of post–breast cancer lymphedema are crucial in preventing complications and improving breast cancer survivors' overall quality of life [Lippi et al., 2023a].

The circumferential method for limb volume measurement is a commonly used approach in clinical practice and research for assessing changes in limb size or volume, particularly in the context of edema, lymphedema, or other pathological conditions. However, several challenges and limitations are associated with the circumferential method, which may impact the accuracy and reliability of the volume measurements. Some of these problems include:

- Assumption of uniform limb shape: The circumferential method often assumes that the limb has a uniform shape along its length, which may not be the case in individuals with irregular limb contours or variations in tissue composition. This assumption can lead to inaccuracies in volume calculations, particularly in cases where there are significant variations in tissue density or composition along the limb.

- Lack of three-dimensional information: The circumferential method relies solely on measuring the circumference of the limb at specific intervals, thereby overlooking the three-dimensional shape and variations in limb geometry. This limitation can result in an underestimation or overestimation of limb volume, especially in cases where the limb shape deviates from a standard cylindrical or conical model.

- Influence of edema or tissue compression: The presence of edema or tissue compression can significantly affect limb circumference measurements, leading to fluctuations in the apparent limb volume. In conditions such as lymphedema or post-surgical swelling, the circumferential method may not accurately capture the true changes in limb volume, as it does not account for the complex interplay between fluid dynamics and tissue properties.

- Inter-observer variability: The accuracy and reliability of the circumferential method can be influenced by inter-observer variability, as different individuals may apply varying degrees of pressure or use different measurement techniques when assessing limb circumference. Inconsistent measurement practices among different observers can introduce errors and discrepancies in the recorded data, compromising the overall reliability of the measurements.

- Difficulty in capturing non–cylindrical limb shapes: The circumferential method may encounter challenges when dealing with non-cylindrical limb shapes, such as irregularly shaped limbs or limbs with indentations or protrusions. In such cases, accurately defining the measurement points and ensuring consistent measurements

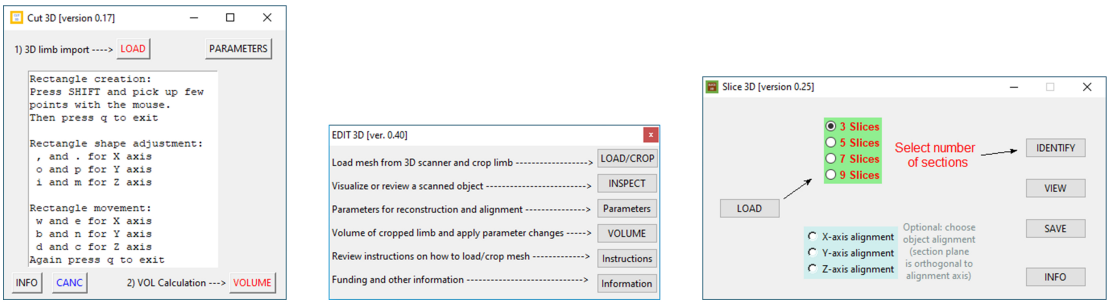

Figure 4.1: The user interface of the three apps composing the software.

along the limb circumference can be challenging, leading to potential errors in volume calculations.

For all these reason, clinician started using complementary techniques, such as water displacement methods, imaging modalities (e.g., MRI or CT scans), or advanced three–dimensional scanning technologies, to obtain a more comprehensive and accurate assessment of limb volume changes. Three–dimensional laser scanning in medicine is a technology that uses lasers to create 3D models or scans of various structures within the human body. It is a non–invasive imaging technique with applications in different medical fields, including surgery, orthopedics, prosthetics, and dentistry. The process typically involves a laser scanner to capture the body's surface topography or a specific body part. The laser rapidly and precisely measures the contours and dimensions of the target area, creating a detailed and accurate 3D representation. The data collected from the laser scanning process are processed using specialized software that analyzes the collected data points and generates a digital 3D model that accurately replicates the scanned body part. Surgeons can use 3D models generated from laser scans to plan and simulate complex surgical procedures, or 3D laser scanning facilitates the creation of custom-fitted prosthetics and orthotic devices. Additionally, 3D computer reconstructions of the patient's limbs could be measured to obtain surface area or volume computed digitally.

In the published manuscript [Nascimben et al., 2022a], we introduced a free–to–use software to calculate limb volume based on 3D laser scans. The software is made up of three apps, each one with peculiar features and computational capabilities. Their interface is shown in Figure 4.1, and it is downloadable from the Zenodo platform at the address `https://zenodo.org/records/7243978`.

Using three-dimensional scanning devices in medicine has significantly enhanced the measurement and quantification of anatomical features and volumes in patients, particularly in pathological conditions. However, the scarcity of freely available software for processing and analyzing the data obtained from these devices has led to challenges in the reproducibility and comparison of studies across different medical centers. To address this gap, a software package comprising three programs has been developed and

released, accompanied by supporting materials, to promote standardized volume assessment and facilitate cross–center comparisons. The article introduces the functions of the software programs and outlines the steps for volume assessment, focusing specifically on the quantification of upper limb volume in a pilot study that compared the digital outcomes to the values calculated with the circumferential method. The primary objective of the study was to evaluate the performance of digital volumes derived from the convex-hull gift-wrapping algorithm and other alternative analysis methods incorporated into the software. Notably, some of the digital volumes generated by the software were found to be dependent on specific parameters, necessitating careful value selection during the analysis process. The pilot study, conducted on a small group of young adults comprising both genders, provided valuable insights into the agreement between the clinical measurements and the digital volumes produced by the software package (Figure 4.2 summarizes the results digital versus CM). The results indicated a strong correlation between the digitial and circumferential sets of measurements, with the coefficient of determination (R^2) ranging from 0.93 to 0.97 and the correlation coefficient (r) ranging from 0.965 to 0.984. Furthermore, the study highlighted the potential influence of gender as a variable in upper limb volume quantification, emphasizing the importance of considering gender-specific models in such analyses. Overall, the development of the software package and the findings of the pilot study demonstrate the potential for standardized, reproducible, and parameter-controlled volume assessment using three-dimensional scanning data. The software package is poised to enhance the comparability of results across different medical centers and improve the accuracy and reliability of volume quantification in clinical practice and research. Even if the published article contains all the information to reproduce the results obtained during the upper limb volume quantification employing the software suite, supplementary materials, video tutorials, and user guides are available on the following website `https://mn-visions.gitbook.io/software-kit-for-3dls-limb-volume-quantification/`.

In conclusion as reported in the narrative review of [Lippi et al., 2023b] offering a comprehensive overview of various methods proposed in the literature for volumetric assessment, highlighting their respective strengths, limitations, and implications in clinical practice, the utilization of various volumetric assessment methods holds great potential in improving patient care, treatment outcomes, and research advancements in the field of upper–limb lymphedema management. By continually refining and validating these techniques, healthcare professionals can make significant strides in providing effective and personalized care for individuals affected by this chronic condition. Each method has unique attributes, with variations in accuracy, reliability, practicality, and cost-effectiveness, making the selection of the most appropriate method crucial in the clinical management of upper-limb lymphedema. Furthermore, factors such as operator experience, equipment availability, and patient population characteristics can significantly influence the choice and efficacy of the volumetric assessment method. Ensuring precise and standardized volumetric assessments is critical for enhancing rehabilitation strategies, patient education, and research outcomes in the context of upper-limb lymphedema management. The integration of emerging technologies is essential to further improve the tailored management of patients with upper-limb lymphedema. Future re-

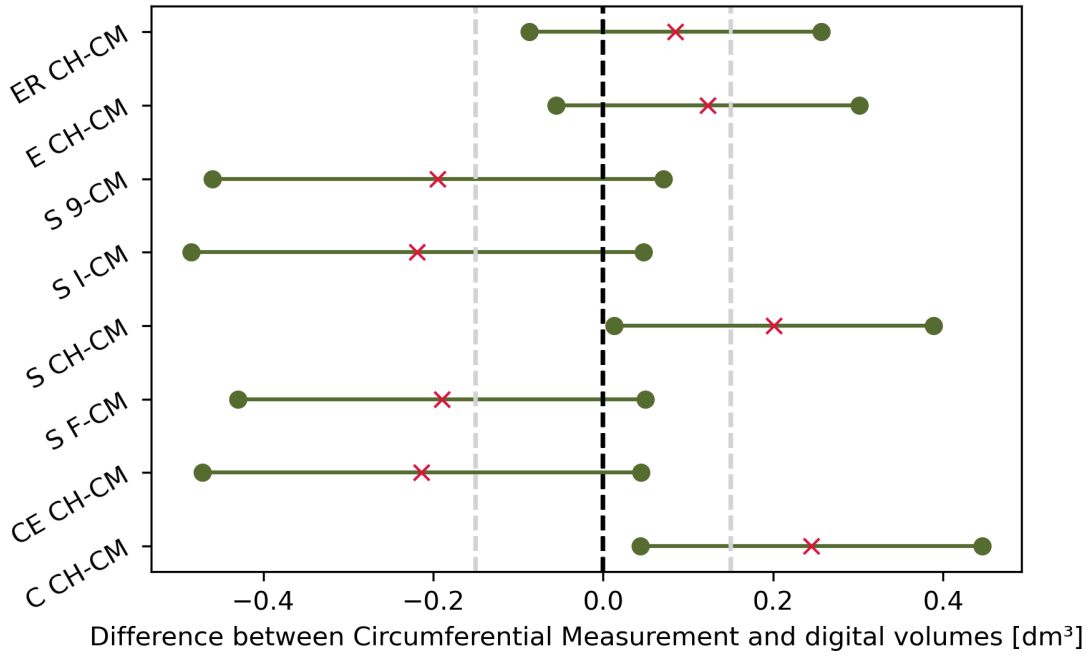

Figure 4.2: The Bland–Altman limits of agreement for the differences between circumferential measurement (CM) and digital volumes obtained from the apps (image as in [Nascimben et al., 2022a]).

search should prioritize the validation and implementation of these innovative solutions on larger patient cohorts to enhance the reproducibility, accuracy, accessibility, and clinical utility of volumetric assessment methods in the complex treatment framework of upper–limb lymphedema.

4.2 Hand volumetry algorithms

Hand volumetry, a technique used to assess the volume of the hand, can be challenging due to its complex structure with numerous small joints, intricate soft tissues, and irregular contours, making it difficult to accurately measure its volume without sophisticated imaging techniques or specialized equipment. The irregular shape of the hand, with variations in finger length, hand width, and palm curvature, presents challenges in obtaining precise and consistent measurements, especially when using manual techniques; for this reason, digital scans for volume estimation might be a valuable alternative. Traditional hand volumetry techniques often involve water displacement methods or circumferential measurements, which may not provide highly accurate or precise volume measurements, especially in cases where there is non-uniform swelling or changes in tissue density. To address these challenges, medical professionals may use advanced imaging techniques, such as 3D scanning, to obtain more accurate and detailed measurements of hand volume. These imaging techniques can provide a comprehensive evaluation of the hand's anatomy and aid in the precise assessment of changes in hand volume over time, especially in conditions like lymphedema, where regular monitoring of hand volume is essential for treatment management.

To address hand volumetry challenges, an ad–hoc study has been proposed and published as a separate work in [Nascimben et al., 2023b]. The comparison of clinical hand volumes computed through water displacement or circumferential measurements with digital volumetry derived from 3D laser scans represents an essential advancement in volume quantification and assessment. Using digital volume quantification algorithms, including applying the gift wrapping concept and cubic tessellation, offers a more precise and comprehensive approach to accurately capturing and analyzing the complex geometry of the human hand (Figure 4.3).

In particular, the gift-wrapping concept in digital volume quantification allows for a 3D hand representation, enabling a more detailed and nuanced analysis of its shape and volume. By utilizing the gift-wrapping concept, the digital volumetry algorithm can effectively enclose the entire hand surface, capturing its intricate features and contours with high precision and accuracy. Additionally, applying the cubic tessellation technique in digital volume quantification provides a parametric approach to volume calculation, allowing for the precise definition of the resolution of the tessellation. This parametric nature of the technique ensures that the digital volume quantification process can be tailored and calibrated according to specific measurement requirements, enhancing the reproducibility and accuracy of the volume calculations. The validation of the calibration methodology for defining the resolution of the tessellation further solidifies the reliability and accuracy of the digital volume quantification approach. By establishing a robust

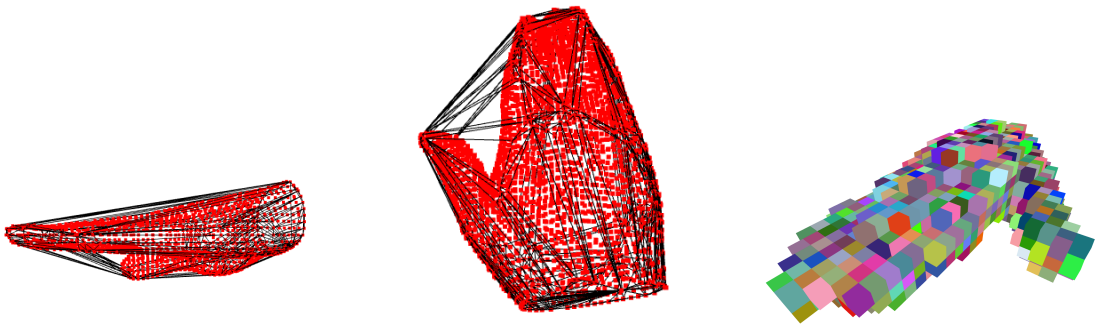

Figure 4.3: Two views of convex envelope edges (black lines) enclosing the scanned hand vertices (red dots) by the gift wrapping algorithm. On the right an example of hand tessellation (from [Nascimben et al., 2023b]).

calibration methodology, researchers can ensure consistency and standardization in the digital volume quantification process, enabling reliable comparisons between different measurement techniques and data sources. Overall, the integration of digital volumetry techniques derived from 3D laser scans, along with the calibration of the tessellation resolution, represents a significant advancement in accurately quantifying and analyzing hand volumes. This approach can potentially enhance our understanding of hand morphology and volume dynamics in various clinical and research settings, facilitating more precise and reliable assessments of hand-related pathologies, functional impairments, and treatment outcomes.

4.3 Algorithm–based post–breast cancer lymphedema risk stratification

Lymphedema risk stratification is crucial after breast cancer treatment because it helps identify patients at a higher risk of developing lymphedema, a chronic condition characterized by the accumulation of lymphatic fluid and subsequent swelling, typically in the arms or legs. Healthcare providers can implement proactive monitoring and preventive measures early in the post-treatment phase by identifying patients at a higher risk of lymphedema. Early detection allows for timely intervention, which can help prevent the progression of lymphedema and improve the overall quality of life for the patients.

In the manuscript published in Cancers journal [Nascimben et al., 2023a], we proposed an example of algorithm–based medicine applied to patient risk stratification for lymphedema. Algorithms provide a standardized approach to medical decision-making, ensuring patients receive consistent, high-quality care regardless of the specific healthcare provider they consult. Additionally, by utilizing algorithms based on the best available evidence, healthcare professionals can make more informed decisions about patient care, leading to improved treatment outcomes and reduced medical errors; another benefit

is that algorithm-based medicine can streamline clinical workflows, allowing healthcare providers to make accurate diagnoses and treatment plans more efficiently. Consequently, improved efficiency can save costs by reducing unnecessary procedures, tests, and hospital stays.

With the increasing availability of large-scale healthcare data, algorithms can analyze and interpret biomarker data to generate insights that help clinicians make informed decisions, leading to more personalized and precise treatments: in complex medical situations, algorithms can guide the best course of action, considering various factors such as patient history, comorbidities, and treatment options, which can be challenging for healthcare providers to manage on their own. Algorithm-based medicine allows for continuous evaluation and updating of protocols based on new evidence and clinical outcomes, enabling the healthcare system to stay current with the latest advancements in medical research. By providing standardized guidelines for medical practice, algorithms can help reduce variability in clinical decision-making, minimizing the potential for disparities in patient care based on individual biases or preferences of healthcare providers.

The mentioned study focuses on the development of a risk stratification model for upper limb unilateral lymphedema (BCRL) in patients with breast cancer. We used data from 294 patients from two hospitals in northern Italy to identify factors associated with the development of BCRL. The multi-centric dataset consisted of twenty-three clinical features from patients who had undergone axillary dissection for breast cancer (BC) and were either presenting with or without upper limb unilateral lymphedema (BCRL). By employing unsupervised low-dimensional data embeddings and clustering, the study aimed to create a prognostic map that divides the patient cohort into three distinct clusters based on specific characteristics. By modeling the patients' clinical variables separately in two distinct embeddings, considering ordinal and binary variables separately. After creating distinct models for the patients' variables, we merged the two models into a bi-dimensional prognostic map. This data fusion helped integrate the insights from both embeddings to provide a comprehensive understanding of the relationships between the clinical features and the presence or absence of BCRL. The use of a Gaussian mixture model, a statistical method for estimating the underlying probability distributions of the data, facilitated the categorization of patients into three distinct clusters based on their specific clinical characteristics and features. By categorizing the patients into three clusters (Figure 4.4), the study likely aimed to identify and delineate the different subgroups of patients based on their clinical profiles and the presence or absence of BCRL. This approach allowed us to uncover patterns and associations between the various clinical features and the development of BCRL, thereby contributing to developing a more comprehensive risk stratification model for this condition. The study's findings have potential implications for developing a more precise risk stratification model for BCRL. By identifying the factors associated with the high-risk cluster, we have uncovered valuable insights that can be used to tailor therapeutic interventions specifically for patients at a higher risk of developing BCRL. Furthermore, the study's results could guide the allocation of healthcare resources, ensuring that patients at high risk receive the necessary attention and targeted interventions to mitigate the onset and progression of BCRL. This personalized approach has the potential to improve patient

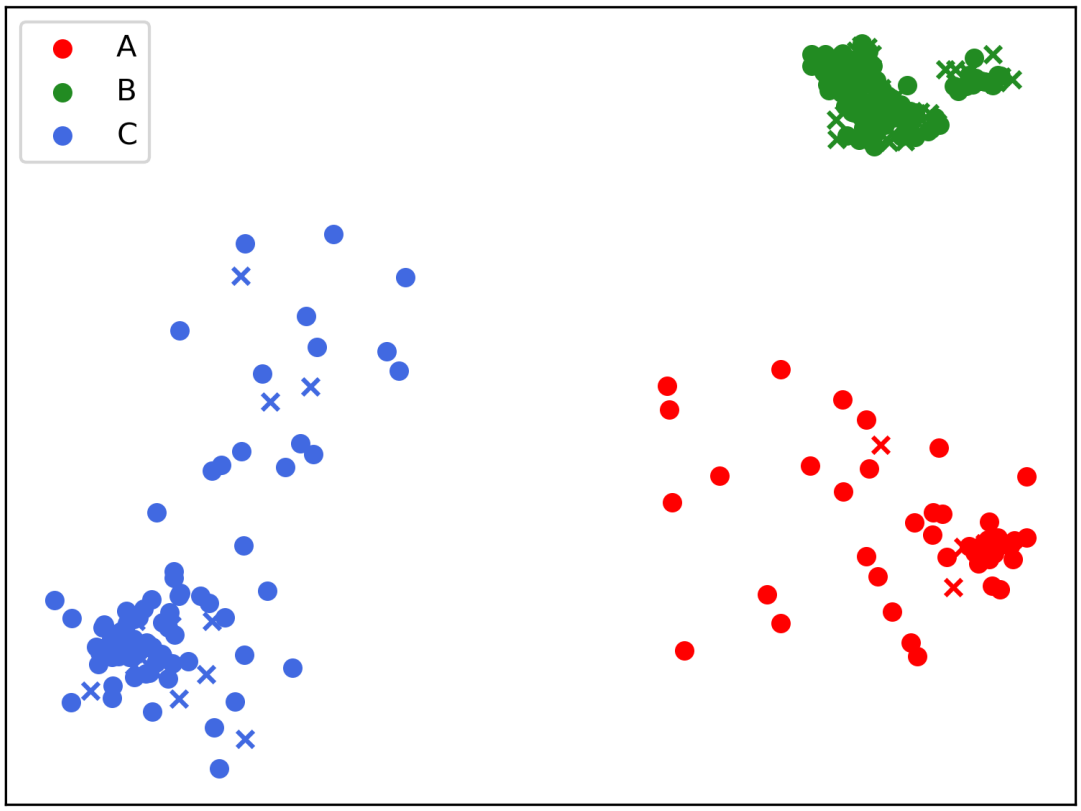

Figure 4.4: Low-dimensional embedding of the patients into a bi-dimensional map: each point is a patient colored according to clustering into the three groups A, B, and C. In the above figure, dots depict patients without BCRL, while crosses represent patients with the disease. (image as published in [Nascimben et al., 2023a]).

outcomes and overall quality of life for breast cancer survivors.

