UNIVERSITY OF SIENA

Department of Biotechnology, Chemistry and Pharmacy

PHD SCHOOL IN BIOCHEMISTRY AND MOLECULAR BIOLOGY, XXXIV CYCLE

PhD coordinator: Prof.ssa Lorenza Trabalzini

Machine learning in Bioinformatics: Novel approaches to Precision Medicine, Life Sciences and Healthcare

Tutor:

Prof.ssa Ottavia Spiga

Co-Tutor:

Prof.ssa Monica Bianchini

PhD student:

Anna Visibelli

ANNA VISIBELLI

Firmato il 01/02/2022

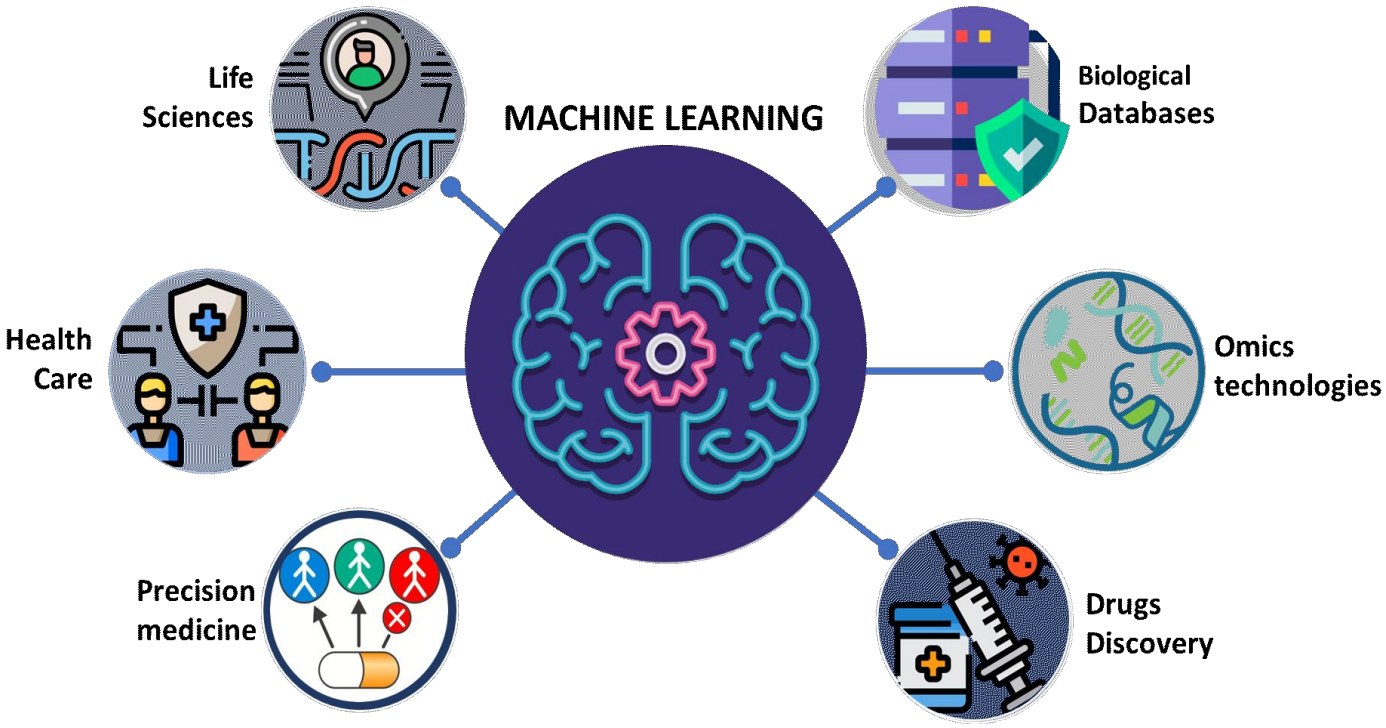
Anno Accademico 2020 – 2021

INDEX

ABSTRACT

In recent years, biological research revolves around huge amounts of data which are extrapolated due to high-throughput techniques. Thanks to the emergence of omics information and big data, the use of computational tools has become crucial to evaluate the efficacy of medical treatments or deeply investigate the correlation between patients and diseases according to their own molecular characteristics. The Precision Medicine approach is widely applied to the healthcare area, in particular to rare diseases with the creation of patient registries leveraging large amounts of data to discover potential links. Harmonizing databases and including disease registries are the major facilitators to understand the complexity of diseases, to conduct clinical trials, to improve the drug development process and to assign the right treatment to the right individual after a reliable patient stratification. Moreover, the application of data mining in healthcare and public health, which has been growing over the last years, allows to systematically identify inefficiencies and best practices that improve care and reduce costs with remarkable economic benefits. In this thesis we focus on the development of new Artificial Intelligence algorithms for a number of important problems in the field of Precision Medicine, Life Sciences and Healthcare. The project demonstrates the power of computational modelling for clinical research, opening up possibilities that would be unimaginable without knowledge of the data. The application of Bioinformatics and Computational biology algorithms together with the creation of digital databases will offer an opportunity to translate new data into actionable information.

GRAPHICAL ABSTRACT



1. INTRODUCTION

1.1. MACHINE LEARNING: A BRIEF OVERVIEW

Artificial Intelligence (AI) is the science that aims to develop machines capable of carrying out tasks related to human perception. Although in 1956 the computer scientist John McCarthy introduced the expression "artificial intelligence", which marked the actual birth of this discipline, already in previous years, other researchers had produced significant results in this area. As early as 1943, Warren McCulloch and Walter Pitt proposed the first artificial neuron to the scientific world [McCulloch et al., 1943], and already in the 1950s the first working prototypes of neural networks were created [Neumann, 1958] [Hebb, 1949]. However, it was the research of Alan Turing, which increased the interest of the public. Turing, attempting to explain how and to what extent computers could actually simulate human behaviour, devised a test - the Turing test [Turing, 1950] – to be able to give a measure of the thinking ability of machines. Initially, computers were better able to successfully solve problems that are intellectually difficult for humans but relatively simple for computers, that is, problems that can be described by a list of mathematical rules [Bengio et al., 2016]. Later, the AI challenge turned out to be the solution of tasks that are easy for people to perform, but difficult to describe formally, because they are related to perceptual skills developed in humans during an evolutionary process of hundreds of thousands of years. The idea was to create machines that were able to acquire the knowledge necessary to solve a problem independently, from experience. The hierarchy of concepts allows the machine to learn complicated notions by building them starting from the simplest ones. For this reason, a stratified form of thinking was necessary to learn complex concepts - Deep Learning [Awad et al., 2015] - which is inspired by the way biological neural networks in the human brain process information. Deep Learning is nothing more than a subcategory of a larger family of Artificial Intelligence methods called Machine Learning.

Machine Learning (ML) was born from the idea that computers are able to learn to perform certain tasks by improving their skills through experience. At the base of machine learning there are a series of different algorithms which, starting from primitive notions, learn to make a specific decision or to perform actions learned over time. The computer is provided with only a set of data (*training set*), which are iteratively examined to extract information, similarly to what happens in human learning. Depending on the way the machine learns data and information, four different learning methods can be distinguished.

- Supervised learning [Cunningham et al., 2008]: the training data are labelled with a target i.e., an "expected result". In this way, after the training phase, the system will be able to use the acquired experience to solve problems which involve the same basic knowledge.
- Unsupervised learning [Ghahramani, 2004]: the training set is not labelled. Learning consists in identifying relationships between data, without any prior knowledge about the data themselves.
- Semi-supervised learning [Zhu, 2005]: the training set is only partially labelled. Particularly useful in cases where the knowledge about the data is partial or the collection and sampling phase of labelled data is too expensive to be carried out comprehensively.
- Reinforcement learning [Russell et al., 2016]: the training set is not labelled, but an example is given with a positive or negative result. This result allows a feedback loop for the algorithm, letting it determine whether the provided solution solves a problem or not. It is therefore the computerized version of human learning by "trial and error".

The Machine Learning process consists of six components regardless of the algorithm adopted as shown in Fig. 1 [Alzubi et al., 2018].

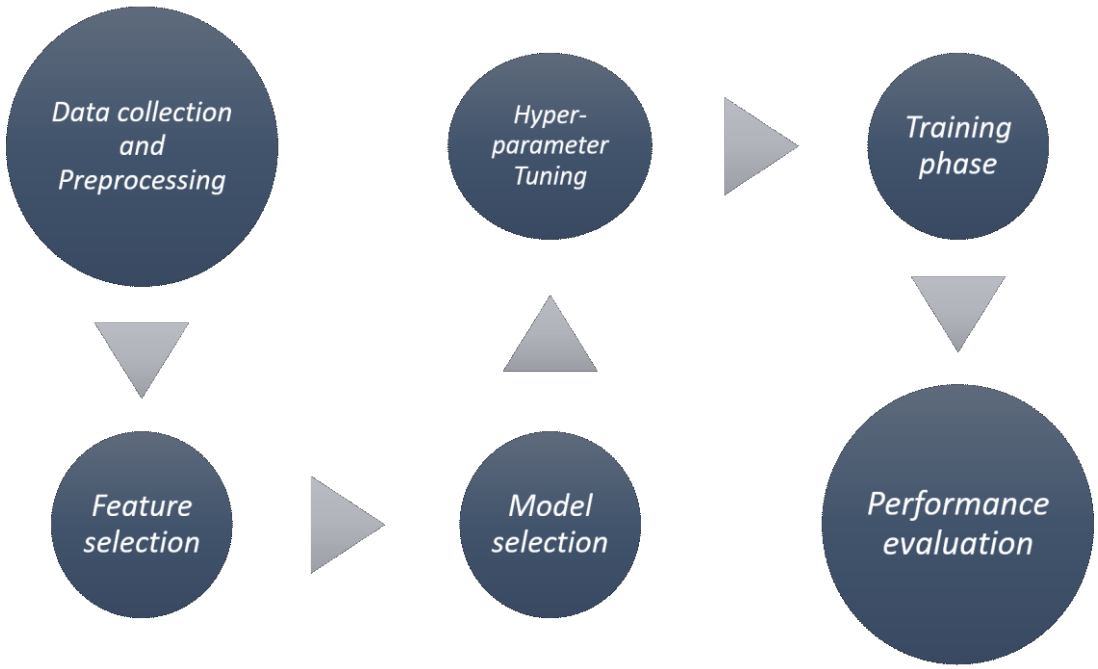


Figure 1. Six main components of the Machine Learning process.

Data collection and pre-processing consist in the preparation of data in a format that can be given as input to the algorithm. Data are unstructured, sparse and contain a lot of irrelevant details as well as they can be redundant, so that they need to be cleaned and pre-processed to a structured format. Since

ML models require all input and output variables to be numeric, categorical data need to be encoded with different strategies. Moreover, learning algorithms benefit from standardization of the dataset to avoid bias in the outcome. The data obtained from the above step may contain numerous features, not all of which are relevant to the learning process.

Feature selection is the process of reducing the number of input variables to both reduce the computational cost of modelling and, in some cases, to improve the performance of the model.

However, even if the dataset is correctly pre-processed, not all ML algorithms are meant for all problems, but certain algorithms are more suited to a particular class of problems. *Selecting the best machine learning algorithm* is imperative in getting the best possible results.

Once obtained the most appropriate model, some strategies will have to be undertaken for setting the most appropriate values of the various parameters. The process of searching for the ideal model architecture is referred to as *hyperparameter tuning*. The same kind of ML model can require different constraints, weights or learning rates to generalize to different data patterns. The ultimate goal for any machine learning model is to learn from examples in such a manner that the model is capable of generalizing the acquired knowledge to new instances which it has not seen yet. The model should then be *trained* on a subset of the total dataset and *tested* against unseen data, to evaluate how much has been learnt using various performance parameters like accuracy, precision and recall.

Furthermore, there are many tasks that a ML tool can perform:

- ***Classification*:** the input data are divided into two or more classes and the learning system aims to produce a model capable of assigning a class among those available to each input.
- ***Regression*:** conceptually similar to classification with the difference that the output belongs to a continuous rather than discrete domain.
- ***Clustering:*** the set of input data is divided into groups about which there is no prior knowledge; unlike the classification, neither the number nor the type (target) of the classes are known.

ML techniques include a vast class of algorithms, starting with decision trees, genetic and boosting algorithms, metric techniques, such as the K-nearest neighbour algorithm (k-NN,) Support Vector Machines, statistical methods, Bayesian networks and Artificial Neural Networks. The selection of the most appropriate ML approach for a specific problem is crucial in order to obtain an ML model that produces robust and reliable results. It depends on many factors, from the problem statement to the type of output desired. The well-known “no free lunch” (NFL) theorem for supervised machine learning [Wolpert et al., 1997] states that all optimization algorithms perform equally well when their performance is averaged across all possible problems. It implies that no single ML algorithm is universally the best-performing algorithm for all problems. Due to the inability to find a single ML

model which outperforms the others, a list of ML algorithms which are best suited to a particular task is listed below in Fig 2.

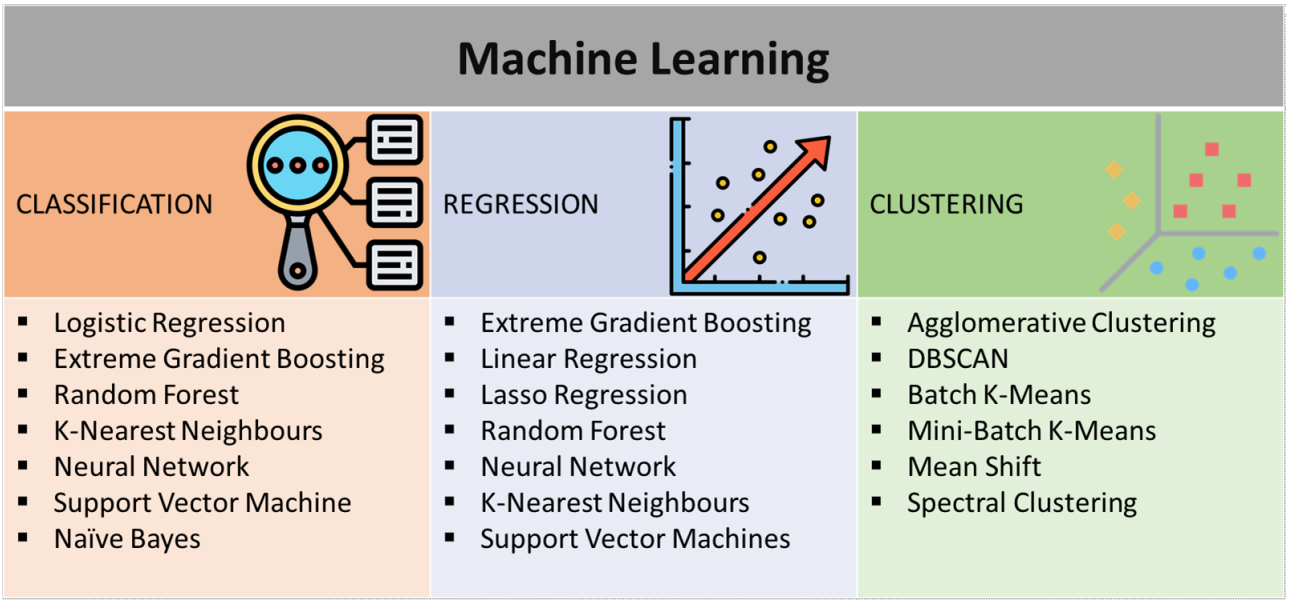


Figure 2. Classification of the ML algorithms more suited to different classes of problems.

Indeed, a detailed methodology for selecting the most appropriate or "best fit" ML architecture remains a challenge to be studied in the future. So far, we have seen the wide range of applicability of Machine Learning, but there are still open questions to be addressed. One of the main obstacles is the *developer deficit*, because there are not a lot of specialists who can develop ML tools. Even a data scientist who has a solid grasp of ML processes very rarely has enough software engineering skills. ML algorithms require large volumes of data to be accurate and efficient, and such amounts are not always available to researchers. Creating a data collection mechanism that adheres to all the rules and standards imposed by governments is a difficult and time-consuming task. In the Healthcare field, acquisition becomes even more difficult, as digital data are scattered and not always accessible, making it difficult to make accurate predictions. Furthermore, even the raw data must be reliable, otherwise the deriving results could be catastrophic.

Finally, the importance of the explainability of ML decision-support systems is evident, especially in contexts that can be very sensitive, such as, for example, in the case of medical diagnoses or autonomous driving systems. In fact, explainability is one of the most debated topics for the application of artificial intelligence in the healthcare. While AI-based systems have been shown to outperform humans in certain analytical tasks, the lack of explainability continues to attract criticism. However, explainability is not a purely technological issue, but invokes a number of medical, legal,

ethical and social issues that require in-depth exploration. Explainable AI approaches are the new frontier of ML applications in healthcare, in order to ensure the understanding, by both clinicians and patients, of the "mental process" followed by the artificial brain to reach a certain decision.

There are also many fields that are still challenging for Machine and Deep Learning algorithms, such as speech understanding, disease detection, drug discovery etc. In the near future, ML is expected to continue to act as a technological innovator with increasingly revolutionary advances. As these technologies continue to grow, they will have more and more impact on the social setting and quality of life.

1.2 ML IN BIOINFORMATICS

In recent years, biological research revolves around huge amounts of data which are extrapolated due to high-throughput techniques. Therefore, the use of computational tools becomes crucial, because they are able to help analyse “big data”, extrapolating their features to populate biological databases. The need for efficient computational tools gave rise to a new field called Bioinformatics, an interface between the field of biological and computational research. Bioinformatics, in combination with Machine Learning, represents a key factor for the development of algorithms and software for the transfer, storage, analysis and development of biological platforms. An increasing number of ML methods have been implemented to address bioinformatics problems in system biology, genomics, structural biology and other relevant bioinformatics domains, as shown in Fig. 3. [Cios et al., 2006].

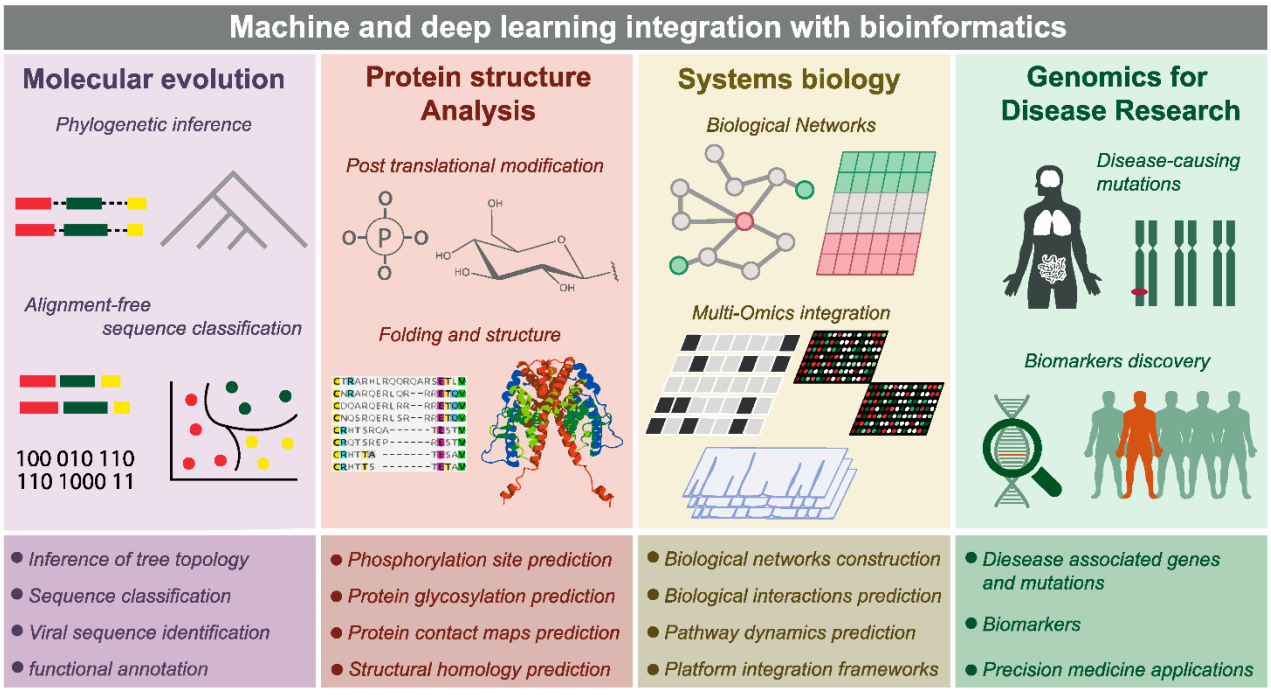


Figure 3. Applications of integrated Machine Learning techniques with Bioinformatics. [Auslander et al., 2021]

ML tasks in bioinformatics include classification, regression, clustering and presentation of data for easy interpretation [Tan et al., 2001]. Such approaches are cheaper and more efficient to handle bioinformatics problems, being able to analyse big amounts of data in just a few seconds to obtain prediction models on biological systems. Some of the most used methods are displayed in Fig. 4.

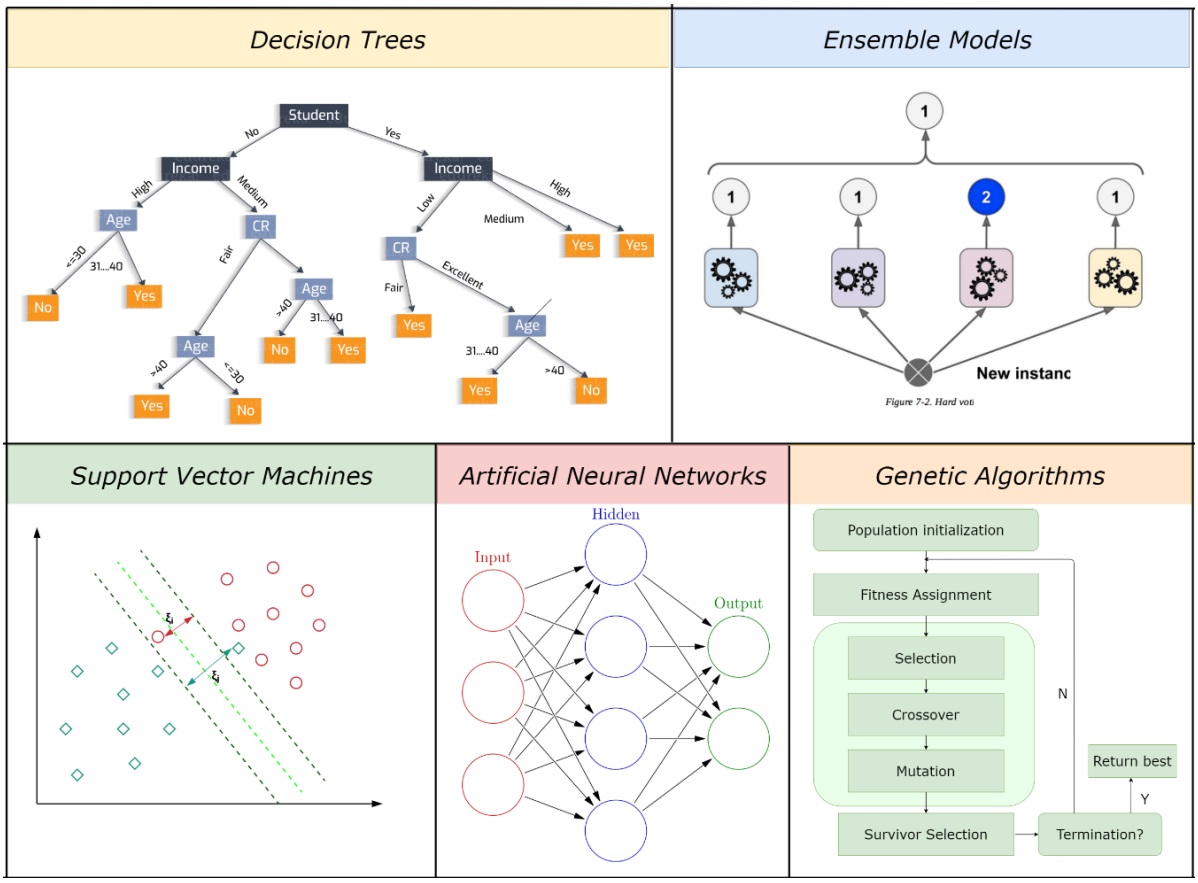


Figure 4. ML algorithms commonly used in Bioinformatics.

- Artificial neural networks (ANNs) [McCulloch et al., 1943] are widely used in bioinformatics thanks to the ability to solve complex real-world problems. ANNs are robust tools that can manage big data identifying key components, thus providing a greater understanding of the biological system being modelled. A neural network consists of an oriented graph formed by nodes (organized in layers) connected by arcs. Each arc is associated with a weight, while nodes are equipped with activation functions that elaborate the inputs to produce the neuron output. Supervised neural network learning is based on a feedback process, called back-propagation, in

which the output of the network is compared with the output it was meant to produce, and the difference between the outputs is used to modify the weights of the connections between the neurons in the network. Moreover, they are usually resistant to noise and errors present in training data. In particular, Recurrent Neural Networks (RNNs) are a powerful and robust type of neural architectures, provided with feedback connections, that produce internal loops. Such loops induce a recursive dynamic within the networks and thus introduce delayed activation dependencies across the processing elements. In doing so, RNNs develop a kind of memory that makes them particularly tailored to process sequential data, such as text, DNA, proteins, etc.

- Decision trees [Wu et al., 2008] are structures resembling a tree in which each internal node represents a decision on an attribute, each leaf node represents a feature label and each branch represents the value of that feature. A decision tree classifies instances by sorting them from the root to some leaf nodes on the basis of feature values. The main goal of decision trees is to arrange different nodes based on valid data, as each hub in a decision tree addresses an item in an occurrence to be sorted, and each branch addresses a value that the hub can accept. While using a decision tree, the focus is on how to decide which attribute is the best classifier at each node level. Statistical measures like information gain, Gini index, Chi-square and entropy are calculated for each node to quantify the worth of that node. For practical applications, decision trees have been used for protein function prediction, protein-protein interaction (PPI) and cancer classification.
- Support Vector Machines (SVMs) [Cortes et al., 1995] are supervised learning tool which can be used for classification as well as regression problems. They represent a method of maximizing the margin to separate two classes so that the trained model generalizes well to test data. Each data item is firstly plotted as a point in a *n*-dimensional space and the model classifies the data into different classes by finding a hyperplane which separates them. Because of their relative simplicity and flexibility to address a variety of problems, SVMs offer generally good predictive performance, with a lower risk of overfitting. For this reason, SVMs have been widely applied to many areas of bioinformatics, including protein function prediction, protease functional site recognition, transcription initiation site prediction, and gene expression data classification.
- Genetic algorithms (GA) [Whitley, 1994] have found popularity in bioinformatics research due to their simplicity. They are heuristic techniques of calculation, which find solutions to problems by using a finite series of standard steps, inspired by the mechanics of natural selection. GAs have been used to determine the structure of DNA using spectrometric data.
- Ensemble learning [Polikar, 2009] is a widely used technique that combines multiple learning algorithms to improve the overall prediction accuracy and reduce the potential overfitting of

training data. A large number of ensemble methods have been applied to biological data, but the three most popular methods are bagging [Breiman, 1996], boosting [Freund et al., 1996], and random forests [Breiman, 2001]. Random forests are used for classification and regression. They use a bagging approach to create a bunch of decision trees with a random subset of data, and the output of all decisions is combined to make the final predictive model. RFs have been applied on a variety of bioinformatic problems, such as gene expression classification, mass spectrum protein expression analysis, biomarker discovery or statistical genetics.

Furthermore, ML in bioinformatics has become even more significant thanks to the birth of Deep Learning due to its capacity to execute feature engineering on its own. A deep learning algorithm will scan the data to search for features that correlate and combine them to enable faster (parallel and distributed) learning without being explicitly told to do so.

Let us see in detail some of the biological fields of application of ML.

- ***Omics:*** The omics sciences refer to all the disciplines that aim to characterize and quantify biological molecule pools, in order to delineate the structure, functions and dynamics of an organism, as shown in Fig. 5.