From the medical records of 294 women with a mean age of 59.823±12.879 years, the patients were grouped into three distinct clusters, each one had a different proportion of subjects affected by upper limb unilateral lymphedema. Specifically, the probability that a patient with BCRL belonged to each cluster was reported as follows:

- Cluster A: 5.71%
- Cluster B: 71.42%
- Cluster C: 22.86%

Evaluating cluster composition, we delved into a comprehensive appraisal of intra- and inter-cluster factors. By examining these factors, we aimed to gain a deeper understand-

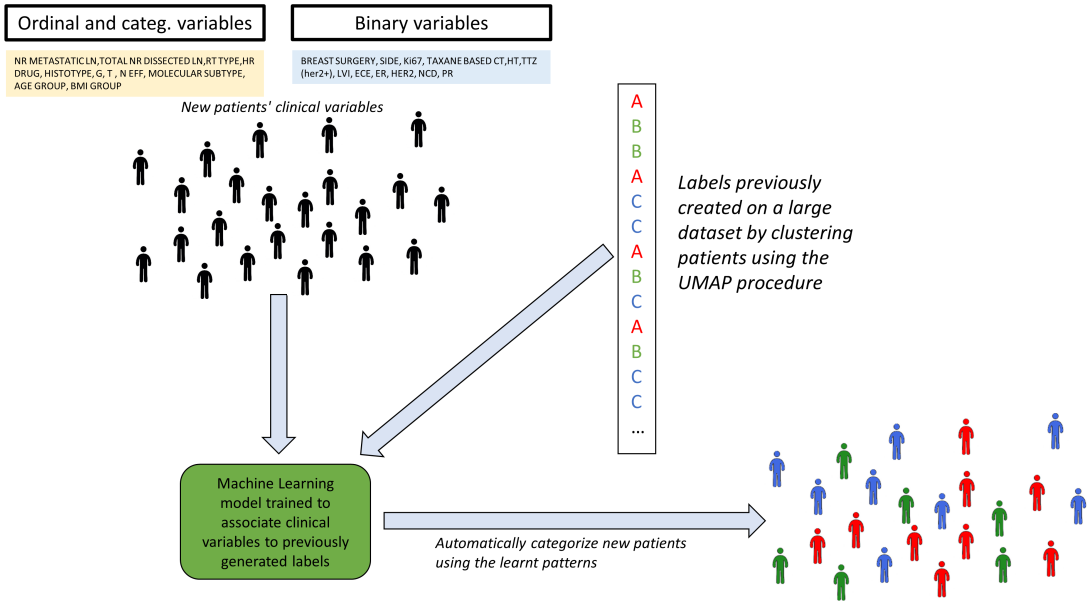

Figure 4.5: Possible usage of the proposed methodology in prospective studies. (image as published in [Nascimben et al., 2023a]).

ing of the characteristics and clinical profiles that were unique to each cluster. Furthermore, the study identified a subset of clinical variables that played a significant role in determining cluster membership; these variables were also found to be significantly associated with the biological hazard of developing BCRL. Identifying these critical clinical variables and their associations with both cluster membership and the biological hazard of BCRL is crucial for enhancing the understanding of the underlying mechanisms and risk factors contributing to the development of this condition. By pinpointing these influential variables, we might have paved the way for developing targeted interventions and personalized treatment strategies for patients at different risk levels within each cluster. This tailored approach could lead to improved patient outcomes and a more effective allocation of healthcare resources for the prevention and management of BCRL, as illustrated in Figure 4.5.

4.4 Final remarks

Analyzing post-breast cancer lymphedema in the upper limb using machine learning can provide several advantages, contributing to a better understanding of the condition, improved diagnosis and assessment, and the development of more effective treatment strategies. We built software algorithms that can facilitate the quantitative assessment and monitoring of lymphedema progression in the upper limb by analyzing changes in limb circumference, volume, and functional impairment. Automated monitoring can

provide clinicians with objective measures for tracking treatment efficacy and disease progression, enabling personalized interventions and timely adjustments to the treatment plan. Moreover, it has been released a paper establishing how machine learning can aid in the identification of specific risk factors and biomarkers associated with disease severity and progression in post-breast cancer lymphedema patients. By stratifying patients based on risk profiles, machine learning models can provide insights into the likelihood of developing complications and guide healthcare providers in implementing preventive measures and targeted interventions to mitigate the risk of disease exacerbation.

5 Bioinformatics' biomarkers analysis

Original contribution to knowledge

This chapter is based upon the articles
Mauro Nascimben, Manolo Venturin, and Lia Rimondini. Double–stage discretization approaches for biomarker-based bladder cancer survival modeling. *Communications in Applied and Industrial Mathematics*, 12(1):29–47, 2021
Mauro Nascimben, Lia Rimondini, Davide Corà, and Manolo Venturin. Polygenic risk modeling of tumor stage and survival in bladder cancer. *BioData Mining*, 15(1):23, 2022b
and the conference presentation
Mauro Nascimben. A machine learning based decision support system in oncology. Parma, Italy, Sept 2021a. University of Parma, 2020+2021 Italian Society of Applied and Industrial Mathematics (SIMAI) Conference

In bioinformatics, biomarker analysis involves identifying, characterizing, and analyzing biomarkers, which are measurable indicators of biological processes, disease states, or pharmacological responses in living organisms. Biomarkers can be molecular, genetic, proteomic, or phenotypic and are used to assess normal biological processes, pathogenic processes, or responses to therapeutic interventions [Durairaj and Ranjani, 2013; Ichimura et al., 2005]. Biomarker analysis in bioinformatics typically includes the following key aspects:

- Biomarker Discovery: Bioinformatics tools and techniques are used to analyze large-scale omics data, such as genomics, transcriptomics, and proteomics data, to identify potential biomarkers associated with specific diseases, biological processes, or drug responses; this involves the application of various statistical and machine-learning methods to detect patterns and correlations in complex biological datasets.
- Biomarker Validation: Once potential biomarkers are identified, bioinformatics approaches are used to validate their clinical relevance and utility; validation includes assessing the robustness of biomarker candidates across different patient cohorts, evaluating their specificity and sensitivity, and determining their predictive value for disease diagnosis, prognosis, or treatment response.
- Pathway and Network Analysis: Biomarker analysis often involves the examination of molecular pathways and biological networks associated with the identified

biomarkers. Bioinformatics tools enable the visualization and analysis of complex molecular interactions and signaling pathways, providing insights into the underlying biological mechanisms and the relationships between biomarkers and disease processes.

- Integration of Multi-Omics Data: Bioinformatics facilitates data integration from multiple omics platforms, allowing researchers to comprehensively analyze and interpret the molecular signatures and interactions associated with biomarkers. Integrative analyses of genomics, transcriptomics, proteomics, and metabolomics data can provide a holistic understanding of disease mechanisms and facilitate the identification of robust and reliable biomarkers.

- Clinical Translation and Application: Bioinformatics is crucial in translating biomarker discoveries into clinical applications by developing computational tools and algorithms for biomarker-based diagnostic tests, patient stratification, and personalized treatment approaches, as well as the design and implementation of clinical trials to evaluate the efficacy and utility of biomarker-driven interventions.

By leveraging bioinformatics for biomarker analysis, researchers can advance our understanding of disease mechanisms, facilitate early detection and diagnosis of diseases, and contribute to processing targeted and personalized therapeutic strategies, ultimately improving patient outcomes and healthcare practices.

The bioinformatic biomarker analysis presented in the subsequent paragraphs will test pipelines enclosing a data discretization step. Data discretization, or binning, involves converting continuous data into discrete intervals or bins. Discretization simplifies complex continuous data by reducing the number of unique values; this simplification makes data more interpretable, especially for non-technical stakeholders who find it easier to understand discrete categories or ranges. In data analysis, granularity is crucial because it affects the level of insight that can be derived from the data: finer granularity provides more detailed information but may also result in larger datasets and increased complexity, whereas coarser granularity simplifies data but may lead to loss of detail [Pal et al., 2017]. Data discretization and granularity are related concepts, as both involve dividing continuous data into distinct intervals or categories. Data discretization is a general term that refers to converting continuous data into discrete intervals or bins; it can be done for various reasons, including simplification, noise reduction, and compatibility with specific algorithms. Discretization involves grouping similar or nearby values to create categories or bins. The number and size of these bins can be determined based on particular criteria or algorithms. Granularity, on the other hand, refers to the level of detail or precision in a dataset. It describes how finely the data is divided into individual units or elements. Granularity can be applied to various aspects of data, such as time, geography, or other dimensions. For example, in time series data, granularity could refer to whether the data is recorded at the level of seconds, minutes, hours, days, or other time units. The relationship between data discretization and granularity lies in the fact that discretization often involves defining the boundaries or categories that determine the granularity of the resulting data. When discretizing continuous data, one must decide on the size and num-

ber of bins, effectively choosing the granularity level in that data's representation. Finer discretization results in smaller bins and higher granularity, while coarser discretization leads to larger bins and lower granularity [Pedrycz, 2000].

5.1 Bladder cancer survival prediction

Bladder cancer is a type of cancer that begins in the cells of the bladder, the organ in the pelvis responsible for storing urine before it is excreted from the body. The most common type of bladder cancer is urothelial carcinoma, which begins in the cells that line the inside of the bladder. Bladder cancer can also develop in other types of cells in the bladder, but these cases are less common. The treatment options for bladder cancer depend on the cancer stage, the patient's overall health, and other factors. Treatment may include surgery to remove the cancerous cells or, in more severe cases, the entire bladder (cystectomy), chemotherapy, radiation therapy, immunotherapy, or a combination of these treatments. Sometimes, a combination of treatments may achieve the best results. Regular check-ups and screening are essential for individuals at risk of bladder cancer, especially those with a smoking history or exposure to certain industrial chemicals. Early detection and timely treatment can significantly improve the prognosis and outcome for individuals diagnosed with bladder cancer.

Several biomarkers have been identified as potential indicators of prognosis and survival in patients with bladder cancer. These biomarkers can help predict the likelihood of disease progression, recurrence, or response to specific treatments. For example, analysis of Cell-Free DNA for specific genetic alterations and mutations, such as alterations in the TERT promoter or FGFR3 mutations, has shown potential in predicting the risk of recurrence and survival outcomes in bladder cancer patients. Also, several molecular markers have been identified as potential prognostic indicators in bladder cancer. Mutations or alterations in genes such as TP53, RB1, FGFR3, and ERBB2 are frequently cited. These markers can provide insights into the aggressiveness of the cancer and the likelihood of response to specific therapies. Other biomarkers related to immune checkpoint pathways, such as PD-L1 expression and tumor-infiltrating lymphocytes, are being increasingly studied as prognostic indicators and predictors of response to immunotherapy in bladder cancer.

The work in [Nascimben et al., 2021] established a novel pipeline to address survival rate prediction employing a dataset of gene expression curated by [Zhang et al., 2020], pre-processed as shown in Figure 5.1. The dataset was composed by hub and seed genes, commonly used in gene expression networks and network biology. Hub genes are genes that play a crucial role in maintaining the connectivity and function of a biological network, such as a gene regulatory network or a protein-protein interaction network. Hub genes are highly connected to other genes in the network and are often involved in crucial regulatory or signaling pathways. Identifying hub genes can provide insights into the central players that control various biological processes and pathways. Seed genes are known to be involved in a particular biological process, disease, or phenotype of interest. Seed genes serve as starting points for constructing gene networks and are used

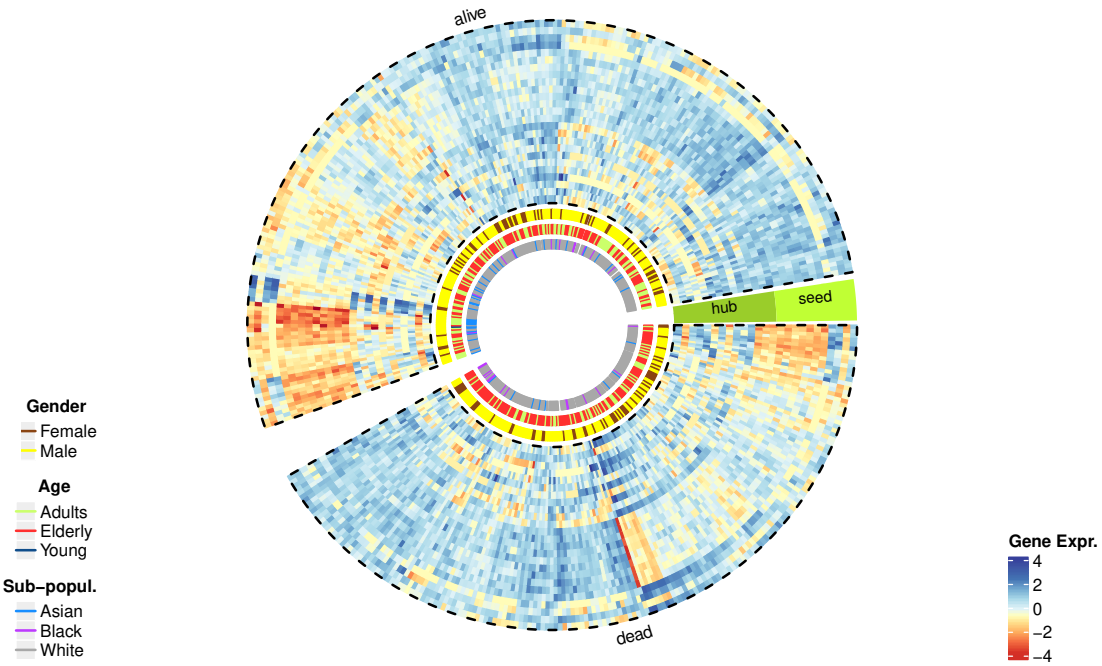

Figure 5.1: Visualization of the gene expression variables including annotations for all patients (from [Nascimben et al., 2021]).

as a reference to identify other genes that may interact with or be functionally related to the seed genes. Analyzing the relationships between seed genes and other genes in the network can help uncover novel pathways or molecular mechanisms associated with the biological process or disease under investigation. Both hub genes and seed genes are critical for understanding the organization and dynamics of gene networks, as well as for identifying potential therapeutic targets or biomarkers for various diseases and biological processes. They play a significant role in systems biology and network-based approaches to studying complex biological systems.

The field of bioinformatics utilizes specialized techniques and analysis pipelines to study gene expression data, aiming to uncover properties, adaptations, and disease outcomes within a given sample population. In this recent investigation focusing on bladder cancer genetic profiles, a comparison of four numerical experiments was conducted, each modeling survival rates. The research findings highlighted the effectiveness of a particular sequence of two discretization phases, showcasing superior performance when contrasted with a conventional approach employing only one discretization of gene expression data (numerical experiments compared together exemplified into Figure 5.2).

The analysis involving two discretization phases comprised an initial discretizer, followed by the refinement or pre–binning of input values before implementing the main discretization scheme. Notably, the research results demonstrated that this two-phase

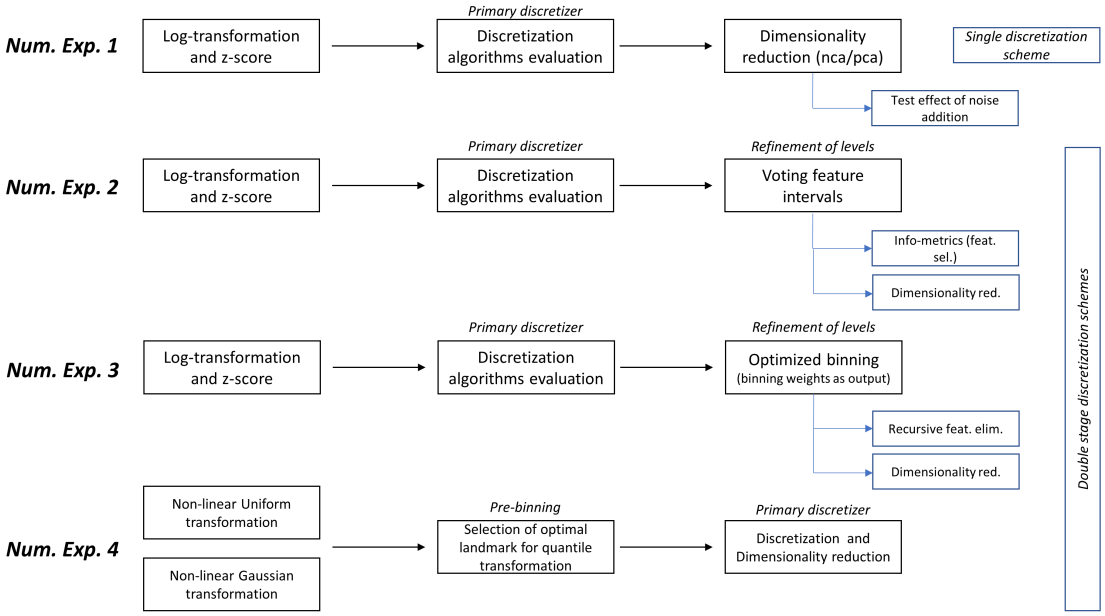

Figure 5.2: Overview of the four numerical experiments for GED analysis in bladder cancer (from [Nascimben et al., 2021]).

approach led to remarkable outcomes, indicating its potential to enhance the accuracy and robustness of survival rate modeling based on bladder cancer genetic profiles. The best-performing model identified in the study involved a sequence of data transformations designed to compensate for skewness in the data. Additionally, the model incorporated a data discretization phase featuring a class-attribute interdependence maximization algorithm [Kurgan and Cios, 2004], which optimized the association between gene expression patterns and survival rates. Furthermore, the final classification process was conducted using a voting feature intervals classifier [Demiröz and Güvenir, 1997], which not only facilitated discrete interval optimization but also contributed to the overall robustness and predictive accuracy of the model (Table 5.1) after 10–Fold cross–validation. Using VFI helped refine the levels created by the primary discretization algorithm. This refinement, in combination with ChiMerge and potentially CACC, led to improved classification scores and high overall accuracy in predicting the disease outcome; this highlights the effectiveness of the combined approach in data preprocessing and classification, thereby emphasizing the importance of the refined feature representation in achieving accurate predictions. These findings underscore the significance of employing advanced and multi-phase data analysis techniques in bioinformatics research, particularly in studying genetic profiles and their associations with disease outcomes. The identification of an optimized data processing pipeline, as demonstrated in this investigation, holds great promise for improving the precision and reliability of survival rate modeling in bladder cancer, thereby facilitating more effective disease prognosis and personalized treatment

Discretizer	AUC	Bal. Accuracy	Original levels	VFI levels
CACC	0.99±0.003	97.63±2.65%	199	189
ChiMerge	0.98±0.01	94.39±3.34%	51	43
CART	0.89±0.08	82.55±10.17%	18	15
AMEVA	0.78±0.09	70.00±7.03%	8	6

Table 5.1: Outcomes of the pipelines identified in the 2nd numerical experiment (from [Nascimben et al., 2021]).

strategies.

5.2 Bladder cancer tumor stage with survival prediction

Predicting the bladder cancer tumor stage involves various diagnostic procedures, including imaging tests, biopsies, and surgical staging. The process aims to accurately determine the extent of the cancer's spread within the bladder and beyond. For these purposes, cystoscopy helps in the direct visualization of the tumor and allows for the collection of tissue samples for biopsy. This procedure involves using a cystoscope, a thin tube with a camera, to visualize the inside of the bladder. Tissue samples obtained through cystoscopy are examined under a microscope to determine the histological type of bladder cancer and assess cellular abnormality and invasiveness. This information is crucial in determining the tumor stage. Accurate tumor staging is crucial for developing an appropriate treatment plan and predicting the prognosis for patients with bladder cancer.

In [Nascimben et al., 2022b], the conducted numerical experiments focused on evaluating the effectiveness of a comprehensive approach combining Gene Expression Data (GED) preprocessing through discretization with tree ensemble embeddings and nonlinear dimensionality reductions (initial data shown in Figure 5.3).