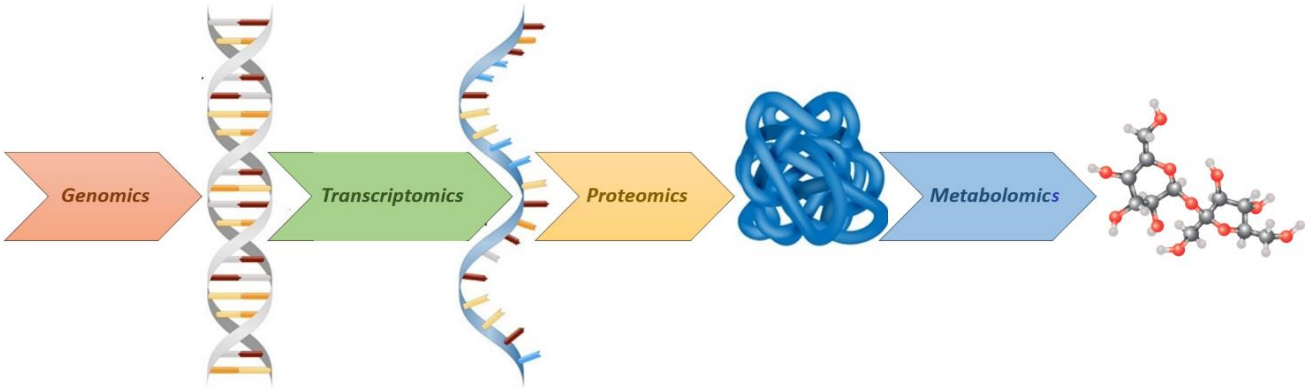


Figure 5. The omics cascade. Metabolomics is the final step in the cascade, because it is closer to the phenotype than the preceding omics.

Individual *omic* as well as the integrated profiles of multiple *omes*, such as the genome, the epigenome, the transcriptome, the proteome, the metabolome, the antibodyome, and other omics information are being used extensively in bioinformatics, thanks to the next-generation sequencing technology which allows the acquisition of big amount of omics data. Sequences, position-specific scoring matrix (PSSM) and biological, physio-chemical and structural properties are often used as inputs for AI algorithms, which are capable of analysing them making it more understandable. Omics technologies have also the potential to transform medicine from traditional symptom-

oriented diagnosis and treatment of diseases towards individualized disease prevention and early diagnostics [Chen et al., 2013]. ML methods have been applied to a wide range of *genomics* problems, ranging from the sequencing of whole genomes [Venter et al., 2001], to the identification of genes and RNA structures [Bernal et al., 2007, Fogel et al., 2002]. Decision trees have been used to identify genes [Lopez-Bigas et al., 2004] and gene-gene interactions [Ritchie et al., 2003] involved in genetic diseases. Genetic Algorithms instead, have been applied for DNA fragment assembly [Rathee et al., 2014]. In *proteomics,* the tertiary structure prediction of proteins represents one of the main challenges for ML methods. Existing methods for resolving PSS usually rely on the primary structure and physio-chemical properties of proteins, making use of specific energy functions that need to be minimized. Lately, graph-based approaches have also been applied, showing that starting from simple geometric properties, graph-based predictions can be as robust as an energy-based score. ANNs have found widespread applications in protein structure prediction [Bidargaddi et al., 2009][Li et al., 2016], functional prediction [Kihara, 2017] and protein classification [Fox et al., 2015]. Moreover, protein structure prediction was also performed using decision trees and support vector machines [He et al., 2006]. ANNs have been increasingly applied to problems in *metabolomics*, which result to be challenging for conventional algorithms. A variety of ML algorithms have been developed for data analysis [Cambiaghi et al., 2016], peak identification and compound identification [Nguyen et al., 2018] in the field of nuclear magnetic resonance and mass spectroscopy-based metabolomics [Puchades-Carrasco et al., 2015]. RFs have been used in metabolomic to determine a set of serum protein and metabolic biomarkers in prostate cancer [Fan et al., 2011].

- ***Medicine:*** Computational approaches can be applied to characterize variations in health and disease, and the outcomes obtained from these models can be used to develop high-performance methods for disease diagnosis and treatment. Understanding the function of highly interconnected molecular networks using AI methods can lead to the discovery of molecular disease networks, discrimination between disease subtypes and prediction of disease progression. ML is applied to distinguish disease phenotypes from genomic data, using ANNs [Khan et al., 2001], RFs [Zhang et al., 2003] and SVMs [Yeang et al., 2001]. ML approaches are able to predict, based on the molecular signatures of the disease, the response to treatment in breast cancer [Weichselbaum et al., 2008], prostate cancer [Zhao et al., 2010] and lung cancer prognosis [Patnaik et al., 2010]. However, the main goal of computational physiological medicine is to develop mechanistic models able to predict emergent behaviours of biological systems in health and disease and how system properties may change over time, and then translate insights gained from these models to improved therapies. Recent cancer studies [Deisboeck et al., 2011] show that computational

models are influencing both diagnosis and treatment. Regarding diabetes, the Artificial Pancreas Project is developing a closed-loop subcutaneous insulin delivery system for the treatment of type 1 diabetes mellitus. In addition, tools such as MRI and Computed Tomography make it possible to model the heart of a single patient. Detailed heart models reconstructed from clinical MRI scans were used to evaluate infarct-related ventricular tachycardia, which can help predict optimal catheter ablation locations in individual patients' hearts [Relan et al., 2011]. The last field of ML application to medicine concerns the comprehension of how ensembles of anatomies differ within and between healthy and diseased states, constructing a global shape model representing typical structures in an ensemble of anatomic image volumes. For example, in a study of structural changes in the brain in Alzheimer's disease [Miller et al., 2009], a model of the hippocampus was constructed from a series of volumes of MR images.

- ***Drug discovery:*** In the past decade, the field of drug discovery and development has been undergoing radical transformations, driven by rapid development of AI. Popular implementations of AI in drug discovery include applications in virtual screening (VS) [Stumpfe et al., 2020], retrosynthesis and reaction prediction [Boström et al., 2020], and de novo protein [Strokach et al., 2018] and drug design [Schneider et al., 2020]. A certain number of computational techniques have been applied for the quantitative structure-activity relationships (QSAR), with the aim to study the biological activity of chemical substances from a set of atomic and molecular descriptors. Moreover, both ANNs and SVMs have been found useful in the field of QSAR [Moss et al., 2012, Molfetta et al., 2008]. AI models also offer technological solutions for drug development. Clinical trials consume the second half of the 10 to 15 year development cycle, to bring a single new drug to market. Thus, a failed trial sinks not only the investment in the trial itself but also the preclinical development costs, making the loss per failed clinical trial from $800 million to $1.4 billion. Two of the key factors causing a clinical trial to fail are patient selection and recruitment mechanisms that fail to bring the most suitable patients to a trial in time, as well as a lack of technical infrastructure to cope with the complexity of conducting a trial [Hwang et al., 2016]. In these cases, AI techniques can be used to reshape key steps of clinical design and trial to increase trial success rates. [Aziz et al., 2019, Berry et al., 2010]

While AI models need to be robust and capable of parsing the data correctly, it is also true that if the dataset is dirty, the accuracy of the classifier decreases. Dirty biological data can be obtained due to many factors, such as errors during experimentation, misinterpretation by biologists, use of non-standard methods. ML approaches must be able to provide optimal decisions by adapting to such

datasets in order to avoid over-fitting. It is therefore essential that the input and output data are analysed and interpreted by specialists in order to maintain the quality of the data and not to run into errors. Moreover, even if the use of ML for solving bioinformatics problems is a relatively new field, randomly selected strategies for data splitting, parameter optimization, dealing with missing data, need to give way to more principled approaches which guarantee statistical validity, generating meaningful results that can be interpretable, repeatable and applicable to practical problems.

1.3 PRECISION MEDICINE, LIFE SCIENCES AND HEALTHCARE

In most developed countries, the healthcare sector comprises over 15% of the economy, making it one of the largest industries in any state. Levels of care are divided into four categories based on the complexity of the medical cases being treated as well as the skills and specialties of the providers.

- **Primary Care** is the first stop for symptom assessment and medical concerns such as some bacterial or viral disease, or any other acute medical problem. Primary care providers may be doctors, pediatricians, or physician assistants which are typically responsible for coordinating the care among specialists. The health system is positively impacted by primary care providers by offering greater access to health services, better health outcomes and a decrease in hospitalizations and use of emergency room visits. [Shi, 2012].
- **Secondary Care** refers to specialists which have more specific expertise in that particular medical problem. Specialists focus either on a specific system of the body or a specific disease or condition. For example, cardiologists focus on the heart and its pumping system while oncologists have a specialty in treating cancers. The main problems with specialty care emerge due to a wrong choice of the kind of specialist. In fact, it is possible that the initial symptoms suggest a certain diagnosis when in reality it is another condition that requires a different specialist. In other cases, problems arise when more than one specialist is consulted but each treats a different condition, providing care that may not be fully coordinated.
- **Tertiary care** is provided after the patient is hospitalized, where higher levels of specialist intervention are necessary. Tertiary care requires highly specialized equipment, procedures and expertise, such as coronary artery bypass surgery, hemodialysis and treatments for severe burns. Not all hospitals can provide this care, so patients need to be transferred to a medical center that provides highly specialized tertiary-level services. [Lo et al., 2016]
- **Quaternary care** is an extension of tertiary care that is not yet widespread since, being so specific, it is not offered by all hospitals or medical centers. Some may only offer quaternary care for particular medical conditions or systems of the body, such as for experimental medicine and procedures and highly uncommon and specialized surgeries.

In every developed country, however, the healthcare system is subject to some constantly changing social trends, including demographic change and the rising of technological innovation [Walshe et al., 2016]. As the number of elderly people is increasing rapidly, there is an incrementing incidence of chronic diseases, a direct result of risk factors such as tobacco use, physical inactivity, and unhealthy diets [World Health Organization 2005]. This also derives from the fact that technological and pharmaceutical innovation reflects an increasing ability to control chronic diseases. Even surgery, diagnostics and telemedicine continue to find new ways to cure diseases with more effective treatments, to slow the progress of the disease or manage its impact. Furthermore, the rapid development of life science industries and high-throughput technologies, have begun to revolutionize healthcare by allowing the examination of biological systems in unprecedented detail. Thanks to the emergence of omics information and big data, it is possible to deeply investigate the correlation between patients and diseases, according to their own molecular characteristics. Moreover, omics technologies have the potential to transform medicine from traditional symptom-oriented diagnosis and treatment of diseases towards individualized disease prevention and early diagnostics. Finally, the application of data mining in healthcare and public health, which has been growing over the last years as shown in Fig. 6 [Alyass et al., 2015, Koh et al., 2005], allows to systematically identify inefficiencies and best practices that improve care and reduce costs with remarkable economic benefits.

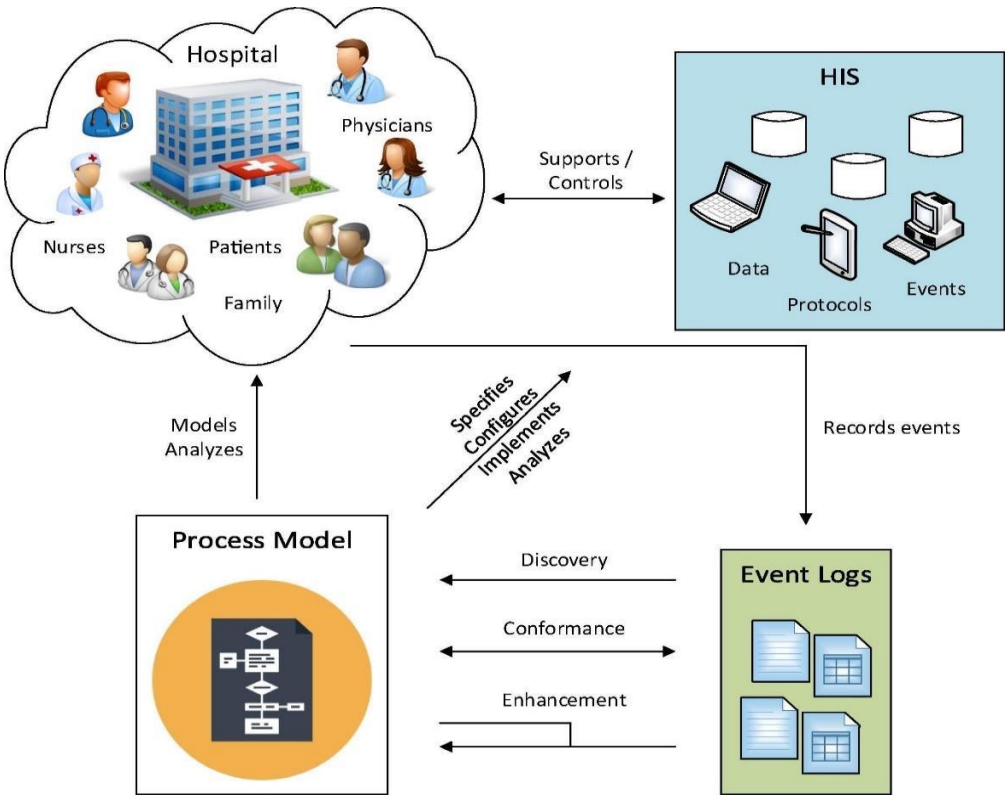


Figure 6. Process mining in healthcare. [Rojas et al., 2016]

However, the most important benefit of ICT in healthcare is understanding the complexity of diseases in the early stages so that they can be treated more easily and effectively, while also managing specific individual factors.

The most common applications involve predictive modelling, which includes traditional statistics, such as multiple discriminant analysis and logistic regression analysis, and non-traditional methods developed in the fields of AI and ML. Data mining applications can be developed to evaluate the efficacy of medical treatments by comparing causes, symptoms and treatment cycles [Milley, 2000]. For example, it is possible to compare the results of groups of patients treated with different drugs for the same disease or pathologic condition to determine which treatments work best and are most affordable. [Kincade, 1998] Similarly, data mining can help identify successful standardized treatments for specific diseases and determine more effective drug compounds for treating differently responding subpopulations to certain drugs [Milley, 2000]. To aid healthcare management, data mining applications can be developed to better identify and monitor chronic disease states and high-risk patients, design appropriate interventions, and reduce the number of hospital admissions [Schuerenberg, 2003]. Moreover, big data analysis allows to develop new techniques for disease prevention and treatment based on the individual molecular characteristics by integrating genomic information, lifestyle data and environmental information [Leyens et al., 2017]. The so called Precision Medicine (PM) was defined by the Horizon 2020 Advisory Group for Societal Challenge of the European Commission as "a medical model using molecular profiling technologies for tailoring the right therapeutic strategy for the right person at the right time and for maximizing the benefit-to-risk ratio; in this way it is possible to determine the predisposition to disease at the population level and to deliver stratified prevention" [European Commission, 2014]. PM not only affords the basis to develop new drugs, but also provides a wide knowledge of the patient, an essential step towards individualized medicine, as shown in Fig. 7. It is therefore essential to collect as much information and data as possible on each patient [Leyens et al., 2017] in order to identify the causes of the different responses to drugs from a pharmacogenomic perspective and to identify biological markers capable of accurately describing the risk signals to develop specific diseases. For this reason, PM approaches are already applied to different health areas such as oncology, cardiology and neurology. PM has the main ambitious goal not to tailor a medical treatment to a single patient, but rather to classify patients into subpopulations based on their sensitivity to a particular disease or their response to a specific

treatment. Patients' stratification aims to identify groups of patients with similar biological features who could respond to the same drug in a similar way [Laifenfeld et al., 2017].

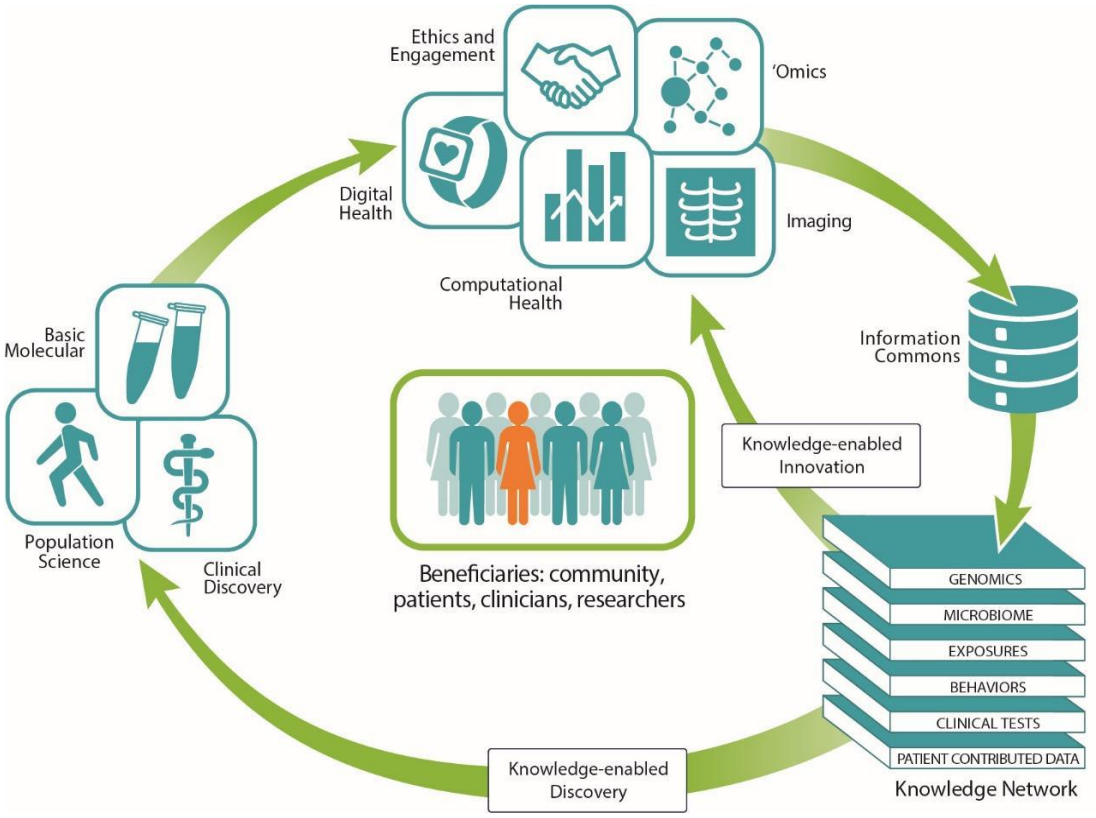


Figure 7. Precision medicine circular process.

The PM approach is also deeply applied to the healthcare area of rare diseases [Schee Genannt Halfmann et al., 2017] by creating patient registries, leveraging large amounts of data to discover potential links and including patients as active partners in this research [Trusheim et al., 2011]. Due to the rarity of these disorders, it is a challenge to convince companies to fund development of effective and affordable treatments, provide programmatic support and facilitate patient interaction. However, harmonizing databases and including the rare disease registries, are the major facilitators to understand the complexity of diseases, to conduct clinical trials, to improve the drug development process and to assign the right treatment to the right individual after a reliable patient stratification. Practical application of PM helps prevent treatment decisions from being guided by empirical practice of medicine, where physicians generally rely on standard models to establish a diagnosis based on a combination of patients' medical history, physical examination, and laboratory data. Such diagnoses can lead to the administration of drugs that only work in some people who suffer from that specific disease or which, even worse, can produce harmful side effects for the patient [Seyhan et al., 2019].

In the field of PM, many challenges still remain open, such as the integration of big data, patient empowerment, the translation of basic research into clinical research, in order to bring innovation to the market and reshape healthcare.

2. AIM OF THE THESIS

In this thesis we focus on the development of new Artificial Intelligence algorithms for a number of important problems in the field of Precision Medicine, Life Sciences and Healthcare. The idea behind the project is the belief that AI could help industries, which generate significant amounts of knowledge, transforming large and complex data into a format that makes them easier to use.

In a PM and Healthcare perspective, we have shown that we can predict the future health of individual patients with highly complex and rare diseases. Our first focus was to advance research on the rare Alkaptonuria disease towards a PM that addresses the complexity of the disease while taking into account individual variability. Moreover, we made two project proposals for ML based applications which support patients affected from rare diseases and their caregivers, resulting in benefits in many aspects of their life. Finally, we performed an intensive pharmacogenetic-oriented study focused to identify genetic markers to personalize cannabinoids treatment.

For the Life Sciences field, we were able to demonstrate the power of ML techniques in extracting information from protein data to make predictions on the protein structural features and for the assessment of their immunogenicity in a vaccine research perspective. We are confident that the development of computational modelling will guide biologists and clinical researchers in realizing the goals offered by this field, introducing possibilities that would be impossible without a deep knowledge on the data.

3. RESULTS AND DISCUSSION

3.1 ML APPLICATIONS TO PRECISION MEDICINE AND HEALTHCARE

As explained in the previous section, PM is an emerging approach which uses molecular profiling technologies for tailoring the right therapeutic strategy for the right person, in order maximize the benefit-to-risk ratio. Although PM is involved in many areas of health, its original purpose was to manage and analyze data related to rare diseases, which suffer from a lack of tools to enable meaningful recognition and understanding of clinical symptoms. We focus particularly on Alkaptonuria (AKU; MIM 203500), an ultra-rare autosomal recessive metabolic disease caused by the loss of the activity of the enzyme Homogentisate 1,2-dioxygenase (HGD; EC1.13.11.5), which affects between 1: 250000 and 1: 1000000 individuals worldwide [Phornphutkul et al., 2002]. Under physiological conditions, HGD is responsible for the conversion of homogentisic acid (HGA) to maleylacetoacetic acid in the tyrosine and phenylalanine pathway. In AKU patients, HGA is not metabolized but partially excreted with urine, where it imparts a characteristic black colour upon oxidation, and partially accumulates in the body where it polymerizes, forming a dark brown ochronotic pigment which is deposited in the connective tissue, resulting in an early onset of severe arthropathy.

In a PM context, rare diseases such as AKU suffer from the problem of data scarcity and sparsity, due to fragmented knowledge and limited number of data and specimens available. It is therefore necessary to implement an ecosystem capable of collecting, integrating and analyzing significant data flows from different research groups. For this reason, a Precision Medicine Ecosystem (PME) dedicated to AKU has been developed, called ApreciseKUre, which is a multidisciplinary, interactive and integrated AKU database (www.bio.unisi.it/aprecisekure/; www.bio.unisi.it/aku-db/) where genetic, biochemical and clinical resources are shared among scientists, clinicians and patients [Aronson et al., 2015]. The ApreciseKUre database allows the organization of data from different research groups in order to make them available and usable for clinicians, it performs the harmonization and standardization of different types of collected data and different sources and builds an easily searchable global reference point for AKU. Computational modelling and database creation can be a useful guide to generate a comprehensive and dynamic picture of a patient with AKU and to identify potential new biomarkers to achieve patient stratification. The creation of a database, which integrates patient-derived information (quality of life), physician-derived information (test results, genotypes) and mutational analysis (molecular modelling of proteins) offers a comprehensive visualization of different information layers to support doctors and researchers in the PM application

to AKU. Furthermore, ApreciseKUre was integrated with data mining models for understanding disease mechanisms. The outcome of these models can open up new opportunities to match therapy to the patient, thus leading to more personalized medicine which maximize the benefit/risk ratio [Rossi et al., 2020]. For instance, in order to address a first patient stratification, a tool for the prediction of the quality of life of AKU patients starting from clinical markers have been developed (see appendix: **Machine Learning application for development of a data-driven predictive model able to investigate Quality of Life scores in a rare disease**, Fig.8).

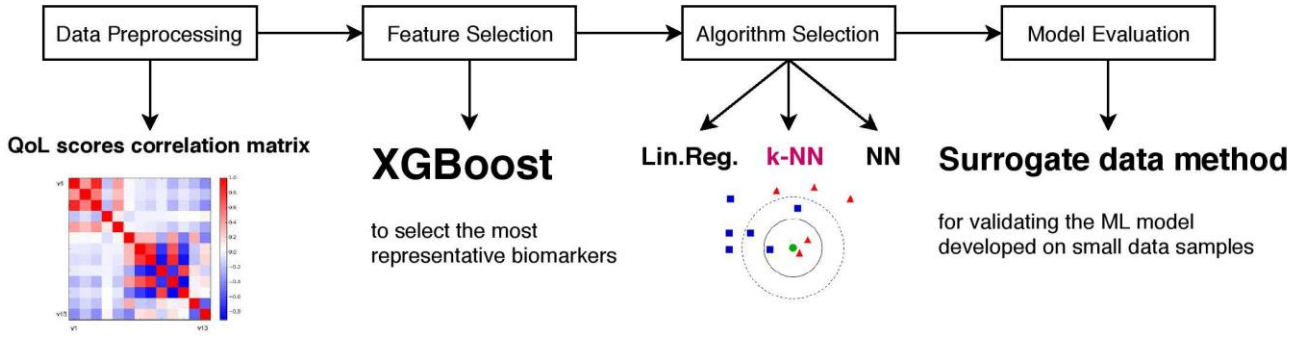


Figure 8. A 4-steps workflow of the ML-based classification model.

After the selection of the most statistically significant biomarkers using XGBoost [Chen et al., 2016], Quality of Life (QoL) score prediction was performed with k-NN. The innovative finding of this work is that, for the first time, we have found an ensemble of multiple complementary biomarkers whose combination produces better k-NN prediction of QoL scores than any single one. Moreover, due to the limited number of data available, the model has been validated using a surrogate data method, in which data were generated from random numbers able to mimic the distribution of the original dataset. They statistically resemble the original data in terms of their mean, standard deviation and range, but they do not maintain the complex relationships between the variables of the real dataset. Therefore, real-data models are consistent if they perform significantly better than the surrogate data models. In conclusion, this framework allowed ML algorithms to successfully predict clinical and QoL score outcomes despite small datasets. Furthermore, another ML approach was implemented with the aim of monitoring the evolution of biomarkers and QoL scores to tailor the treatment to each patient in a typical PM perspective (see appendix: **Towards a Precision Medicine Approach Based on Machine Learning for Tailoring Medical Treatment in Alkaptonuria**, Fig.9).

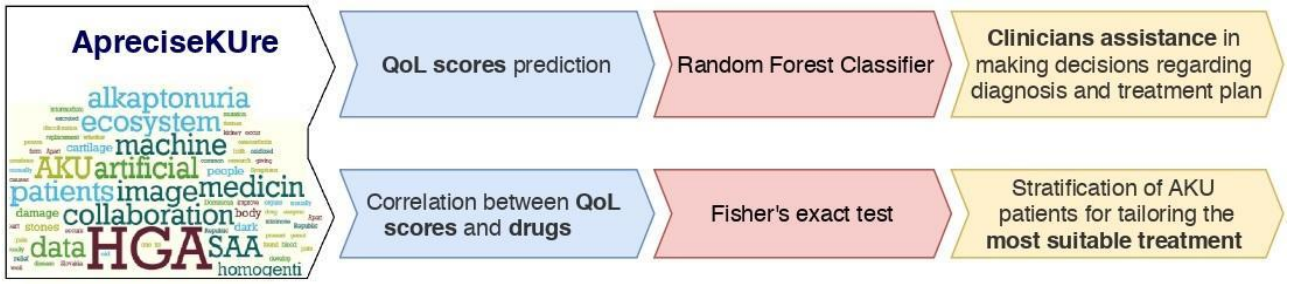


Figure 9. Workflow scheme represented by two stages, 'Quality of Life scores prediction' in the top and 'Correlation between QoL scores and drugs' in the bottom.

In this work, the prediction of the QoL scores, based on both personal and clinical AKU patients' information, was performed using the Random Forest algorithm, which suggested that the Knee injury and Osteoarthritis Outcome Score (KOOS) indicator could be a useful factor to better understand symptoms and difficulties experienced by AKU patients. Indeed, KOOS prediction could be fundamental to assess the main important prognostic biomarkers of AKU and the efficacy of future pharmacological treatments. Moreover, it has been looked for a correlation between the values of the QoL scores and the drugs the patients take. Fisher's exact test was applied on all the combinations QoL score vs. drug, employing the Benjamini-Hochberg procedure to deal with multiple comparisons. In this context, antihypertensive agents could help AKU patients to improve their conditions, as well as FANS and opioid, which resulted to be effective in reducing AKU pain, but also common drugs not related to specific AKU symptoms showed a correlation with some QoL scores. Overall, the validity and effectiveness of the proposed solutions show the potential direct benefits for patient care, treatment and early diagnosis, highlighting the necessity of patient databases for rare diseases. We believe this is not limited to the study of AKU, but it represents a proof of concept that could be applied to other rare diseases, allowing data management, analysis and interpretation. (see appendix: **Machine Learning Approaches in Precision Medicine: Applications to An Integrated Bioinformatics Digital Ecosystem Platform for A Rare Disease**).