The primary objective of the modeling was twofold: to categorize oncological patients by identifying tumor stages and to differentiate survival outcomes. The experiments were conducted under two specific scenarios: one involving complete data embedding and the other simulating partial data embedding, which mimics the addition of new patients to an existing model for rapid disease progression monitoring. To achieve the outlined goals, machine learning procedures were utilized, with a specific emphasis on identifying the most relevant genes that play a crucial role in patient prognosis. The performance of the preprocessed GED was rigorously assessed and compared to that of the untransformed data, particularly in predicting patient conditions and providing insights into disease progression and survival outcomes. Integrating GED preprocessing, tree ensemble embeddings, and nonlinear dimensionality reductions represents a robust approach for enhancing the comprehensive categorization of oncological patients, offering valuable insights into tumor staging and patient prognosis. The incorporation of machine learning techniques facilitated the identification of key genetic markers associated with patient outcomes, thereby enabling a deeper understanding of the underlying biologi-

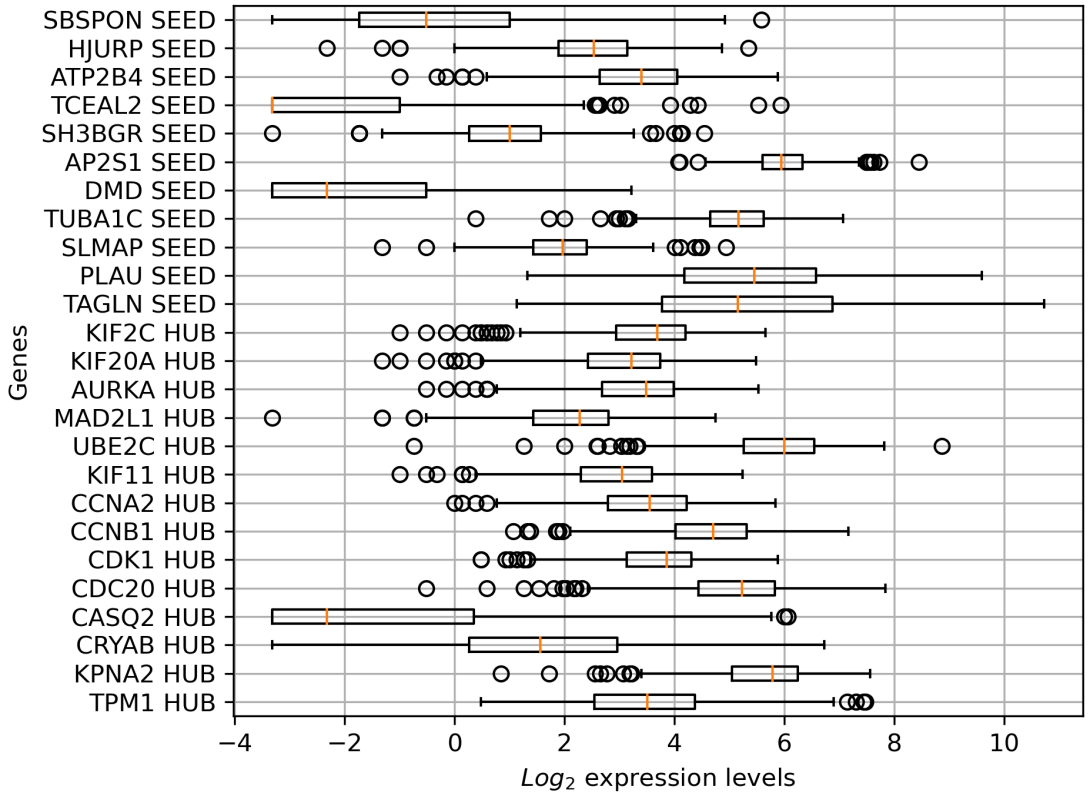

Figure 5.3: The boxplots depict $\log_2$ expression levels for the hub and seed genes before preprocessing (from [Nascimben et al., 2022b], employing the GED identified by [Zhang et al., 2020]).

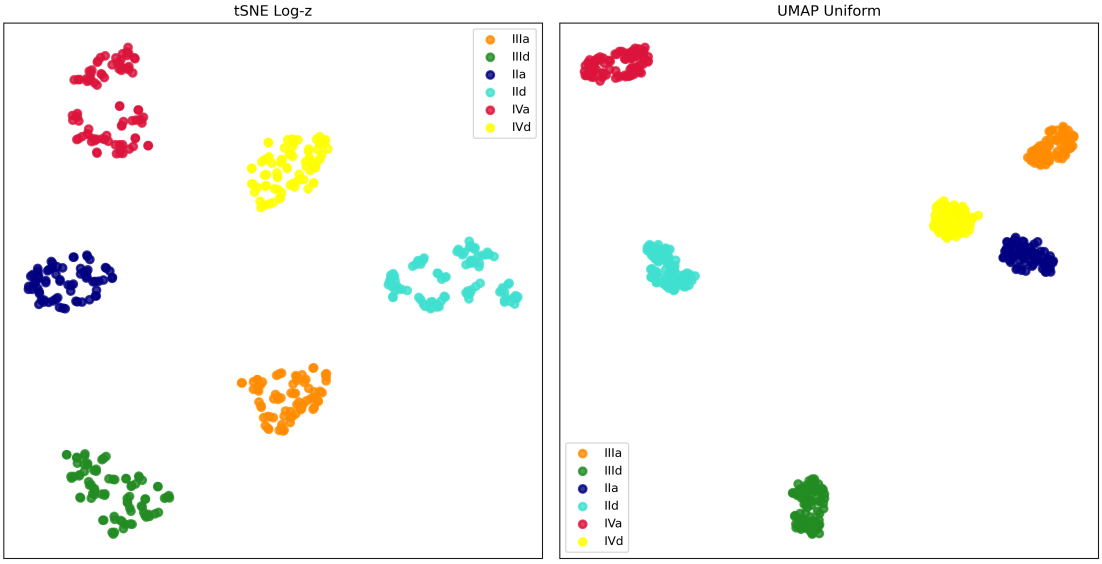

Figure 5.4: On the left: GED full embedding generating prognostic maps using tSNE Log-z values; on the right: prognostic map after Uniform UMAP transformation (from [Nascimben et al., 2022b]).

cal mechanisms and pathways involved in oncological conditions. By evaluating both complete and partial data embedding scenarios, the study underscores the potential utility of the proposed approach in facilitating real-time disease monitoring and prognosis, thereby contributing to developing more effective and personalized treatment strategies for oncological patients. The findings of this research have significant implications for the advancement of precision oncology and the improvement of patient outcomes in oncology and cancer research.

The application of data embedding in conjunction with dimensionality reduction techniques resulted in the generation of robust prognostic maps, revealing well-defined clusters of patients that can facilitate medical decision support (Figure 5.4). The image displays the point clouds representing tumor stages (II, III, IV) and the outcome (alive or dead). A subsequent experiment focused on simulating the addition of new patients to an existing model, employing the Partial Data Embedding approach. This investigation highlighted that the utilization of the Uniform Manifold Approximation and Projection (UMAP, [Dorrity et al., 2020; McInnes et al., 2018]) methodology, coupled with uniform data discretization, yielded more favorable outcomes compared to other analyzed pipelines.

Furthermore, the exploration of the parameter space for both UMAP and t-distributed stochastic neighbor embedding (t-SNE, [Van der Maaten and Hinton, 2008]) techniques emphasized the critical role of tuning a higher number of parameters for UMAP as opposed to t-SNE. This finding underscores the importance of optimizing the configuration of UMAP to effectively capture and represent the underlying patterns and structures

within complex patient data. In addition, two distinct machine-learning experiments were conducted, with the first focusing on identifying a group of genes deemed valuable for partitioning patients through gene relevance analysis. The results of this analysis shed light on the key genetic markers that contribute significantly to patient stratification and prognosis, providing valuable insights for personalized treatment approaches. The second machine learning experiment demonstrated the superior precision achieved by preprocessed data in predicting tumor outcomes for cancer stage and survival rate, particularly in the context of a six-class prediction model; this highlights the crucial role of data preprocessing techniques, such as the application of UMAP and uniform data discretization, in enhancing the accuracy and reliability of prognostic predictions in oncology. Overall, these findings underscore the importance of leveraging advanced data embedding and dimensionality reduction techniques in conjunction with meticulous data preprocessing and machine learning methodologies to improve patient stratification and prognostic modeling in the context of cancer research and clinical decision-making.

The current study developed novel analysis pipelines for modeling disease outcomes based on bladder cancer–related biomarkers. Through comprehensive investigations involving both complete and partial data embedding experiments, it was revealed that pipelines integrating the Uniform Manifold Approximation and Projection (UMAP) technique exhibited superior predictive capabilities. These findings align with recent trends in the literature, highlighting the growing recognition of the efficacy of UMAP in disease modeling and prognostic assessment. However, the study also identified that various UMAP parameters significantly impact the experimental outcomes. As a result, a key recommendation was emphasized for researchers to meticulously consider and optimize the relevant parameters when implementing the UMAP technique. By emphasizing the importance of parameter selection and fine-tuning within the UMAP methodology, the study aims to guide researchers in maximizing the accuracy and reliability of disease outcome predictions based on bladder cancer-related biomarkers. Furthermore, the application of machine learning procedures in the study corroborated the effectiveness of the proposed preprocessing techniques in accurately predicting patients' conditions and disease outcomes. Notably, identifying a specific sub-group of biomarkers deemed significant for forecasting bladder cancer prognosis serves as a crucial step toward improving the understanding of the underlying molecular mechanisms and pathways involved in bladder cancer progression and patient outcomes. The integration of these findings and recommendations underscores the potential of advanced analysis pipelines and machine learning methodologies in enhancing the precision and reliability of disease outcome modeling in the context of bladder cancer research. By leveraging the insights the study provides, researchers can refine their approaches to data analysis and interpretation, ultimately contributing to the development of more effective diagnostic and prognostic tools for improved patient care and treatment outcomes in bladder cancer management.

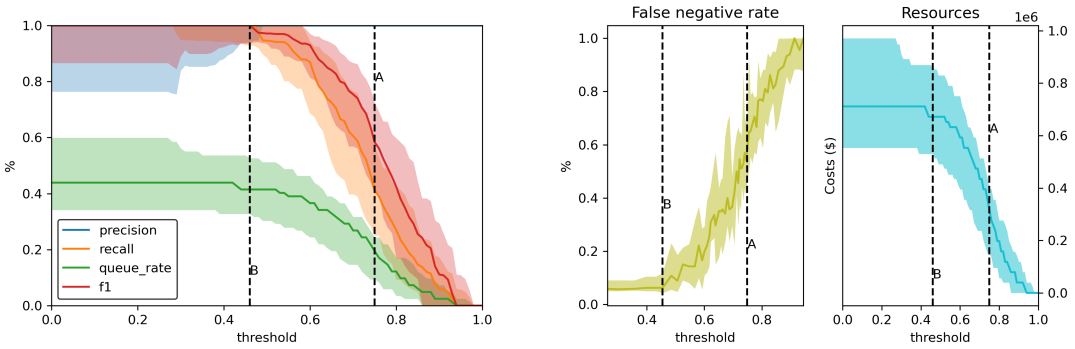

Figure 5.5: On the left: ML model behavior and patient "queue" rate, on the right: forecasted type II errors and operating costs (as presented in [Nascimben, 2021a]).

5.3 Machine learning based decision support system in oncology

In [Nascimben, 2021a], the communication highlights the potential of machine learning in automatic feature learning, particularly in the context of genetics. ML can assist in bridging the gap between the rapid accumulation of genetic data and the slower, more time-consuming process of interpretation. Despite its benefits, one significant challenge lies in effectively communicating how prediction models derived from ML can impact decision-making processes. To enhance interpretability in binary classification problems, the communication suggested identifying a decision threshold that correlates with the outcomes of the confusion matrix. This approach can provide a clearer understanding of how the ML model makes predictions and can aid in decision-making processes, especially in healthcare and insurance industries. The study also involved the development of a theoretical cost-benefit simulation, which healthcare managers and private insurance companies could utilize. This simulation, based on a previous investigation related to bladder cancer survival (from [Nascimben et al., 2021]), was likely designed to demonstrate how ML outcomes can be integrated into real–world decision–making scenarios. By presenting the simulation graphically, the oral communication aimed to offer a practical tool that could facilitate collaboration between ML developers and healthcare managers, fostering a better understanding of how ML insights can be applied in a real-world context. Ultimately, the goal is to improve the communication of ML findings to non-experts, such as healthcare managers and insurance professionals, by providing them with a perspective that aligns more closely with their backgrounds and decision–making processes; this approach can lead to more effective implementation of ML solutions in healthcare and related industries, ultimately benefiting patients and clients.

Gene expression levels (GED) of a patient's tumor can provide prognostic or risk information (including recurrence), assisting healthcare specialists in making decisions.

Furthermore, cancer has extensive costs for its treatment shared between health care systems, insurance companies, and privates. Cancer is the second leading cause of death in the US, but the third is medical errors while seeking treatment. Nowadays, gene expression profiling tests are commercially available, and machine learning models derived from GED can have practical applications in real–world scenarios with the ultimate goal of building a model able to create value for all operators involved in healthcare. However, communicating how prediction models can impact decisions can be challenging. In binary classification problems, a way to address interpretability is to identify a decision threshold linked to the outcomes of the confusion matrix (true positives, false positives, true negatives, and false negatives). Through a threshold, continuous outputs can be translated to a "yes/no" decision. Three scores can be calculated for creating a decision threshold: "queue rate", which represents the number of cases that can be treated by the healthcare system, "recall" also called sensitivity, and "precision" which identifies the fraction of true positive instances among the ones classified as positive. In addition, the "F1" score is a measure derived from the latter two that balances both precision and recall in one single value. In this theoretical cost–benefit simulation shown in Figure 5.5, we applied the VFI model because it showed a good approximation of bladder cancer survival rate to establish a theoretical model for healthcare managers or private insurance companies. To build this model, a few hypothetical assumptions were made: an average observation period of 2.2 years for 405 patients as those included in the original data set and a cost of 4000 dollars for each GED profiling exam (costs derived from breast cancer genetic profiling expenses as in and, usually shared in different percentages between patient and health system or insurance company). Precision and recall values obtained in 15 are the empirical quantiles after running 50 simulations (train/test split proportion 90/10), and the costs simulated in 16 are those an organization could bear at total chance. If managers set a decisional threshold exemplified in "B", the number of type I and type II errors are minimized, and the expenses for screening costs are maximal. These costs include all resources for testing the patients, exemplified by the "queue rate". Suppose the decision threshold is increased until level "A", saving on the resources spent for GED profiling. In that case, the recall score drops to 0.411, causing an increase in type II errors (predicting survival instead of death). According to this theoretical model, saving of resources is followed by the expansion of the false negative rate due to an increase of type II errors. False-negative results concern a health system because patients may feel confident and reassured, with individuals not seeking medical care even if physical conditions deteriorate; these circumstances cause delays in diagnosis and treatment. False negative cancer rates can give a fast metric to understand mortality without running a randomized clinical study. Future works will improve this model with a more accurate forecast of the costs, including information coming from the meta-data (age and tumor stage). Interpretation of this theoretical simulation could provide a tool for deriving solutions to bridge between developers and healthcare managers by combining the outcomes of the machine learning model with a real–world decision-making scenario.

5.4 Final remarks

Analyzing gene expression data using machine learning in the context of bladder cancer can provide several advantages that contribute to a better understanding of the disease, improved diagnostic capabilities, and the development of more effective treatments. Machine learning can aid in identifying specific gene expression patterns that serve as potential biomarkers for early detection, prognosis, and treatment response prediction in bladder cancer. The chapter demonstrated how, by analyzing large-scale gene expression datasets, machine learning algorithms can identify genes or molecular signatures associated with different stages of bladder cancer, facilitating the development of sensitive and specific biomarkers. Moreover, Machine learning models can help in predicting patient response to various treatment modalities, including chemotherapy, immunotherapy, and targeted therapy, based on individual gene expression profiles. This personalized approach can optimize treatment selection, minimize adverse effects, and improve overall patient outcomes in bladder cancer management. For subtype classification and prognosis prediction, we showed how machine learning algorithms can categorize bladder cancer into distinct cancer stages based on gene expression patterns. This enables more accurate prognosis prediction and tailored treatment strategies for patient subgroups. One future goal employing merged data from other sources, machine learning analysis of gene expression data, can provide valuable insights into the molecular mechanisms and signaling pathways involved in bladder cancer development and progression. By uncovering key genes and regulatory networks associated with tumor growth, invasion, and metastasis, machine learning can contribute to a more comprehensive understanding of the complex biology underlying bladder cancer, paving the way for developing novel therapeutic targets and interventions. Indeed, integrating multi-omics data, including genomic, transcriptomic, epigenomic, and proteomic data, creates a holistic view of the molecular landscape of bladder cancer. Integrative analysis can reveal intricate interactions between different molecular layers, providing a comprehensive understanding of the complex interplay between genetic alterations and phenotypic changes in bladder cancer. Finally, we demonstrated how to leverage machine learning models for public health management.

6 Biostatistics: equivalence analysis

Original contribution to knowledge

This chapter is based upon the article
Mauro Nascimben and Lia Rimondini. Visually enhanced Python functions for clinical equality of measurement assessment. *Annals of Computer Science and Information Systems*, 32:241–249, 2022

In biostatistics, comparative statistical and equivalence tests serve different purposes and address distinct research questions. Comparative statistical tests are used to assess whether there are significant differences between two or more groups or treatments. These tests help researchers determine whether the observed differences in means, proportions, or other relevant parameters are statistically significant. Some common examples of comparative statistical tests include t-tests, ANOVA, chi-square tests, and nonparametric tests like the Mann-Whitney U test or the Kruskal-Wallis test. Comparative tests are essential for identifying and quantifying differences between groups or treatments under investigation. On the other hand, equivalence tests in biostatistics are used to determine whether the difference between two groups or treatments is within a predetermined range that is considered practically insignificant. Equivalence tests are particularly relevant when researchers aim to demonstrate that the effects of different treatments or interventions are similar or when they want to establish the non-inferiority or similarity of a new treatment compared to an existing standard. Equivalence tests typically involve setting up bounds of equivalence and testing whether the observed effect falls within these bounds. Common methods for conducting equivalence tests include TOST (Two One-Sided Tests, [Walker and Nowacki, 2011]), confidence interval approaches, and Bayesian methods. In summary, while comparative statistical tests focus on identifying significant differences between groups, equivalence tests aim to establish whether the differences between groups are within a predefined range of equivalence. Both tests play crucial roles in biostatistics, helping researchers draw meaningful conclusions about the effects of different treatments or interventions in clinical and biomedical research.

Equivalence analysis is crucial in medical practice for assessing the comparability or similarity of two or more medical interventions, treatments, or formulations. This type of analysis is essential in various clinical and research settings for the following reasons:

- Comparative Effectiveness Research: Equivalence analysis helps determine whether two treatments or interventions have similar efficacy and safety profiles.Evaluating the relative benefits and risks of different treatment options and informing clinical

decision-making is essential to ensure that patients receive the most appropriate and effective interventions.

- Generic Drug Evaluation: Equivalence analysis is commonly used to evaluate the bioequivalence of generic drugs compared to their brand-name counterparts. It helps to determine whether generic versions of a drug produce similar blood concentrations and therapeutic effects as the original product, ensuring the safety and efficacy of generic drug substitution.

- Clinical Trials and Research Studies: Equivalence analysis is essential in clinical trials and research studies to establish the comparability of different interventions or treatments. It enables researchers to assess whether a new treatment is as effective and safe as an existing standard of care, allowing for the evaluation of novel therapies and interventions in a rigorous and standardized manner.

- Regulatory Approval and Drug Development: Equivalence analysis is critical in the regulatory approval process for new drugs and medical devices. It helps regulatory agencies determine whether a new product is therapeutically equivalent to existing treatments and meets the required standards for safety and efficacy before it can be approved for use in clinical practice.

- Pharmacovigilance and Post-Market Surveillance: Equivalence analysis is valuable for monitoring medical products' post-market safety and effectiveness. It allows for the ongoing assessment of the equivalence of different formulations or variations of a drug to ensure that any changes in manufacturing processes or product formulations do not compromise its safety or efficacy.

By conducting rigorous and comprehensive equivalence analyses, healthcare professionals, regulators, and researchers can ensure that medical interventions and treatments meet established standards for efficacy, safety, and quality, promoting evidence-based decision-making and enhancing patient care and safety.

6.1 equiv_med: a library for equivalence assessment

While Python has gained significant traction in the scientific community, some specialized areas of biostatistics may have more comprehensive support in other programming languages, particularly those that have been historically popular in the field, such as R and MATLAB.