Rare disease patients, therefore, face uncommon, severe and debilitating conditions, often characterized by poor prognosis and limited treatment options. In some cases, people lose the ability to speak and can only use their hands as their disease progresses, which can be emotionally devastating. That is because the ability to communicate is strongly associated with patients' QoL and communication is seen as crucial for the adaptation to terminal diseases. Therefore, while verbal as well as nonverbal communication abilities deteriorate, augmentative and alternative communication (AAC) strategies and technologies become more and more important. However, in the more severe cases AACs are not able to support communication, to express personality and feelings. Here the

necessity to implement a device, EMMA, addressed to severely disabled patients in order to reduce their psychological distress. EMMA is the acronym for Emotional Multimodal Assistant, a virtual assistant able to recognize and communicate emotions for people with rare diseases with impaired communication abilities. It has been already established that emotion recognition can provide a scientific basis for monitoring emotional health. Emotions are not only expressed through behavioural gestures, but also through a series of physiological signals [Xiefeng et al., 2019], which can be measured using electrocardiogram (ECG) [Nardelli et al., 2015], sweating detection and blood volume pulse. Among these changes, the heart rate variability extracted from an ECG, results to be one of the most important indicators of emotion recognition [Agrafioti et al., 2012, Gaetano et al., 2012]. By relying on this information, through advanced AI techniques, EMMA will be able to capture the guttural sounds emitted by the subject together with biometric data such as stress levels, saturation, heart rate and sleep quality and associate them with the emotions felt at that moment.
It will be therefore possible to develop an Artificial Neural Network which automatically learns how to associate the emotion experienced with the inputs acquired by EMMA. Integrating the application simply in a wearable device it will be possible to process the inputs and communicate in real time, through audio-visual feedback, the emotional state of the patients, as shown in Fig. 10.

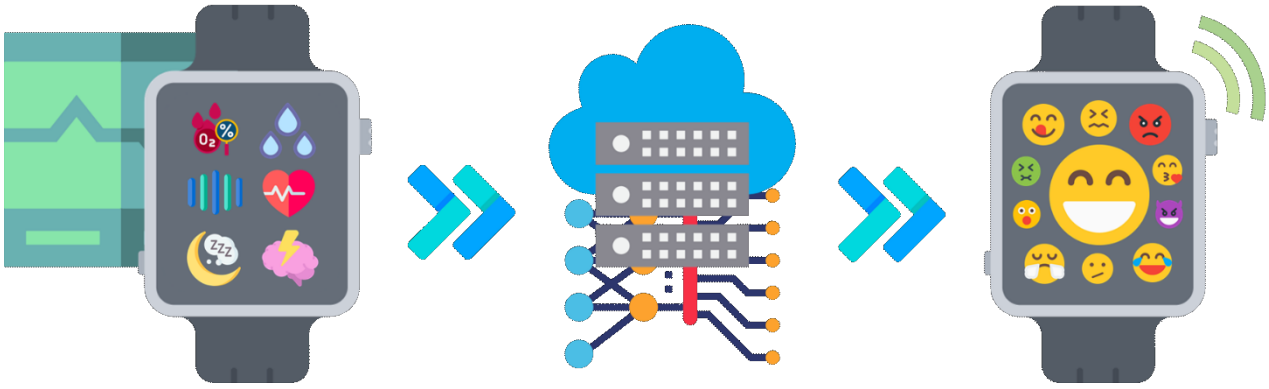


Figure 10. EMMA prototype. From the left: capture of the guttural sounds emitted by the subject together with biometric data such as stress levels, saturation, heart rate and sleep quality. Association with the emotions felt at that moment. Communication of the emotional state of the patient in real time, through audio-visual feedback.

EMMA brings numerous benefits to both the patient and the caregiver, constituting a palliative care tool which provides immediate feedback facilitating diagnostic, therapeutic choices and improving quality of life. This project was awarded with the first prize at the Rare Disease Hackathon contest.

This positively affects caregivers, who play an important role in supporting people with rare diseases. Caregivers provide daily assistance to people with impairments caused by ageing, chronic diseases, infirmities, etc. The constant attention to the patient's needs, and the social isolation that the role of being caregivers entails are at the basis of the obstacles they have to deal with in the daily assistance [Navaie-Waliser et al., 2002]. In a rare disease, where diagnosis is often a slow and difficult process [Thevenon et al., 2016] which can lead to sudden changes in the life of a patient, it is often challenging for a caregiver to give immediately the appropriate support to the patient. A further obstacle is represented by the fact that rare diseases-dedicated associations are generally dispersed around the world. This makes it difficult for caregivers and their patients to communicate with specialized centers, resulting in the lack of psychological and practical support. In order to cope with the issues of isolation and poor communication with healthcare professionals, a network of caregivers is extremely valuable [Munsell et al., 2011]. Here the necessity to develop a cross–platform application, CaregiverMatcher, which facilitates communication between caregivers, patients' associations and specialists. (see appendix: **CaregiverMatcher: graph neural networks for connecting caregivers of rare disease patients**, Fig. 11).

A direct communication channel between caregivers is realized by means of a graph neural network (GNNs [Scarselli et al., 2009]), which performs a matching between similar caregivers based on information regarding the assisted patient. Consequently, CaregiverMatcher would give the opportunity to caregivers to establish direct contact with other people that face similar issues in daily assistance. Moreover, this platform offers a section of easily understandable information material on rare diseases curated by doctors, associations and health professionals, with useful links to get in touch with them, as well as to external websites or to additional material. This multi-purpose platform, enriched with artificial intelligence tools, could pave the way for the development of other platforms for the exchange of information between scientists, doctors and patients in the field of rare diseases.

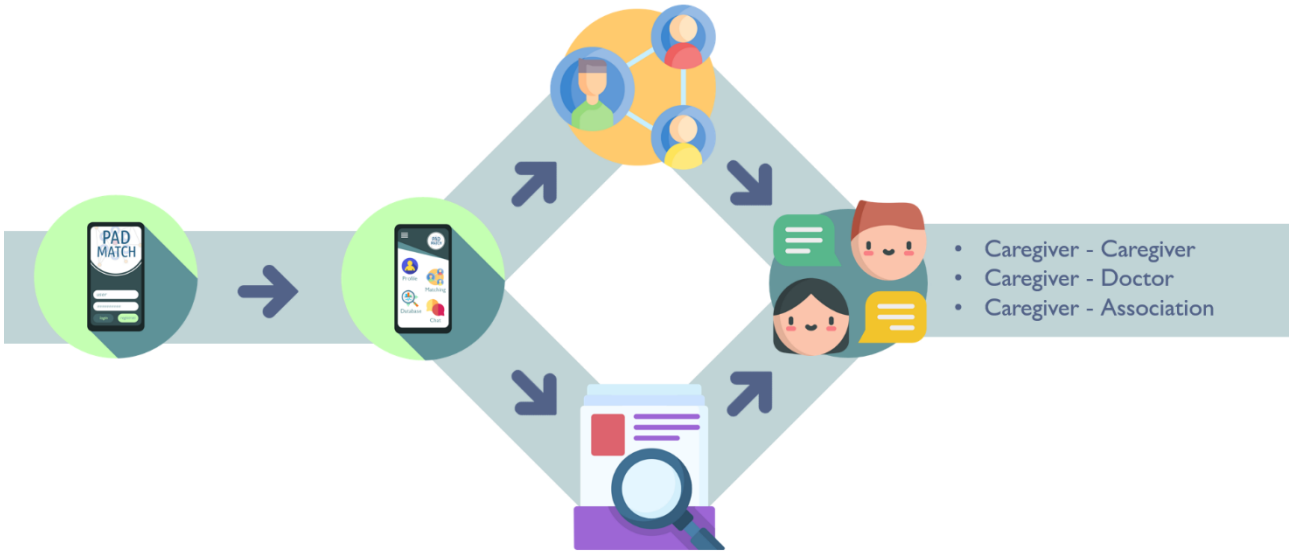


Figure 11. General architecture of CaregiverMatcher mobile app. From the left: to access the platform, caregivers log in with username and password. Four sections are available in the home page: *Profile*, to manage personal and patient data; *Chat*, where all messages and chat conversations are stored; *Get Informed* to retrieve rare diseases information as well as associations or doctors contacts; *Match* to start the matching process. As a result, caregivers can then connect with patient associations, specialized clinicians and other caregivers.

In the general framework just described, designed to define an innovative approach to personalized medicine and to the targeted support of the patient during the whole evolution of the disease, and thanks to advances in genetics and the increasing availability of health data, PM increases the ability of physicians to use patients' genetic and molecular information as part of routine medical care. In cancer, for example, each tumour is involved in interactions with various non-cancer elements such as gene-environment interactions (GxE), transcriptional regulation and gene co-expression [De Anda-Jáuregui et al., 2020]. The application of genetic data integration and analysis, as well as the use of molecular modelling algorithms, allows to formulate many predictions of drug-target interactions to greatly facilitate the guided development of personalized drugs [Tolios et al., 2020]. (see appendix: **Multi-Omics Model Applied to Cancer Genetics**, Fig. 12).

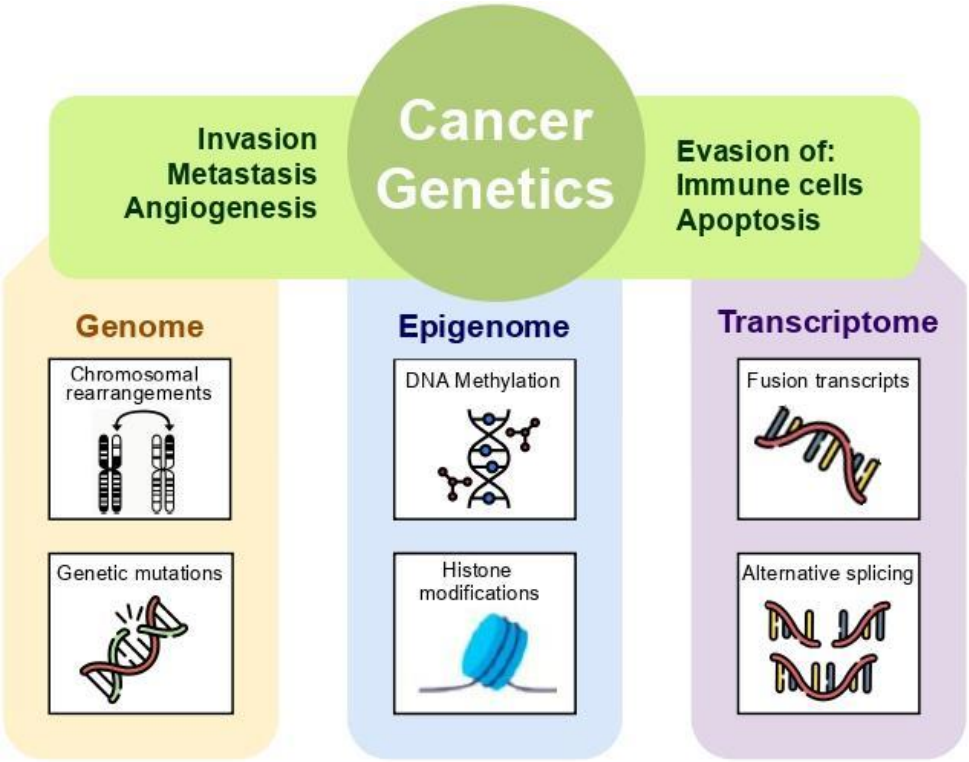


Figure 12. Levels of interactions found in a cancer system, that can be measured via the different omics technologies, such as genomics, epigenomics, transcriptomics, and proteomics.

Furthermore, from a pharmacogenetic point of view, many researchers focus on using the individual's genome to prescribe the safest and most effective drug for a specific patient [Drew, 2016]. For example, the patient's response to cannabinoid treatment may have a genetic background, depending on gene polymorphisms involved in the metabolism of these substances in the organism. Different variants may determine different therapeutic effects or the occurrence of possible side effects [Hryhorowicz et al., 2018]. Starting from this assumption, a genetic-based precision medicine approach was developed for patients treated with cannabinoids.

First of all, a dataset of patients, containing both static and dynamic features, was created. Static features consisted of personal, genetic and pathological data, while dynamic characteristics included data that change over the treatment period, such as the daily dose of cannabinoids taken, the side effects of the therapy, and the Visual Analogue Scale (VAS). VAS is a one-dimensional measure of pain intensity widely used in the adult population, with values in a "0–10" scale, from no pain to extreme one. Moreover, although there are more than 100 cannabinoids isolated from Cannabis, we concentrated on the prediction of Δ9-tetrahydrocannabinol (THC) and cannabidiol (CBD). THC is the most studied of these cannabinoids and also the most psychoactive, while CBD shouldn't have

any intoxicating or psychoactive effects, but could be a potential treatment for a wide range of diseases.

Once the dataset has been pre-processed, a prediction model based on Gradient Boosting Regression was implemented, with an incredible precision in predicting the daily dose of THC and CBD. Moreover, features' importance in the prediction was calculated, resulting in a significant importance of genetic polymorphisms, demonstrating that they have a crucial role in the response to cannabinoids. In the next few years, pharmacogenetics could therefore provide answers regarding the medical use of cannabis, with the identification of future genetic markers to personalize cannabinoids treatment.

3.2 ML APPLICATIONS TO LIFE SCIENCES

The rapid development of life science industries, high-throughput technologies, computational frameworks, have begun to revolutionize healthcare by allowing the examination of biological systems in unprecedented detail. Artificial intelligence methods are able to process the enormous amounts of data coming from the Life Science industries at an unprecedented speed, revealing information and patterns hidden within. As an example, the ability to predict protein structure is making a revolution in biology as it allows us to better understand how all the information maintenance/transformation mechanisms of cells work. Providing high-quality 3D structures lets structural biologists focus their work on applications related to human health, for example tackling some of the most serious diseases by predicting the structures of the proteins involved, characterising how they interact, and understanding how they cause a certain disease. New proteins could be designed for novel vaccines or biological therapies to modulate diseases, and new candidate drugs can be identified more effectively. However, determining the 3D structure of proteins is one of the most challenging tasks in computational biology. Although gradual developments have been made in predicting the 3D conformation, the results obtained are generally of lower quality than experimental techniques – apart from the very recent AlphaFold from Google [Deepmind, 2020], that anyway presupposes an enormous computational power –, which actually constitutes the ground-truth for evaluating the performance of predictive methods. Thanks to these tests, it has been discovered that protein conformations can be established mainly on the basis of the sequence of its amino acids [Anfinsen, 1973], a fundamental hypothesis which could help the development of novel protein folding prediction techniques. This is further corroborated by the rapid advances in genomics and proteomics, which have seen the discovery of millions of protein sequences that can be processed in a reasonable time through computational approaches for the prediction of the protein structure. Moreover, the protein secondary structure prediction could provide a significant first step toward the

tertiary structure prediction, also yielding information about protein activity, relationships, and functions. This means that, given a protein sequence, the first step towards the prediction of the three-dimensional native configuration consists in determining which backbone regions are likely to form helices, strands and turns. Secondary structure prediction algorithms employ a variety of computational techniques, including neural networks [Wang et al., 2016], hidden Markov models [Aydin et al., 2007], clustering techniques and genetic algorithms [Thanh et al., 2015]. In the present study, based on the intuition that signals should exist, in the form of particular amino acid concentrations, which determine the formation of secondary structures and define their extension, we carried out a statistical analysis of the amino acid concentrations in the vicinity of α-helices, revealing that informative patterns can be evidenced at the beginning and at the end of amino acid sequences representing α-helices (see appendix: **A deep attention network for predicting amino acid signals in the formation of α-helices**, Fig. 13).

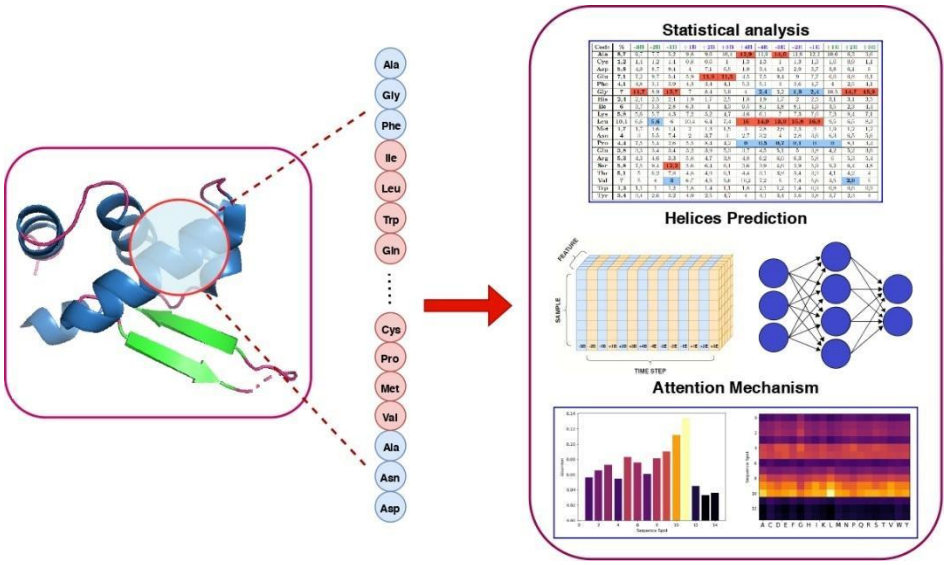


Figure 13. Workflow scheme. From the top: Statistical analysis, helices prediction and Attention Mechanism implementation.

In order to validate this assumption, three ML models, specifically built for the task of amino acid signal identification were implemented and compared. Each of them was equipped with an attention module, which measures the importance given by the model to each feature in each sequence position, allowing an interpretation of its behaviour. The experimental results show how all the models focus on the most important information, suggesting that the amino acids located at the sequence boundaries are fundamental in determining the occurrence of α-helices. It is a matter of future research to extend the proposed approach to the prediction of signals defining other common secondary structures,

namely β-sheets and *U*-turns. Indeed, the combination of the existing sophisticated ML techniques with a deeper knowledge of the primary and secondary structure information content will play a significant role for the prediction of the structure of new proteins.

Next to protein structure prediction, as one of the most successful advances in modern medicine, vaccination is still facing the difficulty of developing safe and effective vaccines against many infectious diseases such as tuberculosis, HIV and Salmonella. Salmonella in particular is a bacterium that causes life-threatening diseases in adults and children. Typhoid fever, paratyphoid fever and invasive non-typhoidal Salmonella infections (iNTS) have a high incidence worldwide and coexist in many geographical areas. Current treatment for Salmonella infections is insufficiently effective [Hohmann, 2001; Kariuki et al., 2015], taking Salmonella on the WHO antimicrobial resistance high-priority list [World Health Organization, 2017]. The possibility to deliver multiple antigens and to confer protection against multiple Salmonella serovars is therefore becoming increasingly important. We exploited the potential of machine learning methods for the identification of protective immunogens towards improved vaccine design and development. First, AI techniques were applied to classify proteins as immunogenic or non-immunogenic from a vaccine research perspective, information that will later be supplemented with the use of the modified outer membrane vesicle (mOMV) platform, to test its effectiveness. in in vitro/in vivo models. For this project, we developed a prediction model of bacterial immunogens based on XGBoost, showing an outstanding ability to identify Salmonella immunogens. The potential deriving from SHASI-ML and the later combination AI/mOMV would automatically lead to a number of advantages in the area of vaccine product development. Using a new vaccine for Salmonella as an example, there would be immediate advantages in terms of cost and ease of production, safety and flexibility but also significant social repercussions, both in terms of increasing global health security and in providing assistance to low- and middle-income countries. The use of this methodology for the development of a universal vaccine against Salmonella constitutes an ideal test bed for its applicability to research and development processes of any other vaccine. Indeed, being able to identify vaccines and curative drugs more quickly for a new and emerging infectious disease would have a very significant impact on a global level.

Fig. 14 represents a map showing all the above-mentioned projects and publications concerning Machine Learning application to Precision Medicine, Life Sciences and Healthcare from my PhD.

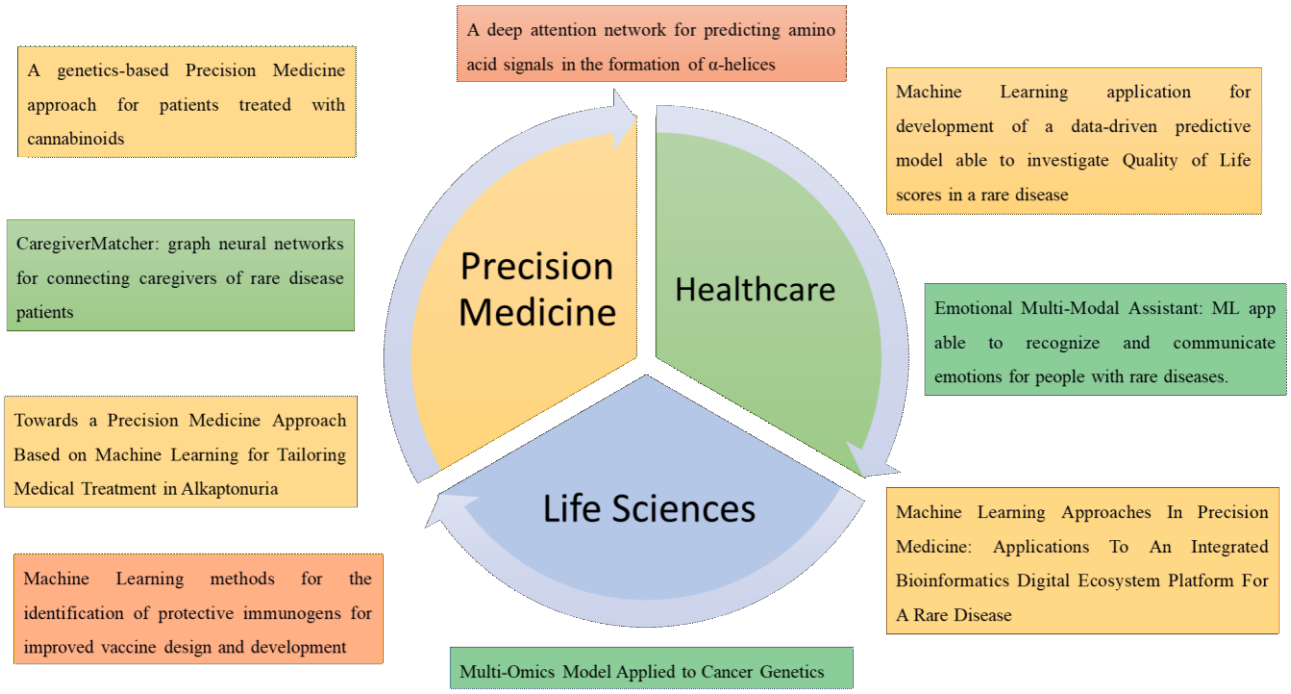


Figure 14. Map of projects and publications arising from my PhD.

4. CONCLUSIONS AND FUTURE PERSPECTIVES

The rapid development of life science industries and high-throughput technologies have begun to revolutionize biological research by allowing the examination of biological systems in an extraordinary detail. The use of computational tools has become ever more crucial, because they are able to process enormous amounts of data at an unprecedented speed, revealing information and patterns hidden within. Therefore, Bioinformatics in combination with Machine Learning represent a key factor for the development of algorithms and software for the transfer, storage, analysis and development of biological platforms. Data mining applications can be developed to evaluate the efficacy of medical treatments by comparing causes, symptoms and treatment cycles. Similarly, computational methods can help identify successful standardized treatments for specific diseases and determine more effective drug compounds for treating differently responding subpopulations to certain drugs. Moreover, the new Precision Medicine techniques make it possible to develop models for disease prevention and treatment based on the individual molecular characteristics, by integrating genomic information, lifestyle data and environmental information. The Precision Medicine approach is deeply applied to the healthcare area, and in particular to rare diseases, with the creation of patient registries leveraging large amounts of data to discover potential links. Based on the idea that AI could help industries which generate significant amounts of knowledge, in this thesis we have focused on the development of novel Artificial Intelligence algorithms for a number of important problems in the field of Precision Medicine, Life Sciences and Healthcare. The application of Bioinformatics and Computational biology algorithms together with the creation of digital databases will offer an opportunity to translate new data into actionable information, thus allowing earlier diagnosis and precise treatment options. Indeed, our results are very encouraging, and suggest continuing using ML approaches in the biological field. Many challenges still remain open, requiring the development of alternative strategies to complement/improve existing techniques.

5. REFERENCES

1. W.S. McCulloch and W. Pitts, A logical calculus of the ideas immanent in nervous activity, Bulletin of Mathematical Biophysics 5:115_133, 1943.
2. J.V. Neumann, The Computer and the Brain, Yale University Press, USA, 1958.
3. D.O. Hebb, The organization of behavior; a neuropsychological theory, Wiley, 1949.
4. A.M. Turing, Computing Machinery and Intelligence, MIND 59:433_460, 1950.
5. Y. Bengio, A. Courville, I. Goodfellow, Deep Learning, MIT Press, 2016.
6. M. Awad, R. Khanna, Machine learning. Efficient learning machines. Apress, Berkeley, CA, 2015.
7. P. Cunningham, M. Cord, S.J. Delany, Supervised Learning. In: M. Cord, P. Cunningham, Machine Learning Techniques for Multimedia. Cognitive Technologies. Springer, Berlin, Heidelberg, 2008.
8. Z. Ghahramani, Unsupervised Learning. In: O. Bousquet, U. von Luxburg, G. Rätsch (eds) Advanced Lectures on Machine Learning. Lecture Notes in Computer Science, vol 3176. Springer, Berlin, Heidelberg, 2004.
9. X.J. Zhu, Semi-supervised learning literature survey, 2005.
10. S.J. Russell, P. Norvig, Artificial intelligence: a modern approach. Malaysia, Pearson Education Limited, 2016.
11. J. Alzubi et al., Machine Learning from Theory to Algorithms: An Overview J. Phys.: Conf. Ser. 1142 012012, 2018.
12. D.H. Wolpert, W.G. Macready et al., No free lunch theorems for optimization. IEEE transactions on evolutionary computation, 1(1):67–82, 1997.
13. K. Cios, H. Mamitsuka, T. Nagashina, Computational intelligence techniques in bioinformatics (special issue). Artificial Intelligence in Medicine, 35 (1—2), 2005.
14. N. Auslander, A.B. Gussow, E.V. Koonin, Incorporating Machine Learning into Established Bioinformatics Frameworks, *Int. J. Mol. Sci.*, *22*, 2903, 2021.
15. A.C. Tan, D. Gilbert, Machine learning and its application to bioinformatics: an overview, 2001.
16. X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, et al., Top 10 algorithms in data mining. Knowledge and Information Systems, 14(1), 1–37, 2008.
17. C. Cortes, V. Vapnik, Support-vector networks. Machine Learning, 20(3), 273–297, 1995.
18. D. Whitley, A genetic algorithm tutorial. Statistics and Computing, 4(2), 65–85, 1994.
19. R. Polikar, Ensemble learning. Scholarpedia, 4(1):2776, 2009.

20. L. Breiman, Bagging predictors. Machine Learning, 26 (2): 123–140, 1996.
21. Y. Freund, R. Schapire, Experiments with a new boosting algorithm, Proceedings of the Thirteenth National Conference on Machine Learning, pp. 148–156, 1996.
22. L. Breiman, Random Forests. Machine Learning, 45 (1): 5–32, 2001.
23. R. Chen, M. Snyder, Promise of personalized omics to precision medicine. Wiley interdisciplinary reviews. Systems biology and medicine, 5(1), 73–82, 2013.
24. J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, et al. The sequence of the human genome, Science, 291(5507):1304—51, 2001.
25. A. Bernal, K. Crammer, A. Hatzigeorgiou, F. Pereira, Global discriminative learning for higher-accuracy computational gene prediction. PLoS Computational Biology, 3(3), 2007
26. G. Fogel, W. Porto, D. Weekes, Prediction of the phenotypic effects of non synonymous single nucleotide polymorphisms using structural and evolutionary information. Nucleic Acid Research, 30(23):5310—7, 2002.
27. N. Lopez-Bigas, C. Ouzounis, Genome-wide identification of genes likely to be involved in human genetic diseases. Nucleic Acid Research, 32(10):3108—14, 2004.
28. M. Ritchie, B.C. White, J.S. Parker, L.W. Hahn, J.H. Moore. Optimization of neural network architecture using genetic programming improves detection and modelling of gene—gene interactions in studies of human diseases. BMC Bioinformatics;4(28), 2003.
29. M. Rathee, T.V. Vijay Kumar, DNA fragment assembly using multi-objective genetic algorithms. Int J Appl Evol Comput 5(3):84–108, 2004.
30. N. Bidargaddi, M. Chetty, J. Kamruzzaman, Combining segmental semi-Markov models with neural networks for protein secondary structure prediction, Neurocomput 72:3943– 3950, 2009.
31. Z. Li, Y. Yu, Protein secondary structure prediction using cascaded convolutional and recurrent neural networks, Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, pp. 2560–2567, 2016.
32. K., Daisuke. Protein Function Prediction: Methods and Protocols. 2017.
33. N.K. Fox et al., The value of protein structure classification information-Surveying the scientific literature, Proteins vol. 83,11, 2025-38, 2015.
34. J. He, H.J. Hu, R. Harrison, P.C. Tai, Y. Pan, Rule generation for protein secondary structure prediction with support vector machines and decision tree. IEEE TransNanoBiosci 5(1):46–53, 2006.
35. R.N. Nijil, T. Mahalekshmi, Multilabel classification of membrane protein in human by decision tree (DT) approach. Biomed Pharmacol J 11(1), 2018.

36. A. Cambiaghi, M. Ferrario, M. Masseroli, Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration Briefings Bioinf, 18 (3), pp. 498-510, 2016
37. D.H. Nguyen, C.H. Nguyen, H. Mamitsuka, Recent advances and prospects of computational methods for metabolite identification: a review with emphasis on machine learning approaches Briefings Bioinf, 20 (6), pp. 2028-2043, 2018.
38. L. Puchades-Carrasco *et al.,* Bioinformatics tools for the analysis of NMR metabolomics studies focused on the identification of clinically relevant biomarkers Briefings Bioinf, 17 (3), pp. 541-552, 2015.
39. Y. Fan, T.B. Murphy, J.C. Byrne, et al., Applying random forests to identify biomarker panels in serum 2D-DIGE data for the detection and staging of prostate cancer. *J Proteome Res.*;10(3):1361–73, 2011.
40. J. Khan, J. S. Wei, M. Ringnér, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, P. S. Meltzer, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med. 7, 673–679, 2001.
41. H. Zhang, C. Y. Yu, B. Singer, Cell and tumor classification using gene expression data: Construction of forests. Proc. Natl. Acad. Sci. U.S.A. 100, 4168–4172, 2003.
42. C.H. Yeang, S. Ramaswamy, P. Tamayo, S. Mukherjee, R.M. Rifkin, M. Angelo, M. Reich, E. Lander, J. Mesirov, T. Golub, Molecular classification of multiple tumor types. Bioinformatics 17 (Suppl. 1), S316–S322, 2001.
43. R.R. Weichselbaum, H. Ishwaran, T. Yoon, D.S.A. Nuyten, S.W. Baker, N. Khodarev, A.W. Su, A.Y. Shaikh, P. Roach, B. Kreike, B. Roizman, J. Bergh, Y. Pawitan, M.J. van de Vijver, A.J. Minn, An interferon-related gene signature for DNA damage resistance is a predictive marker for chemotherapy and radiation for breast cancer. Proc. Natl. Acad. Sci. U.S.A. 105, 18490–18495, 2008.
44. H. Zhao, C.J. Logothetis, I.P. Gorlov, Usefulness of the top-scoring pairs of genes for prediction of prostate cancer progression. Prostate Cancer Prostatic Dis. 13, 252–259, 2010.
45. S.K. Patnaik, E. Kannisto, S. Knudsen, S. Yendamuri, Evaluation of microRNA expression profiles that may predict recurrence of localized stage I non–small cell lung cancer after surgical resection. Cancer Res. 70, 36–45, 2010.
46. T.S. Deisboeck, Z. Wang, P. Macklin, V. Cristini, Multiscale cancer modeling. Annu. Rev. Biomed. Eng. 13, 127–155, 2011.

47. J. Relan, P. Chinchapatnam, M. Sermesant, K. Rhode, M. Ginks, H. Delingette, C.A. Rinaldi, R. Razavi, N. Ayache, Coupled personalization of cardiac electrophysiology models for prediction of ischaemic ventricular tachycardia. Interface Focus 1, 396–407, 2011.
48. M.I. Miller, A. Qiu, The emerging discipline of computational functional anatomy, Neuroimage 45, S16–S39, 2009.
49. D. Stumpfe, J. Bajorath, Current trends, overlooked issues, and unmet challenges in virtual screening. Journal of chemical information and modelling, 60, 4112-4115, 2020.
50. P. Schneider, W.P. Walters, A.T. Plowright, N. Sieroka, J. Listgarten, R.A. Goodnow, J. Fisher, J. M. Jansen, J. S Duca, T.S. Rush et al., Rethinking drug design in the artificial intelligence era. Nature Reviews Drug Discovery, 19, 353-364, 2020.
51. J. Boström, D.G. Brown, R.J. Young, G.M. Keserü, Expanding the medicinal chemistry synthetic toolbox. Nature Reviews Drug Discovery 2018, 17, 709-727
52. A. Strokach, D. Becerra, C. Corbi-Verge, A. Perez-Riba, P.M. Kim, Fast and flexible protein design using deep graph neural networks. Cell Systems, 11, 402-411, 2020.
53. G.P. Moss, A.J. Shah, R. G. Adams, N. Davey, S.C. Wilkinson, W.J. Pugh, Y. Sun, The application of discriminant analysis and Machine Learning methods as tools to identify and classify compounds with potential as transdermal enhancers. Eur. J. Pharm. Sci., 45(1-2), 116-127, 2021.
54. A. Molfetta, W.F.D. Angelotti, R.A.F Romero, C.A. Montanari, A.B.F. da Silva, A neural networks study of quinone compounds with trypanocidal activity. J. Mol. Model., 14(10), 975-985, 2008.
55. T.J. Hwang, D. Carpenter, J.C. Lauffenburger, B. Wang, J. M. Franklin, A.S. Kesselheim, Failure of investigational drugs in late-stage clinical development and publication of trial results. JAMA Internal Med, 176(12):1826–33, 2016.
56. M. Aziz, E. Kaufmann, M.K. Riviere, On Multi-Armed Bandit Designs for Phase I Clinical Trials, arXiv preprint arXiv:1903.07082, 2019.
57. S.M. Berry, B. P. Carlin, J.J. Lee, P. Muller, Bayesian adaptive methods for clinical trials CRC Press, 2010.
58. L. Shi, The impact of primary care: a focused review. Scientifica (Cairo), 2012:432892, 2012.
59. C. Lo, D. Ilic, H. Teede et al., Primary and tertiary health professionals' views on the healthcare of patients with co-morbid diabetes and chronic kidney disease - a qualitative study. BMC Nephrol, 17(1):50, 2016.
60. K. Walshe, J. Smith, Healthcare management. 3rd ed. Open University Press, 2016.

61. World Health Organization. WHO Framework Convention on Tobacco Control. Geneva, Switzerland: WHO, 2005.
62. A. Alyass, M. Turcotte, D. Meyre, From big data analysis to personalized medicine for all: challenges and opportunities, BMC Med Genet 8(1):33, 2015.
63. H.C. Koh, G. Tan, Data mining applications in healthcare. J Healthc Inf Manag. Spring;19(2):64-72, 2005.
64. A. Milley, Healthcare and data mining. Health Management Technology, 21 (8), 44-47, 2000.
65. K. Kincade, Data mining: digging for healthcare gold. Insurance & Technology, 23 (2), IM2-IM7, 1998.
66. E. Rojas, J. Munoz-Gama, M. Sepúlveda, D. Capurro, Process mining in healthcare: A literature review, Journal of Biomedical Informatics, Volume 61, Pages 224-236, ISSN 1532-0464, 2016.
67. B.K. Schuerenberg, An information excavation. Health Data Management, 11 (6), 80-82, 2003.
68. L. Leyens, M. Reumann, N. Malats, A. Brand, Use of big data for drug development and for public and personal health and care, Genet Epidemiol 41(1):51–60, 2017.
69. European Commission, Directorate-General for Health and Consumers, Unit D3 eHealth and Health Technology Assessment, The Use of Big Data in Public Health Policy and Research, 2014.
70. D. Laifenfeld, D.A. Drubin, N.L. Catlett, J.S. Park, A.A. Van Hooser, B.P. Frushour, D. de Graaf, D.A. Fryburg, R. Deehan, Early patient stratification and predictive biomarkers in drug discovery and development: a case study of ulcerative colitis anti-TNF therapy. Adv Exp Med Biol 736:645-53, 2012.
71. S. Schee Genannt Halfmann, L. Mählmann, L. Leyens, M. Reumann, A. Brand, Personalized Medicine: What's in it for Rare Diseases?. Adv Exp Med Biol 1031:387-404, 2017.
72. M.R. Trusheim, B. Burgess, S. Xinghua Hu, T. Long, S.D. Averbuch, A.A. Flynn, A. Lieftucht, A. Mazumder, A. Milloy, A.P. Shaw, D. Swank, J. Wang, E.R. Berndt, F. Goodsaid, M.C. Palmer, Quantifying factors for the success of stratified medicine. Nature Reviews Drug Discovery 10:11, 2011.
73. A. Seyhan, C. Carini, Are innovation and new technologies in precision medicine paving a new era in patients centric care? Journal of Translational Medicine 17:114, 2019.
74. C. Phornphutkul, W.J. Introne, M.B. Perry, I. Bernardini, M.D. Murphey, D.L. Fitzpatrick, P.D. Anderson, M. Huizing, Y. Anikster, L.H. Gerber, W.A. Gahl, Natural history of alkaptonuria. N Engl J Med 347(26):2111-21, 2002.

75. S.J. Aronson, H.L. Rehm, Building the foundation for genomics in Precision Medicine. Nature 526(7573):336-42, 2015.
76. Rossi A, Giacomini G, Cicaloni V, Galderisi S, Milella MS, Bernini A, Millucci, Lia, Spiga O, Bianchini M, Santucci A. (2020). "AKUImg: A database of cartilage images of Alkaptonuria patients." Computers in Biology and Medicine. 122. 103863.
77. T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). New York, NY, USA: ACM, 2016.
78. C. Xiefeng, Y. Wang, S. Dai, et al., Heart sound signals can be used for emotion recognition. Sci Rep 9, 6486, 2019.
79. M. Nardelli et al., Recognizing Emotions Induced by Affective Sounds through Heart Rate Variability. IEEE Transactions on Affective Computing. 6, 385–394, 2015.
80. F. Agrafioti, D. Hatzinakos, A.K. Anderson, ECG Pattern Analysis for Emotion Detection. IEEE Transactions on Affective Computing. 3, 102–115, 2012.
81. V. Gaetano et al., Dominant Lyapunov exponent and approximate entropy in heart rate variability during emotional visual elicitation. Frontiers in Neuroengineering. 5, 3, 2012.
82. M. Navaie-Waliser, P.H. Feldman, D. A. Gould, C. Levine, A.N. Kuerbis, K. Donelan, When the Caregiver Needs Care: The Plight of Vulnerable Caregivers. American Journal of Public Health, 92(3):409–413, ISSN 0090-0036, 1541-0048, 2002.
83. J. Thevenon, Y. Duffourd, A. Masurel-Paulet, M. Lefebvre, F. Feillet, S. El Chehadeh-Djebbar, J. St-Onge, A. Steinmetz, F. Huet, M. Chouchane, V. Darmency-Stamboul, P. Callier, C. Thauvin-Robinet, L. Faivre, and J.B. Rivire, Diagnostic odyssey in severe neurodevelopmental disorders: toward clinical whole-exome sequencing as a first-line diagnostic test. Clinical Genetics, 89(6), 2016. ISSN 00099163, 2016.
84. E. Palamaro Munsell, R. P. Kilmer, J. R. Cook, C. L. Reeve, The effects of caregiver social connections on caregiver, child, and family wellbeing. American Journal of Orthopsychiatry, 82(1):137–145, ISSN 1939-0025, 0002-9432, 2012.
85. F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, The graph neural network model. IEEE Transactions on Neural Networks, 20(1):61–80, 2009.
86. G. De Anda-Jáuregui, E. Hernández-Lemus, Computational Oncology in the Multi-Omics Era: State of the Art. Front. Oncol, 2020.
87. A. Tolios, J. De Las Rivas, E. Hovig, P. Trouillas, A. Scorilas, T. Mohr, Computational approaches in cancer multidrug resistance research: Identification of potential biomarkers, drug targets and drug-target interactions. Drug Resist. Updat., 48, 100662, 2020.

88. L. Drew, Pharmacogenetics: The right drug for you. Nature 537, S60–S62, 2016.
89. S. Hryhorowicz, M. Walczak, O. Zakerska-Banaszak, R. Słomski, M. Skrzypczak-Zielińska, Pharmacogenetics of Cannabinoids. Eur J Drug Metab Pharmacokinet, 43(1):1-1, 2018.
90. Deepmind. "AlphaFold", Retrieved 30 November 2020.
91. C.B. Anfinsen, Principles that govern the folding of protein chains, Science 181:223–230, 1973.
92. Z. Aydin, Y. Altunbasak, M. Borodovsky, Protein secondary structure prediction for a single-sequence using hidden semi-Markov models, BMC Bioinf. 7 (1), 178, 2006.
93. Y. Wang, H. Mao, Z. Yi, Protein Secondary Structure Prediction by Using Deep Learning Method Knowledge-Based Systems, 2016.
94. T. Nguyen, A. Khosravi, D. Creighton et al., Multi-output interval type-2 fuzzy logic system for protein secondary structure prediction, Int. J. Uncertainty Fuzziness Knowledge Based Syst. 23 (05) 735–760, 2015.
95. Hohmann EL. 2001. Nontyphoidal salmonellosis. Clin Infect Dis 32:263-269.
96. S. Kariuki, M.A. Gordon, N. Feasey, C.M. Parry, Antimicrobial resistance and management of invasive Salmonella disease. Vaccine 33 Suppl 3:C21-C29, 2015.
97. World Health Organization. Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics, 2017.

6. APPENDIX

Spiga et al. Orphanet Journal of Rare Diseases (2020) 15:46
https://doi.org/10.1186/s13023-020-1305-0

Orphanet Journal of Rare Diseases

RESEARCH Open Access

Machine learning application for development of a data-driven predictive model able to investigate quality of life scores in a rare disease

Ottavia Spiga[1*†], Vittoria Cicaloni[1,2†], Cosimo Fiorini[3], Alfonso Trezza[1], Anna Visibelli[1,4], Lia Millucci[1], Giulia Bernardini[1], Andrea Bernini[1], Barbara Marzocchi[1,5], Daniela Braconi[1], Filippo Prischi[6] and Annalisa Santucci[1]

Abstract

Background: Alkaptonuria (AKU) is an ultra-rare autosomal recessive disease caused by a mutation in the homogentisate 1,2-dioxygenase (HGD) gene. One of the main obstacles in studying AKU, and other ultra-rare diseases, is the lack of a standardized methodology to assess disease severity or response to treatment. Quality of Life scores (QoL) are a reliable way to monitor patients' clinical condition and health status. QoL scores allow to monitor the evolution of diseases and assess the suitability of treatments by taking into account patients' symptoms, general health status and care satisfaction. However, more comprehensive tools to study a complex and multi-systemic disease like AKU are needed. In this study, a Machine Learning (ML) approach was implemented with the aim to perform a prediction of QoL scores based on clinical data deposited in the ApreciseKUre, an AKU-dedicated database.

Method: Data derived from 129 AKU patients have been firstly examined through a preliminary statistical analysis (Pearson correlation coefficient) to measure the linear correlation between 11 QoL scores. The variable importance in QoL scores prediction of 110 ApreciseKUre biomarkers has been then calculated using XGBoost, with K-nearest neighbours algorithm (k-NN) approach. Due to the limited number of data available, this model has been validated using surrogate data analysis.

Results: We identified a direct correlation of 6 (age, Serum Amyloid A, Chitotriosidase, Advanced Oxidation Protein Products, S-thiolated proteins and Body Mass Index) out of 110 biomarkers with the QoL health status, in particular with the KOOS (Knee injury and Osteoarthritis Outcome Score) symptoms (Relative Absolute Error (RAE) 0.25). The error distribution of surrogate-model (RAE 0.38) was unequivocally higher than the true-model one (RAE of 0.25), confirming the consistency of our dataset. Our data showed that inflammation, oxidative stress, amyloidosis and lifestyle of patients correlates with the QoL scores for physical status, while no correlation between the biomarkers and patients' mental health was present (RAE 1.1).

Conclusions: This proof of principle study for rare diseases confirms the importance of database, allowing data management and analysis, which can be used to predict more effective treatments.

Keywords: Rare disease, Alkaptonuria, Machine learning, QoL scores, Precision medicine

* Correspondence: ottavia.spiga@unisi.it
†Ottavia Spiga and Vittoria Cicaloni contributed equally to this work.
[1]Department of Biotechnology, Chemistry and Pharmacy, University of Siena, Via A., 53100 Siena, Italy
Full list of author information is available at the end of the article

Background

Alkaptonuria (AKU) was described by Garrod in 1908 [1] as the first disorder to conform with the principles of Mendelian recessive inheritance. The estimated incidence of AKU is 1 case in 250.000–1.000.000 births in most ethnic groups [2], with about 950 patients reported in 61 countries [3]. AKU patients carry homozygous or compound heterozygous mutations of the *HGD* gene leading to a deficiency of the enzyme homogentisate 1,2-dioxygenase (HGD), which is involved in the catabolic pathway of tyrosine [4, 5]. Such dysfunction causes accumulation of homogentisic acid (HGA). Most of HGA is excreted through the urine, resulting in the characteristic darkening-upon-standing, but smaller HGA amounts can also accumulate in connective tissues, where HGA polymerizes forming a dark brown melanin-like pigment (ochronotic pigment). Ochronosis affects skin, sclera and ears (presenting with blue-black discolouration), spine and joints (causing a dramatic degeneration and chronic inflammation), heart valves (leading to stenosis), and kidneys (where stones may develop) [2]. Ochronosis is also the main cause of arthropathy early onset, severely reducing patients' quality of life and causing pain and deficiency in locomotion [6]. HGA has also been found to trigger oxidative stress in AKU [7–10]. Since oxidized lipids are cytotoxic and responsible for initiating inflammatory reactions, a strict correlation between cytotoxicity of the ochronotic pigment and inflammation has be suggested [11]. It has been shown that useful biomarkers for oxidative stress and inflammation in AKU are the Advanced Oxidation Protein Products (AOPP), the products of the oxidation reaction between plasma proteins and oxidizing agents [12–14].

Recent studies have classified AKU as a secondary amyloidosis [11, 15–18], characterised by deposition of serum amyloid A (SAA) fibers, which in its soluble form is a circulating protein produced during chronic inflammatory processes. Studies on AKU patients' samples (cartilage, salivary glands, chondrocytes and synoviocytes) showed that ochronotic pigment and amyloid fibers share the same location, confirming that SAA is associated with the ochronotic pigment derived from HGA [15]. Under normal conditions SAA is found at low concentrations in plasma (4–6 mg/L), while inflammatory stimulus or tissue damage increase SAA plasma levels 100–1000 times [19], making SAA a sensitive biomarker of inflammation [19]. On top of SAA deposition, SAA plasma level have also been reported to be high in AKU patients ([11, 12, 15–18, 20].

Chitotriosidase (CHIT1) is a chitinase mainly expressed in the differentiated and polarized macrophages [21]. CHIT1 serum concentration correlates with the progression or the severity of several diseases (sarcoidosis, rheumatoid arthritis, ankylosing spondylitis, uveitis, idiopathic pulmonary fibrosis, scleroderma-associated interstitial lung diseases, and chronic obstructive lung diseases), suggesting a potential use of CHIT1 as an AKU biomarker [20, 21].

The major obstacle in carrying out clinical research on AKU is the lack of a standardized methodology to assess disease severity and response to treatment [22], which is complicated by the fact that AKU symptoms differ from an individual to another and no correlation between specific *HGD* mutations and disease severity has been observed so far [5, 23]. A reliable way to monitor patients' clinical condition and overall health status is the use in clinical practice and research of measures of quality of life (QoL) [20, 24]. QoL allows to observe the evolution of diseases from acute to chronic, and to assess the suitability of the therapeutic interventions considering patients' symptoms, general health status and care satisfaction [24].

Our previous studies showed that, in a rare and multisystemic disease like AKU, QoL scores help to identify health needs and to evaluate the impact of disease [20, 25], suggesting the presence of a correlation between QoL and the clinical data deposited in the ApreciseKUre database, which could be instrumental in shading light on AKU complexity. Here we have developed a machine learning application that perform a prediction of the QoL scores based on clinical data deposited in the ApreciseKUre. We believe this approach can be turned into a best practice model also for other rare diseases and can be useful for overcoming the obstacles in small dataset management and analysis.

Materials and methods

Patient data

The ApreciseKUre contains data from 203 patients, but only 129 have a complete and comprehensive set of information, which have been used in this study [26–28]. ApreciseKUre contains information about biomarkers and replies to questionnaires (for a full description of data deposited in ApreciseKUre see [20]. Patients data are classified according with 11 QoL scores: (i) physical health score, (ii) mental health score, (iii) AKU Severity Score Index (AKUSSI) joint pain, (iv) AKUSSI spinal pain, (v) Knee injury and Osteoarthritis Outcome Score (KOOS) pain, (vi) KOOS symptoms, (vii) KOOS daily living, (viii) KOOS sport, (ix) KOOS QOL, (x) Health Assessment Questionnaire Disability Index (HAQ-DI) and (xi) global pain visual analog scale (hapVAS). (for more details see Additional file 1).

Statistical analysis and machine learning

– Preliminary statistical analysis

The input data were firstly examined through a preliminary statistical analysis. A correlation matrix based on Pearson correlation coefficient was calculated to measure the linear correlation between QoL scores:

$$-1\frac{\sigma_{xy}}{\sigma_x\sigma_y}=\frac{\sum_{i=1}^{n}(x_i-\mu_x)\left(y_i-\mu_y\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i-\mu_x\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i-\mu_y\right)^2}}\leq+1$$

where σ_{xy} is the covariance of the two variables x and y, σ_x and σ_y are the variances of x and y, respectively, and μ_x and μ_y are the mean values.

Application of different ML algorithms

Machine learning (ML) is an algorithm-based novel modeling technique that has been introduced recently to select key behavior features (biomarkers) and predict risk levels [29]. ML methods are more precise and accurate in terms of prediction abilities compared with traditional statistical methods, because complex inter-variable interactions are taken into account in ML only [30]. There are several key steps of the machine learning-based classification model: data preprocessing, feature selection, algorithm selection and model evaluation. Our workflow is described in Fig. 1.

In this study, to select the most representative predictors (among biomarkers included in ApreciseKUre) for QoL scores we have applied Extreme Gradient Boosting (XGBoost). It is a key algorithm in the processes of clustering evaluation, resampling evaluation, feature selection and prediction, [31] able to calculate variable importance defined as the statistical significance of each variable with respect to its effect on the generated model [32]. Starting from selected biomarkers, QoL score prediction is then evaluated comparing the performance of three other different ML techniques: (i) Linear Regression [33], (ii) Neural networks [34], and (iii) K-nearest neighbours algorithm (k-NN) [35].). Finally, we applied a surrogate data method [36].

Results

QoL scores statistical correlation

In the present study, a machine learning algorithm was implemented with the aim to perform a prediction of QoL scores based on 129 patients' clinical data deposited in the ApreciseKUre database [26, 27]. QoL scores were firstly examined through a preliminary statistical analysis in order to evaluate the degree of correlation among pairs of variables (Fig. 2).

It is interesting to notice the presence of correlation among AKUSSI, KOOS and HAQ scores. Specifically, KOOS pain, KOOS symptoms, KOOS daily living and KOOS sport have a high correlation with AKUSSI joint pain and spinal pain, and with hapVAS and HAQ-DI. Differently, the mental health score correlation with all the other QoL scores is not statistically significant (between − 0.3 and 0.3). Taken together, these data suggest that the mental health score, the only one assessing the psychological status of the patient, is independent from other QoL scores linked to the individual's physical status. Surprisingly, this finding shows that the patients' psychological experience, based on the evaluation of levels of anxiety and depression, is not directly related with their actual physical and clinical status.

AKU biomarkers selection using XGBoost

Selection of the most representative predictors for QoL scores was performed by Extreme Gradient Boosting. XGBoost reveals that the most statistically significant variables among 110 biomarkers included in ApreciseKUre [27] are: age, SAA, CHIT1, AOPP, RSSP, BMI. Variable importance scores of the above mentioned six best biomarkers, with respect to every QoL score, are reported in Fig. 3.

ML algorithm selection

Based on these preliminary analyses, different ML models (Linear Regression, Neural networks and k-NN) were implemented to improve the correlation analysis of biomarkers and QoL score. The ML models were compared based on RAE (Relative Absolute Error) indicator (Table 1) and R^2 score (Coefficient of determination):

As such, k-NN resulted to be the most accurate algorithm to predict QoL scores. Therefore, we performed a

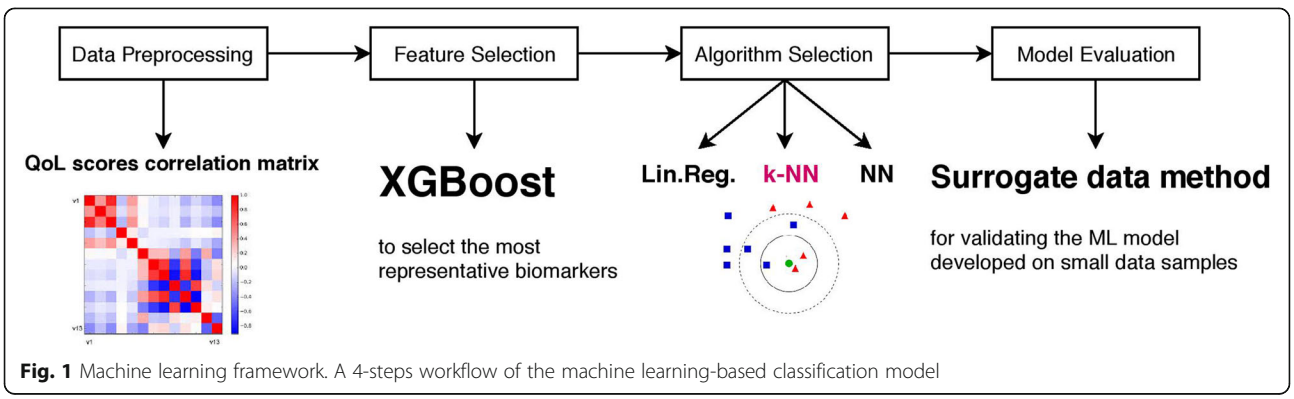


Fig. 1 Machine learning framework. A 4-steps workflow of the machine learning-based classification model

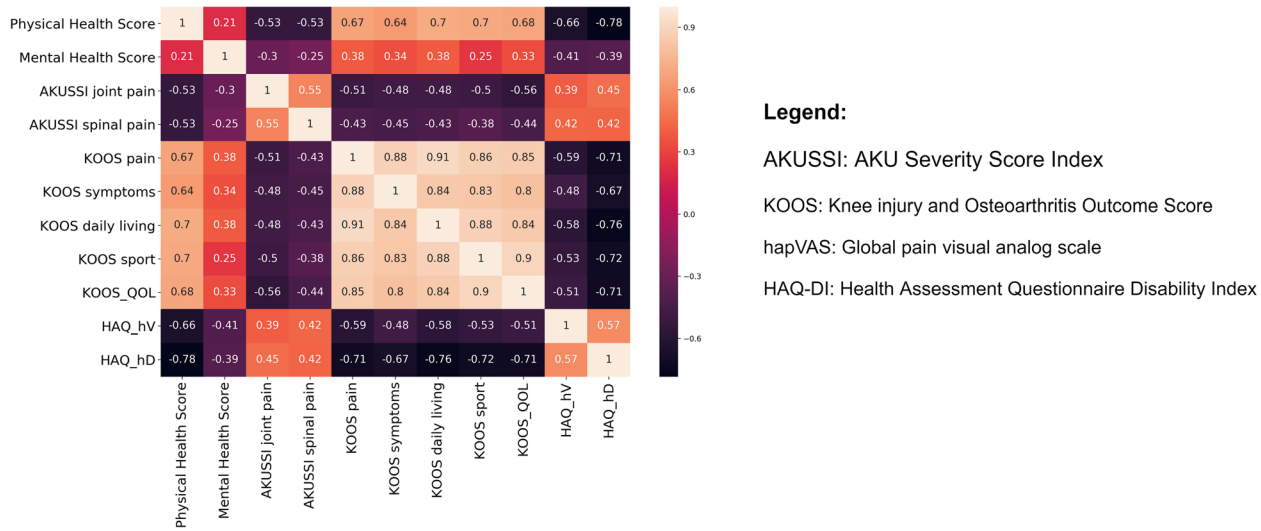


Fig. 2 Correlation matrix of health survey questionnaires. In this correlation matrix all QoL scores are correlated to each other. In black statistically significant inverse correlation, in light-pink statistically significant direct correlation, in red or purple not statistically significant correlations

k-NN on each of the 11 QoL scores and KOOS symptoms score showed the most accurate prediction (lowest RAE: 0.25) (Fig. 4). Conversely, mental health scores might not be predicted with a sufficient accuracy (highest RAE: 1.1), indicating limited or no connection with age, SAA, CHIT1, AOPP, RSSP, BMI values, which is in line with our preliminary statistical analysis.