In the conference paper [Nascimben and Rimondini, 2022], a novel Python library targeting equivalence testing has been proposed to the Python users community. This library is freely downloadable at the following address `github.com/m89p067/equiv_med`. An overview of all the funcionality provided by the library is shown in the following scheme in Figure 6.1. TOST, which stands for Two One-Sided Tests, is a method used in statistical hypothesis testing, particularly in equivalence testing. The first column of Figure 6.1 is dedicated to the methods developed to run the TOST procedure which

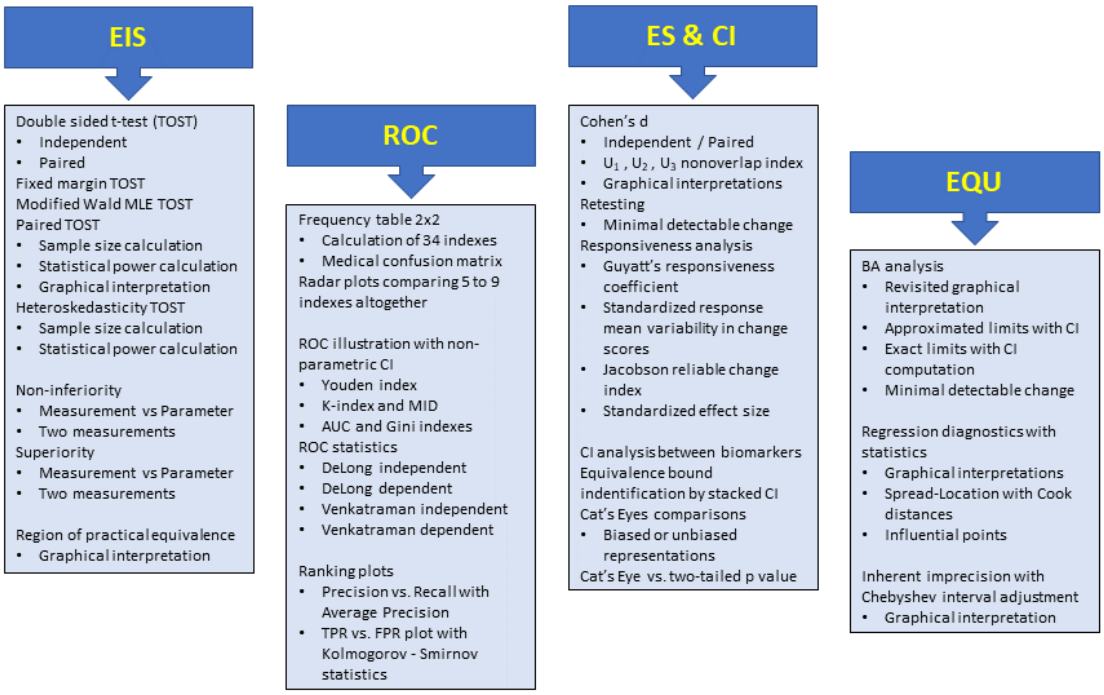

Figure 6.1: Overview of the four numerical experiments for GED analysis in bladder cancer (from [Nascimben et al., 2021]).

determines whether the difference between two groups or treatments is within a specified range that is considered practically insignificant or clinically unimportant. This method is commonly used in clinical trials and other biomedical research to assess whether a new treatment is not substantially different from a standard treatment, i.e., whether the new treatment is equivalent to the standard. The TOST procedure involves conducting two separate one-sided hypothesis tests. First, a lower equivalence bound is set, and a one-sided test is performed to determine whether the observed effect is greater than this lower bound. Next, an upper equivalence bound is set, and another one-sided test is conducted to ascertain whether the observed effect is less than this upper bound. If the results of both tests suggest that the observed effect is greater than the lower bound and less than the upper bound, the null hypothesis of equivalence is accepted. This method helps researchers establish the equivalence of treatments by testing whether the difference between them is within a predefined range that is deemed practically irrelevant. TOST provides a rigorous approach to assessing equivalence and is widely used in evaluating new treatments, drugs, or interventions in biostatistics and clinical research.

Receiver Operating Characteristic (ROC) curves are typically used to assess the performance of a binary classification model, such as a diagnostic test or a predictive model, by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings. While ROC curves are primarily used to evaluate the dis-

criminatory power of a test, they are not commonly used to establish equivalence directly. However, ROC curves can still be indirectly helpful in assessing the equivalence of diagnostic tests or classifiers, especially in cases where the focus is on determining whether two tests have similar discriminatory abilities. In medical analysis, several indexes can be calculated from a 2x2 frequency table, which is commonly used to summarize the results of binary classification tests. A 2×2 table typically includes information on the presence or absence of a condition or disease and the results of a diagnostic test or screening tool. From this table, various performance metrics and indices can be derived to assess the accuracy, reliability, and effectiveness of the diagnostic test. Some important indexes calculated from a 2x2 frequency table in medical analysis include:

1. Sensitivity: Sensitivity (also known as the true positive rate) is the proportion of actual positive cases correctly identified by the test. It is calculated as true positives divided by the sum of true positives and false negatives.

2. Specificity: Specificity (also known as the true negative rate) is the proportion of actual negative cases correctly identified by the test. It is calculated as true negatives divided by the sum of true negatives and false positives.

3. Positive predictive value (PPV): PPV is the probability that a positive test result indicates the presence of the condition. It is calculated as true positives divided by the sum of true and false positives.

4. Negative predictive value (NPV): NPV is the probability that a negative test result indicates the absence of the condition. It is calculated as true negatives divided by the sum of true and false negatives.

5. Accuracy: Accuracy is the overall proportion of correct classifications the test makes. It is calculated as the sum of true positives and true negatives divided by the total number of cases.

6. Likelihood ratios: Likelihood ratios, including the positive likelihood ratio and the negative likelihood ratio, provide information on how much a test result will change the odds of having the condition compared to not having the condition.

7. Youden's J statistic: Youden's J statistic is the sum of sensitivity and specificity minus one. It is used to determine the optimal cut-off point for a diagnostic test.

These indexes help assess the performance of a diagnostic test or screening tool in correctly identifying the presence or absence of a particular condition or disease. They play a crucial role in evaluating the effectiveness and reliability of medical tests in clinical practice and research.

Cohen's d is a measure of effect size that quantifies the standardized difference between two means. It is commonly used when comparing the means of two groups or treatments. While Cohen's d is not typically used directly for assessing equivalence, it can provide information about the magnitude of the difference between two measurements or groups, which can be relevant in determining whether the difference is practically significant.

When considering equivalence between measurements using Cohen's d, one would typically assess whether the magnitude of the difference falls within a predefined range that is considered practically negligible. Using confidence intervals to assess equivalence between measurements involves comparing the intervals to a predefined equivalence margin. The general approach is explained below:

1. Calculate the confidence intervals: Calculate the confidence intervals for the measurements or groups under consideration. Typically, this involves computing the confidence intervals for the means or differences between the means.

2. Define the equivalence margin: Determine the range or margin of equivalence considered practically significant or acceptable in the study context. This margin should reflect the level of difference that is deemed negligible or insignificant.

3. Compare the intervals with the equivalence margin: Assess whether the confidence intervals fall entirely within the predefined equivalence margin. If the intervals are entirely contained within the equivalence margin, this suggests that the measurements or groups are practically equivalent.

4. Interpret the results: Based on the comparison between the confidence intervals and the equivalence margin, make conclusions regarding the equivalence of the measurements or groups. If the confidence intervals fall within the equivalence margin, one can infer that the measurements are equivalent within the predefined range.

It is important to note that the choice of the equivalence margin is critical and should be determined based on the context of the study and the specific criteria for establishing practical equivalence. Additionally, using confidence intervals for assessing equivalence is more indirect than dedicated equivalence testing procedures like the TOST (Two One-Sided Tests) approach. However, confidence intervals can provide valuable information about the precision of the estimates and can offer insights into whether the measurements or groups are practically equivalent.

Another pillar of medical equivalence analysis is the Bland–Altman visualization [Sedgwick, 2013], which is primarily used to assess the agreement or the level of agreement between two quantitative measurement methods. While it is not typically used for formal equivalence testing, it can provide insights into the degree of agreement between two measurement techniques, which could be relevant for evaluating the equivalence of the methods in specific contexts. The code required to perform Bland–Altman is shown below and the visual output in Figure 6.2; additionally, it could perform exact limits of agreements and sample size evaluation as in [Jan and Shieh, 2018].

```
from equiv_med.EQU import eq_BA
my_BA=eq_BA.BA_analysis(var1,var2) # var1,var2 are two measurements from two distinct devices
#Bland-Altman plot
my_BA.run_analysis() # default 95% of the difference will lie in this interval [revised plot]
# Evaluate sample size and assurance probability of exact agreement limist
#Exact limits of agreement sample size
out1=my_BA.exact_Bound_sample_size(mu1,sigma1,len(var1),95,0.05)
```

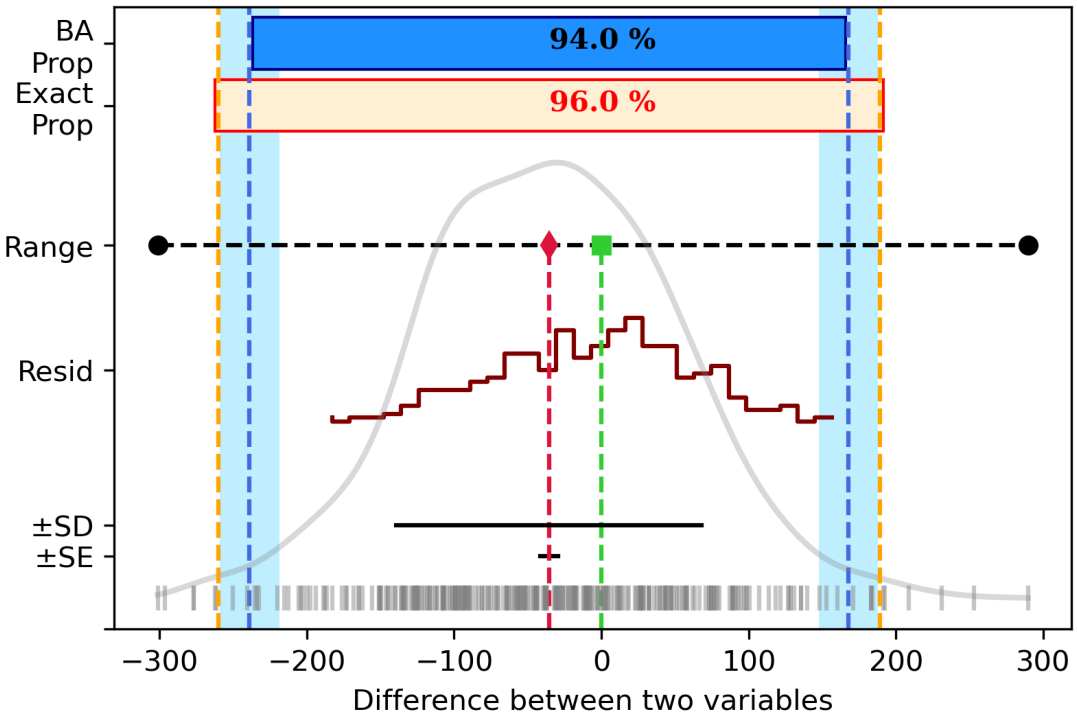

Figure 6.2: Alternative design of the Bland-Altman plot as found in the `BA_analysis` class, using function `run_analysis`.

```
out2=my_BA.exact_Bound_sample_size(mu2,sigma2,len(var2),95,0.05)
#Exact limits of agreement assurance
out3=my_BA.exact_Bound_assurance(mu1,sigma1,len(var1),95,0.05,0.9)
out4=my_BA.exact_Bound_assurance(mu2,sigma2,len(var2),95,0.05,0.9)
#In case of repeated measures
my_BA.minimal_detectable_change() #output also Minimal Detectable Change
```

Indeed, visual interpretation enhances the overall statistical analysis process by facilitating data exploration, pattern recognition, effective communication, quality control, model assessment, and hypothesis generation. It allows analysts to uncover meaningful insights from data and communicate those insights clearly and compellingly. For this reason, all Python functions were coded to produce novel graphs to enhance the display of the statistical outcomes.

The Table details 6.1 the composition of each folder in the GitHub repository, and the operations the Python functions could perform. Among the developed functions, two of them allow the user to perform statistics directly on the ROC curves. DeLong's method is a statistical approach used to compare the areas under two correlated ROC curves [Sun and Xu, 2014]. It is commonly applied when there is a need to assess whether two diagnostic tests or biomarkers have significantly different discriminatory abilities. DeLong's method considers the correlation between the ROC curves and provides a

statistical test for comparing the areas under the curves. Moreover, Venkatraman's method is a variation of the DeLong, and it is particularly relevant in survival analysis or studies where time-to-event data is involved and the ROC curves are based on time-dependent outcomes [Venkatraman and Begg, 1996].

Folder	Operations	Characteristics
EIS	TOST	Independent or paired two one-sided t-tests
EIS	TOST	Fixed margin $\delta = f \times \sigma_{Ref}$ equivalence test
EIS	TOST	Modified Wald with maximum likelihood estimator
EIS	TOST	Paired inputs, sample size and statistical power dedicated functions
EIS	TOST	Heteroscedastic inputs, sample size and statistical power dedicated functions
EIS	Non-inferiority	Measurement vs. parameter or two-sample statistics
EIS	Superiority	Measurement vs. parameter or two-sample statistics
EIS	ROPE	Region of practical equivalence by equally-tailed or highest density intervals
ROC	2×2 freq. table	Confusion matrix and derivation of 34 performance indexes
ROC	Radar plots	Circular graphs to compare cross-table's indexes (single/paired/bars) between two treatments
ROC	ROC visual interpr.	ROC computation, Youden index, k-index, MID, non-parametric CI
ROC	ROC Statistics	DeLong and Venkatraman independent or dependent methods
ROC	Ranking plots	False positive rate vs. true positive rate with statistics and precision vs. recall with AUC
ES	Cohen's d	Independent or paired inputs, non-overlapping indexes
ES	Re-testing	Repeated measurements' minimal detectable change
ES	Responsiveness	Guyatt coeff., standardized response mean, effecct size, normalized ratio, reliable change index
ES	Exploratory stats	Equivalence estimation based on CI analysis between biomarkers
CI	Margin value study	Stacked CI representations to study the optimal equivalence margin
CI	Paired Cat's Eyes	Biased or unbiased representation of two biomarkers
CI	Car's Eye vs. p value	Single biomarker visual two-tailed analysis of CI vs. p significance
EQU	BA analysis	Revisited graphical interpretation, approximated and exact limits, min. detectable change
EQU	Regress. diagnostics	Residuals interpretation, spread-location plot with Cook distances, influential points graphs
EQU	Inherent imprecision	Graphical inherent imprecision with Chebyshev interval adjustment

Table 6.1: Overview of the methodologies implemented in the Python functions

Another unusual aspect the Python library contains is the implementation of a Bayesian approach to statistical equivalence. Bayesian statistics and traditional (frequentist) statistics are two distinct approaches to statistical inference and data analysis, each with its own set of principles, methodologies, and interpretations. The key differences between the two are highlighted below:

- **Probability Interpretation:**
  - Bayesian Statistics: In Bayesian statistics, probability is interpreted as a degree of belief or subjective uncertainty. Prior knowledge or beliefs about the parameters of interest are combined with the observed data to form a posterior probability distribution, representing updated beliefs after data analysis.
  - Traditional Statistics: In traditional statistics, probability is interpreted as the long-run frequency of events based on repeated sampling. It does not involve the notion of prior beliefs, and the focus is primarily on the observed data and the likelihood of the data given the parameters.
- **Parameter Estimation:**
  - Bayesian Statistics: Bayesian inference involves the use of prior distributions, likelihood functions, and Bayes' theorem to update prior beliefs into posterior distributions, which represent the updated knowledge about the parameters of interest.
  - Traditional Statistics: Traditional statistics primarily rely on point estimates (e.g., maximum likelihood estimates) and confidence intervals derived from the observed data without considering prior information.
- **Uncertainty Representation:**
  - Bayesian Statistics: Bayesian analysis explicitly quantifies uncertainty through probability distributions, allowing for a more comprehensive representation of uncertainty in parameter estimation and model predictions.
  - Traditional Statistics: Traditional statistics generally focuses on point estimates and standard errors, which provide limited information about the estimates' uncertainty.
- **Hypothesis Testing:**
  - Bayesian Statistics: Bayesian hypothesis testing involves comparing the relative probabilities of different hypotheses based on the observed data and prior beliefs, often using measures such as Bayes factors or posterior probabilities.
  - Traditional Statistics: Traditional hypothesis testing relies on p-values and significance levels, which indicate the strength of evidence against a null hypothesis under the assumption that the null hypothesis is true.

The Python library offers the calculations of the "region of pratical equivalence" (ROPE) using the formulation proposed by [Kruschke and Liddell, 2018]; the following code snippet could be used to run such Bayesian statistical approach for equivalance between measurements assessment.

```
from equiv_med.EIS import ROPE_test
out2=ROPE_test.ROPE(var1, rope_range=[-0.3,0.3]) #user-defined ROPE region [-0.3,0.3]
out2.rope_calc()
```

6.2 Final remarks

The introduction of a Python library for visual understanding of medical-related statistical tests targeting various aspects of bioequivalence is a valuable contribution to medical research and analysis. By providing a free alternative to commercial software, this library enables researchers, practitioners, and analysts to access advanced visualization tools and automated functions for interpreting output parameters in bioequivalence studies. The inclusion of minimal working examples further enhances the usability of the library and supports the reproducibility of results, contributing to the transparency and rigor of scientific research. The emphasis on producing enhanced graphs to facilitate the interpretation of output parameters is particularly commendable, as effective visualization can significantly aid in understanding complex statistical analyses and communicating findings to diverse audiences, including clinicians, researchers, and stakeholders. The commitment to continuous improvement and expansion of the methodologies implemented in the library promises ongoing advancements in bioequivalence research and analysis. By focusing on generating visual insights, the library is poised to play a crucial role in advancing medical understanding and decision-making processes. The availability of open-source tools that support the visual exploration of statistical results is essential for promoting transparency, collaboration, and innovation within the scientific community. This Python library can potentially empower researchers and analysts to gain deeper insights into bioequivalence, contributing to the advancement of medical knowledge and the development of evidence-based practices.

7 Regenerative medicine: biomaterials production tracking

Original contribution to knowledge

This chapter is based upon the article
Mauro Nascimben, Ilijana Kovrlija, Janis Locs, Dagnija Loca, and Lia Rimondini. Fusion and classification algorithm of octacalcium phosphate production based on xrd and ftir data. *Scientific Reports*, 14(1):1489, 2024

Regenerative medicine is a multidisciplinary field that focuses on developing techniques to replace, repair, or regenerate human cells, tissues, or organs to restore or establish normal function. The goal of regenerative medicine is to harness the body's natural healing processes or to create new solutions, such as tissue engineering and stem cell therapy, to treat diseases and injuries that currently have limited or no effective treatment options. Key components and approaches within regenerative medicine include:

- Stem Cell Therapy: Involves using stem cells, which can differentiate into various cell types, to repair or replace damaged or diseased tissues. Stem cell therapy holds promise for treating conditions such as spinal cord injuries, heart disease, and neurodegenerative disorders.
- Tissue Engineering: Focuses on creating functional biological substitutes to replace or repair damaged tissues or organs. Tissue engineering combines cells, biomaterials, and biochemical factors to construct artificial organs, tissues, or scaffolds that can integrate with the patient's body and promote natural regeneration.
- Biomaterials and Scaffold Design: Involves the development and use of biocompatible materials, such as synthetic polymers, hydrogels, and biodegradable scaffolds, to support cell growth, tissue regeneration, and organ transplantation. These materials provide structural support and promote cell attachment and growth during regeneration.
- Organ Transplantation and Replacement: Focuses on developing techniques for transplanting or replacing damaged or diseased organs with healthy donor organs, bioengineered organs, or artificially created organ constructs. Regenerative medicine aims to address the shortage of donor organs and improve the outcomes of organ transplantation procedures.

- Gene Therapy and Cellular Reprogramming: Involves using genetic engineering techniques to modify or reprogram cells to restore their normal function or promote tissue regeneration. Gene therapy can be applied to correct genetic defects, promote therapeutic protein production, or stimulate damaged tissue regeneration.

In regenerative medicine, tracking the production of biomaterials is essential for ensuring the quality, safety, and efficacy of the materials used in tissue engineering, organ transplantation, and other regenerative therapies. Several vital reasons highlight the importance of biomaterial production tracking in regenerative medicine:

1. Quality Control and Assurance: Tracking the production process of biomaterials allows rigorous quality control measures to be implemented at each manufacturing stage. High quality helps ensure that the biomaterials meet established standards for purity, biocompatibility, mechanical properties, and sterility, reducing the risk of adverse reactions or complications in patients.

2. Traceability and Accountability: The ability to trace the production history of biomaterials enables manufacturers to identify and address any issues or discrepancies that may arise during the manufacturing process. It also facilitates accountability in cases where product defects or failures need to be investigated to prevent future occurrences.

3. Compliance with Regulatory Standards: Biomaterials used in regenerative medicine are subject to strict regulatory requirements to ensure patient safety and product efficacy. Production tracking allows manufacturers to demonstrate compliance with regulatory standards and facilitates the timely submission of documentation and data required for product approval and market authorization.

4. Batch Consistency and Reproducibility: Tracking the production parameters and variables for biomaterial batches enables manufacturers to achieve consistent product quality and reproducibility across multiple production runs. It ensures uniformity in product performance and therapeutic outcomes, enhancing regenerative medicine interventions' reliability and predictability.

5. Product Development and Optimization: Detailed production tracking data can provide insights into the impact of various manufacturing parameters on the characteristics and performance of biomaterials. This information can be leveraged to optimize production processes, improve product design, and enhance the functionality and biocompatibility of biomaterials for specific regenerative medicine applications.