Differently from other scores (AKUSSI, KOSS, HAQ, hapVAS), mental health score is measured across eight domains (vitality, physical functioning, bodily pain, general health perception, physical role functioning, social functioning, emotional role functioning, mental health), thus it is not unexpected that there is not a correlation with age and other AKU biomarkers. This observation, in line with [20], confirms a not infrequent disability paradox in inherited/chronic disease, underlying the difference between the physical and mental impact on disease severity, which may underestimate overall mental state.

The obtained results demonstrated the power of ML techniques in extrapolating information from a biomarkers dataset to make predictions of QoL scores. ML, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer

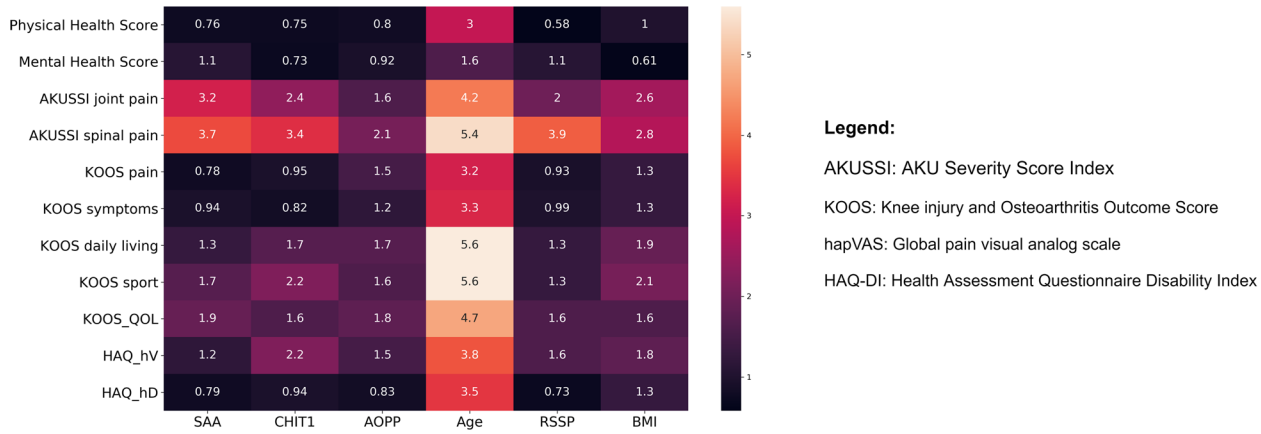


Fig. 3 Variable importance Xgboost for each QoL score. In the matrix are reported all the most representative indicators (X axes) with respect to Qol scores (Y axes) for scores prediction with their corresponding variable importance. Color scale goes from the lower value (in black) to highest value (light pink)

Table 1 ML algorithm performance comparison

Model	RAE	R^2
Linear Regression	0.34	0.87
Neural networks	0.28	0.91
k-NN	0.25	0.94

Comparison based on RAE and R^2 score among different ML models. K-NN resulted to have the lowest RAE, thus the best performance

techniques. For instance, in Fig. 3, age, SAA, CHIT1, AOPP, RSSP, BMI related to AKUSSI spinal pain and AKUSSI joint pain scores assumed the highest variable importance, suggesting the hypothesis they would have been the best QoL indicators. However, as shown in Fig. 4, AKUSSI spinal score and AKUSSI joint pain RAE for k-NN prediction resulted to be higher in comparison with KOOS symptom. Additionally, HAQ hapVAS and HAQ-DI showed high RAE despite the biomarkers variable importance is not different from KOOS symptom score. In view of this, based on the k-NN prediction, KOOS symptoms can be considered as a useful guide for better understanding symptoms and difficulties experienced by patients.

In conclusion, a k-NN based on the combination of parameters like age, SAA, CHIT1, AOPP, RSSP and BMI was able to predict with low RAE the value of KOOS symptoms. Taken singularly these features are not predictive and it is already well known that parameters like age, SAA, CHIT1 are linked with disease severity. The innovative finding of the present work is that, for the first time, we have found an ensemble of multiple complementary features (SAA, CHIT1, AOPP, RSSP, related with inflammation, oxidative stress, amyloidosis; age and BMI, linked with lifestyle) whose combination produce better k-NN prediction results than any single one.

Validating ML models using surrogate data

Small dataset conditions and the associated random effects make validation of ML models a challenging task. For these reasons, to validate the obtained model, we applied a surrogate data method, which has been previously shown to be the most suitable method for small dataset [36]. In this approach, the surrogate data were generated from random numbers able to mimic the distribution of the original dataset independently for each component of the input. They statistically resemble the original data in terms of their mean, standard deviation and range, but they do not maintain the complex relationships between the variables of the real dataset (Table 2).

Therefore, real-data models are expected to perform significantly better than the surrogate data models [36]. The same k-NN algorithm was applied to both datasets, which were randomly split into 80–20% for, respectively, the training and test sets. Each model was trained and validated on 1000 different runs, each using a different training sets, selecting a 10% of the training set to validate the model. The performances of the model, in terms of RAE and R^2 score, were calculated as the average over the runs.

The models trained on our real biochemical and clinical dataset achieve an increase in the average of predictive performance than analogous models trained on the surrogate controls. Indeed, the error distribution of surrogate-model (RAE 0.38) was unequivocally higher than the true-model one (RAE of 0.25) confirming the consistency of our dataset. Thus, it is possible to conclude that the obtained predictive method is not biased or resulting from an overfitting of the model on a small-sized dataset (Fig. 5). This framework allowed ML algorithms to successfully predict clinical and QoL scores outcomes despite small datasets.

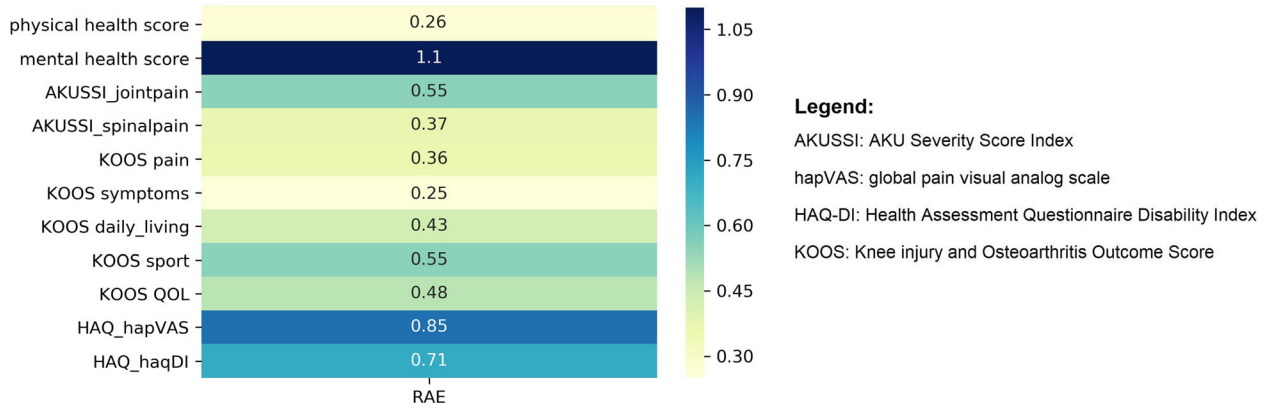


Fig. 4 Performance for each QoL Score. Representation of model accuracy (RAE) for each QoL score, scale from the lower value (in light green) to highest value (blue)

Table 2 Correlation matrix of original and surrogate dataset

ORIGINAL	Pearson correlation coefficient						
	Variables	SAA	CHIT1	AOPP	RSSP	age	BMI
	SAA	1.00	−0.01	− 0.01	0.15	0.02	0.23
	CHIT1	−0.01	1.00	0.00	0.28	0.40*	−0.01
	AOPP	−0.01	0.00	1.00	0.06	0.09	0.17
	RSSP	0.15	0.28	0.06	1.00	0.38*	0.09
	Age	0.02	0.40*	0.09	0.38*	1.00	0.14
	BMI	0.23	−0.01	0.17	0.09	0.14	1.00
	p-value						
	Variables	SAA	CHIT1	AOPP	RSSP	age	BMI
	SAA	0.00	0.56	1.00	0.11	0.57	0.01
	CHIT1	0.56	0.00	0.87	0.00	0.00	0.86
	AOPP	1.00	0.87	0.00	0.69	0.45	0.10
	RSSP	0.11	0.00	0.69	0.00	0.00	0.59
	Age	0.57	0.00	0.45	0.00	0.00	0.28
	BMI	0.01	0.86	0.10	0.59	0.28	0.00
SURROGATE	Pearson correlation coefficient						
	Variables	SAA	CHIT1	AOPP	RSSP	age	BMI
	SAA	1.00	−0.16	0.02	0.22	−0.02	−0.16
	CHIT1	−0.16	1.00	−0.03	−0.06	−0.08	0.06
	AOPP	0.02	−0.03	1.00	−0.12	0.06	−0.01
	RSSP	0.22	−0.06	−0.12	1.00	−0.18	0.09
	Age	−0.02	−0.08	0.06	−0.18	1.00	−0.10
	BMI	−0.16	0.06	−0.01	0.09	−0.10	1.00
	p-value						
	Variables	SAA	CHIT1	AOPP	RSSP	age	BMI
	SAA	0.00	0.72	1.00	0.23	0.57	0.10
	CHIT1	0.72	0.00	0.88	0.02	0.00	1.00
	AOPP	1.00	0.88	0.00	0.66	0.61	0.20
	RSSP	0.23	0.02	0.66	0.00	0.02	0.58
	Age	0.57	0.00	0.61	0.02	0.00	0.28
	BMI	0.10	1.00	0.20	0.58	0.28	0.00

The first table shows the Pearson correlations coefficients and the *p*-values of our original dataset, the second table shows the Pearson correlations coefficients and the *p*-values of surrogate dataset
*indicates statistically significant values

Discussion

The limited number of AKU patients spread around the world represent a major obstacle for generating a standardized strategy to assess disease stage and progression. While several biomarkers for AKU have been identified, a clear connection between biomarkers levels and disease severity (QoL score) is still missing. Here, we implemented an ML method from which QoL of AKU patients can be predicted based on age, oxidative stress (AOPP and RSSP), amyloidosis (SAA) inflammation (CHIT1) biomarkers and BMI, while HGA appears to be extremely variable and unrelated with disease severity. An intricate and complex pattern of oxidative stress, amyloidosis and inflammation is evidently the main important indicator of patients' health status.

Moreover, QoL scores worsen progressively with the age. Aging is associated to decrease antioxidant defenses (for instance the age-related decline in plasma glutathione (GSH) and low molecular weight thiols) and increase ROS production, allowing oxidatively damaged macromolecules to accumulate [37]. AKU subjects undergo a significant decrease in serum free protein thiols and a significant increase in low molecular weight mixed-protein thiols with aging [38].

Our ML model suggested that KOOS indicators could be used to better understanding symptoms and difficulties experienced by AKU patients.

KOOS is a valid, reliable and responsive tool to evaluate both short-term and long-term consequences of knee injury and primary OA. It is a patient-reported outcome measurement, developed to assess the opinion of patients about their knees and associated problems, and it is routinely used for follow-up evaluations [39]. Multiple studies in patients with knee injury and knee OA report that the KOOS demonstrates expected convergent and divergent construct validity, with the KOOS more strongly correlated with subscales of the ShortForm- 36 (SF-36) that measure similar constructs [40]. This is the reason why KOOS prediction could be potentially useful to assess consequences of primary OA, to evaluate changes from week to week induced by treatment (medication, surgery, physical therapy) or over the years due to a primary knee injury, posttraumatic OA or primary OA [39], to identify the main important prognostic biomarkers of AKU, to help the clarification of physiopathological mechanisms of AKU and ochronosis, and to assess the efficacy of future pharmacological treatments.

Similarly, AOPP and RSSP, indicators of oxidative stress and inflammation, have shown to influence the k-NN model. This is not surprising since AKU patients undergo a significant increase in RSSP with aging [38]. Such a trend suggests that progression of AKU symptoms could be related to impaired anti-oxidant status [10]. HGA induces a significant oxidation of a number of serum and chondrocyte proteins. Further investigations allowed highlighting how HGA-induced proteome alteration, lipid peroxidation, thiol depletion, and amyloid production could contribute to oxidative stress generation and protein oxidation in AKU [7]. Furthermore, this is in line with our findings that SAA can be considered as an AKU biomarker for amyloidosis [15]. In fact, a chronic inflammatory status paralleled by inadequate antioxidant defenses is known to promote the aberrant production of amyloidogenic proteins, ultimately leading to secondary amyloid deposition [7]. SAA-amyloidosis

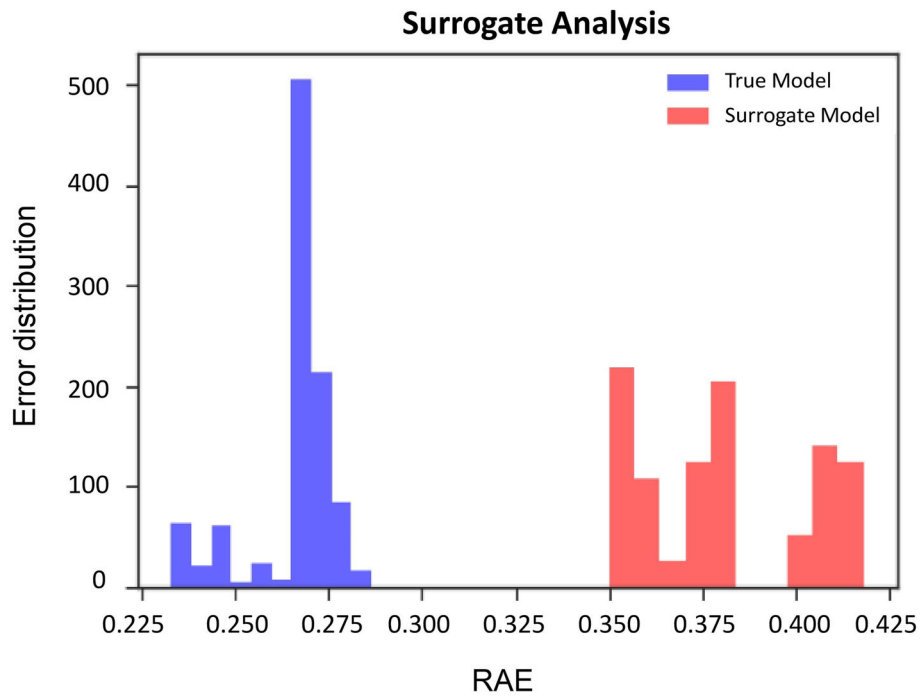


Fig. 5 Surrogate Test Analysis. Comparison of performance based on RAE values, between k-NNs trained on surrogate data (red) and original dataset (blue)

colocalizes with ochronotic pigment as well as with tissue calcification, lipid oxidation, macrophages infiltration, cell death, and tissue degeneration [11, 16, 17].

One of the most striking results is that, differentially from the physical QoL scores based on bodily pain scales and general factor of physical health, mental health status is not predictable by k-NN using the biomarkers listed above. It is measured across eight domains: vitality, physical functioning, bodily pain, general health perception, physical role functioning, social functioning, emotional role functioning, mental health. Surprisingly, in line with the study of [20], the level of biomarkers reported to be directly linked to physical status and pain are not influencing social functioning, role-emotional, levels of depression and anxiety [20]. In conclusion, the outcome of our work was that, for the first time, we have found a biomarkers combination which, in accordance with literature, was able to produce reliable k-NN prediction results. Thanks to this ML algorithm, we will be able to correctly predict KOOS symptoms of a new AKU patient just relying on clinical and lifestyle data.

Current study limitations and future perspective

There are several challenges in studying an ultra-rare and complex disease like AKU, and specifically (i) the paucity of specimens and available data, and (ii) the lack of a standardized method able to objectively assess disease severity or response to treatment. For this reason we developed ApreciseKUre database, aiming to collect as many AKU patients' data as possible, and to use QoL scores to monitor patients' clinical condition and health status, although the database does not yet include objective disease severity findings (i.e. imaging, cardiac valve or calcification, radiographic severity score, treatment modalities, time to surgery, etc). We believe that this study could be a starting point for a better investigation of the utility and reliability of QoL scores, which are becoming increasingly popular, and their correlation to biochemical and clinical biomarkers. For example, the AKUSSI score, which incorporates into a single score multiple clinically meaningful AKU outcomes, medical photography imaging investigations and detailed questionnaires, performed poorly in the model based on the selected biomarkers (AKUSSI joint pain RAE 0.37 and AKUSSI spinal pain RAE 0.55). However, as shown in Fig. 3, parameters like age, SAA, CHIT1, AOPP, RSSP, BMI were the 6 variables with the highest importance values. In literature, these 6 variables have been already used as biomarkers for AKU. In fact, there is an intimate connection between HGA and the ochronotic process, SAA and amyloidosis, inflammation and oxidative stress in AKU, as demonstrated by structural co-localization of ochronotic pigment and SAA-amyloid and co-localization of SAA with crucial cytoskeletal proteins in AKU chondrocytes [20]. As described in [12], some AKU patients, who underwent joint replacement surgery

and complained about articular disorders, arthropathy and joint pain together with other co-morbidities, showed pathological levels of SAA and AOPP above the reference value. Moreover, serum concentration of SAA [41, 42] and CHIT1 activity [43, 44] are markers of disease severity in several rheumatic conditions, and in [20] was provided the evidence that AKU patients present significantly high SAA and chitotriosidase activity in comparison with controls. Some objective disease severity findings, such as cardiac valve calcification and treatment modalities, are strictly linked with amyloidosis, inflammation and oxidative stress. For example, in [11, 16, 17], SAA deposition was detected by immunofluorescence technique in AKU aortic valve and it was tested that low dose methotrexate can down-regulate inflammation and lower SAA production in AKU [20].

In a complex disease like AKU, also lifestyle parameters like BMI are not neglectable. As shown in Table 2, SAA and AOPP have a weak direct correlation with BMI (*p*-value respectively 0.01 and 0.10), which in turn increases with age. It has been previously shown that oxidative stress increase with a rising BMI, as a consequence of an impaired antioxidant status [20, 45] through various biochemical mechanisms, such as superoxide generation from NADPH oxidases, oxidative phosphorylation and glyceraldehyde auto-oxidation [46]. Moreover, in line with [20], a positive association was found between SAA and BMI, since in obesity (where low-grade inflammation is found) adipose tissue is the major source of SAA, which can be considered an obesity-related inflammatory protein [47, 48].

Age is an important driving factor for the prediction of QoL scores and it is a common observation that clinical symptoms might worse with aging. In fact, as shown in Table 2, CHIT1 and RSSP correlate with age (p-value 0.0 for both biomarkers). This is confirmed by the fact that when age is removed from the set of six biomarkers (SAA, CHIT1, AOPP, RSSP, BMI) able to predict QoL scores, the k-NN RAE of KOOS symptoms jump to 0.31. Unfortunately, it is not easy to gather data of very young patients, since people start showing AKU symptoms in their 30/40s, even if the dark discoloration of the urine is present from birth. The systematic use of the ApreciseKUre database will increase the number of patients and will allow us to develop an upgraded version of our algorithm to include an adjustment for the age of the patients.

It is important to specify that this study was based on baseline biochemical and clinical analysis, since the very limited number of information regarding the longitudinal changes, changes during the acute phase, medication effects, differences after joint replacement did not produce robust statistical results. Being AKU a chronic but not lethal disease, the future direction of our study will aim at collecting more AKU follow-up patients' data before and after treatments, in order to evaluate the effectiveness of different therapies. This will be an essential point for a typical precision medicine approach, in which each patient is closely monitored over time and several types of information are collected to understand the uniqueness of each individual. This predictive system will allow for the easy monitoring of AKU disease evolution and it will help clinicians in the selection of the most appropriate treatment, and evaluate its efficacy by observing the trend of QoL scores and biomarkers. In summary, this cost-effective computational method will be beneficial in supporting experimental and clinical studies and, at the same time, will help patients by identifying the most promising treatments.

Conclusion

In conclusion, the combination of a ML to analyse and reinterpret data available in the ApreciseKUre shows the potential direct benefits for patient care and treatments, highlighting the necessity of patient databases for rare diseases, like ApreciseKUre. We believe this is not limited to the study of AKU, but it represents a proof of principle study that could be applied to other rare diseases, allowing data management, analysis and interpretation.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13023-020-1305-0.

Additional file 1. In Additional file 1 a more detailed description of QoL scores is provided. Moreover, informational layers, data and features included in ApreciseKUre are collected and listed.

Abbreviations

aimAKU: Italian Association of Alkaptonuric patients; AKU: Alkaptonuria; AKUSSI: AKU Severity Score Index; AOPP: Advanced Oxidation Protein Products; BMI: Body Mass Index; CHIT1: Chitotriosidase; GSH: glutathione; hapVAS: global pain visual analog scale; HAQ-DI: Health Assessment Questionnaire Disability Index; HGA: homogentisic acid; HGD: Homogentisate 1,2-dioxygenase; IL-1: Interleukin-1; IL-6: Interleukin-6; k-NN: K-nearest neighbors algorithm; KOOS: Knee injury and Osteoarthritis Outcome Score; ML: Machine learning; OA: Osteoarthritis; QoL: Quality of life; RAE: Relative Absolute Error; RSSP: S-thiolated proteins; SAA: Serum amyloid A; SF-36: Short Form-36 questionnaire; SOFIA: Subclinical Ochronotic Features In Alkaptonuria; SONIA1: Suitability Of Nitisinone In Alkaptonuria 1; SONIA2: Suitability of Nitisinone in Alkaptonuria 2; TNF: Tumor necrosis factor; XGBoost: Extreme Gradient Boosting

Acknowledgements

Many thanks are due to Energy Way srl.

Authors' contributions

OS designed the experiments. VC conceived and performed the experiments, analyzed the data, contributed with reagents/materials/analysis tools, wrote the paper. CF analyzed data (information technology expert). LM acquired and analyzed data (AKU expert). GB acquired and analyzed data (AKU expert). AB analyzed data (information technology expert). BM supervisor of the research and AKU clinical data expert. AT analyzed bioinformatic data. AV analyzed ML data. DB acquired and analyzed data, reviewed the paper (AKU expert). FP supervisor of computational approach, reviewed the paper. AS supervisor of the research and scientific-technical AKU expert, reviewed the paper. All authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
The datasets generated and/or analysed during the current study are available in the ApreciseKUre repository, [http://www.bio.unisi.it/aku-db/]. The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Ethics approval and consent to participate
Procedures were approved by Siena University Hospital and national Ethics (Comitato Etico Policlinico Universitario di Siena, number GGP10058, date 21/07/2010) in accordance with 1975 Helsinki Declaration, revised in 2000 (52nd WMA General Assembly, Edinburgh, Scotland, October 2000). Informed written consent was obtained from the patient.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Biotechnology, Chemistry and Pharmacy, University of Siena, Via A., 53100 Siena, Italy. [2]Toscana Life Sciences Foundation, Siena, Italy. [3]Energy way, Modena, Italy. [4]Department of Information Engineering and Mathematics, University of Siena, Siena, Italy. [5]UOC Patologia Clinica, Azienda Ospedaliera Senese, Siena, Italy. [6]School of Life Sciences, University of Essex, Colchester CO4 3SQ, UK.

Received: 26 July 2019 Accepted: 14 January 2020

References

1. Garrod A. Croonian lectures on inborn errors of metabolism, lecture II: alkaptonuria. Lancet. 1908;2:73–9.
2. Phornphutkul CIW, Anderson P, Huizing M, Anikster Y, Gerber L, Gahl W. Natural history of alkaptonuria. N Engl J Med. 2002;347(26):2111–21.
3. Nemethova M, Radvanszky J, Kadasi L, Ascher D, Pires D, Blundell T, Porfirio B, Mannoni A, Santucci A, Milucci L, Sestini S, Biolcati G, Sorge F, Aurizi C, Aquaron R, Alsbou M, Lourenço CM, Ramadevi K, Ranganath LR, Gallagher JA, van Kan C, Hall AK, Olsson B, Sireau N, Ayoob H, Timmis OG, Sang KH, Genovese F, Imrich R, Rovensky J, Srinivasaraghavan R, Bharadwaj SK, Spiegel R, Zatkova A. Twelve novel HGD gene variants identified in 99 alkaptonuria patients: focus on 'black bone disease' in Italy. Eur J Hum Genet. 2016;24(1):66–72.
4. La Du B, Zannoni V, Laster L, Seegmiller J. The nature of the defect in tyrosine metabolism in alcaptonuria. J Biol Chem. 1958;230:251–60.
5. Ascher DB, Spiga O, Sekelska M, Pires DEV, Bernini A, Tiezzi M, Kralovicova J, Borovska I, Soltysova A, Olsson B, Galderisi S, Cicaloni V, Ranganath L, Santucci A, Zatkova A. Homogentisate 1,2-dioxygenase (HGD) gene variants, their analysis and genotype–phenotype correlations in the largest cohort of patients with AKU. Eur J Hum Genet. 2019;27(6):888–902.
6. Milch R. Studies of alcaptonuria: inheritance of 47 cases in eight highly inter-related Dominican kindreds. Am J Hum Genet. 1960;12(1):76–85.
7. Braconi D, Millucci L, Bernardini G, Santucci A. Oxidative stress and mechanisms of ochronosis in alkaptonuria. Free Radic Biol Med. 2015;88:70–80.
8. Braconi D, Laschi M, Amato L, Bernardini G, Millucci L, Marcolongo R, Cavallo G, Spreafico A, Santucci A. Evaluation of anti-oxidant treatments in an in vitro model of alkaptonuric ochronosis. Rheumatology. 2010a;49(10):1975–83.
9. Braconi D, Laschi M, Taylor A, Bernardini G, Spreafico A, Tinti L, Gallagher JA, Santucci A. Proteomic and redox-proteomic evaluation of homogentisic acid and ascorbic acid effects on human articularchondrocytes. J Cell Biochem. 2010b;111(4):922–32.
10. Braconi D, Bianchini C, Bernardini G, Laschi M, Millucci L, Spreafico A, Santucci A. Redox-proteomics of the effects of homogentisic acid in an in vitro human serum model of alkaptonuric ochronosis. J Inherit Metab Dis. 2011;34(6):1163–76.
11. Millucci L, Ghezzi L, Bernardini G, Braconi D, Lupetti P, Perfetto F, Orlandini M, Santucci A. Diagnosis of secondary amyloidosis in alkaptonuria. Diagn Pathol. 2014a;9:185.
12. Braconi D, Bernardini G, Paffetti A, Millucci L, Geminiani M, Laschi M, Frediani B, Marzocchi B, Santucci A. Comparative proteomics in alkaptonuria provides insights into inflammation and oxidative stress. Int J Biochem Cell Biol. 2016;81(Pt B):271–80.
13. Bay-Jensen A, Wichuk S, Byrjalsen I, Leeming D, Morency N, Christiansen C, Maksymowych W. Circulating protein fragments of cartilage and connective tissue degradation are diagnostic and prognostic markers of rheumatoid arthritis and ankylosing spondylitis. PLoS One. 2013;1(e54504). https://doi.org/10.1371/journal.pone.0054504.
14. Gibson D, Rooney M, Finnegan S, Qiu J, Thompson D, Labaer J, Pennington SR, Duncan M. Biomarkers in rheumatology, now and in the future. Rheumatology (Oxford). 2012;51(3):423–33.
15. Millucci L, Spreafico A, Tinti L, Braconi D, Ghezzi L, Paccagnini E, Bernardini G, Amato L, Laschi M, Selvi E, Galeazzi M, Mannoni A, Benucci M, Lupetti P, Chellini F, Orlandini M, Santucci A. Alkaptonuria is a novel human secondary amyloidogenic disease. Biochim Biophys Acta. 2012;1822(11):1682–91.
16. Millucci L, Ghezzi L, Braconi D, et al. Secondary amyloidosis in an alkaptonuric aortic valve. Int J Cardiol. 2014c;172:e121–3.
17. Millucci L, Ghezzi L, Paccagnini E, Giorgetti G, Viti C, Braconi D, Laschi M, Geminiani M, Soldani P, Lupetti P, Orlandini M, Benvenuti C, Perfetto F, Spreafico A, Bernardini G, Santucci A. Amyloidosis, inflammation, and oxidative stress in the heart of an alkaptonuric patient. Mediat Inflamm. 2014b:(2014):258471.
18. Millucci L, Braconi D, Bernardini G, Lupetti P, Rovensky J, Ranganath L, Santucci A. Amyloidosis in Alkaptonuria. J Inherit Metab Dis. 2015;38(5):797–805.
19. Gabay C, Kushner I. Acute-phase proteins and other systemic responses to inflammation. Engl J Med. 1999;340:448–54.
20. Braconi D, Giustarini D, Marzocchi B, Peruzzi L, Margollicci M, Rossi R, Bernardini G, Millucci L, Gallagher JA, Le Quan Sang KH, Imrich R, Rovensky J, Al-Sbou M, Ranganath LR, Santucci A. Inflammatory and oxidative stress biomarkers in alkaptonuria: data from the DevelopAKUre project. Osteoarthr Cartil. 2018;26(8):1078–86.
21. Cho S, Weiden MD, Lee C. Chitotriosidase in the pathogenesis of inflammation, interstitial lung diseases and COPD. Allergy Asthma Immunol Res. 2015;7(1):14–21.
22. Ranganath L, Cox T. Natural history of alkaptonuria revisited: analyses based on scoring systems. J Inherit Metab Dis. 2011;34(6):1141–51.
23. Vilboux T, Kayser M, Introne W, Suwannarat P, Bernardini I, Fischer R, Suwannarat P, Bernardini I, Fischer R, O'Brien K, Kleta R, Huizing M, Gahl WA. Mutation spectrum of homogentisic acid oxidase (HGD) in alkaptonuria. Hum Mutat. 2009;30:1611–9.
24. Clivio, L. (2005). Qualità della vita e stato di salute. Tratto da Unità di Informatica per la Ricerca Clinica - Laboratorio per la ricerca Traslazionale e Outcome Research, Dipartimento di Oncologia: http://crc.marionegri.it/qdv/index.php?page=sf36].
25. Cicaloni V, Zugarini A, Rossi A, Zazzeri M, Santucci A, Bernini A, Spiga O. Towards an integrated interactive database for the search of stratification biomarkers in Alkaptonuria. PeerJ Preprints; 2016;4:e2174v1. https://doi.org/10.7287/peerj.preprints.2174v1.
26. Spiga O, Cicaloni V, Bernini A, Zatkova A, Santucci A. ApreciseKUre: an approach of Precision Medicine in a Rare Disease. BMC Med Inform Decis Making. 2017;17:42.
27. Spiga O, Cicaloni V, Zatkova A, Millucci L, Bernardini G, Bernini A, Marzocchi B, Bianchini M, Zugarini A, Rossi A, Zazzeri M, Trezza A, Frediani B, Ranganath L, Braconi D, Santucci A. A new integrated and interactive tool applicable to inborn errors of metabolism: application to alkaptonuria. Comput Biol Med. 2018;103:1–7.
28. Cicaloni V, Spiga O, Dimitri GM, Maiocchi R, Millucci L, Giustarini D, Bernardini G, Bernini A, Marzocchi B, Braconi D, Santucci A. Interactive alkaptonuria database: investigating clinical data to improve patient care in a rare disease. FASEB J. 2019;33(11):12696–703.
29. Mondal P, Yirinec A, Midya V, Sankoorikal B, Smink G, Khokhar A, Abu-Hasan M, Bascom R. Diagnostic value of spirometry vs impulse oscillometry: a comparative study in children with sickle cell disease. Pediatr Pulmonol. 2019. https://doi.org/10.1002/ppul.24382.
30. Cleophas TJ, Zwinderman AH. Machine learning in medicine. The Netherlands: Springer; 2013.
31. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016. p. 785–94.
32. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29:1189–232.