6. Post–Market Surveillance and Safety Monitoring: Continuous tracking of biomaterials production data supports post-market surveillance efforts, allowing manufacturers to monitor product performance, detect potential adverse events, and implement corrective actions or product recalls if necessary. This proactive approach to safety monitoring helps ensure patient well-being and fosters confidence in using regenerative medicine products.

By implementing robust tracking systems for biomaterials production, manufacturers can uphold the highest quality, safety, and efficacy standards in regenerative medicine, thereby promoting the advancement and widespread adoption of these innovative therapies for patient care.

7.1 Octacalcium phosphate production

Octacalcium phosphate (OCP, [Kovrlija et al., 2021]) is a calcium phosphate compound that is an intermediate phase in hydroxyapatite formation, the main mineral component of human bone. OCP has a chemical formula of $Ca_8H_2(PO_4)_6 \cdot 5H_2O$ and is a precursor to hydroxyapatite formation during bone mineralization and tooth development. It is considered a bioresorbable material, meaning it can be broken down and resorbed by the body over time. In biomaterials and biomedical engineering, Octacalcium phosphate has gained attention due to its biocompatibility and potential applications in bone regeneration, dental materials, and as a biomimetic material for the repair and restoration of bone defects. The properties that make it a valuable material in these applications include its ability to promote osteoconductivity, its biodegradability, and its similarity to the mineral composition of natural bone. Octacalcium phosphate (OCP) can be synthesized through various methods, including precipitation, hydrothermal synthesis, and other chemical processes. The specific manufacturing process for Octacalcium phosphate may vary depending on the desired characteristics of the final product and the intended applications. Quality control measures are essential throughout the manufacturing process to ensure the production of high-quality and consistent Octacalcium phosphate. To produce OCP, two precursor solutions are needed (calcium and phosphate-containing solutions), and they are processed separately. The calcium solution is typically prepared by dissolving a calcium salt in water, such as calcium nitrate or calcium hydroxide. The phosphate solution is prepared in water by dissolving a phosphate salt, such as ammonium phosphate or sodium phosphate. The prepared calcium and phosphate solutions are then mixed under controlled conditions, typically at specific temperatures and pH levels. The mixing process allows the ions to react and form the desired Octacalcium phosphate compound. Coupling the calcium and phosphate solutions leads to the precipitation of Octacalcium phosphate crystals. The control of reaction conditions, including temperature, pH, and stirring rate, is essential to ensure the formation of pure and uniform OCP crystals. The resulting OCP crystals are separated from the solution through filtration. The collected solid product is then washed thoroughly to remove any impurities or unreacted chemicals from the surface of the crystals. The washed OCP crystals are dried under controlled conditions to remove any remaining moisture. Depending on the intended application, the dried OCP powder may undergo further processing, such as milling or sieving, to achieve the desired particle size and morphology.

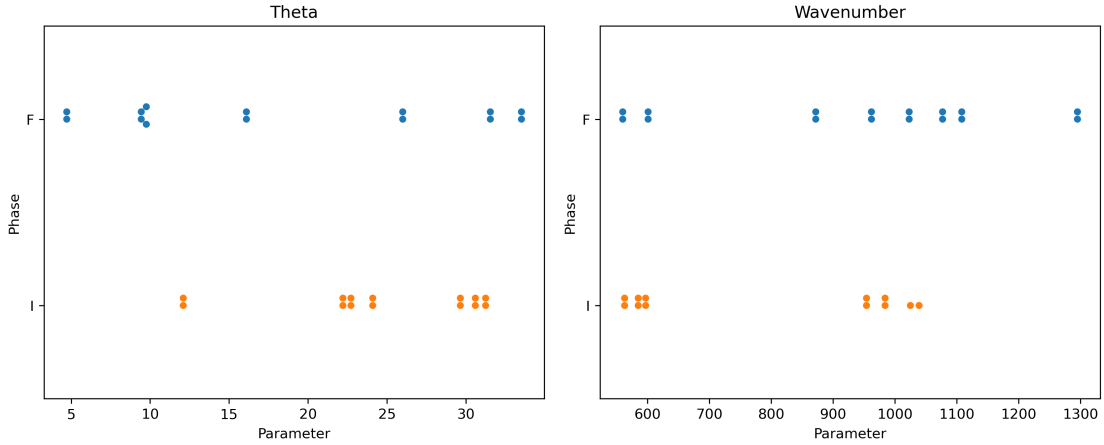

Figure 7.1: On the left: Class membership of XRD θ angles, on the right: Class membership of FTIR wavenumber values.

7.2 ML for OCP production tracking

The paper [Nascimben et al., 2024] describes the implementation of an automated analysis sequence aimed at developing a decision support system for tracking the synthesis of octacalcium phosphate (OCP) from alpha-tricalcium phosphate (α–TCP) over time. The process involves the fusion of X-ray diffraction (XRD) and Fourier-transform infrared (FTIR) signals from a scaled-up hydrolysis of OCP from [Kovrlija et al., 2023], followed by curve fitting based on established maxima from the literature and the extraction of nine features from the fitted shapes. X–ray diffraction is a powerful analytical technique used in material science to study the structure of crystalline materials. When X–rays interact with a crystalline material, they undergo constructive and destructive interference, resulting in a diffraction pattern that provides valuable information about the material's atomic arrangement and crystallographic properties. Fourier–transform infrared spectroscopy (FTIR) is an analytical technique widely used in materials science for the identification and characterization of various materials based on their molecular composition. FTIR spectroscopy measures the absorption of infrared light by a sample, providing information about the functional groups and chemical bonds present within the material. In the manuscript, we provided a way to merge features extracted from the XRD and FTIR signals to build a more robust model to track OCP production phases. The relation between OCP production phases and XRD or FTIR values of each detected peak is shown in Figure 7.1. In the image "phase" stands for initial or final OCP synthesis, whereas "parameter" means XRD angle or FTIR wavenumber.

The sequence of operations enclosed machine learning techniques for feature ranking, spatial filtering, and dimensionality reduction, aiding in automatically recognizing different synthesis stages (Figure 7.2). Another innovative aspect to aid class separability was the application of an ad–hoc spatial filtering technique. Spatial filters are a preprocessing technique used in machine learning and signal processing to enhance the separation

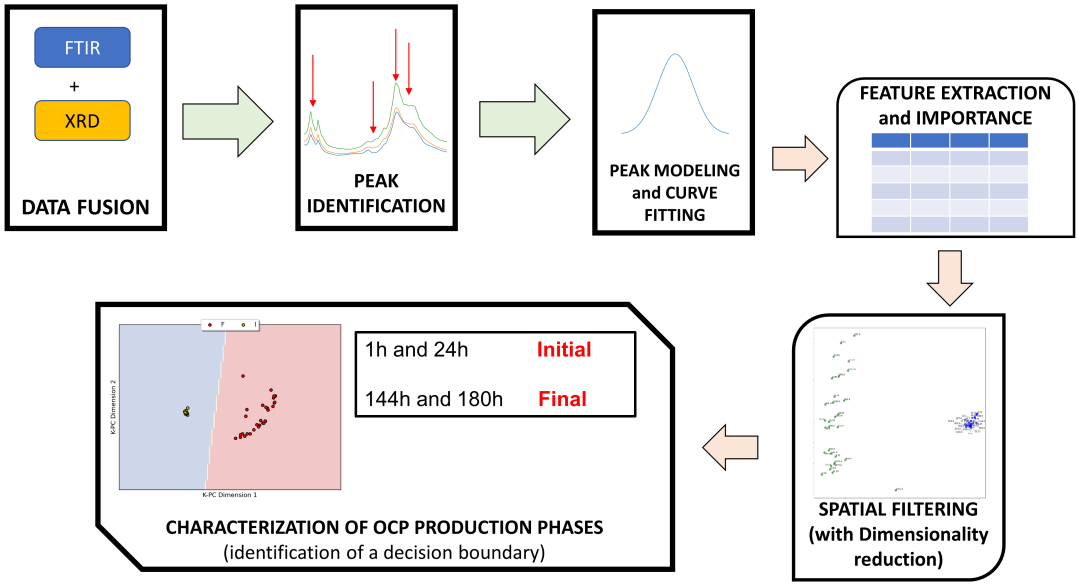

Figure 7.2: Overview of the proposed pipeline in [Nascimben et al., 2024].

between different classes or patterns in data. Spatial filters aim to extract features or patterns relevant to discrimination or classification; by enhancing the separability of different classes or patterns in the data, spatial filters can significantly improve the performance of various machine learning models. The reduced variability due to the application of spatial filtering transforming one class helps produce two clusters representing the distinct production phases of OCP. Indeed, the two distributions overlap without spatial filtering, and it could be very difficult (or impossible) for an algorithm to discriminate the two OCP production phases. On the contrary, after applying spatial filtering in a one–versus–rest configuration, one class shows reduced variability over the Cartesian plane compared to the not-filtered other; this expedient facilitates the discrimination of the two OCP production stages, and one algorithm could be trained to learn the two stages and score the XRD or FTIR samples automatically applying a theoretical decision boundary.

7.3 Final remarks

The proposed analysis pipeline for OCP identification represents a promising foundation for a decision support system explicitly tailored for OCP synthesis monitoring. This methodology in the future could be leveraged for tracking OCP production over time, including the intermediary phases involved in OCP formation. Additionally, another prospective aim is to explore the integration of biological variables with biomaterial properties to construct a comprehensive model of tissue response to the implant. Overall, the

manuscript highlights the potential of the developed analysis pipeline in facilitating the monitoring and understanding of OCP synthesis, paving the way for enhanced decision-making in the context of biomaterial development and tissue engineering applications.

8 Proteomics: anomaly expression identification

Original contribution to knowledge

This chapter is based upon the article (submitted)
Mauro Nascimben, Hugo Abreu, Marcello Manfredi, Annalisa Chiocchetti, and Lia Rimondini. Latent expression of extracellular vesicles proteins in doped bioactive glasses through machine learning–based mass–spectometry data analysis. International Journal of Molecular Sciences, Submitted
and the conference presentation
Mauro Nascimben. Anomaly detection of EV-related protein expression in doped bioactive glasses. Novara, Italy, Oct 2023b. Italian Chemical Society, 3rd International Proteomics And Metabolomics Conference

Protein anomaly detection in proteomics is crucial for identifying abnormal or atypical protein behaviors, expressions, or modifications that may indicate underlying pathological conditions, disease states, or cellular dysfunctions [Tiwari et al., 2022; Tadepalli et al., 2020]. Several reasons highlight the importance of protein anomaly detection in proteomics:

- Biomarker Discovery: Anomaly detection helps identify proteins that exhibit significant deviations from their normal expression patterns, providing insights into potential disease biomarkers. By detecting abnormal protein expressions or modifications, researchers can identify candidate biomarkers indicating specific diseases or physiological changes, facilitating early disease detection and personalized treatment approaches.

- Disease Pathogenesis Understanding: Anomaly detection in proteomics contributes to a better understanding of the molecular mechanisms underlying various diseases and disorders. By identifying aberrant protein activities or expressions associated with specific pathological conditions, researchers can unravel the complex signaling pathways and molecular interactions involved in disease pathogenesis, aiding in developing targeted therapeutic interventions.

- Drug Target Identification: Protein expression or function anomalies can highlight potential drug targets for developing novel therapeutics. By pinpointing proteins

that play critical roles in disease progression or pathophysiological processes, proteomics anomaly detection can guide the selection of specific molecular targets for drug discovery and the design of precision medicine approaches tailored to individual patient profiles.

- Personalized Medicine: Proteomics anomaly detection enables the identification of patient-specific protein profiles and aberrant molecular signatures, facilitating the customization of treatment strategies and therapeutic interventions based on an individual's unique proteomic profile. This personalized medicine approach allows selecting targeted therapies more likely to be effective and well-tolerated by patients, leading to improved treatment outcomes and patient care.

- Diagnostic and Prognostic Applications: Anomaly detection in proteomics can be instrumental in producing diagnostic tests and prognostic indicators for various diseases. By detecting specific protein anomalies associated with disease onset, progression, or response to treatment, proteomics can aid in the development of reliable diagnostic tools and prognostic markers that enable early disease detection, accurate disease staging, and the prediction of treatment outcomes.

By leveraging advanced proteomics techniques and anomaly detection algorithms, researchers can gain a deeper understanding of the complex protein dynamics underlying health and disease, paving the way for obtaining innovative diagnostic tools, targeted therapies, and personalized treatment approaches in clinical practice.

8.1 Biomaterials' proteomics in extracellular vesicles

Extracellular vesicles (EV, [Abreu et al., 2021]) play essential roles in mediating intercellular communication and the transfer of bioactive molecules between cells. When exposed to biomaterials, such as those used in medical devices, implants, or tissue engineering scaffolds, EVs can interact with the material surfaces, leading to various biological responses. Several factors can influence these responses, including the type of biomaterial, its surface properties, and the specific cell types involved. Some of the key ways in which extracellular vesicles may react to biomaterials include:

- Adhesion and Uptake: EVs can adhere to the surfaces of biomaterials and may be taken up by cells in the local microenvironment. The surface properties of the biomaterial can influence the adhesion and internalization of EVs, potentially affecting the cellular response to the biomaterial.

- Biocompatibility and Inflammation: Interaction with biomaterials can trigger the release of EVs from immune cells and other cell types, leading to the modulation of the local inflammatory response. The biocompatibility of the biomaterial can influence the nature and extent of the inflammatory response, which in turn may affect the behavior of EVs in the surrounding tissue.

- Regulation of Cellular Processes: EVs released from cells in response to biomaterials can carry various bioactive molecules, including proteins, nucleic acids, and lipids, which can regulate cellular processes such as proliferation, differentiation, and tissue regeneration. The cargo carried by EVs can influence the cellular response to the biomaterial and contribute to tissue repair and regeneration.

- Immune Modulation: EVs released from immune cells in response to biomaterials can mediate immune modulation by transferring immunomodulatory molecules and antigens to other cells. This process can affect the local immune response to the biomaterial and may influence the overall biocompatibility and long-term integration within the host tissue.

Understanding how extracellular vesicles react to biomaterials is crucial for developing biocompatible materials, designing improved medical devices, and advancing tissue engineering approaches. By elucidating the intricate interplay between EVs and biomaterials, researchers can develop strategies to optimize the biocompatibility of biomaterials, promote tissue regeneration, and minimize adverse immune reactions, ultimately leading to improved clinical outcomes for patients receiving biomaterial-based interventions. Proteomics in extracellular vesicles involves the study of the proteins carried within these small, membrane-bound vesicles released by cells into the extracellular environment. Extracellular vesicles play crucial roles in cell-to-cell communication, signal transduction, and the transfer of biological molecules between cells. Proteomics studies focused on extracellular vesicles aim to characterize the protein content within these vesicles, allowing for a better understanding of their roles in various physiological and pathological processes.

8.2 Application of anomaly detection to EV-related protein expression

The investigation in [Nascimben, 2023b] involved the analysis of extracellular vesicle protein content derived from mesenchymal stem cells cultured on various bioactive glasses using mass spectrometry. Mass spectrometry in proteomics has revolutionized the study of complex protein mixtures, allowing researchers to identify and quantify proteins with high sensitivity and specificity. During the proteomics analysis, the researchers primarily relied on statistical analysis based on p–values to evaluate the significance of differences in protein expression levels between the experimental groups. The p–value, in this context, represents the probability of observing a test statistic as extreme as, or more extreme than, the one calculated from the actual data, assuming there is no real difference between the compared groups. In simpler terms, the p-value helps determine the likelihood that observed differences in protein expression are due to random chance alone. Lower p-values indicate that the observed differences are less likely to result from chance and are more likely to result from a genuine effect. However, it is essential to note that a small p-value does not confirm the presence of a significant difference between groups; it merely suggests that the observed data is improbable if the null hypothesis (no difference in protein

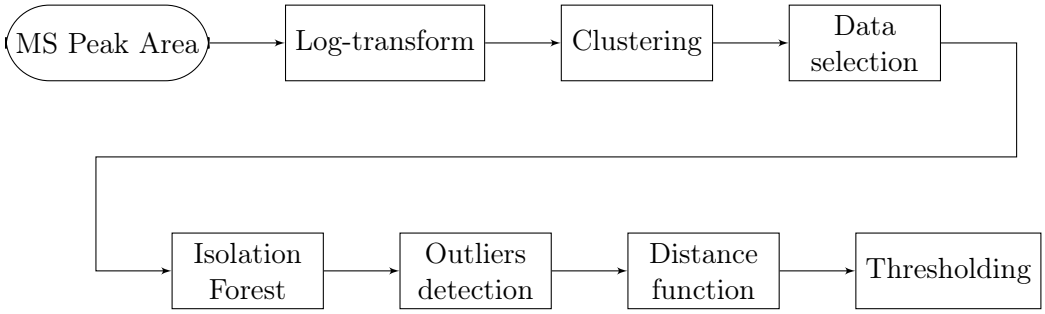

Figure 8.1: The proposed procedure workflow

expression between the control and treatment groups) were true. Moreover, statistical significance is a concept not related to biological or clinical significance [Ranganathan et al., 2015]. Statistical significance is indeed influenced by the study's sample size, among other factors. When conducting hypothesis tests or analyzing data, researchers typically use statistical significance to determine whether an observed effect is genuine or simply due to random chance. Sample size plays a crucial role in determining statistical significance because it affects the power of a statistical test [Ioannidis, 2008]. Power refers to the probability that a test will correctly reject the null hypothesis when it is false. A larger sample size generally increases the power of a test, making it more likely to detect a true effect if one exists. In contrast, smaller sample sizes can lead to broader confidence intervals and less precise estimates, making detecting small or subtle effects harder; this can fail to see statistically significant differences, even if they truly exist in the population. The problem could be connected with inadequate statistical power to detect meaningful differences or excessive statistical power to detect differences that are not biologically meaningful [Bhardwaj et al., 2004].

To offer an alternative approach to the traditional statistical analysis, due to the limited effect size demonstrable with a small sample [De Winter, 2019; Ioannidis, 2005], we considered anomaly detection techniques. These techniques are aimed at identifying a restricted set of EV-related proteins that exhibit substantial changes in behavior compared to the majority of the data. Proteins displaying anomalous behavior that contradicts most of the data might serve as potential indicators of underlying biological phenomena occurring between the experimental conditions. Using anomaly detection techniques in this context allowed us to pinpoint specific proteins that deviate significantly from the norm, providing insights into potentially crucial biological processes or responses under experimental conditions. By focusing on these anomalies, the ultimate goal was to gain a deeper understanding of the complex mechanisms and interactions underlying the effects of different bioactive glasses on the protein content of extracellular vesicles derived from mesenchymal stem cells.

8.2.1 Wet-lab experimental conditions

The experimental setup consisted of the culture of 5000 cells of each of the three donors independently, on top of several different bioactive glasses (and the respective control conditions), at 37°C, 5%CO_2, for seven days. At the endpoint, the supernatants were collected for EVs isolation through ultracentrifugation at 100000 x g for 2 hours at 4°C. The pellet enriched in EVs was then resuspended in 500uL of Phosphate Buffer Saline (PBS 1X), and the EVs protein content was evaluated through mass spectrometry. The initial data was from three donors and contained the mass spectrum peak area for the samples of each participant. The following experimental conditions were tested:

- cell cultures on "SBA2", "SBA3", and "STe0" are not modified bioactive glasses (i.e., controls or ctrl).
- cell cultures on "AgSBA2", "CuSBA3", and "STe5" modified bioactive glasses doped with silver, copper, and tellurium, respectively (i.e., doped).
- cell culture on "Plastic", a baseline condition without the presence of biomaterials (i.e., plast)

The laboratory experiments aimed at establishing protein content modifications: those accurring between the doped glasses and the "plastic" condition could be a consequence of the presence of the bioactive glass. Furthermore, protein expression altered between the doped glasses and the respective control glass should be due to the metal ion doping [Taye, 2022].

It should be remarked that there were no experimental differences between donors (culture conditions, number of wells, time-point, cell density, etc.).