33. Neter J, Wasserman W, Kutner MH. Applied linear statistical models. Homewood: Irwin; 1985.
34. Haykin, S. (1998). Neural networks: a Comprehensive Foundation. 2nd prentice Hall PTR upper Saddle River, NJ, USA ©1998.
35. Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat. 1992;46(3):175–85.
36. Shaikhina T, Lowe D, Daga S, Briggs D, Higgins R, Khovanova N. Machine learning for predictive Modelling based on small data in biomedical engineering. IFAC-PapersOnLine. 2015;48(20):469–74.
37. Braconi D, Millucci L, Ghezzi L, Santucci A. Redox proteomics gives insights into the role of oxidative stress in alkaptonuria. Expert Rev Proteomics. 2013; 10(6):521–35.
38. Giustarini D, Dalle-Donne I, Lorenzini S, Selvi E, Colombo G, Milzani A, Fanti P, Rossi R. Protein thiolation index (PTI) as a biomarker of oxidative stress. Free Radic Biol Med. 2012;53(4):907–15.
39. Roos E, Lohmander L. The knee injury and osteoarthritis outcome score (KOOS): from joint injury to osteoarthritis. Health Qual Life Outcomes. 2003;1:64.
40. Collins N, Misra D, Felson D, Crossley K, Roos E. Measures of knee function: international knee documentation committee (IKDC) subjective knee evaluation form, knee injury and osteoarthritis outcome score (KOOS), knee injury and osteoarthritis outcome score physical function short form (KOOS-PS), knee Ou. Arthritis Care Res. 2011;63:S208–28.
41. Cantarini L, Giani T, Fioravanti A, Iacoponi F, Simonini G, Pagnini I, et al. Serum amyloid a circulating levels and disease activity in patients with juvenile idiopathic arthritis. Yonsei Med. 2012;J53:1045e8.
42. Jung SY, Park M-C, Park Y-B, Lee S-K. Serum amyloid a as a useful indicator of disease activity in patients with ankylosing spondylitis. Yonsei Med J. 2007;48:218e24.
43. Brunner KH Jr, Scholl-Bürgi S, Hossinger D, Wondrak P, Prelog M, Zimmerhackl LB. Chitotriosidase activity in juvenile idiopathic arthritis. Rheumatol Int. 2008;28:949e50.
44. Basok IB, Kucur M, Kizilgul M, Yilmaz I, Ekmekci BO, Uzunlulu M, Isman KF. Increased chitotriosidase activities in patients with rheumatoid arthritis: a possible novel marker? J Med Biochem. 2014;33:245–51.
45. Yang RL, Shi YH, Hao G, Li W, Le GW. Increasing oxidative stress withprogressive hyperlipidemia in human: relation between mmalondialdehyde and aiatherogenic index. J Clin Biochem Nutr. 2008;43:154–8.
46. Ramos LF, Shintani A, Ikizler TA, Himmelfarb J. Oxidative stress and inflammation are associated with adiposity in moderate to severe CKD. J Am Soc Nephrol. 2008;19:593–9.
47. Christenson K, Bjorkman L, Ahlin S, Olsson M, Sjoholm K, Karlsson A, et al. Endogenous acute phase serum amyloid a lacks pro-inflammatory activity, contrasting the two recombinant variants that activate human neutrophils through different receptors. Front Immunol. 2013;4:92.
48. Wang Z, Nakayama T. Inflammation, a link between obesity and cardiovascular disease. Mediat Inflamm. 2010;(2010):535918.

Publisher's Note

International Journal of
Molecular Sciences

MDPI

Article

Towards a Precision Medicine Approach Based on Machine Learning for Tailoring Medical Treatment in Alkaptonuria

Ottavia Spiga [1,*,†], Vittoria Cicaloni [2,†], Anna Visibelli [1], Alessandro Davoli [3], Maria Ausilia Paparo [3], Maurizio Orlandini [1], Barbara Vecchi [3] and Annalisa Santucci [1]

[1] Department of Biotechnology, Chemistry and Pharmacy, University of Siena, 53100 Siena, Italy; anna.visibelli@student.unisi.it (A.V.); maurizio.orlandini@unisi.it (M.O.); annalisa.santucci@unisi.it (A.S.)
[2] Toscana Life Sciences Foundation, 53100 Siena, Italy; v.cicaloni@toscanalifesciences.org
[3] Hopenly s.r.l., 41058 Vignola, Italy; alessandrodavoli@hopenly.com (A.D.); ausiliapaparo@hopenly.com (M.A.P.); barbara@hopenly.com (B.V.)
* Correspondence: ottavia.spiga@unisi.it
† These authors contributed equally to this work.

Abstract: ApreciseKUre is a multi-purpose digital platform facilitating data collection, integration and analysis for patients affected by Alkaptonuria (AKU), an ultra-rare autosomal recessive genetic disease. It includes genetic, biochemical, histopathological, clinical, therapeutic resources and quality of life scores that can be shared among registered researchers and clinicians in order to create a Precision Medicine Ecosystem (PME). The combination of machine learning application to analyse and re-interpret data available in the ApreciseKUre shows the potential direct benefits to achieve patient stratification and the consequent tailoring of care and treatments to a specific subgroup of patients. In this study, we have developed a tool able to investigate the most suitable treatment for AKU patients in accordance with their Quality of Life scores, which indicates changes in health status before/after the assumption of a specific class of drugs. This fact highlights the necessity of development of patient databases for rare diseases, like ApreciseKUre. We believe this is not limited to the study of AKU, but it represents a proof of principle study that could be applied to other rare diseases, allowing data management, analysis, and interpretation.

Keywords: alkaptonuria; rare disease; machine learning; precision medicine; data analysis; QoL scores

Citation: Spiga, O.; Cicaloni, V.; Visibelli, A.; Davoli, A.; Paparo, M.A.; Orlandini, M.; Vecchi, B.; Santucci, A. Towards a Precision Medicine Approach Based on Machine Learning for Tailoring Medical Treatment in Alkaptonuria. *Int. J. Mol. Sci.* **2021**, *22*, 1187. https://doi.org/10.3390/ijms22031187

Academic Editor: Alessandro Desideri
Received: 20 November 2020
Accepted: 22 January 2021
Published: 26 January 2021

1. Introduction

Precision medicine (PM) is an emerging approach for disease prevention, diagnosis and treatment that takes into account individual variability in genes, environment, proteomics, metabolomics and lifestyle [1]. The capacity to collect, harmonize and analyse data streams is the core for developing a "Precision Medicine Ecosystem" (PME) in which biochemical and clinical resources are shared between researchers, clinicians and patients [2] and can constitute useful guides to generate an exhaustive and dynamic picture of the individual, to identify new potential biomarkers and to tailor a medical treatment suitable for every patient. In PM context, multimedia data management plays a key role not only for common pathologies, but especially for rare disorders, where patients are scattered around the world.

In particular, Alkaptonuria (AKU) is an ultra-rare autosomal recessive metabolic disease [3] with a very low prevalence (1:1,000,000–250,000) [4], caused by mutation in the structure of homogentisate 1,2-dioxygenase (HGD) [4], an enzyme involved in the metabolism of tyrosine and phenylalanine. The deficient activity of HGD enzyme leads to the accumulation of Homogentisic Acid (HGA), which undergoes oxidation and polymerization, forming a dark-brown pigmentation in different connective tissues with a phenomenon called "ochronosis". Such pigmentation involves mostly the osteoarticular

tissues leading to a serious arthropathy with tissues degeneration, chronic inflammation and oxidative stress [5]. The deposition of the dark pigment involves skin, salivary glands [5], brain [6] and cardiac system [7,8], but the most damaged tissues are bone and cartilage [9]. Moreover, recent studies have classified AKU as a secondary amyloidosis [7,8], characterised by deposition of serum amyloid A (SAA) fibers, which is a circulating protein produced at high levels (100–1000 times the normal plasmatic condition of about 4–6 mg/L) in chronic inflammation, making SAA a sensitive biomarker of inflammation. Another marker linked to chronic inflammation is chitotriosidase (CHIT1), a chitinase mainly expressed in the differentiated and polarized macrophages. Therefore, in AKU, besides inflammation, patients also suffer from significant oxidative stress caused by high systemic levels of HGA and its products. In this context, Protein Thiolation index (PTI) interestingly denotes and summarizes the oxidative state of AKU patients. One of the main problems in carrying out clinical research on AKU is the lack of a standardized methodology to assess disease severity and response to treatment, which is complicated by the large variety of AKU symptoms from an individual to another. A reliable way to monitor patients' clinical condition and overall health status is the use in clinical practice and research of measures of Quality of Life (QoL) scores.

To overcome the limitations due to the scarcity of specimens and data available for AKU and the wide range of AKU symptoms, we have recently established a comprehensive digital ecosystem, ApreciseKUre, that integrates patient-derived information (QoL scores, lifestyle), clinician-derived information (urine, blood, plasma analysis), mutational analysis (genotypes, protein stability) and therapeutic treatments offering an exhaustive visualization of different informative layers, to support clinicians and researchers in a PM approach to AKU [10–15]. The ApreciseKUre database can be a good starting point for the creation of a new clinical management tool in AKU, which will lead to the development of a deeper knowledge network on the disease and will advance its treatment [10–12].

The integration of quality of life scores with clinical and therapeutic data will have a central role in order to create a complete PME, supporting clinicians to tailor a medical treatment to every AKU patient. AKU can be treated symptomatically during the early stages (generally using anti-inflammatories, painkillers, low protein diet and vitamin C) whereas, for end stages, total joint and heart valve replacements may be required. Currently, there is no specific therapy for AKU, although a clinical trial with nitisinone is in progress. Moreover, it has been already proved that both methotrexate and anti-oxidants have an excellent efficacy to inhibit the production of amyloid in AKU model chondrocytes [16,17]. Our integrated platform, jointly with a machine learning analysis, described in this study, will be useful to achieve an AKU patients stratification and in monitoring the evolution of biomarkers and QoL scores to tailor the most suitable treatment to each patients sub-group.

The workflow of our study is summarized in Figure 1. The first goal of this work was the prediction of the QoL scores based on both personal and clinical AKU patients' information collected in ApreciseKUre. A fine-tuned scoring system can indeed assist clinicians in making sound decisions regarding diagnosis and treatment plan. Then, it was better investigated the correlation between the values of the QoL scores and the drugs the patients take. This could pave the way to stratify AKU patients and to tailor the most suitable treatment to each patient sub-group in a typical PM perspective. Tailoring treatment to the patient has become a promising approach for maximizing efficacy and minimizing drug toxicity and it is not trivial in an ultra-rare disease like AKU. We believe that this AKU-dedicated preliminary study can represent a proof of principle applicable not only to other rare diseases, but it could be also valuable to larger research communities with an increasing number of affected patients.

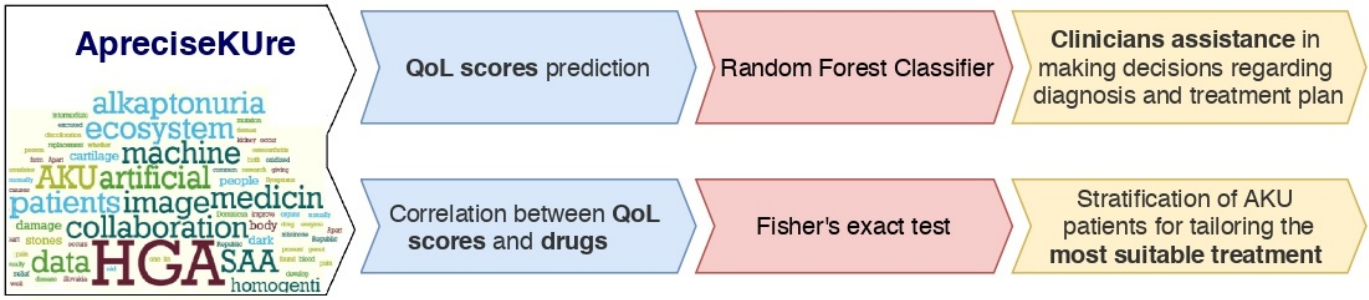


Figure 1. Workflow scheme represented by two stages, 'Quality of Life (QoL) scores prediction' in the top and 'Correlation between QoL scores and drugs' in the bottom.

2. Materials and Methods

2.1. Dataset

The ApreciseKURe (http://www.bio.unisi.it/aku-db/) contains data from 203 patients, of whom 129 do not contain missing data (for a full description of ApreciseKURe see Supplementary Materials S1). Each patient in the ApreciseKURe database is characterized by more than 100 features (for the complete list see supplementary materials S1), describing biochemical (i.e., SAA, CHIT1 and PTI), clinical, genotypic information and replies to questionnaires evaluating QoL scores. It has been performed patients assessment involving 11 QoL scores: (i) physical health score (PHS), (ii) mental health score (MHS); (iii) AKU Severity Score Index (AKUSSI) for joint pain (AJP) and (iv) AKUSSI spinal pain (ASP); (v) Knee injury and Osteoarthritis Outcome Score (KOOS) pain (KOOSp), (vi) KOOS symptoms (KOOSs), (vii) KOOS daily living (KOOSdl), (viii) KOOS sport (KOOSsp), (ix) KOOS QOL; (x) Health Assessment Questionnaire Disability Index (HAQ-DI) and (xi) global pain visual analog scale (hapVAS) (for more details about each score, see supplementary materials in [14]). Moreover, it includes information about drugs taken. We decide to divide the drugs in painkillers, anti-inflammatories and others; then, we group them in several sub-categories:

1. painkillers: opioid, paracetamol, metamizole;
2. anti-inflammatories: Non-steroidal anti-inflammatory drugs (FANS), corticosteroid;
3. others: antiacid, antiarhythmic, antiasthma, antibiotic, anticoagulant, anticonvulsant, antidepressant, antiglaucoma, antigout, antihistamine, antihyperglycemic, antihypertensive, antimuscarinic, antiosteoporotic, antiparkinson, antipsychotic, antireumatic, antiviral, calcium, cholesterol-lowering medication, corticosteroid, diuretic, hormone, methotrexate, proton pump inhibitor, skeletal muscle relaxant, sodium chloride, thyroid hormones, vitamins.

The amount of data used in the analysis varies according to the information available for each QoL score: in particular, we have 134 to 138 rows of data at our disposal, depending on the particular QoL score we are focusing on.

2.2. Machine Learning Classification

The first goal of this work has been the prediction of the QoL scores based on different patients information collected in ApreciseKURe. Because of the small amount of available data, we decided to turn these scores into categorical variables; for each of them, in particular, we divided its range in three equally spaced regions denoted by 0, 1 and 2, corresponding to decreasing severity of health conditions. Given a specific QoL score, we defined y as the vector representing its values (one for each patient): the k-th element y_k could then take value 0, 1 or 2. The prediction was performed with a one-vs-all approach:

one of three classes was chosen (let i represent its value), and the new vector $y^{(i)}$ was defined such that its k-th element is:

$$y_k^{(i)} \equiv \begin{cases} 1 & \text{if } y_k = i \\ 0 & \text{if } y_k \neq i \end{cases} . \tag{1}$$

The prediction for $y^{(i)}$ turned out to be a standard binary classification, which was carried out using the Random Forest (RF) algorithm [18,19], an ensemble classifier that uses multiple decision trees to obtain a better prediction performance. It creates many classification trees and a bootstrap sample technique is used to train each tree from the set of training data.Finally, to evaluate the performance of the model, we defined the usual elements of the confusion matrix, i.e., true positive (TP), true negative (TN), false positive (FP) and false negative (FN) as:

$$\mathrm{TP}^{(i)} \equiv \sum_k \delta_{\hat{y}_k^{(i)}, y_k^{(i)}} \, \delta_{y_k^{(i)}, 1} \tag{2a}$$

$$\mathrm{TN}^{(i)} \equiv \sum_k \delta_{\hat{y}_k^{(i)}, y_k^{(i)}} \, \delta_{y_k^{(i)}, 0} \tag{2b}$$

$$\mathrm{FN}^{(i)} \equiv \sum_k \left[1 - \delta_{\hat{y}_k^{(i)}, y_k^{(i)}} \right] \delta_{y_k^{(i)}, 1} \tag{2c}$$

$$\mathrm{FP}^{(i)} \equiv \sum_k \left[1 - \delta_{\hat{y}_k^{(i)}, y_k^{(i)}} \right] \delta_{y_k^{(i)}, 0} , \tag{2d}$$

where δ is the Kronecker delta and $\hat{y}_k^{(i)}$ is the prediction for $y_k^{(i)}$. Once the elements of the confusion matrix were computed, we introduced other standard metrics such as:

1. accuracy:

$$\mathrm{acc}^{(i)} \equiv \frac{\mathrm{TP}^{(i)} + \mathrm{TN}^{(i)}}{\mathrm{TP}^{(i)} + \mathrm{TN}^{(i)} + \mathrm{FP}^{(i)} + \mathrm{FN}^{(i)}} ; \tag{3}$$

2. recall:

$$\mathrm{recall}^{(i)} \equiv \frac{\mathrm{TP}^{(i)}}{\mathrm{TP}^{(i)} + \mathrm{FN}^{(i)}} ; \tag{4}$$

3. precision:

$$\mathrm{prec}^{(i)} \equiv \frac{\mathrm{TP}^{(i)}}{\mathrm{TP}^{(i)} + \mathrm{FP}^{(i)}} ; \tag{5}$$

4. F_1 score:

$$F_1^{(i)} \equiv \frac{2\mathrm{TP}^{(i)}}{2\mathrm{TP}^{(i)} + \mathrm{FP}^{(i)} + \mathrm{FN}^{(i)}} ; \tag{6}$$

5. Matthews correlation coefficient (MCC) [20]:

$$\mathrm{MCC}^{(i)} \equiv \frac{\mathrm{TP}^{(i)} \times \mathrm{TN}^{(i)} - \mathrm{FP}^{(i)} \times \mathrm{FN}^{(i)}}{\sqrt{(\mathrm{TP}^{(i)} + \mathrm{FP}^{(i)})(\mathrm{TP}^{(i)} + \mathrm{FN}^{(i)})(\mathrm{TN}^{(i)} + \mathrm{FP}^{(i)})(\mathrm{TN}^{(i)} + \mathrm{FN}^{(i)})}} . \tag{7}$$

Among these, the most appropriate metric to be considered was the MCC, as it is the least sensitive to the case of imbalanced classes [21,22]; by definition, it varies over the range $(-1, 1)$, with the value 0 corresponding to random guess.

Given that the index i takes three values, $i = 0, 1, 2$, we ended up with three values for each of these metrics, corresponding to the number of 2-combinations of three elements. In order to derive a single value, different ways of computing a mean value were possible

(e.g., macro-, micro- and weighted-average); in particular, we used a weighted average, and defined:

$$m \equiv \sum_{i=1}^{3} m^{(i)}\, w_i \quad , \quad w_i \equiv \frac{\mathrm{TP}^{(i)} + \mathrm{FN}^{(i)}}{\dim y}\,, \tag{8}$$

where m generically stands for one of the metrics in Equations (3)–(7); each class, then, was weighted by the number of positive instances with respect to the total.

2.3. Techniques in Determining Correlation

The second goal of the present work has been to look for a correlation between the values of the QoL scores and the drugs the patients take. Clearly, it is not uncommon for these 33 drugs to be taken in different combinations: for this reason, we treated each of them as a binary variable (with value equal to 1 if it is taken by the patient, to 0 otherwise), and each record has been characterized by a 33-dimensional vector representing if the patient takes a particular drug or not. The problem of looking for a possible correlation between the drugs and the QoL scores then became that of studying the correlation between two categorical variables. Therefore various methodologies were accessed and compared (see Supplementary Section S2 for a detailed discussion).

3. Results

3.1. Quality of Life Scores Prediction

The first goal of this work has been the prediction of the QoL scores based on different patients information collected in ApreciseKUre, both personal (e.g., date of birth, gender, country of origin, etc.) and clinical (e.g., inflammation biomarkers, results from blood tests, etc.).

For this purpose, we have considered all the QoL scores with the exception of PHS and MHS. Each score takes real values, with KOOS being the only one where large values correspond to good health conditions (absence of pain).

In order to carry out the classification, we used the RF algorithm [18,19]; by comparing its performance against that of logistic regression (LR) and support vector machine (SVM) [23], it turned out the be the one giving the best results.

The hyperparameters of the RF were optimized with the Python library Hyperopt in order to maximize (the absolute value of) the MCC, with a training-validation spitting of 0.8–0.2. We optimized the following hyperparameters: max depth (d_{max}), max features (f_{max}), min samples leaf (sl_{min}), min samples split (ss_{min}), number of estimators ($N_{\mathrm{estim.}}$); we report in Table 1 the results for the different QoL scores.

Table 1. List of optimized hyperparameters for RF used in the analysis, for the different QoL scores.

QoL Score	d_{max}	f_{max}	sl_{min}	ss_{min}	N_{estim}
AKU joint pain	10	0.718	6	8	53
AKU spinal pain	23	0.990	27	14	72
KOOS pain	1	0.718	5	17	51
KOOS symptoms	10	0.609	21	17	94
KOOS daily living	2	0.554	25	44	78
KOOS sport	23	0.663	23	52	56
KOOS QOL	6	0.554	10	25	80
hapVAS	24	0.554	24	13	38

Given the limited amount of data, we adopted the following procedure for training and testing: once the hyperparameters had been optimized, we performed $M = 50$ independent trainings and tests, each time with a different training-test splitting, with the training size randomly chosen between 0.7 and 0.8; we then computed an average on all the metrics obtained in each iteration. The results of the prediction are given in Table 2, together with

the number of records available for each QoL score; in particular, for each metric, both the mean value (μ) and the standard deviation (σ) are shown.

Table 2. QoL scores prediction with RF; the last column represents the number of records available for that particular QoL score.

	Accuracy		**Precision**		**Recall**		F_1		**MCC**		N
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	
AKU joint pain	0.669	0.052	0.530	0.082	0.593	0.061	0.530	0.071	0.180	0.095	138
AKU spinal pain	0.589	0.046	0.327	0.111	0.440	0.065	0.342	0.084	0.037	0.101	138
KOOS pain	0.648	0.064	0.487	0.084	0.547	0.077	0.495	0.085	0.204	0.127	134
KOOS symptoms	0.657	0.070	0.543	0.111	0.585	0.089	0.542	0.102	0.235	0.147	134
KOOS daily living	0.718	0.044	0.553	0.064	0.623	0.061	0.578	0.061	0.346	0.089	134
KOOS sport	0.689	0.049	0.415	0.086	0.546	0.073	0.464	0.079	0.275	0.096	130
KOOS QOL	0.662	0.050	0.463	0.129	0.509	0.076	0.460	0.090	0.232	0.112	134
hapVAS	0.571	0.054	0.371	0.136	0.359	0.086	0.325	0.098	0.066	0.127	136
HAQ-DI	0.624	0.084	0.624	0.104	0.624	0.084	0.596	0.096	0.163	0.183	138

As can be seen, the prediction algorithm performs best for KOOS daily living, KOOS sport, KOOS symptoms and KOOS QOL: despite the rather limited amount of data, about 70% of the records where correctly classified for these QoL scores. If, in the future, information from new patients is recorded, we expect these results to improve significantly.

3.2. Correlation between Drugs and Quality of Life Scores

The second goal of the present work has been to look for a correlation between the values of the QoL scores and the drugs the patients take, grouped in the sub-categories, as explained before. We decided to perform Fisher's exact test on all the combinations QoL score vs. drug, using the software R, employing the Benjamini-Hochberg procedure to deal with multiple comparisons. Out of the 33 drugs we considered in the analysis, it turns out that 8 of them showed significant correlation with at least one QoL score; we report a summary of the results in Table 3, where we simply indicate whether a given drug is correlated with a given QoL score. It is important to notice that "no" does not mean that the drug and the QoL score are uncorrelated, but simply that there is not a significant evidence of correlation; in the future, with a larger amount of data available, it is possible that those drug will turn out to be correlated.

Table 3. Evidence of correlation between the drugs considered in this analysis and the QoL scores; while "yes" means that the correlation is significant, "no" indicates that with the available data there is no evidence of correlation.

	FANS	**Antiarry-Thmic**	**Antihi-Stamine**	**Antihyper-Tensive**	**Cholesterol-Lowering**	**Opioid**	**Proton Pump in.**	**Vitamins**
AKUSSI joint pain	no	no	no	yes	no	no	no	no
AKUSSI spinal pain	no	no	no	yes	no	no	no	no
KOOS pain	yes	yes	yes	yes	no	yes	yes	yes
KOOS symptoms	no	no	no	yes	no	no	no	no
KOOS daily living	yes	no	no	yes	yes	no	yes	no
KOOS sport	yes	no	no	yes	no	yes	yes	no
KOOS QOL	yes	no	no	yes	no	yes	yes	no
HAQ-DI	yes	no	no	no	no	yes	yes	no
hap-VAS	yes	no	no	no	no	yes	yes	no

The full results of the Fisher's exact test can be found in Supplementary Table S3, together with the threshold (shown in the last column) used to accept or reject the null hypothesis, computed according to the Benjamini-Hochberg procedure with a false discovery rate Q set to $Q = 0.2$; the drugs which show a significant correlation are highlighted with bold characters. Moreover, a dense representation is shown in Figure 2.