8.2.2 Mass spectrum summary

Sample processing for MS analysis and data collection was conducted at the Mass Spectrometry unit of the University of Piemonte Orientale (Novara, Italy). Proteins extracted from uEVs were quantified using BCA assay (Pierce BCA protein assay kit; ThermoFisher Scientific). Samples were denaturated with TFE, reduced in DTT 200 mM, and alkylated with IAM 200 mM before complete tryptic digestion with 2 mg of Trypsin/Lys-C (Promega, Madison, WI, USA). Digested peptides were desalted on the Discovery® DSC-18 solid phase extraction (SPE) 96-well Plate (25 mg/well) (Sigma-Aldrich Inc., St. Louis, MO, USA) and vacuum evaporated to be reconstituted with 20 mL of 0.05% formic acid in water. Trypsin-digested sample proteins were analyzed with a microLC Eksigent Technologies (Eksigent Technologies, Dublin, CA, USA) system that included a micro LC200 Eksigent pump with flow module 5-50 µL, interfaced with a 5600+ TripleTOF system (Sciex, Concord, ON, Canada) equipped with DuoSpray Ion Source and CDS (Calibrant Delivery System). The stationary phase was a Halo C18 column (0.5 x 100 mm, 2.7 μm; Eksigent Technologies, Dublin, CA, USA). The mobile phase was a mixture of 0.1% (v/v) formic acid in water (A) and 0.1% (v/v) formic acid in acetonitrile (B), eluting at a flowrate of 15.0 μL min^{-1} at an increasing concentration of solvent B

from 2% to 40% in 30 min. For identification purposes, the samples were subjected to a data-dependent acquisition (DDA): the mass spectrometer analysis was performed using a mass range of 100-1500 Da (TOF scan with an accumulation time of 0.25 s), followed by an MS/MS product ion scan from 200 to 1250 Da (accumulation time of 5.0 ms) with the abundance threshold set at 30 cps (35 candidate ions can be monitored during every cycle). For the label-free quantification, the samples were subjected to cyclic data independent analysis (DIA) of the mass spectra using a 25-Da window: the mass spectrometer was operated such that a 50-ms survey scan (TOF-MS) was performed, and subsequent MS/MS experiments were performed on all precursors. These MS/MS experiments were performed cyclically using an accumulation time of 40 ms per 25-Da swath (36 swaths) for a total cycle time of 1.5408 s. By using the rolling collision energy, the ions were fragmented for each MS/MS experiment in the collision cell. The MS data were acquired with Analyst TF 1.7 (Sciex, Concord, ON, Canada). Two DDA and three DIA acquisitions were performed. The DDA files were searched using Protein Pilot software v. 4.2 (Sciex, Concord, ON, Canada) and Mascot v. 2.4 (Matrix Science Inc., Boston, MA, USA). The UniProt Swiss-Prot reviewed database containing human proteins (version 01/02/2018, containing 42271 sequence entries) was used, and a target-decoy database search was performed. The probability of peptide assignments was corrected with False Discovery Rate set at 1%.

Mass spectrum peak area

In mass spectrometry, the peak area measures the intensity of a mass spectral peak, summarizing the number of ions contributing to that peak. It is proportional to the abundance of the ions with a specific mass-to-charge ratio in the sample, allowing for quantitative analysis. A larger peak area indicates a higher abundance of ions with that particular mass-to-charge ratio in the sample, while a smaller peak area indicates a lower abundance.

8.2.3 Dry-lab experimental sequence

The proposed workflow starting from the raw mass spectrum peak area involved the following steps:

1. The raw values are **log-transformed**

2. The log-transformed values were clustered (using **OPTICS**) and the values of the same cluster taken to ensure analysis on similar data representing the same biological phenomena

3. Each cluster value from the 3 donors was labelled as *outlier* (potential "anomality" or extreme variation) or not applying **Isolation Forest** (machine learning technique)

4. Selected the common proteins marked as *outliers* inside each condition and computed the **metric distance** to identify abnormal variations in the EV-related protein expression

as exemplified on Figure 8.1.

Proposed sequence: log-transformation

The log transformation of the raw peaks was performed because mass spectrometry data can have a wide range of intensities, and some peaks might be much larger than others due to various factors such as instrument variability, sample concentration, and ionization efficiency. Indeed, log transformation helps normalize the data by compressing the dynamic range and making smaller peaks more visible. Additionally, log transformation can reduce the impact of random noise in the data. Noise often contributes more to the lower intensity peaks, and by taking the logarithm, the noise is dampened, making it easier to distinguish valid signals from noise.

Proposed sequence: clustering

OPTICS (Ordering Points To Identify the Clustering Structure) is a data clustering algorithm used in machine learning to identify natural clusters and their hierarchies in a dataset [Ankerst et al., 1999]. It is handy for datasets with varying densities, irregular shapes, and noise. OPTICS is an extension of the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm, which aims to discover clusters based on the density of data points. In OPTICS, two main parameters are to be evaluated: core (minPts) and reachability distances (epsilon). The concept of "reachability distance" means that for a data point P, the reachability distance to another data point Q is defined as the maximum distance between P and Q, such that P can be directly reached from Q while staying within a predefined neighborhood size. Instead, a data point's core distance is the smallest such that there are at least a certain number of points within that distance, forming a dense region around the point. In the current investigation, a reachability parameter of 0.05 and a minPts parametere of 50 were applied.

The effect of clustering all log-transformed peak values is shown in Figure 8.2. By employing only values found in the blue cluster, the analysis focused on finding aberrant proteins inside a group with similar expression profiles, excluding proteins markedly belonging to other clusters that probably portray different biological phenomena (or artifacts). Indeed, among the different biological activities depicted in Figure 8.2, the values inside the blue cluster might represent the more relevant biological phenomena. However, the proposed pipeline could be employed over other clusters in the same way.

Proposed sequence: outlier detection by Isolation Forest

Isolation Forest is a machine-learning algorithm for anomaly detection and outlier identification [Liu et al., 2008]. Isolation Forest conceptualizes that anomalies are usually rare

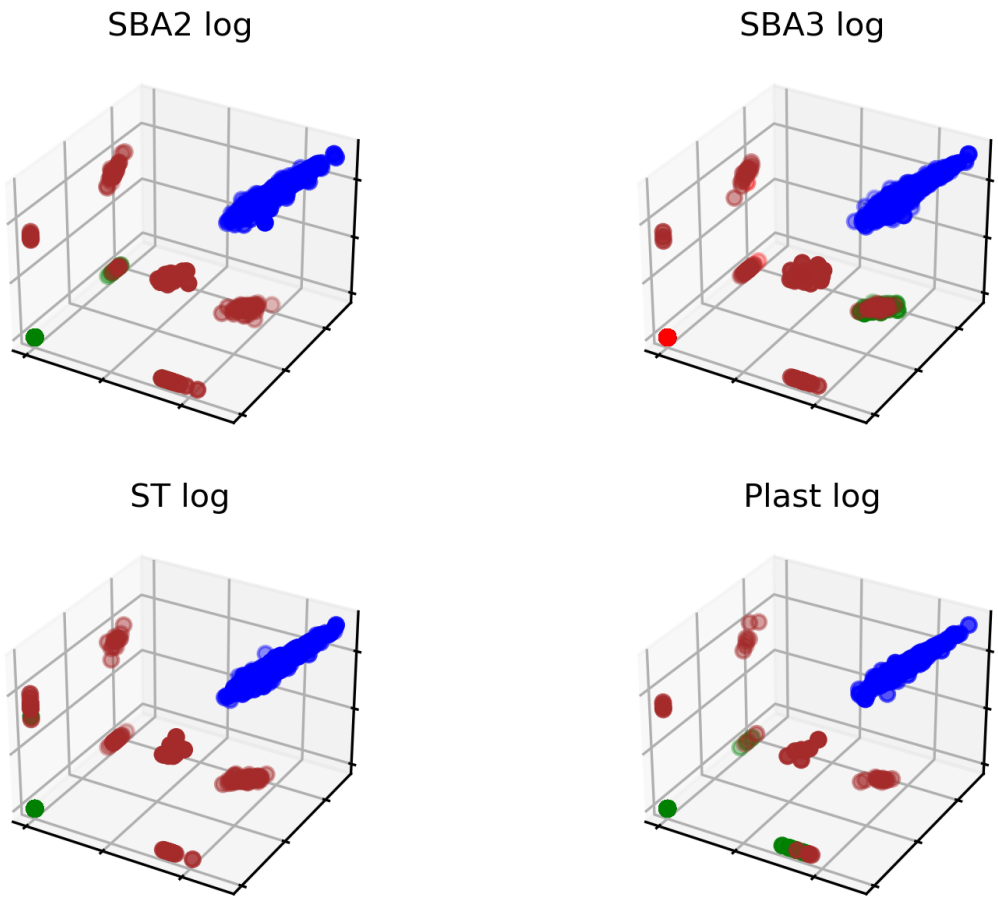

Figure 8.2: All data from the three donors and each experimental conditions underwent automatic labeling to retain only uniformly distributed values for further analysis. Only the values in the blue cluster were used in the next steps of the experimental sequence.

instances that can be "isolated" more quickly than regular instances. The algorithm constructs a binary tree-like structure in which each internal node represents a feature and a split point. In contrast, each leaf node represents an isolated instance or an anomaly. To detect anomalies, the algorithm calculates the path length from the root of the tree to the leaf where a data point resides. Anomalies are expected to have shorter paths because they are isolated more quickly. The average path length of a data point across all trees in the forest is used as a score of atypical expression. Smaller average path lengths indicate higher anomaly scores.

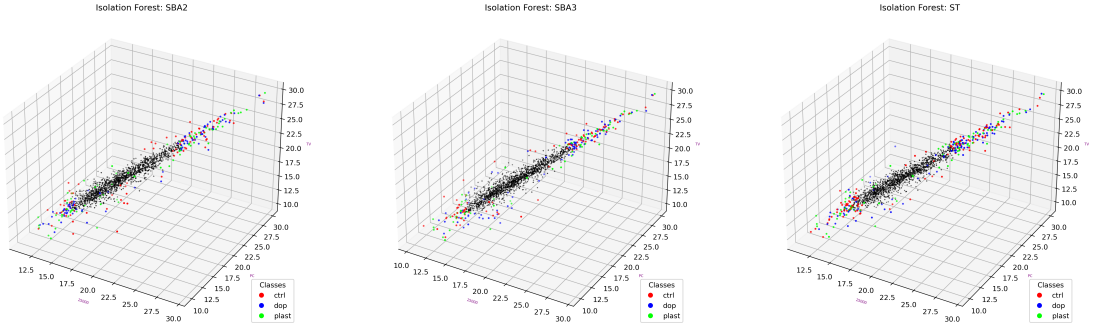

Figure 8.3: Values from the three donors marked as black dots were considered *inliers*, thus close to each other, in the three experimental conditions (plast, control, and doped) by the Isolation Forest algorithm, whereas colored points were those showing more relevant changes (aka “outliers”).

As depicted in the Figure 8.3, the Isolation Forest identified a set of core values in the distribution that are close to each other: these values were marked as black dots, and could be considered those with similar peak area. To find proteins showing extremely changing behavior between experimental conditions, only the colored values were retained for being considered *outliers* by the algorithm. Among the *outliers*, the possibility to find proteins with unusual expression in the three experimental conditions might be high. Note that the “plast” condition was the same in all the three graphs.

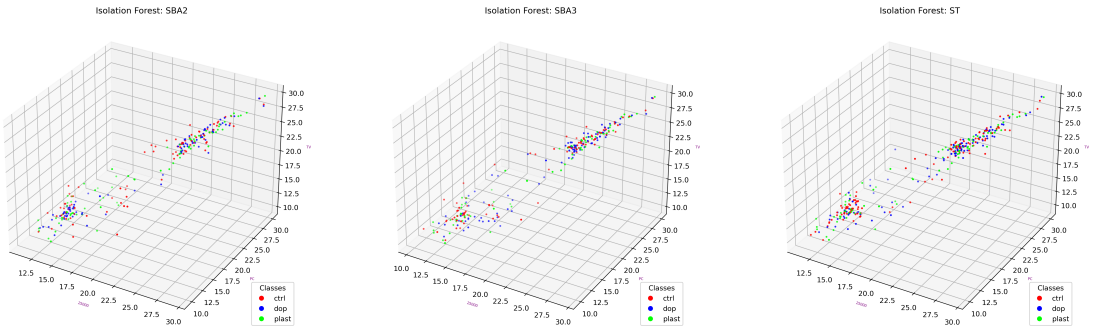

Figure 8.4: Retaining only the values considered *outliers* by Isolation Forest. The three axes represent the three donors.

The Figure 8.4 reports only the values of protein expression kept for further analysis.

Proposed sequence: Distance function

In [Nascimben, 2023b], the metric function was the Mahalanobis distance as general framework. It can be thought of as a measure of how many standard deviations away a particular point is from the mean of a distribution after adjusting for correlation among variables. Mathematically, the Mahalanobis distance D between a point x and a distribution with mean μ and covariance matrix Σ is given by:

$$D^2 = (x - \mu)^T \Sigma^{(-1)} (x - \mu)$$

It measures the similarity or dissimilarity between the point and the distribution: if the distance is small, the point is close to the distribution whereas if it is large, it is far away. One of the main advantages of using Mahalanobis distance is that it considers the relationships among variables, which can be particularly useful in cases where variables are correlated. Among all proteins marked as outliers, the Mahalanobis distances were computed to identify abnormal variations in the EV-related protein expression, and values thresholded considering proteins whose variations were above the average±one standard deviation the Mahalanobis distance of the whole set of proteins; these results gathered proteins based on this threshold.

However in [Nascimben et al., Submitted] (a forthcoming manuscript) the procedure has been changed to employ an Euclidean metric, more suitable for a three-dimensional space made by three subjects or donors. This is the sample size typically encountered in biological experiments involving data from a few cell lines [Lazic et al., 2018]. Euclidean distance might be a good option when operating on a three-dimensional donor space to evaluate the single subject's actual values.

8.3 Results

In [Nascimben, 2023b], Venn diagrams were created to summarize the common proteins in the three experimental conditions. A Venn diagram is a graphical representation of the relationships between different groups or sets of items. It consists of overlapping circles, each representing a set, and the overlapping regions show the common elements of those sets. Figure 8.5 contains the proteins arranged as a word cloud, thus with font size proportional to the magnitude of the protein expression changes. The character size on the right panel of Figure 8.5 was fixed for visualization purposes and does not reflect protein expression modifications. The Control versus doped condition could not find shared proteins, meaning each experimental condition activated a peculiar set of aberrant proteins with specific characteristics. Instead, the Plastic versus doped condition shared a few proteins, with one highly over- or under-expressed in all laboratory preparations (right art of Figure 8.5), the protein P22413 linked to bone mineralization was identified in all experimental conditions as highly active.

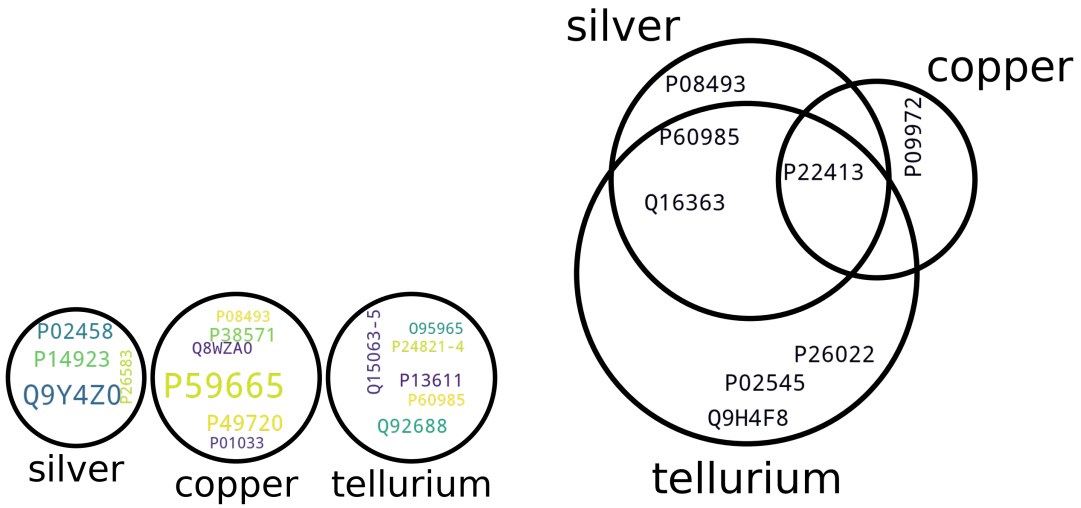

Figure 8.5: On the left: Doped vs. Ctrl, on the right: Doped vs. Plast

8.4 Final remarks

Machine learning techniques can be applied to proteomic data for anomaly detection, including supervised and unsupervised learning. Supervised algorithms can be trained on labeled data to classify proteins as normal or abnormal, while unsupervised algorithms can detect patterns or outliers without predefined labels. Moreover, unsupervised clustering methods can help identify groups of proteins with similar expression profiles. Anomalies may manifest as proteins that do not fit into the expected clusters. The reported work introduces a possible analysis sequence targeting abnormal protein expression between experimental conditions.

While in [Nascimben, 2023b], a general procedure has been proposed, a forthcoming analysis ([Nascimben et al., Submitted]) will be tested specifically to work on cell lines coming from three donors, which is a common sample size in biological experiments.

9 Conclusions and future perspectives

Machine learning plays a significant role in precision medicine, offering data-driven insights and predictive models that can tailor medical treatments to individual patients. By incorporating machine learning into precision medicine, healthcare professionals can leverage data-driven insights to tailor therapies and interventions to individual patients, leading to more effective, targeted, and personalized healthcare solutions. Machine learning is used to analyze and interpret genomic data, including DNA sequencing, gene expression profiles, and epigenetic information; it allows researchers and clinicians to identify genetic mutations, variations, and molecular signatures associated with diseases and drug responses. Another critical task ML models can execute is helping in categorizing patients into subgroups based on their genetic, clinical, and molecular profiles. This patient stratification enables the identification of patient-specific treatment options and the optimization of therapeutic strategies. Through ML algorithms, biomarkers discovery could be achieved to identify potential biomarkers and molecular signatures associated with disease risk, progression, or response to treatment. These biomarkers can guide early diagnosis, prognosis, and targeted therapy selection. Machine learning is also used to predict drug–target interactions, drug-drug interactions, and the potential of existing drugs for new therapeutic purposes; it accelerates drug discovery and the identification of personalized treatment options. As evaluated also in the current thesis, machine learning assists healthcare providers in making data-driven clinical decisions. It can help predict patient outcomes, assess treatment responses, and recommend personalized treatment plans, all while considering individual patient characteristics. As part of the radiomics field, ML is applied to medical imaging, such as radiological images and histopathology slides. It aids in the automated detection, classification, and quantification of anomalies, enhancing diagnostic accuracy and treatment planning. Modern bio–banks offer large amounts of electronic health records that ML models can analyze to extract valuable clinical insights, identify risk factors, predict disease trajectories, and support decision-making in real-time clinical settings. In so-called smart-hospital environments, ML can be used for remote patient monitoring, wearable device data analysis, and early warning systems to improve disease management, adherence to treatment plans, and patient engagement. Another powerful aspect is the possibility of Data Integration through Machine learning to merge and harmonize diverse healthcare data sources, including genomic data, electronic health records, imaging data, and patient-reported data: it facilitates a comprehensive view of patients' health profiles. Machine learning has the potential to optimize clinical trials by assisting in the design and execution of clinical trials to identify eligible patient populations, predicting patient recruitment rates, and optimizing trial protocols for more efficient and cost-effective research. A last remark is about ethical and privacy considerations. Machine learning in precision medicine must address ethi-

cal concerns related to patient consent, data privacy, transparency, and the responsible use of AI models. Ethical considerations are critical to building trust and safeguarding patient rights.

The works illustrated throughout this thesis demonstrate how machine learning offers numerous benefits in precision medicine, revolutionizing the approach to healthcare and patient treatment. It could be possible to develop personalized treatment plans because machine learning models can analyze large–scale patient data, including genomic information and clinical records, to tailor treatment plans to each patient's unique characteristics, optimizing therapeutic efficacy and minimizing adverse effects. Examples have been provided in Section 5: one study established a novel pipeline to predict survival rate employing a dataset of gene expression (hub and seed) related to bladder cancer [Nascimben et al., 2021]. The numerical experiments tested pipelines including a prebinning followed by a primary discretizer or a primary discretizer followed by refinement or optimization of the levels. Also, machine learning algorithms can detect subtle patterns in patient data that indicate the early onset of diseases, allowing for timely interventions and proactive disease management, potentially leading to better treatment outcomes. In [Nascimben et al., 2022b], a novel pipeline enclosed tree embedding and manifold dimensionality reduction to produce graph-like forecasts exposing peculiar patterns suitable for extending patient categorization into six classes (three grades of tumor severity and two classes for overall survival) in an unsupervised fashion. The inclusion of cancer staging supports medical decisions regarding prognosis and treatment. As explained in Section 4 and 3, ML could be exploited for targeted therapy selection by analyzing patient-specific data; ML can help identify the most effective treatment options, including targeted therapies and precision drugs that are tailored to the molecular profiles of individual patients, thereby improving treatment response rates. Furthermore, the same techniques could enhance drug development, accelerating the drug discovery process by predicting the efficacy and safety of potential drug candidates, identifying new drug targets, and facilitating the repurposing of existing drugs for novel therapeutic purposes, ultimately developing more effective and targeted medications. The works published as [Nascimben and Rimondini, 2023] explored public-domain toxicological datasets, each one evaluated in separate numerical experiments through specific SNNs. All SNNs had in common the neuronal model, the leaky integrate-and-fire, and received the molecules' structures encoded as binary fingerprints. Toxicity of the compounds was determined by the number of spikes fired by the last two neurons, assigning it to the neuron that fired more spikes. Improved versions of the original SNN were applied to bioacculumation prediction in [Nascimben et al., 2023c] for a three class prediction, and in [Nascimben, 2023a] investigating other neuron models. As shown in Section 4, machine learning in precision medicine can contribute to improved patient outcomes by enabling more accurate disease prognosis, better treatment planning, and the early identification of potential complications, thereby reducing the overall burden of the disease on patients and healthcare systems. The study in [Nascimben et al., 2023a] merged clinical factors from 294 women who underwent axillary dissection for breast cancer, gathering 23 clinical variables and metadata from anonymized subjects to predict the risk of developing upper-arm lymphedema. The variables included patient characteristics, macroscopical

cancer features, anatomopathological cancer attributes, surgical outcomes, and medical therapies. In the proposed approach, the ordinal and the binary variables were modeled separately in two distinct UMAP models. After obtaining the two UMAP models, the final representation was a single low-dimensional embedding that merged the two UMAP graphs by intersection. UMAP tries to preserve the local and global information contained in the input variables and supports merging distinct models by intersection, union, or subtraction. The procedure associated each patient to a low or high risk of BCRL occurrence. Also, in Section 4, software was tested for faster clinical decision-making with tools providing rapid and data-driven insights that assist healthcare providers in making timely and informed clinical decisions, facilitating proactive and personalized patient care, especially in critical care and emergency medicine scenarios. The papers [Nascimben et al., 2022a, 2023b] offered free-to-use computational methods to aid healthcare professionals for lymphedema management. The upper arm volumetry methods were deployed as a software the calculates limb volumes and surfaces based on 3D laser scans; it is made up of three apps, each one with peculiar features and computational capabilities downloadable from Zenodo. Section 5 also discussed how ML models can be employed for cost efficiency and resource optimization [Nascimben, 2021a]. By optimizing treatment plans, minimizing unnecessary interventions, and reducing trial-and-error approaches, machine learning in precision medicine can lead to cost savings for healthcare providers and improved allocation of healthcare resources, ensuring that resources are used where they are most needed. The general benefits of ML approaches extend to the analysis of complex biological and clinical data, leading to a better understanding of disease mechanisms, identifying novel biomarkers, and developing innovative research approaches, thereby advancing our knowledge of various diseases and their treatments. The Section 6, a copyright-free Python library was released to allow scientists perform analysis targeting bioequivalence [Nascimben and Rimondini, 2022]. The source code of the functions has been uploaded to GitHub and archived on Zenodo. Installation of the package was made possible directly from GitHub through pip. Comparative statistical tests could not address the interchangeability of measurements obtained from different laboratory devices or the similarity between two treatments. Analyzing the equivalence means reversing the null hypothesis of standard biostatistical testing by validating the alternative hypothesis of no difference between measurements. The importance of this topic is particularly relevant for the medical sector, especially for the biopharmaceutical industry, with guidelines for therapeutic equivalence between drugs established by regulatory agencies like USA Food and Drug Administration. In Chapters 7 and 8, ML has been employed on wet laboratory data. Analyzing wet laboratory data with ML involves extracting meaningful insights and predictions from experimental data generated in a laboratory setting. The Chapter 7, the analysis pipeline has been studied for categorizing samples into “initial” or “final” stages of OCP production. This substance is well-equipped to act as a bone replacement in regenerative medicine, and the approach proposed enclosed several innovative aspects. It merged two heterogeneous sources of data: XRD and FTIR. XRD is beneficial for investigating the crystalline structure of materials, providing information about the arrangement of atoms in the crystal lattice. On the other hand, FTIR is valuable for analyzing the chemical composition of mate-

rials, focusing on identifying functional groups and molecular bonds. Both techniques are complementary and often used in material science to comprehensively understand the properties and characteristics of a wide range of materials. A spatial filter has been created to categorize OCP production stages from the merged dataset to reduce the variance of one class versus the other; this solution improved the discriminative power of the algorithm, easing the selection of a decision boundary. In Section 8, an analysis sequence has been conceived to identify unusual or unexpected observations in the protein expression data [Nascimben et al., 2024]. If the sample size of a proteomic dataset is limited, it is crucial to manage statistics accurately. A novel analysis sequence was proposed in [Nascimben, 2023b; Nascimben et al., Submitted] to assess protein expression from the subjects' actual values by utilizing machine learning anomaly detection techniques. This proposed procedure may support and aid researchers in evaluating findings when using fewer cell lines in their experiments by detecting abnormal protein behavior in the dataset. The researcher's focus on individual proteins in various ways beyond standard statistical testing can offer additional evidence or uncover concealed aspects of the experimental design. The methodology has been validated in an experimental environment where the EV protein content of MSC cultured on three bioactive glasses, doped or not with metallic ions, has been examined. The process identified a subset of proteins that displayed highly varying behavior between experimental conditions: the peculiar set of abnormal proteins each metal activates describes the effect of ion doping. Conversely, comparing doped biomaterials and the baseline plastic scaffold led to a common set of proteins. Some proteins were statistically significant at the t-test, while others had a high variance pattern between experimental conditions. This technique's additional information on the data being examined may provide a more in-depth understanding of the experiment and the results it produces.

9.1 Future perspectives

The future perspectives of machine learning in precision medicine are promising and point towards transformative advancements in healthcare delivery, patient outcomes, and disease management. Some innovative future perspectives include:

- Integration of Multi-Modal Data: Machine learning will enable the seamless integration and analysis of diverse healthcare data, including genomics, proteomics, metabolomics, imaging data, and real-time patient-generated data, facilitating a comprehensive and holistic understanding of individual health profiles.

- Real–Time Predictive Analytics: Machine learning models will evolve to provide real-time predictive analytics and decision support, enabling healthcare providers to anticipate patient health risks, intervene proactively, and prevent disease progression, leading to more effective and timely interventions.

- Enhanced Patient Engagement and Empowerment: Future machine learning applications will focus on enhancing patient engagement and empowerment by provid-

ing personalized health insights, predictive risk assessments, and self-management tools, fostering a proactive approach to disease prevention and health maintenance.

- AI-Driven Drug Discovery and Development: Machine learning will revolutionize the drug discovery process by enabling the rapid screening of vast chemical libraries, predicting drug-target interactions, and accelerating the development of novel therapeutics and precision medicines tailored to specific patient populations.

- Augmented Clinical Decision–Making: Machine learning will assist healthcare professionals by providing augmented intelligence tools for complex clinical decision-making, treatment planning, and risk prediction, facilitating more informed and personalized patient care delivery.

- Population Health Management: Machine learning will play a crucial role in population health management by enabling the analysis of large-scale health data to identify disease trends, public health risks, and healthcare disparities, thereby informing public health policies and interventions to improve community health outcomes.

- AI–Enabled Telemedicine and Remote Monitoring: Machine learning will drive the development of AI-enabled telemedicine platforms and remote patient monitoring systems, enabling the delivery of virtual healthcare services, remote diagnostics, and continuous patient monitoring, especially in underserved or remote areas.

- Ethical and Regulatory Framework Development: Future perspectives of machine learning in precision medicine will also involve the development of robust ethical guidelines, privacy protection measures, and regulatory frameworks to ensure responsible and ethical use of AI-driven technologies, safeguarding patient rights and data privacy.

- Advancements in Computational Biology and Systems Medicine: Machine learning will advance computational biology and systems medicine by modeling complex biological systems, network analysis, and simulating disease processes, fostering a deeper understanding of the molecular mechanisms underlying health and disease.

By embracing these future perspectives, the integration of machine learning in precision medicine will pave the way for a patient-centric, data-driven healthcare paradigm that prioritizes personalized, proactive, and evidence-based approaches to disease prevention, diagnosis, and treatment.

The works presented throughout the current thesis demonstrated the broad applicability of ML principles and techniques to the various fields of precision medicine, contributing to developing this research topic.

9.2 Personal Bibliography

9.2.1 Chemoinformatics

Mauro Nascimben and Lia Rimondini. Molecular toxicity virtual screening applying a quantized computational SNN–based framework. *Molecules*, 28(3):1342, 2023

Mauro Nascimben, Silvia Spriano, Lia Rimondini, and Manolo Venturin. Molecular fingerprint based and machine learning driven QSAR for bioconcentration pathways determination. In Gabriella Bretti, Roberto Natalini, Pasquale Palumbo, and Luigi Preziosi, editors, *Mathematical Models and Computer Simulations for Biomedical Applications*, pages 193–215, Cham, 2023c. Springer Nature Switzerland. ISBN 978-3-031-35715-2

Mauro Nascimben. Molecular fingerprint based and machine learning driven QSAR for bioconcentration pathways determination. Rome, Italy, Sept 2021b. National Research Council of Italy, Virtual Workshop of Mathematical Modelling and Control for Healthcare and Biomedical Systems

Mauro Nascimben. Quantized computational QSAR framework for molecular toxicity virtual screening. Rome, Italy, July 2022. Sapienza University of Rome, 3rd Molecules Medicinal Chemistry Symposium - Shaping Medicinal Chemistry for the New Decade

Mauro Nascimben. Virtual screening by spiking neural networks: a case study on cytochrome P450. Padua, Italy, Sept 2023a. University of Padua, 18th Conference on Computational Intelligence Methods for Bioinformatics and Biostatistics

Mauro Nascimben. Low–power or resource–constrained environments for virtual screening and quantitative structure-activity relationship analysis for in silico precision medicine. Rome, Italy, Sept 2023c. Italian Association for Industrial Research, NanoInnovation Conference and Exhibition

9.2.2 Clinical precision medicine

Mauro Nascimben, Lorenzo Lippi, Nicola Fusco, Marco Invernizzi, and Lia Rimondini. A software suite for limb volume analysis applicable in clinical settings: upper limb quantification. *Frontiers in Bioengineering and Biotechnology*, 10:863689, 2022a

Mauro Nascimben, Lorenzo Lippi, Nicola Fusco, Alessandro de Sire, Marco Invernizzi, and Lia Rimondini. Technical aspects and validation of custom digital algorithms for hand volumetry. *Technology and Health Care*, 31(5):1835–1854, 2023b

Mauro Nascimben, Lorenzo Lippi, Alessandro De Sire, Marco Invernizzi, and Lia Rimondini. Algorithm–based risk identification in patients with breast cancer–related lymphedema: A cross-sectional study. *Cancers*, 15(2):336, 2023a

Lorenzo Lippi, Alessio Turco, Stefano Moalli, Mauro Nascimben, Claudio Curci, Alessandro de Sire, Lia Rimondini, and Marco Invernizzi. Quantitative assessment of upper-limb volume: Implications for lymphedema rehabilitation? *Applied Sciences*, 13(17):9810, 2023b

Lorenzo Lippi, Mauro Nascimben, Alessandro de Sire, Arianna Folli, Nicola Fusco, Lia Rimondini, and Marco Invernizzi. A novel free–to–use software for upper limb volume quantification in breast cancer related lymphedema: implementing cutting–edge technology in the individualized therapeutic approaches of breast cancer survivors. *Cancer Research*, 83(Supplement 5):P5–08–18–P5–08–18, 03 2023a. ISSN 0008–5472. doi: 10.1158/1538-7445.SABCS22-P5-08-18. URL `https://doi.org/10.1158/1538-7445.SABCS22-P5-08-18`

9.2.3 Bioinformatics

Mauro Nascimben, Manolo Venturin, and Lia Rimondini. Double–stage discretization approaches for biomarker-based bladder cancer survival modeling. *Communications in Applied and Industrial Mathematics*, 12(1):29–47, 2021

Mauro Nascimben, Lia Rimondini, Davide Corà, and Manolo Venturin. Polygenic risk modeling of tumor stage and survival in bladder cancer. *BioData Mining*, 15 (1):23, 2022b

Mauro Nascimben. A machine learning based decision support system in oncology. Parma, Italy, Sept 2021a. University of Parma, 2020+2021 Italian Society of Applied and Industrial Mathematics (SIMAI) Conference

9.2.4 Biostatistics

Mauro Nascimben and Lia Rimondini. Visually enhanced Python functions for clinical equality of measurement assessment. *Annals of Computer Science and In-*

formation Systems, 32:241–249, 2022

9.2.5 Regenerative medicine

Mauro Nascimben, Ilijana Kovrlija, Janis Locs, Dagnija Loca, and Lia Rimondini. Fusion and classification algorithm of octacalcium phosphate production based on xrd and ftir data. *Scientific Reports*, 14(1):1489, 2024

9.2.6 Proteomics

Mauro Nascimben, Hugo Abreu, Marcello Manfredi, Annalisa Chiocchetti, and Lia Rimondini. Latent expression of extracellular vesicles proteins in doped bioactive glasses through machine learning–based mass–spectometry data analysis. International Journal of Molecular Sciences, Submitted

Mauro Nascimben. Anomaly detection of EV-related protein expression in doped bioactive glasses. Novara, Italy, Oct 2023b. Italian Chemical Society, 3rd International Proteomics And Metabolomics Conference

Mauro Nascimben
February 2024

Acknowledgements

I am grateful to Professor Lia Rimondini for her invaluable supervision, support, and tutelage during the course of my Ph.D. degree. My gratitude extends to the European Commission for the funding opportunity to undertake my studies under grant No. 860462, Horizon 2020 Research and Innovation program. Additionally, I greatly appreciate Dr. Manolo Venturin's treasured support, which influenced my experimental methods and helped me discuss my results. I also thank Prof. Andrea Cochis, Prof. Davide Corà, and Prof. Marco Invernizzi for their mentorship. I want to thank Prof. Janis Locs, Premurosa project mates (in particular Hugo Abreu, and Ilijana Kovrlija for the fruitful collaboration), colleagues at Enginsoft, and the UPO research team for a cherished time working together on our research goals (my sincere appreciation to Dr Lorenzo Lippi for the extraordinary teamwork), and in social settings. My appreciation includes my family and friends for their encouragement and support throughout my studies.

Bibliography

Hugo Abreu, Elena Canciani, Davide Raineri, Giuseppe Cappellano, Lia Rimondini, and Annalisa Chiocchetti. Extracellular vesicles in musculoskeletal regeneration: modulating the therapy of the future. *Cells*, 11(1):43, 2021.

Gary An, Qi Mi, Joyeeta Dutta-Moscato, and Yoram Vodovotz. Agent-based models in translational systems biology. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 1(2):159–171, 2009.

Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2):49–60, 1999.

Associazione italiana oncologia medica. I numeri del cancro in italia. `https://www.aiom.it/i-numeri-del-cancro-in-italia/`, 2023.

A. M. Attanayake, S. S. Perera, and S. Jayasinghe. Phenomenological modelling of covid-19 epidemics in sri lanka, italy, the united states, and hebei province of china. *Computational and Mathematical Methods in Medicine*, 2020:1–15, 2020. doi: 10.1155/2020/6397063.

Noam Auslander, Ayal B Gussow, and Eugene V Koonin. Incorporating machine learning into established bioinformatics frameworks. *International journal of molecular sciences*, 22(6):2903, 2021.

Rohit Batra, Le Song, and Rampi Ramprasad. Emerging materials intelligence ecosystems propelled by machine learning. *Nature Reviews Materials*, 6(8):655–678, 2021.

Brett K Beaulieu-Jones, William Yuan, Gabriel A Brat, Andrew L Beam, Griffin Weber, Marshall Ruffin, and Isaac S Kohane. Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? *NPJ digital medicine*, 4(1):62, 2021.

Sachin S Bhardwaj, Fabian Camacho, Amy Derrow, Alan B Fleischer, and Steven R Feldman. Statistical significance and clinical relevance: the importance of power in clinical trials in dermatology. *Archives of Dermatology*, 140(12):1520–1523, 2004.

Jacob G Calcei and Scott A Rodeo. Orthobiologics for bone healing. *Clinics in sports medicine*, 38(1):79–95, 2019.

Yue Cao, Thomas Andrew Geddes, Jean Yee Hwa Yang, and Pengyi Yang. Ensemble deep learning in bioinformatics. *Nature Machine Intelligence*, 2(9):500–508, 2020.

Paula Carracedo-Reboredo, Jose Liñares-Blanco, Nereida Rodríguez-Fernández, Francisco Cedrón, Francisco J Novoa, Adrian Carballal, Victor Maojo, Alejandro Pazos, and Carlos Fernandez-Lozano. A review on machine learning approaches and trends in drug discovery. *Computational and structural biotechnology journal*, 19:4538–4558, 2021.

Hongming Chen, Thierry Kogej, and Ola Engkvist. Cheminformatics in drug discovery, an industrial perspective. *Molecular Informatics*, 37(9-10):1800041, 2018.

Wei-Hao Chen, Win-San Khwa, Jun-Yi Li, Wei-Yu Lin, Huan-Ting Lin, Yongpan Liu, Yu Wang, Huaqiang Wu, Huazhong Yang, and Meng-Fan Chang. Circuit design for beyond von neumann applications using emerging memory: From nonvolatile logics to neuromorphic computing. In *2017 18th International Symposium on Quality Electronic Design (ISQED)*, pages 23–28. IEEE, 2017.

Elizabeth A Chrischilles, Danielle Riley, Elena Letuchy, Linda Koehler, Joan Neuner, Cheryl Jernigan, Brian Gryzlak, Neil Segal, Bradley McDowell, Brian Smith, et al. Upper extremity disability and quality of life after breast cancer treatment in the greater plains collaborative clinical research network. *Breast cancer research and treatment*, 175:675–689, 2019.

Carl F Craver. When mechanistic models explain. *Synthese*, 153(3):355–376, 2006.

Joost CF De Winter. Using the student's t-test with extremely small sample sizes. *Practical Assessment, Research, and Evaluation*, 18(1):10, 2019.

Gülşen Demiröz and H Altay Güvenir. Classification by voting feature intervals. In *European Conference on Machine Learning*, pages 85–92. Springer, 1997.

Michael W Dorrity, Lauren M Saunders, Christine Queitsch, Stanley Fields, and Cole Trapnell. Dimensionality reduction by umap to visualize physical and genetic interactions. *Nature communications*, 11(1):1537, 2020.

Manoj Durairaj and Veera Ranjani. Data mining applications in healthcare sector: a study. *International journal of scientific & technology research*, 2(10):29–35, 2013.

Jean-Loup Faulon and Léon Faure. In silico, in vitro, and in vivo machine learning in synthetic biology and metabolic engineering. *Current opinion in chemical biology*, 65: 85–92, 2021.

Thomas Fox and Jan M Kriegl. Machine learning techniques for in silico modeling of drug metabolism. *Current Topics in Medicinal Chemistry*, 6(15):1579–1591, 2006.

Kaitlyn M Gayvert, Neel S Madhukar, and Olivier Elemento. A data-driven approach to predicting successes and failures of clinical trials. *Cell chemical biology*, 23(10): 1294–1301, 2016.

Ophira Ginsburg, Cheng-Har Yip, Ari Brooks, Anna Cabanes, Maira Caleffi, Jorge Antonio Dunstan Yataco, Bishal Gyawali, Valerie McCormack, Myrna McLaughlin de Anderson, Ravi Mehrotra, et al. Breast cancer early detection: A phased approach to implementation. *Cancer*, 126:2379–2393, 2020.

Enrico Glaab, Armin Rauschenberger, Rita Banzi, Chiara Gerardi, Paula Garcia, and Jacques Demotes. Biomarker discovery studies for patient stratification using machine learning analysis of omics data: a scoping review. *BMJ open*, 11(12):e053674, 2021.

Aldo Glielmo, Brooke E Husic, Alex Rodriguez, Cecilia Clementi, Frank Noé, and Alessandro Laio. Unsupervised learning methods for molecular simulation data. *Chemical Reviews*, 121(16):9722–9758, 2021.

Joe G Greener, Shaun M Kandathil, Lewis Moffat, and David T Jones. A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, 23(1):40–55, 2022.

Rachel A Greenup, Christel Rushing, Laura Fish, Brittany M Campbell, Lisa Tolnitch, Terry Hyslop, Jeffrey Peppercorn, Stephanie B Wheeler, S Yousuf Zafar, Evan R Myers, et al. Financial costs and burden related to decisions for breast cancer surgery. *Journal of oncology practice*, 15(8):e666–e676, 2019.

Francesca Grisoni, Viviana Consonni, Marco Vighi, Sara Villa, and Roberto Todeschini. Investigating the mechanisms of bioconcentration through qsar classification trees. *Environment international*, 88:198–205, 2016.

Maja Hadzic, Darshan Dillon, and Tharam Dillon. Use and modeling of multi-agent systems in medicine. In *2009 20th International Workshop on Database and Expert Systems Application*, pages 303–307. IEEE, 2009.