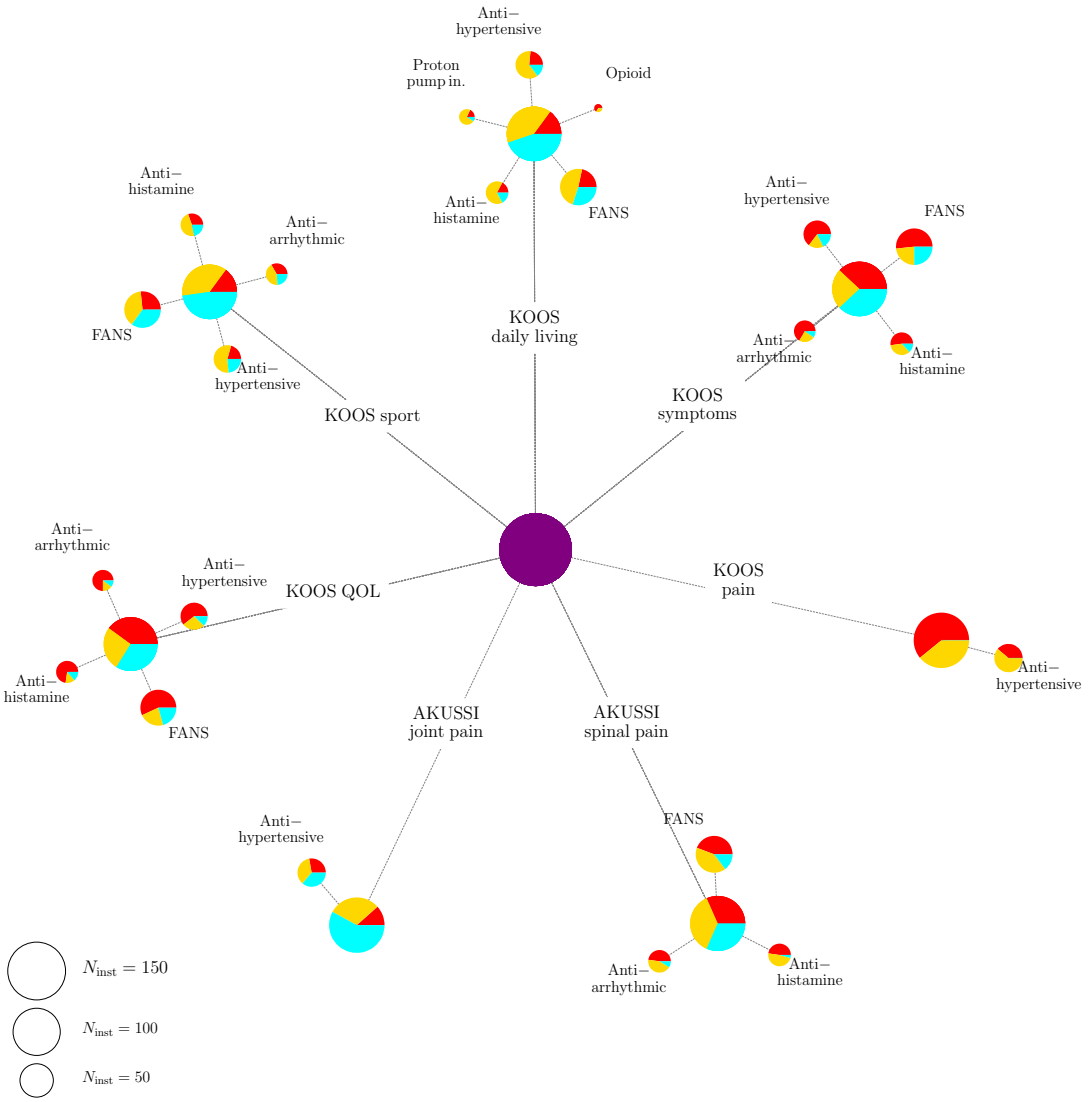


Figure 2. Dense results of the Fisher's exact test. For each QoL score, a first level of pie charts is shown, representing the psycho-physical state of the patients (from red to cyan, corresponding to progressively better health conditions); the area of each circle is proportional to the number of patients for which the information about that QoL score is available. A second level of pie charts, then, shows the impact of drugs on that particular QoL score, with the same conventions as before. As a reference, we also show three benchmark circles whose sizes correspond to the case where the number of patients is 150, 100 and 50, respectively.

In Figure 2, for each QoL score a first pie chart is represented, whose dimension is proportional to the number of patients for which there is information for that given QoL score. The colours are divided according to the psycho-physical state of the patient: from red (bad health conditions) to cyan (absence of pain). In the second level of pie charts, only the drugs for which evidence of correlation has been found are shown. The area of the circle is proportional to the number of patients taking that drug for that given QoL score. As a reference, we also show the size of the circles corresponding to three benchmark values for the number of patients, i.e., 150, 100 and 50.

4. Discussion

The first goal of this work is the prediction of QoL scores in AKU patients. Our previous studies showed that, in a rare and multisystemic disease like AKU, QoL scores help to identify health needs and to evaluate the impact of disease, suggesting the presence of a correlation between QoL and the clinical data deposited in the ApreciseKUre database, which could be instrumental in shading light on AKU complexity. Here, we have developed machine learning applications that perform a prediction of the QoL scores based on data deposited in the ApreciseKUre. In particular, it is based on information about the patients, both personal (date of birth, gender, country of origin, etc.), biochemical and clinical (e.g., amyloidosis, oxidative stress and inflammation biomarkers, results from blood and urine tests, etc.). In this analysis, we consider 9 QoL scores: AKUSSI joint pain, AKUSSI spinal pain, KOOS pain, KOOS symptoms, KOOS daily living, KOOS sport, KOOS QOL, HAQ-DI and hapVAS. Because of the small amount of available data, we decide to turn these scores into three categorical variables (0, 1 and 2) corresponding to decreasing severity of health conditions (i.e., 0 is the worst condition and 2 is the best condition). The classification was carried out using the RF algorithm and comparing its performance against LR and SVM in order to obtain the best result which were then validated. In accordance with our previous study, [14], the algorithm prediction performs best for KOOS daily living, KOOS sport and KOOS symptoms. In fact, despite the rather limited amount of data, about 70% of the records where correctly classified. Thus, our model suggested that KOOS indicator could be a useful tool to better understand symptoms and difficulties experienced by AKU patients. Indeed, KOOS is a valid, reliable and responsive instrument to evaluate both short-term and long-term consequences of knee injury and primary osteoarthritis (OA). It is a patient-reported outcome measurement, developed to assess the opinion of patients about their knees and associated problems, and it is routinely used for follow-up evaluations [24]. KOOS prediction could be important to assess consequences of primary OA, to evaluate changes from week to week induced by treatment (such as medication, surgery, physical therapy etc.) or over the years due to a primary knee injury, post-traumatic OA or primary OA [24], to identify the main important prognostic biomarkers of AKU, to help the clarification of physio-pathological mechanisms of AKU and ochronosis, and to assess the efficacy of future pharmacological treatments. The second goal of this study is the investigation of the correlation between QoL scores and drugs taken by AKU patients. Similarly to the majority of rare genetic diseases, the existing state-of-the-art treatment for AKU is unsatisfactory. To date AKU has no licensed therapy and treatment is symptomatic. Generally, for end-stages joint and heart valve, replacement surgery is required. Previously suggested approaches included a low protein diet for reducing the amount of tyrosine and phenylalanine intake and hence HGA production. Thanks to this attitude, lower values of HGA in blood and urines have been detected especially for children [25].However, in AKU the low-tyrosine dietary strategy was found not always effective, only palliative and also difficult to follow without the supervision of a specialist and it cannot be performed for prolonged times. The idea of adapting diet or treatment according to "personal" factors (such as age, gender, physiological state, or physical activity and QoL scores) and to pathological features (need to follow a low level-protein diet), as well as to special conditions (such as risk of disease) is common today. We believe that our tool could be effective to investigate the most suitable therapy in accordance with QoL scores, which indicates changes in quality of life of patients before/after a specific treatment. Being AKU related to chronic inflammation, oxidative stress and amyloidosis, symptomatic treatments are based on anti-inflammatories (FANS, corticosteroid, FANS+corticosteroid), anti-oxidant (such as Vitamin C) and painkillers (opioid, paracetamol and metamizole). AKU is also linked to cardiovascular ochronosis [26]. Ochronosis is associated with aortic valve stenosis but mitral and pulmonary valves can be affected as well. Numerous case reports have suggested that cardiovascular calcification and stenosis may be associated with pigment deposition in the aortic and mitral valves, endocardium, pericardium, aortic intima, and coronary arteries. In this context, antiarrhythmic and antihypertensive agents

could help AKU patients to improve AKU conditions, as obtained by the application of our method. As well as FANS and opioid resulted to be particularly effective in reducing AKU pain as suggested by a high correlation with KOOS scores, HAQ-DI, hap-VAS. Also, common drugs not related to specific AKU symptoms, such as cholesterol lowering and proton pomp inhibitors, showed a correlation with some QoL scores. In the case of vitamins, they resulted to be effective in the only case of KOOS pain evaluation.

5. Conclusions

In conclusion, our study could be summarized in two main goals

1. Prediction of the QoL scores based on both personal and clinical AKU patients' information collected in ApreciseKUre.
2. The investigation of the correlation between the values of the QoL scores and the drugs the patients take.

The previously described bioinformatics approach could pave the way to achieve AKU patient stratification and to tailor the most suitable treatment to each patient sub-group in a typical PM perspective. This AKU-dedicated preliminary study can represent a proof of principle useful not only to other rare diseases, but it could be also valuable to more common diseases with a larger cohort of patients.

Supplementary Materials: The following are available online at https://www.mdpi.com/1422-0067/22/3/1187/s1.

Author Contributions: Conceptualization, O.S. and V.C.; Data curation, V.C.; Formal analysis, A.D. and M.A.P.; Methodology, A.D. and M.A.P.; Project administration, O.S.; Resources, B.V.; Supervision, M.O., B.V. and A.S.; Validation, A.V.; Writing—original draft, V.C. and A.V.; Writing—review & editing, M.O. and A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: ApreciseKUre Database available at: http://www.bio.unisi.it/aku-db/.

Acknowledgments: We would like to thank Andrea Casonati for his insightful contribution to this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bahcall, O. Precision medicine. *Nature* **2015**, 335. [CrossRef] [PubMed]
2. Aronson, S.; Rehm, H. Building the foundation for genomics in precision medicine. *Nature* **2015**, *526*, 336–342. [CrossRef] [PubMed]
3. Nemethova, M.; Radvanszky, J.; Kadasi, L.; Ascher, D.B.; Pires, D.E.V.; Blundell, T.L.; Porfirio, B.; Mannoni, A.; Santucci, A.; Milucci, L.; et al. Twelve novel HGD gene variants identified in 99 alkaptonuria patients: Focus on 'black bone disease' in Italy. *Eur. J. Hum. Genet.* **2016**, *24*, 66–72. [CrossRef] [PubMed]
4. Ascher, D.; Spiga, O.; Sekelska, M.; Pires, D.; Bernini, A.; Tiezzi, M.; Kralovicova, J.; Borovska, I.; Soltysova, A.; Olsson, B.; et al. Homogentisate 1,2-dioxygenase (HGD) gene variants, their analysis and genotype-phenotype correlations in the largest cohort of patients with AKU. *Eur. J. Hum. Genet.* **2019**, *27*, 888–902. [CrossRef]
5. Millucci, L.; Bernardini, G.; Spreafico, A.; Orlandini, M.; Braconi, D.; Laschi, M.; Geminiani, M.; Lupetti, P.; Giorgetti, G.; Viti, C.; et al. Histological and Ultrastructural Characterization of Alkaptonuric Tissues. *Calcif. Tissue Int.* **2017**, *101*, 50–64. [CrossRef]
6. Bernardini, G.; Laschi, M.; Geminiani, M.; Braconi, D.; Vannuccini, E.; Lupetti, P.; Manetti, F.; Millucci, L.; Santucci, A. Homogentisate 1,2 dioxygenase is expressed in brain: Implications in alkaptonuria. *J. Inherit. Metab. Dis.* **2015**, *38*, 807–814. [CrossRef]
7. Millucci, L.; Ghezzi, L.; Paccagnini, E.; Giorgetti, G.; Viti, C.; Braconi, D.; Laschi, M.; Geminiani, M.; Soldani, P.; Lupetti, P.; et al. Amyloidosis, inflammation, and oxidative stress in the heart of an alkaptonuric patient. *Mediat. Inflamm.* **2014**, *2014*, 258471. [CrossRef]

8. Millucci, L.; Ghezzi, L.; Braconi, D.; Laschi, M.; Geminiani, M.; Amato, L.; Orlandini, M.; Benvenuti, C.; Bernardini, G.; Santucci, A. Secondary amyloidosis in an alkaptonuric aortic valve. *Int. J. Cardiol.* **2014**, *172*, 121–123. [CrossRef] [PubMed]
9. Braconi, D.; Millucci, L.; Bernardini, G.; Santucci, A. Oxidative stress and mechanisms of ochronosis in alkaptonuria. *Free. Radic. Biol. Med.* **2015**, *88*, 70–80. [CrossRef]
10. Cicaloni, V.; Zugarini, A.; Rossi, A.; Zazzeri, M.; Santucci, A.; Bernini, A.O.S. Towards an integrated interactive database for the search of stratification biomarkers in Alkaptonuria. *PeerJ Prepr.* **2016**. [CrossRef]
11. Spiga, O.; Cicaloni, V.; Bernini, A.; Zatkova, A.; Santucci, A. ApreciseKUre: An approach of Precision Medicine in a Rare Disease. *BMC Med Inform. Decis. Mak.* **2017**, *17*, 42. [CrossRef] [PubMed]
12. Spiga, O.; Cicaloni, V.; Zatkova, A.; Millucci, L.; Bernardini, G.; Bernini, A.; Marzocchi, B.; Bianchini, M.; Zugarini, A.; Rossi, A.; et al. A new integrated and interactive tool applicable to inborn errors of metabolism: Application to alkaptonuria. *Comput. Biol. Med.* **2018**, *103*, 1–7. [CrossRef] [PubMed]
13. Cicaloni, V.; Spiga, O.; Dimitri, G.M.; Maiocchi, R.; Millucci, L.; Giustarini, D.; Bernardini, G.; Bernini, A.; Marzocchi, B.; Braconi, D.; et al. Interactive alkaptonuria database: Investigating clinical data to improve patient care in a rare disease. *FASEB J.* **2019**, *33*, 12696–12703. [CrossRef]
14. Spiga, O.; Cicaloni, V.; Fiorini, C.; Trezza, A.; Visibelli, A.; Millucci, L.; Bernardini, G.; Bernini, A.; Marzocchi, B.; Braconi, D.; et al. Machine learning application for development of a data-driven predictive model able to investigate quality of life scores in a rare disease. *Orphanet J. Rare Dis.* **2020**, *15*, 46. [CrossRef] [PubMed]
15. Rossi, A.; Giacomini, G.; Cicaloni, V.; Galderisi, S.; Milella, M.S.; Bernini, A.; Millucci, L.; Spiga, O.; Bianchini, M.; Santucci, A. AKUImg: A database of cartilage images of Alkaptonuria patients. *Comput. Biol. Med.* **2020**, *122*, 103863. [CrossRef] [PubMed]
16. Spreafico, A.; Millucci, L.; Ghezzi, L.; Geminiani, M.; Braconi, D.; Amato, L.; Chellini, F.; Frediani, B.; Moretti, E.; Collodel, G.; et al. Antioxidants inhibit SAA formation and pro-inflammatory cytokine release in a human cell model of alkaptonuria. *Rheumatology* **2013**, *52*, 1667–1673. [CrossRef] [PubMed]
17. Millucci, L.; Spreafico, A.; Tinti, L.; Braconi, D.; Ghezzi, L.; Paccagnini, E.; Bernardini, G.; Amato, L.; Laschi, M.; Selvi, E.; et al. Alkaptonuria is a novel human secondary amyloidogenic disease. *Biochim. Biophys. Acta Mol. Basis Dis.* **2012**, *1822*, 1682–1691. [CrossRef]
18. Ho, T.K. Random decision forests. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14–16 August 1995; Volume 1, pp. 278–282.
19. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
20. Matthews, B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta Protein Struct.* **1975**, *405*, 442–451. [CrossRef]
21. Chicco, D. Ten quick tips for machine learning in computational biology. *BioData Min.* **2017**, *10*, 35. [CrossRef]
22. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*. [CrossRef] [PubMed]
23. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297.:1022627411411. [CrossRef]
24. Roos, E.M.; Lohmander, L.S. The Knee injury and Osteoarthritis Outcome Score (KOOS): From joint injury to osteoarthritis. *Health Qual. Life Outcomes* **2003**, *1*, 1–8.
25. de Haas, V.; Weber, E.C.; De Klerk, J.; Bakker, H.; Smit, G.; Huijbers, W.; Duran, M. The success of dietary protein restriction in alkaptonuria patients is age-dependent. *J. Inherit. Metab. Dis.* **1998**, *21*, 791–798. [CrossRef] [PubMed]
26. Thakur, S.; Markman, P.; Cullen, H. Choice of valve prosthesis in a rare clinical condition: Aortic stenosis due to alkaptonuria. *Hear. Lung Circ.* **2013**, *22*, 870–872. [CrossRef]

frontiers

Machine Learning Approaches for Precision Medicine: Applications To An Integrated Bioinformatics Digital Ecosystem Platform For A Rare Disease An integrated database for Alkaptonuria

Anna Visibelli[1], Vittoria Cicaloni[2*], Ottavia Spiga[1], Annalisa Santucci[1]

[1]Department of Biotechnology, Chemistry and Pharmacy, University of Siena, Italy, [2]Toscana Life Sciences, Italy

Submitted to Journal:
Frontiers in Molecular Medicine

Specialty Section:
Bioinformatics and Artificial Intelligence for Molecular Medicine

Article type:
Review Article

Manuscript ID:
827340

Received on:
01 Dec 2021

Journal website link:
www.frontiersin.org

Conflict of interest statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest

Author contribution statement

VC and AV conceived experiments and wrote the paper. OS AKU expert, reviewed the paper.
AS supervisor of the research and scientific-technical AKU expert, reviewed the paper.

Keywords

Alkaptonuria, rare disease, machine learning, precision medicine, data analysis, bioinformatics

Abstract

Word count: 348

Alkaptonuria (AKU) is an ultra-rare autosomal recessive disease caused by a mutation in the homogentisate 1,2-dioxygenase gene. One of the main obstacles in studying AKU and other ultra-rare diseases, is the lack of a standardized methodology to assess disease severity or response to treatment. Based on that, a multi-purpose digital platform, called ApreciseKUre, was implemented to facilitate data collection, integration and analysis for patients affected by AKU. It includes genetic, biochemical, histopathological, clinical, therapeutic resources and Quality of Life (QoL) scores that can be shared among registered researchers and clinicians to create a Precision Medicine Ecosystem. The combination of machine learning applications to analyse and re-interpret data available in the ApreciseKUre clearly indicated the potential direct benefits to achieve patients' stratification and the consequent tailoring of care and treatments to a specific subgroup of patients. Computational modelling and database building can be a useful guide to generate an exhaustive and dynamic picture of the individual and to identify potential new biomarkers, opening new opportunities to match therapy to patients, and thus leading to a more personalized medicine for maximizing the benefit-to-risk ratio. In this work, different Machine Learning implemented approaches were described:
predictive model for the estimation of oxidative status trend of each AKU patient based on different biochemical predictors [Cicaloni et al., 2019].
prediction of QoL scores based on clinical AKU patients' clinical data to perform patients'stratification [Spiga et al., 2020].
a tool able to investigate the most suitable treatment in accordance with AKU patients' QoL scores [Spiga et al., 2021 A].
the comparison of different algorithms to explore the phenotype–genotype relationships unknown in AKU so far [Spiga et al., 2021 B].
We also implemented an ApreciseKUre plugin, called AKUImg [Rossi et al., 2021], dedicated to the storage and analysis of AKU histopathological slides, where images can be shared to extend the AKU knowledge network. The outcomes of these predictions highlight the necessity of development databases for rare diseases like ApreciseKUre. We believe this is not limited to the study of AKU, but it could be applied to other rare diseases, allowing data management, analysis, and interpretation.

Contribution to the field

To favor implementation of Precision Medicine (PM) approach for a rare disease, Alkaptonuria (AKU), we created the ApreciseKUre-database that represent suitable "PM Ecosystem" in which genetic, biochemical and clinical AKU patient's data are shared. All the information obtained from different AKU research lines were used to populate our growing digital ecosystem. Including updated case-data and samples from clinicians and patients, the researchers benefit from new information sources and contribute to improve and increase the knowledge of the disease through data analysis. Potential applications include getting a deeper knowledge of AKU, eventually advancing its treatment, and identifying exploitable prognostic biomarkers for a reliable clinical monitoring. Moreover, overcoming the classical idea of a database as a storage tool, ApreciseKUre is also able to perform computational analysis allowing a more complete biological understanding of a complex disease like AKU. Overall, the validity and effectiveness of the proposed models shows the potential direct benefits for patient care, treatment and early diagnosis, highlighting the necessity of patient databases for rare diseases. We believe this is not limited to the study of AKU, but it represents a proof of principle study that could be applied to other rare diseases, allowing data management, analysis and interpretation.

Machine Learning Approaches for Precision Medicine: Applications To An Integrated Bioinformatics Digital Ecosystem Platform For A Rare Disease

An integrated database for Alkaptonuria

Anna Visibelli[1ǂ], Vittoria Cicaloni[2ǂ*], Ottavia Spiga[1], Annalisa Santucci[1]

[1]Department of Biotechnology, Chemistry and Pharmacy, University of Siena, ITALY

[2]Toscana Life Sciences Foundation, Siena, ITALY

*** Correspondence:**
Vittoria Cicaloni
v.cicaloni@toscanalifesciences.org

ǂ These authors contributed equally to this work.

Keywords: Alkaptonuria, Rare disease, Machine Learning, Precision Medicine, Data analysis, Bioinformatics.

Abstract

Alkaptonuria (AKU) is an ultra-rare autosomal recessive disease caused by a mutation in the homogentisate 1,2-dioxygenase gene. One of the main obstacles in studying AKU and other ultra-rare diseases, is the lack of a standardized methodology to assess disease severity or response to treatment. Based on that, a multi-purpose digital platform, called ApreciseKUre, was implemented to facilitate data collection, integration and analysis for patients affected by AKU. It includes genetic, biochemical, histopathological, clinical, therapeutic resources and Quality of Life (QoL) scores that can be shared among registered researchers and clinicians to create a Precision Medicine Ecosystem. The combination of machine learning applications to analyse and re-interpret data available in the ApreciseKUre clearly indicated the potential direct benefits to achieve patients' stratification and the consequent tailoring of care and treatments to a specific subgroup of patients. Computational modelling and database building can be a useful guide to generate an exhaustive and dynamic picture of the individual and to identify potential new biomarkers, opening new opportunities to match therapy to patients, and thus leading to a more personalized medicine for maximizing the benefit-to-risk ratio. In this work, different Machine Learning implemented approaches were described:

- predictive model for the estimation of oxidative status trend of each AKU patient based on different biochemical predictors [Cicaloni et al., 2019].
- prediction of QoL scores based on clinical AKU patients' clinical data to perform patients'stratification [Spiga et al., 2020].

- a tool able to investigate the most suitable treatment in accordance with AKU patients' QoL scores [Spiga et al., 2021 A].
- the comparison of different algorithms to explore the phenotype–genotype relationships unknown in AKU so far [Spiga et al., 2021 B].

We also implemented an ApreciseKUre plugin, called AKUImg [Rossi et al., 2021], dedicated to the storage and analysis of AKU histopathological slides, where images can be shared to extend the AKU knowledge network. The outcomes of these predictions highlight the necessity of development databases for rare diseases like ApreciseKUre. We believe this is not limited to the study of AKU, but it could be applied to other rare diseases, allowing data management, analysis, and interpretation.

Introduction

Although evidence-based medicine (EBM) has been the main guide for medical treatment over the last decades, this approach does not consider the individual molecular characteristics of the patients, which are of great importance for the efficacy and safety of therapies. Indeed, the decision-making process in medical practice that considers only the most reliable scientific information combined with the individual expertise of the clinician [Bereczki et al., 2012], cannot be generalized for all patients. It is well known that not all people respond to therapies and drugs in the same way [Hafen et al., 2014; Lehrach, 2015; Roden, 2015] for their differences in genomic, epigenomic and metabolomic profile [Leyens et al., 2014] and other several factors including diet, comorbidities, age and weight [Haga SB. et al., 2017]. In fact, it is possible that patients do not improve their condition after taking the drugs recognized as the 'best' for that pathology, or even suffer from more serious complications due to the accompanying side effects such as adverse drug reactions (ADRs). To maximize the benefit/risk ratio, pharmaceutical interventions and dosage should be specifically tailored for individual patient on their disease risk and expected response. This aspect will become crucial especially in multisystemic, multifactorial and complex disorders, like Alkaptonuria (AKU).

To address this problem, a new approach called Precision Medicine (PM) has become a reality in recent years. This recent technique focuses on different individual parameters, such as genes variability, environment, lifestyle, and various biological markers [www.nih.gov/precision-medicine-initiative-cohortprogram] for the prevention and treatment of diseases.

Biomarkers for example are biological indicators that could have a specific molecular, anatomic, physiologic, or biochemical character, which can be accurately detected and evaluated [Biomarkers Definition Working Group, 2001]. They play a key role as indicators of an ordinary or pathogenic biological process, having a specific physical characteristic or biological change produced. Thanks to PM it is possible primarily to promote research and understanding a wide range of diseases, but also to identify the causes of the different responses to drugs commonly used to treat different patients. Patients can be "stratified" [Laifenfeld et al., 2012] according to their susceptibility to a particular disease or their response to a specific treatment. The PM approach is already profitably applied in various health areas such as oncology, cardiology, nutrition, and in particular rare diseases [Schee Genannt Halfmann et al., 2017]. Thanks to this approach patients' registries can be implemented, exploiting large amounts of data to uncover potential links, and including patients as active partners in this research [Trusheim et al., 2011].

1. Precision Medicine in an ultra-rare disease

While the PM has focused on large amounts of data to study more common diseases, the data obtained from rare diseases are often limited and sparse. This lack of information makes the ability to collect, integrate and analyze data an extremely difficult but necessary effort. Therefore, to overcome this obstacle, PM in rare diseases focuses on creating patients' registries, leveraging the largest amounts of data available to discover potential connections. A process of data harmonization in rare disease registries allows to conduct clinical studies to understand the complexity of diseases, allowing a more accurate classification based on their genetic characteristics [Ogino et al., 2012]. It is also possible to improve the drug development process and assign the right treatment, as well as the most suitable dosage and posology, to the right individual after reliable patients' stratification that implies a process of patients' classification into new subcategories of common diseases.

An obstacle in the creation of such registers is that they are often created at the national or local level, to map rare diseases in certain areas and to gather information on their incidence and prevalence in those selected areas. Data for such disease registries are mostly obtained on a voluntary basis, observational studies, and clinical data. It would be desirable that such registers could be also strengthened by expanding data thanks to the implementation of PM in health systems across the EU [Schee Genannt Halfmann et al., 2017].

In this review, we focused our attention on the application of Artificial Intelligence techniques to analyze and re-interpret data on Alkaptonuria (AKU), an ultra-rare disease characterized by no apparent genotype-phenotype relationship and no prognosis. Our overall goal was to advance research

on rare orphan AKU disease towards a PM approach that addresses disease complexity while considering individual variability.