Lynette Hirschman, Jong C Park, Junichi Tsujii, Limsoon Wong, and Cathy H Wu. Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18(12):1553–1561, 2002.

Takumi Ichimura, Shinichi Oeda, Machi Suka, Akira Hara, Kenneth J Mackin, and Yoshida Katsumi. Knowledge discovery and data mining in medicine. *Advanced techniques in knowledge discovery and data mining*, pages 177–210, 2005.

Gabriel Idakwo, Joseph Luttrell, Minjun Chen, Huixiao Hong, Zhaoxian Zhou, Ping Gong, and Chaoyang Zhang. A review on machine learning methods for in silico toxicity prediction. *Journal of Environmental Science and Health, Part C*, 36(4):169–191, 2018.

Costas Ioannides and David F V Lewis. Cytochromes p450 in the bioactivation of chemicals. *Current topics in medicinal chemistry*, 4(16):1767–1788, 2004.

John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2 (8):e124, 2005.

John PA Ioannidis. Why most published research findings are false. *New Doctor*, (88): 21–28, 2008.

Kevin Maik Jablonka, Daniele Ongari, Seyed Mohamad Moosavi, and Berend Smit. Big-data science in porous materials: materials genomics and machine learning. *Chemical reviews*, 120(16):8066–8129, 2020.

Show-Li Jan and Gwowen Shieh. The bland-altman range of agreement: Exact interval procedure and sample size determination. *Computers in Biology and Medicine*, 100:247–252, 2018. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2018.06.020. URL `https://www.sciencedirect.com/science/article/pii/S0010482518301677`.

Doo Seok Jeong and Cheol Seong Hwang. Nonvolatile memory materials for neuromorphic intelligent machines. *Advanced Materials*, 30(42):1704729, 2018.

Lei Jia and Hua Gao. Machine learning for in silico admet prediction. *Artificial Intelligence in Drug Design*, pages 447–460, 2022.

Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

Richard S Judson, Keith A Houck, Robert J Kavlock, Thomas B Knudsen, Matthew T Martin, Holly M Mortensen, David M Reif, Daniel M Rotroff, Imran Shah, Ann M Richard, et al. In vitro screening of environmental chemicals for targeted testing prioritization: the toxcast project. *Environmental health perspectives*, 118(4):485–492, 2010.

Alexandr A Kalinin, Gerald A Higgins, Narathip Reamaroon, Sayedmohammadreza Soroushmehr, Ari Allyn-Feuer, Ivo D Dinov, Kayvan Najarian, and Brian D Athey. Deep learning in pharmacogenomics: from gene regulation to patient stratification. *Pharmacogenomics*, 19(7):629–650, 2018.

Md Rezaul Karim, Tanhim Islam, Md Shajalal, Oya Beyan, Christoph Lange, Michael Cochez, Dietrich Rebholz-Schuhmann, and Stefan Decker. Explainable ai for bioinformatics: Methods, tools and applications. *Briefings in bioinformatics*, 24(5):bbad236, 2023.

Jihoon Kim, Bon-Kyoung Koo, and Juergen A Knoblich. Human organoids: model systems for human biology and medicine. *Nature Reviews Molecular Cell Biology*, 21 (10):571–584, 2020.

Ara Ko and Sherry M Wren. Advances in appropriate postoperative triage and the role of real-time machine-learning models: The goldilocks dilemma. *JAMA Network Open*, 4(11):e2133843–e2133843, 2021.

Lefteris Koumakis. Deep learning models in genomics; are we there yet? *Computational and Structural Biotechnology Journal*, 18:1466–1473, 2020.

Ilijana Kovrlija, Janis Locs, and Dagnija Loca. Octacalcium phosphate: Innovative vehicle for the local biologically active substance delivery in bone regeneration. *Acta Biomaterialia*, 135:27–47, 2021.

Ilijana Kovrlija, Ksenia Menshikh, Olivier Marsan, Christian Rey, Christèle Combes, Janis Locs, and Dagnija Loca. Exploring the formation kinetics of octacalcium phosphate from alpha-tricalcium phosphate: Synthesis scale-up, determination of transient phases, their morphology and biocompatibility. *Biomolecules*, 13(3):462, 2023.

Olga Krivorotko, Mariia Sosnovskaia, Ivan Vashchenko, Cliff Kerr, and Daniel Lesnic. Agent-based modeling of covid-19 outbreaks for new york state and uk: Parameter identification algorithm. *Infectious Disease Modelling*, 7(1):30–44, 2022. ISSN 2468-0427. doi: https://doi.org/10.1016/j.idm.2021.11.004. URL `https://www.sciencedirect.com/science/article/pii/S2468042721000798`.

John K Kruschke and Torrin M Liddell. The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychonomic bulletin & review*, 25:178–206, 2018.

Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. The sider database of drugs and side effects. *Nucleic acids research*, 44(D1):D1075–D1079, 2016.

Lukasz A Kurgan and Krzysztof J Cios. Caim discretization algorithm. *IEEE transactions on Knowledge and Data Engineering*, 16(2):145–153, 2004.

Kun Lan, Dan-tong Wang, Simon Fong, Lian-sheng Liu, Kelvin KL Wong, and Nilanjan Dey. A survey of data mining and deep learning in bioinformatics. *Journal of medical systems*, 42:1–20, 2018.

Stanley E Lazic, Charlie J Clarke-Williams, and Marcus R Munafò. What exactly is 'n'in cell culture and animal experiments? *PLoS biology*, 16(4):e2005282, 2018.

Paola Lecca. Machine learning for causal inference in biological networks: perspectives of this challenge. *Frontiers in Bioinformatics*, 1:746712, 2021.

Zhong Li, Shiqi Xiang, Eileen N Li, Madalyn R Fritch, Peter G Alexander, Hang Lin, and Rocky S Tuan. Tissue engineering for musculoskeletal regeneration and disease modeling. *Organotypic Models in Drug Development*, pages 235–268, 2021.

Xin Liang, Lizi Luo, Shiying Hu, and Yuke Li. Mapping the knowledge frontiers and evolution of decision making based on agent-based modeling. *Knowledge-Based Systems*, 250:108982, 2022. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2022.108982. URL `https://www.sciencedirect.com/science/article/pii/S0950705122004762`.

Lorenzo Lippi, Mauro Nascimben, Alessandro de Sire, Arianna Folli, Nicola Fusco, Lia Rimondini, and Marco Invernizzi. A novel free–to–use software for upper limb

volume quantification in breast cancer related lymphedema: implementing cutting–edge technology in the individualized therapeutic approaches of breast cancer survivors. *Cancer Research*, 83(Supplement 5):P5–08–18–P5–08–18, 03 2023a. ISSN 0008–5472. doi: 10.1158/1538-7445.SABCS22-P5-08-18. URL https://doi.org/10.1158/1538-7445.SABCS22-P5-08-18.

Lorenzo Lippi, Alessio Turco, Stefano Moalli, Mauro Nascimben, Claudio Curci, Alessandro de Sire, Lia Rimondini, and Marco Invernizzi. Quantitative assessment of upper-limb volume: Implications for lymphedema rehabilitation? *Applied Sciences*, 13(17): 9810, 2023b.

Chuang Liu, Yifang Ma, Jing Zhao, Ruth Nussinov, Yi-Cheng Zhang, Feixiong Cheng, and Zi-Ke Zhang. Computational network biology: data, models, and applications. *Physics Reports*, 846:1–66, 2020.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.

Yu-Chen Lo, Stefano E Rensi, Wen Torng, and Russ B Altman. Machine learning in chemoinformatics and drug discovery. *Drug discovery today*, 23(8):1538–1546, 2018.

Charles M. Macal. *Agent-Based Modeling and Artificial Life*, pages 725–745. Springer US, New York, NY, 2020. ISBN 978-1-0716-0368-0. doi: 10.1007/978-1-0716-0368-0_7. URL https://doi.org/10.1007/978-1-0716-0368-0_7.

Sarah J MacEachern and Nils D Forkert. Machine learning for precision medicine. *Genome*, 64(4):416–425, 2021.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Dane Morgan and Ryan Jacobs. Opportunities and challenges for machine learning in materials science. *Annual Review of Materials Research*, 50:71–103, 2020.

Mauro Nascimben. A machine learning based decision support system in oncology. Parma, Italy, Sept 2021a. University of Parma, 2020+2021 Italian Society of Applied and Industrial Mathematics (SIMAI) Conference.

Mauro Nascimben. Molecular fingerprint based and machine learning driven QSAR for bioconcentration pathways determination. Rome, Italy, Sept 2021b. National Research Council of Italy, Virtual Workshop of Mathematical Modelling and Control for Healthcare and Biomedical Systems.

Mauro Nascimben. Quantized computational QSAR framework for molecular toxicity virtual screening. Rome, Italy, July 2022. Sapienza University of Rome, 3rd Molecules Medicinal Chemistry Symposium - Shaping Medicinal Chemistry for the New Decade.

Mauro Nascimben. Virtual screening by spiking neural networks: a case study on cytochrome P450. Padua, Italy, Sept 2023a. University of Padua, 18th Conference on Computational Intelligence Methods for Bioinformatics and Biostatistics.

Mauro Nascimben. Anomaly detection of EV-related protein expression in doped bioactive glasses. Novara, Italy, Oct 2023b. Italian Chemical Society, 3rd International Proteomics And Metabolomics Conference.

Mauro Nascimben. Low–power or resource–constrained environments for virtual screening and quantitative structure-activity relationship analysis for in silico precision medicine. Rome, Italy, Sept 2023c. Italian Association for Industrial Research, NanoInnovation Conference and Exhibition.

Mauro Nascimben and Lia Rimondini. Visually enhanced Python functions for clinical equality of measurement assessment. *Annals of Computer Science and Information Systems*, 32:241–249, 2022.

Mauro Nascimben and Lia Rimondini. Molecular toxicity virtual screening applying a quantized computational SNN–based framework. *Molecules*, 28(3):1342, 2023.

Mauro Nascimben, Manolo Venturin, and Lia Rimondini. Double–stage discretization approaches for biomarker-based bladder cancer survival modeling. *Communications in Applied and Industrial Mathematics*, 12(1):29–47, 2021.

Mauro Nascimben, Lorenzo Lippi, Nicola Fusco, Marco Invernizzi, and Lia Rimondini. A software suite for limb volume analysis applicable in clinical settings: upper limb quantification. *Frontiers in Bioengineering and Biotechnology*, 10:863689, 2022a.

Mauro Nascimben, Lia Rimondini, Davide Corà, and Manolo Venturin. Polygenic risk modeling of tumor stage and survival in bladder cancer. *BioData Mining*, 15(1):23, 2022b.

Mauro Nascimben, Lorenzo Lippi, Alessandro De Sire, Marco Invernizzi, and Lia Rimondini. Algorithm–based risk identification in patients with breast cancer–related lymphedema: A cross-sectional study. *Cancers*, 15(2):336, 2023a.

Mauro Nascimben, Lorenzo Lippi, Nicola Fusco, Alessandro de Sire, Marco Invernizzi, and Lia Rimondini. Technical aspects and validation of custom digital algorithms for hand volumetry. *Technology and Health Care*, 31(5):1835–1854, 2023b.

Mauro Nascimben, Silvia Spriano, Lia Rimondini, and Manolo Venturin. Molecular fingerprint based and machine learning driven QSAR for bioconcentration pathways determination. In Gabriella Bretti, Roberto Natalini, Pasquale Palumbo, and Luigi Preziosi, editors, *Mathematical Models and Computer Simulations for Biomedical Applications*, pages 193–215, Cham, 2023c. Springer Nature Switzerland. ISBN 978-3-031-35715-2.

Mauro Nascimben, Ilijana Kovrlija, Janis Locs, Dagnija Loca, and Lia Rimondini. Fusion and classification algorithm of octacalcium phosphate production based on xrd and ftir data. *Scientific Reports*, 14(1):1489, 2024.

Mauro Nascimben, Hugo Abreu, Marcello Manfredi, Annalisa Chiocchetti, and Lia Rimondini. Latent expression of extracellular vesicles proteins in doped bioactive glasses through machine learning–based mass–spectometry data analysis. International Journal of Molecular Sciences, Submitted.

Anuraj Nayarisseri, Ravina Khandelwal, Poonam Tanwar, Maddala Madhavi, Diksha Sharma, Garima Thakur, Alejandro Speck-Planche, and Sanjeev K Singh. Artificial intelligence, big data and machine learning approaches in precision medicine & drug discovery. *Current drug targets*, 22(6):631–655, 2021.

Serena Nembri, Francesca Grisoni, Viviana Consonni, and Roberto Todeschini. In silico prediction of cytochrome p450-drug interaction: Qsars for cyp3a4 and cyp2c9. *International journal of molecular sciences*, 17(6):914, 2016.

Sarfaraz K Niazi and Zamara Mariam. Recent advances in machine-learning-based chemoinformatics: a comprehensive review. *International Journal of Molecular Sciences*, 24(14):11488, 2023.

John W Nichols, Mark Bonnell, Sabcho D Dimitrov, Beate I Escher, Xing Han, and Nynke I Kramer. Bioaccumulation assessment using predictive approaches. *Integrated Environmental Assessment and Management: An International Journal*, 5(4):577–597, 2009.

Frank Noé, Alexandre Tkatchenko, Klaus-Robert Müller, and Cecilia Clementi. Machine learning for molecular simulation. *Annual review of physical chemistry*, 71:361–390, 2020.

Emin Orhan. The leaky integrate-and-fire neuron model. *no*, 3:1–6, 2012.

Ryosaku Ota and Fumiyoshi Yamashita. Application of machine learning techniques to the analysis and prediction of drug pharmacokinetics. *Journal of Controlled Release*, 352:961–969, 2022.

Sankar K Pal, Shubhra S Ray, and Avatharam Ganivada. *Granular neural networks, pattern recognition and bioinformatics*. Springer, 2017.

Hazel R. Parry. *Agent-Based Modeling, Large-Scale Simulations*, pages 913–926. Springer US, New York, NY, 2020. ISBN 978-1-0716-0368-0. doi: 10.1007/978-1-0716-0368-0_9. URL `https://doi.org/10.1007/978-1-0716-0368-0_9`.

Witold Pedrycz. *Granular computing: an introduction*. Springer, 2000.

Darren Plant and Anne Barton. Machine learning in precision medicine: lessons to learn. *Nature Reviews Rheumatology*, 17(1):5–6, 2021.

Priya Ranganathan, CS Pramesh, and Marc Buyse. Common pitfalls in statistical analysis: Clinical versus statistical significance. *Perspectives in clinical research*, 6(3):169, 2015.

Alexander D Rast, Francesco Galluppi, Xin Jin, and Steve B Furber. The leaky integrate-and-fire neuron: A platform for synaptic model exploration on the spinnaker chip. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2010.

Parminder S Reel, Smarti Reel, Ewan Pearson, Emanuele Trucco, and Emily Jefferson. Using machine learning approaches for multi-omics data analysis: A review. *Biotechnology Advances*, 49:107739, 2021.

Christopher A Reilly and Garold S Yost. Metabolism of capsaicinoids by p450 enzymes: a review of recent findings on reaction mechanisms, bio-activation, and detoxification processes. *Drug metabolism reviews*, 38(4):685–706, 2006.

Ann M Richard, Ruili Huang, Suramya Waidyanatha, Paul Shinn, Bradley J Collins, Inthirany Thillainadarajah, Christopher M Grulke, Antony J Williams, Ryan R Lougee, Richard S Judson, et al. The tox21 10k compound library: collaborative chemistry advancing toxicology. *Chemical Research in Toxicology*, 34(2):189–216, 2020.

Nicolas Rodrigue and Hervé Philippe. Mechanistic revisions of phenomenological modeling strategies in molecular evolution. *Trends in Genetics*, 26(6):248–252, 2010. doi: 10.1016/j.tig.2010.04.001.

Pedro Sanchez, Jeremy P Voisey, Tian Xia, Hannah I Watson, Alison Q O'Neil, and Sotirios A Tsaftaris. Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, 9(8):220638, 2022.

Philip Sedgwick. Limits of agreement (bland-altman method). *Bmj*, 346, 2013.

K Aditya Shastry and HA Sanjay. Machine learning for bioinformatics. *Statistical modelling and machine learning principles for bioinformatics techniques, tools, and applications*, pages 25–39, 2020.

Vivek Subbiah. The next generation of evidence-based medicine. *Nature medicine*, 29 (1):49–58, 2023.

Xu Sun and Weichao Xu. Fast implementation of delong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters*, 21(11):1389–1393, 2014.

Sivani Tadepalli, Nasrin Akhter, Daniel Barbará, and Amarda Shehu. Anomaly detection-based recognition of near-native protein structures. *IEEE Transactions on NanoBioscience*, 19(3):562–570, 2020.

Doron Tal and Eric L Schwartz. Computing with the leaky integrate-and-fire neuron: logarithmic computation and multiplication. *Neural computation*, 9(2):305–318, 1997.

Amirhossein Tavanaei, Masoud Ghodrati, Saeed Reza Kheradpisheh, Timothée Masquelier, and Anthony Maida. Deep learning in spiking neural networks. *Neural networks*, 111:47–63, 2019.

Mannie Belay Taye. Biomedical applications of ion-doped bioactive glass: A review. *Applied Nanoscience*, 12(12):3797–3812, 2022.

Anamika Tiwari, Vikrant Bansode, and Anurag S Rathore. Application of advanced machine learning algorithms for anomaly detection and quantitative prediction in protein a chromatography. *Journal of Chromatography A*, 1682:463486, 2022.

Mark K Transtrum and Peng Qiu. Bridging mechanistic and phenomenological models of complex biological systems. *PLoS computational biology*, 12(5):e1004915, 2016.

George Tzanis. Biological and medical big data mining. *International Journal of Knowledge Discovery in Bioinformatics (IJKDB)*, 4(1):42–56, 2014.

Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, et al. Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6):463–477, 2019.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Ennapadam S Venkatraman and Colin B Begg. A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika*, 83(4):835–848, 1996.

Jean-Philippe Vert and Laurent Jacob. Machine learning for in silico virtual screening and chemical genomics: new strategies. *Combinatorial chemistry & high throughput screening*, 11(8):677–685, 2008.

Yoram Vodovotz and Gary An. Agent-based models of inflammation in translational systems biology: A decade later. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 11(6):e1460, 2019.

Nikil Wale. Machine learning in drug discovery and development. *Drug Development Research*, 72(1):112–119, 2011.

Esteban Walker and Amy S Nowacki. Understanding equivalence and noninferiority testing. *Journal of general internal medicine*, 26:192–196, 2011.

Anthony Yu-Tung Wang, Ryan J Murdock, Steven K Kauwe, Anton O Oliynyk, Aleksander Gurlo, Jakoah Brgoch, Kristin A Persson, and Taylor D Sparks. Machine learning for materials scientists: an introductory guide toward best practices. *Chemistry of Materials*, 32(12):4954–4965, 2020.

S. L. Waters, L. J. Schumacher, and A. J. El Haj. Regenerative medicine meets mathematical modelling: Developing symbiotic relationships. *npj Regenerative Medicine*, 6 (1), 2021. doi: 10.1038/s41536-021-00134-2.

Gerhard-Wilhelm Weber, Süreyya Özöğür-Akyüz, and Erik Kropat. A review on data mining and continuous optimization applications in computational biology and medicine. *Birth Defects Research Part C: Embryo Today: Reviews*, 87(2):165–181, 2009.

Craig R White and Dustin J Marshall. Should we care if models are phenomenological or mechanistic? *Trends in ecology & evolution*, 34(4):276–278, 2019.

Daniel S Wigh, Jonathan M Goodman, and Alexei A Lapkin. A review of molecular representation in the age of machine learning. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12(5):e1603, 2022.

Jack Wilkinson, Kellyn F Arnold, Eleanor J Murray, Maarten van Smeden, Kareem Carr, Rachel Sippy, Marc de Kamps, Andrew Beam, Stefan Konigorski, Christoph Lippert, et al. Time to reality check the promises of machine learning-powered precision medicine. *The Lancet Digital Health*, 2(12):e677–e680, 2020.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Li Xiaoxue, Zhang Xiaofan, Yi Xin, Liu Dan, Wang He, Zhang Bowen, Zhang Bohan, Zhao Di, and Wang Liqun. Review of medical data analysis based on spiking neural networks. *Procedia Computer Science*, 221:1527–1538, 2023.

Aimin Yang, Wei Zhang, Jiahao Wang, Ke Yang, Yang Han, and Limin Zhang. Review on the application of machine learning algorithms in the sequence data mining of dna. *Frontiers in Bioengineering and Biotechnology*, 8:1032, 2020.

Chuan Zhang, Mandy Berndt-Paetz, and Jochen Neuhaus. Bioinformatics analysis identifying key biomarkers in bladder cancer. *Data*, 5(2):38, 2020.