From a PM perspective, a digital platform dedicated to AKU called *ApreciseKUre* was created, containing data collected from all over the world from different information levels (biochemical, clinical, genetic, lifestyle, pharmacological and quality of life data). The exploitation of multidisciplinary data of AKU patients and the integration of heterogeneous information feed and refine ApreciseKUre to stratify patients in a typical PM approach. The ApreciseKUre platform was not created as a simple registry, but rather as a PME in which genetic, biochemical, and clinical resources are shared between researchers, clinicians, and patients [Aronson et al., 2015] in order to promote a better understanding of the pathophysiological mechanisms of AKU and related comorbidities. Thanks to the integration of sophisticated computational techniques, it may be possible to identify potential biomarkers to assess AKU severity and progression (so far impossible to be determined), providing data with prognostic value (AKU prognosis is undeterminable as well), and supporting the identification of new drug targets for the design and in vitro and ex vivo experimentation of potentially therapeutic molecules, opening new perspectives for the treatment of AKU.

Alkaptonuria (AKU)

AKU in an ultra-rare autosomal recessive disease caused by the mutations of the Homogentisate 1,2-dioxygenase (HGD) gene which leads to a deficiency of the HGD enzyme [Ascher et al., 2019, La Du et al., 1958] producing accumulation of the unprocessed toxic catabolite homogentisic acid (HGA), especially in connective tissues. AKU was the first disorder to conform with the principles of Mendelian recessive inheritance [Garrod et al., 1908] with an estimated incidence of 1 case in 250.000 – 1.000.000 births in most ethnic groups [Phornphutkul et al., 2002] and around 1300 patients around the world [Zatkova et al., 2020, Ascher et al., 2019]. At a structural level, the active form of the HGD enzyme is a complex hexamer [Titus et al., 2000] with a low tolerance to mutations including missense variants (about 65% of all known AKU substitutions) which can cause a harmful effect on proteins folding stability and, consequently a possible alteration of HGA accumulation [Nemethova et al., 2016]. While the HGA excess is mostly eliminated through the urine, the remaining part contributes to the production of an ochronotic pigment deposited in cartilage [Milch et al., 1961; Braconi et al., 2015; Bernardini et al., 2019; Bernini et al. 2021; Braconi et al., 2020], which contributes in arthropathy early onset, responsible for reducing patients' quality of life and causing severe pain and deficit in locomotion [Milch et al., 1961; Braconi et al., 2015; Spiga et al., 2020]. Oxidative stress and chronic inflammation are also triggered by the HGA accumulation

[Braconi et al., 2015, Braconi et al., 2010, Braconi et al., 2011, Millucci et al., 2014] in different organs, making AKU a complex multisystemic disease. Lately, AKU has been classified as a secondary amyloidosis [Millucci et al., 2014, Millucci et al., 2012, Millucci et al., 2015], characterized by the deposition of serum amyloid A (SAA) fibers, a circulating protein produced at high levels in chronic inflammation, making SAA a sensitive biomarker of inflammation [Gabay et al., 1999], confirmed by elevated SAA plasma levels also in AKU patients [Millucci et al., 2014, Millucci et al., 2012, Millucci et al., 2015, Braconi et al., 2016 , Braconi et al., 2018]. Moreover, both ochronotic pigment and SAA-amyloid share the same location in human cartilage and other tissues [Millucci et al., 2012]. Another marker linked to chronic inflammation is chitotriosidase (CHIT1), a chitinase mainly expressed in the differentiated and polarized macrophages [Cho et al., 2014]. Serum concentration of CHIT1 is linked with sarcoidosis, rheumatoid arthritis, ankylosing spondylitis and chronic obstructive lung diseases, suggesting a probable involvement of CHIT1 as an AKU biomarker [Braconi et al., 2018, Cho et al., 2014]. Therefore, in AKU, besides inflammation, patients also suffer from significant oxidative stress caused by high systemic levels of HGA and its products. Proteomics revealed oxidative-stress relayed alterations in AKU patients and showed interesting similarities with other rheumatic diseases [Braconi et al., 2016]. In this context, Protein Thiolation index (PTI) interestingly denotes and summarizes the oxidative state of AKU patients, as revealed by ApreciseKUre tools and experimentally confirmed [Cicaloni et al., 2019]. Moreover, one of the main obstacles in carrying out clinical research on AKU is the lack of a standardized methodology to assess disease severity and response to treatment [Ranganath et al., 2011]. The large variety of AKU symptoms from an individual to another [Ascher et al., 2019, Vilboux et al., 2009] needs a reliable way to monitor patients' clinical conditions and overall health status. A way to help to identify health needs and to evaluate the impact of the disease is represented by the measure of Quality of Life (QoL) [Braconi et al., 2018] whose correlation with the clinical data deposited in the ApreciseKUre database mat help to effectively face AKU complexity [Spiga et al., 2020].

3. ApreciseKUre digital ecosystem platform

As shown in Fig 1, ApreciseKUre (www.bio.unisi.it/aprecisekure/; www.bio.unisi.it/aku-db/) is a digital platform populated by heterogeneous data from different research groups around the world.

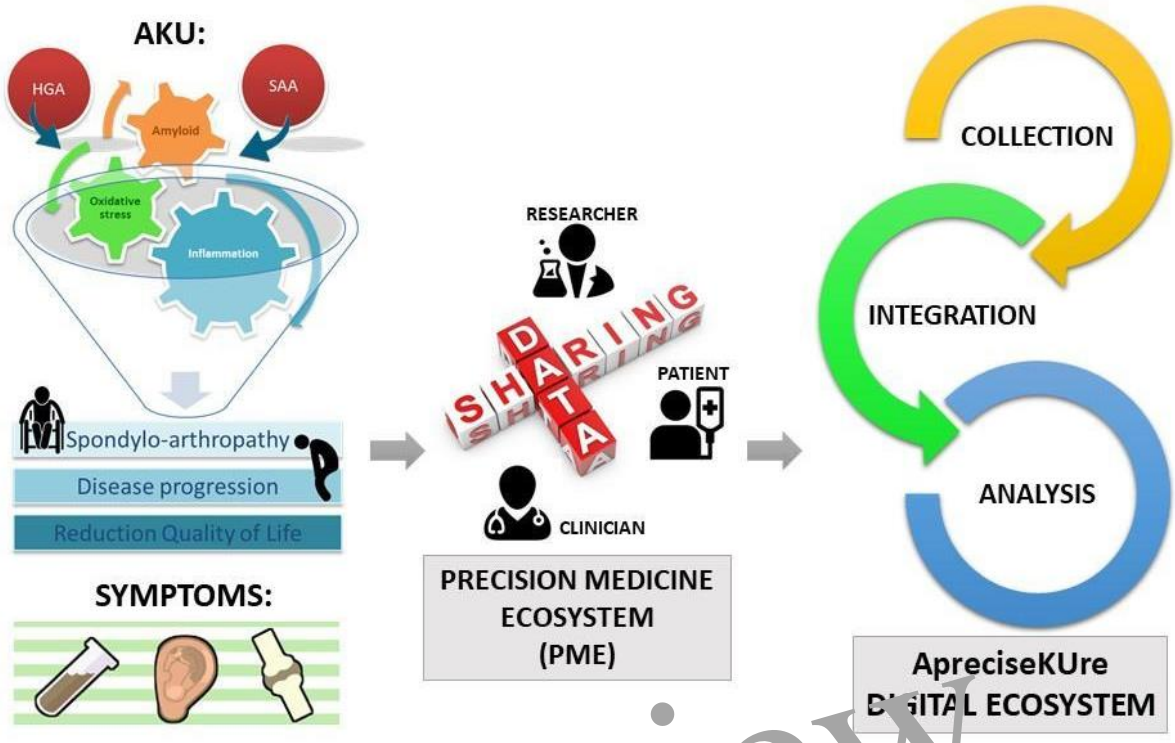


Figure 1. ApreciseKUre digital ecosystem. AKU dedicated Precision Medicine Ecosystem (PME).

The aim was to develop an AKU-PME in which patient-derived information (QoL), clinician-derived information (test results, genotypes), and mutational analysis (protein molecular modeling) can be collected, integrated and shared between scientists, clinicians and patients [Spiga 2017 and 2018], to build a worldwide easily consultable reference point for AKU. In detail, AKU patients' data have been collected and divided into different levels such as genetic, protein, biochemical, histopathologic, clinical, lifestyle and habitual, as shown in Fig.2. All this kind of information is accessible to clinicians which are also able to insert new data, refreshing or replacing previous entries thanks to an easy graphical user interface (GUI).

Currently, ApreciseKUre [Spiga and Cicaloni et al., 2021 A and B] incorporates data of over 210 subjects with AKU, 119 more than its original version [Spiga et al. 2018, Spiga et al., 2017, Cicaloni et al., 2016] which is an exceptional result considering the rarity of AKU. The total number of fields making up each record is 110, with 82 numeric attributes and eight Booleans; the remaining fields are categorical values or concise notes (for the complete list see supplementary material by Spiga and Cicaloni et al., 2021 A and B).

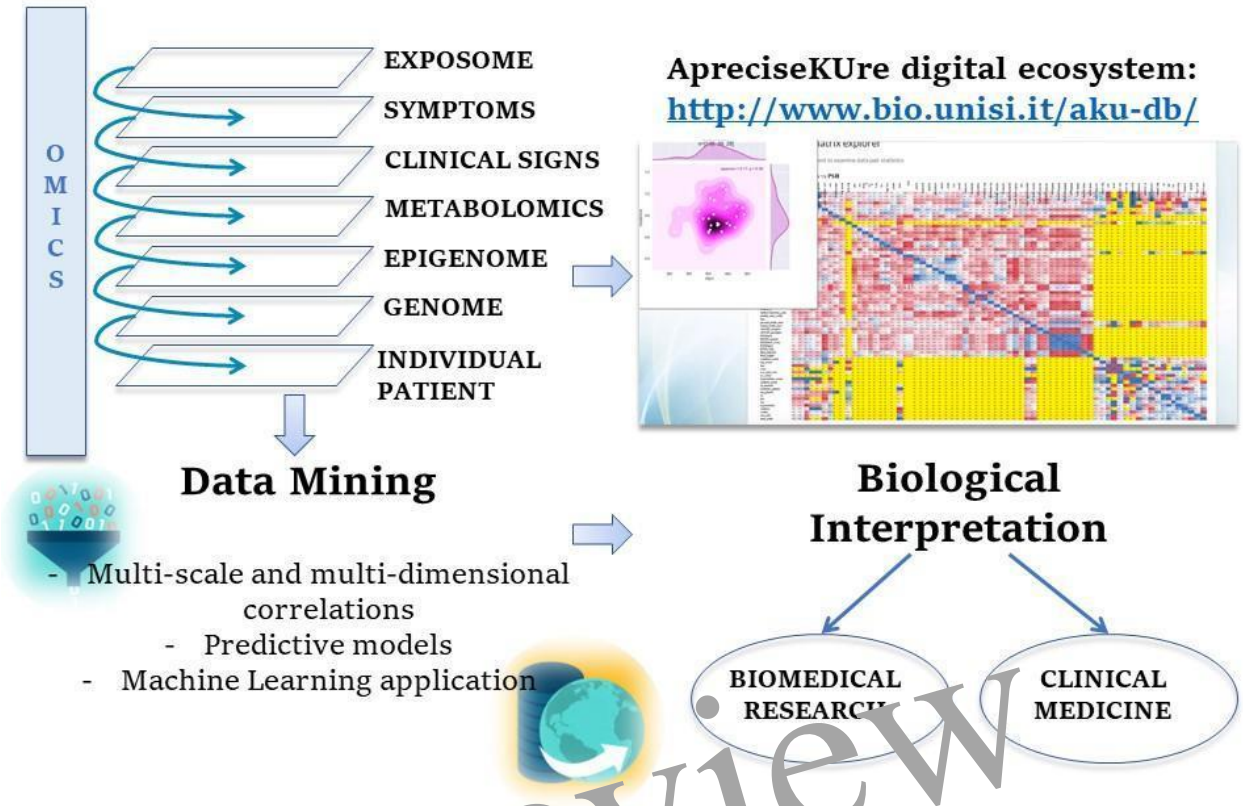


Figure 2. **ApreciseKUre structure.** Data are stratified into various levels: genetic, protein, biochemical, histopathologic, and clinical. Lifestyle and habitual information are also included. Researchers, clinicians, and patients could both easily access all the current information, as well as being able to insert new data, refreshing or replacing previous entries.

Moreover, the development of databases and a proper data analysis can constitute useful tools to generate an exhaustive and dynamic picture of an AKU patient and to identify potential new biomarkers in order to achieve patient stratification. Based on that, different data mining techniques were implemented to discover potential biomarkers, opening new opportunities to match therapy to patients, possibly single therapy to single group of patients, thus leading to a more personalized medicine for maximizing the benefit to risk ratio. The outcomes obtained from these models could be useful not only to advance the treatment of AKU, but also to serve as a model for other rare diseases. In Figure 3, all the data analysis techniques are summarized, ranging from more common statistical data mining to deeper ML models.

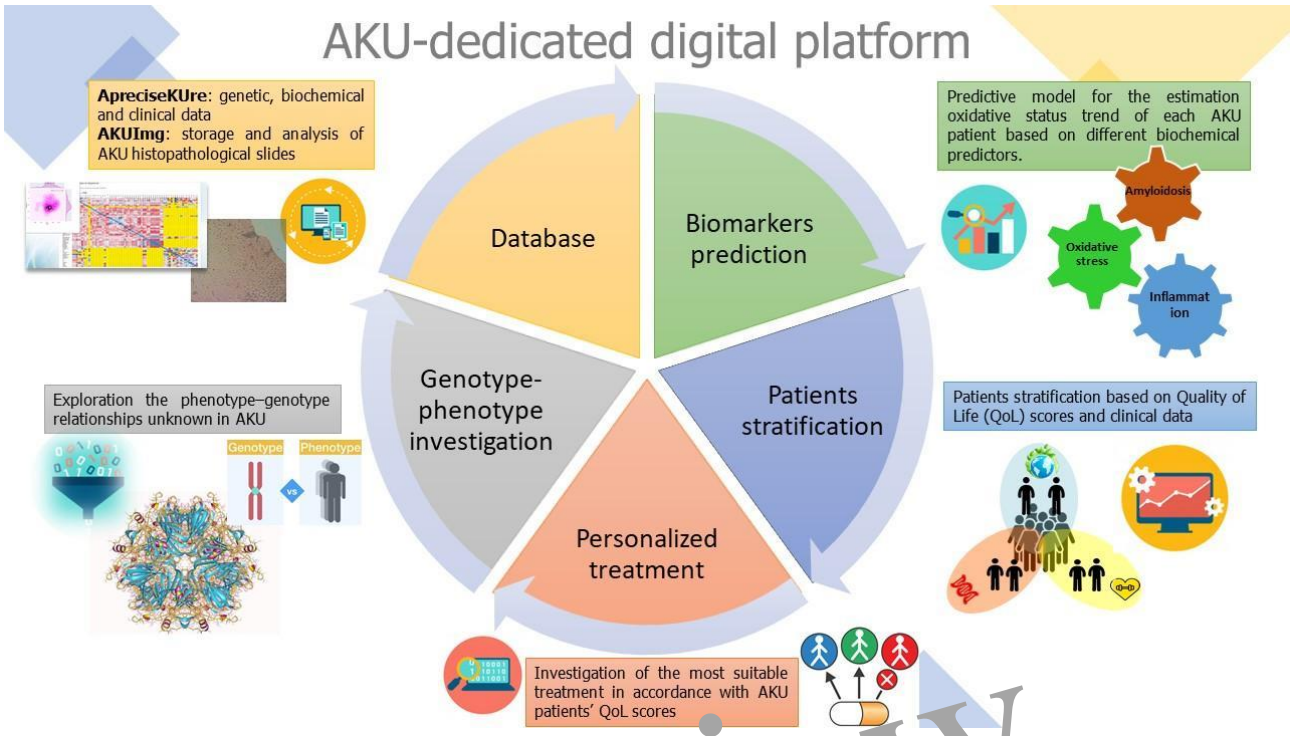


Figure 3. Data mining techniques. All the outcomes derived from statistical and computational approaches included in ApreciseKUre are displayed.

1. **Data analysis by a refreshable correlation matrix**

The first analytic method developed is based on Pearson's correlation coefficient and P value that generates a refreshable correlation matrix where every numeric datum included in ApreciseKUre is correlated with all the others. The modelling correlations have significant implications for early diagnosis, monitoring and therapeutic intervention in AKU, revealing that some clinically used biomarkers might not be suitable prognostic biomarkers in AKU.

For instance, our automated method was able to figure out an inverse statistical correlation between Cystatin C (CysC) and Cathepsin D (CatD) [Cicaloni et al., 2016, Spiga et al, 2017; Spiga et al., 2018]. On the one hand, CysC is a marker for monitoring kidney function: if the kidney function and glomerular filtration rate (GFR) decline, blood levels of CysC increase [Croda-Todd et al., 2007; Randers et al., 1998]. On the other hand, CatD levels are particularly elevated in muscular dystrophy and rheumatic diseases [Khalkhali-Ellis et al., 2014]. The main function of CatD is to degrade proteins like CysC [Lenarčič et al., 1991]. Ochronotic manifestations in AKU gradually lead to kidney stones and nephrolithiasis [Faria et al., 2012]. Even though AKU patients often suffer from kidney dysfunction, in 40 AKU subjects for whom CatD and CysC were tested we did not observe increased CysC levels, whereas CatD showed higher values in AKU subjects compared to controls [Braconi et

al., 2018]. Thanks to omni-comprehensive approaches like such correlation matrix, starting from a statistical observation, it was possible to biologically suggest that CysC might not be a suitable marker to measure GFR in AKU, since overexpression of CatD in AKU might lead to degradation of CysC, making it no longer detectable.

This first data-mining approach revealed the amount of hidden information which can be extrapolated from computational models, in order to acquire a deeper knowledge of the AKU and to identify prognostic biomarkers that can be exploited for a reliable clinical monitoring. Furthermore, given the chronic nature of AKU, it is necessary to monitor the clinical condition of each patient over time and implement a correlation system able to compare biomarkers at different times with follow-up studies, providing a guideline for newly diagnosed patients and estimation of socio-economic costs for affected people and health institutions.

2. Predictive model for the estimation oxidative status

After this preliminary model, a prognostic method based on linear regression able to investigate oxidative stress status of AKU patients, starting from easily measurable clinical parameters [Cicaloni et al., 2019] was integrated in APreciseKURe. This predictive system could help clinicians to easily monitor the oxidative stress evolution in single patients, with the consequent most appropriate antioxidant treatment prescription for each of them. It has already emerged from the correlation matrix that PTI is a reliable biomarker to monitor oxidative stress in AKU [Giustarini et al., 2017]. A linear regression model was then implemented, revealing the most influential biomarkers for PTI prediction, and consequently, for oxidative stress estimation. Such biomarkers are parameters easily measured in AKU clinical analysis and they are related to inflammation, amyloidosis, and lifestyle. They are Body Mass Index (BMI), SAA, HGA, cholesterol, and CTH1. The outcome obtained, not only could help clinicians and researchers to monitor the trend of oxidative stress in an AKU-affected individual, but also could be used as a model for other research groups for improving the AKU-knowledge network.

3. Prediction of QoL scores based on clinical AKU patients' clinical data to perform patients' stratification

Patients' stratification is one of the main goals that computational modelling together with databases can achieve. A K-nearest neighbors algorithm (k-NN) was then modelled, capable of outperforming other predictive models in predicting quality of life scores starting from clinical markers [Spiga et al., 2020]. Moreover, due to the limited number of data available in a rare disease, it is essential to develop methods that would cope with the limited data size. The model has been therefore validated using

surrogate data, because small dataset conditions and the associated random effects make validation of ML models for regression tasks impractical. Conventional methods, such as cross-validation, may become unreliable when the number of independent test samples is limited. The surrogate data method consists in the generation of a so-called "surrogate dataset" generated from random numbers and able to mimic the distribution of the original dataset independently for each component of the input vector. While resembling the original data statistically, it will not retain the intricate relationships between the variables of the real dataset, so it is expected to perform worse than the real-data one.

The validated and effective proposed solution identified a direct correlation between different significant clinical markers and QoL scores, making it addressable to several open issues in AKU with a strong clinical impact on early diagnosis, prediction of disease and of treatment outcome.

4. **A tool able to investigate the most suitable treatment in accordance with AKU patients' QoL scores**

It has been already studied that QoL scores could identify health needs and to evaluate the impact of disease.

QoL of AKU patients was assessed through the following validated questionnaires [Braconi et al., 2010]:

- Knee injury and Osteoarthritis Outcome Score (KOOS) [Roos et al., 2003], evaluating both short- and long-term consequences of knee injury. It contains 5 subscales: pain, other symptoms, function in daily living, function in sport and recreation, and knee-related QoL. Scores are normalized to a "0–100" scale, from extreme knee problems to no knee problems.
- Health Assessment Questionnaire (HAQ), including a disability index (haqDI) and a global pain visual analog scale (hapVAS). Scores are normalized to a "0–3" scale, from no difficulties to extreme ones.
- AKUSSI, incorporating clinically meaningful AKU outcomes combined with medical photography imaging investigations, and detailed questionnaires into a single score.

In this study [Spiga et al., 2021], starting from the idea that there is a correlation between QoL and the clinical data deposited in the ApreciseKUre database, we have developed a ML model that performs a prediction of the QoL scores based on both personal, biochemical and clinical patients data. In this analysis, we considered the following QoL scores: AKUSSI joint pain, AKUSSI spinal pain, KOOS pain, KOOS symptoms, KOOS daily living, KOOS sport, KOOS QOL, HAQ-DI and hapVAS. All these QoL scores were standardized into three categorical variables (0, 1 and 2)

corresponding to decreasing severity of health conditions (i.e., 0 is the worst condition and 2 is the best condition), to face the problem of data scarcity.

The classification was carried out using the RF algorithm, revealing that KOOS indicator could be a useful tool to better understand symptoms and difficulties experienced by AKU patients, as already discovered in [Spiga et al., 2020]. KOOS prediction could be important to assess consequences of primary osteoarthritis (OA), to evaluate weekly changes induced by treatment, to identify the main important prognostic biomarkers of AKU, to help the clarification of physio-pathological mechanisms of AKU and ochronosis, and to assess the efficacy of future pharmacological treatments. Similarly, to most rare genetic diseases, the existing state-of-the-art treatment for AKU is unsatisfactory. With the only exception of Nitisinone, that resulted in reducing urinary excretion of HGA, in decreasing ochronosis and in improving clinical signs with a slower disease progression, there is still no other licensed therapy [Ranganath et al., 2020]. Symptomatic treatments with anti inflammatories and painkillers are generally taken by AKU patients. The idea of personalizing the treatment according to "personal" and pathological features, as well as to special conditions could be the right approach to follow. For that reason, we performed a correlation between QoL scores and drugs taken by AKU patients, believing that our tool could be effective to investigate the most suitable therapy in accordance with QoL scores. Antiarrhythmic and antihypertensive agents, as well as FANS and opioid, resulted to be particularly effective in reducing AKU pain as suggested by a high correlation with KOOS scores, HAQ-DI, hap-VAS. Also, common drugs not related to specific AKU symptoms, such as cholesterol lowering and proton pomp inhibitors, showed a correlation with some QoL scores. In conclusion, vitamins resulted to be effective in the only case of KOOS pain evaluation.

5. Comparison of different algorithms to explore the phenotype–genotype relationship

By taking advantage of the dataset containing the highest number of AKU patients ever considered, it is also possible to apply more sophisticated ML methods to achieve a first and preliminary AKU patient stratification based on phenotypic and genotypic data. Our contribution [Spiga et al. 2021 B] started from a preliminary statistical analysis based on Pearson Correlation Coefficient to evaluate the relationship between pairs of clinical data, biochemical parameters and QoL scores. While all the QoL scores resulted to be statistically correlated, biomarkers of chronic inflammation and amyloidosis like CHIT1 and SAA did not result strongly correlated with disease severity. PTI instead was correlated with Knee injury and Osteoarthritis Outcome Score (KOOS) scores and age. After this preliminary analysis, we performed an innovative strategy applying both K-means and Hierarchical Clustering to stratify AKU population into subgroups with similar features. The experiment was conducted using three different stratification sizes and the resulting clusters were then grouped

according to the severity of the AKU disease, by considering age, levels of oxidative stress, inflammation, and amyloidosis biomarkers and QoL scores. Cluster evaluation was performed by applying the Kruskall-Wallis ranking non-parametric test. Additionally, we computed the Silhouette Score with the aim to test the consistency within elements which have been assigned to the same cluster. Once AKU stratification and cluster validation were performed, we investigated the HGD mutations distribution across the obtained clusters. We paid attention to G161R mutation, responsible for a dramatic reduction of HGD activity [Rodríguez et al., 2000], which occurred in higher percentage in the most phenotypically severe clusters, and M368V and A122V mutations, in which enzymatic activity of HGD is conserved for more than 30%, [Rodríguez et al., 2000] and the trend described a higher percentage in less severe phenotypic sub-groups.

6. AKUImg

Starting from the assumption that bio-imaging technologies are increasingly impacting on life sciences and sharing of image data is required to enable innovative future research, an ApreciseKUre plugin, called AKUImg [Rossi et al., 2020], was created. AKUimg is the first AKU-dedicated image repository. It is dedicated to the storage and analysis of AKU histopathological slides where images can be shared among registered researchers and clinicians to extend the AKU knowledge network. It allows to extend the recognition and reading of slides in the scientific community for an ultra-rare disease, like AKU by supporting clinicians and researchers with a user-friendly online tool able to distinguish between AKU or control cartilage slides. As a matter of fact, the plugin is also integrated with an accurate predictive model based on a standard image processing approach, namely histogram comparison, able to discriminate the presence of AKU by comparing histopathological images. Deep learning (DL) and convolutional neural networks (CNNs) have shown impressive results in many image-processing tasks. However, despite their popularity, they generally require huge datasets to reach good performance. Although we could divide each acquired image in patches, our dataset was not that big. To overcome the obstacle of the paucity of images available, the model we created has been a simple but effective binary classification of the knee cartilage. It performs a comparative analysis of the color histograms of the three channels revealing that AKU and healthy cartilages are easily distinguishable. Therefore, it has been calculated and stored color histograms for all the images in the dataset. For each new image to be classified, it has been evaluated the intersection region between the related histogram and all the histograms in the dataset. Finally, the test image has been assigned to the class with the largest intersection region. In conclusion, the algorithm can perform image classification with a high accuracy, making it a useful guide for non-AKU researchers and clinicians.

4. Conclusion

Bioinformatics is an interdisciplinary field combining biology, computer science, information engineering, mathematics and statistics that develops methods and software tools to analyze and interpret biological data. Bioinformatics is taking a key role in big data analysis especially in healthcare, public health and in PM for a new understanding of the complexity of diseases and for tailoring the most appropriate treatment. PM is an innovative approach which aims to build a knowledge base network that can better guide individualized patient care, giving benefits in terms of health and quality of life. In this review, we focused on its application to an ultra-rare disease named Alkaptonuria, characterized by no apparent genotype-phenotype relationship, no prognosis, and no therapy.

To develop an AKU-dedicated PME, clinical and experimental data have been collected and integrated in ApreciseKUre, a multi-purpose digital platform containing information of more than 200 AKU subjects, uniquely identified based on an anonymous key. Data are stratified into different layers related to genotype, biomarkers, environment, lifestyle, habit, histopathologic, social functioning, clinical and therapies of patients. Including updated case-data and samples from clinicians and patients, the researchers benefit from new information sources and can contribute to get a deeper knowledge of AKU.

However, ApreciseKUre is more than a data storage, as it also integrates computational predictive models able to map highly non-linear input and output and to investigate the health status of AKU patient patterns even when mechanistic relationships between model variables could not be determined. The main ML goal are listed below:

- Estimation of oxidative status trend of each AKU patient based on different biochemical predictors.
- Patients' stratification based on QoL scores and clinical data
- Investigation of the most suitable treatment in accordance with AKU patients' QoL scores
- Exploration of the phenotype–genotype relationships unknown in AKU

Overall, the combination of a ML to analyze and re-interpret data available in the ApreciseKUre shows the potential direct benefits for patient care and treatments, highlighting the necessity of patient databases for rare diseases, like ApreciseKUre. We believe this is not limited to the study of AKU, but it represents a proof of principle study that could be applied to other rare diseases, allowing data management, analysis, and interpretation. Thanks to our sufficiently populated and organized dataset,

it was possible, for the first time, to extensively explore the phenotype-genotype distribution in a typical PM perspective

Lists of abbreviations

AKU: Alkaptonuria

ML: machine learning

HGD: Homogentisate 1,2-dioxigenase

PM: Precision Medicine

PME: Precision Medicine Ecosystem

OA: osteoarthritis

GUI: Graphical user interface

SAA: Serum Amyloid A

HGA: Homogentisic Acid

CTH1: Chitotriosidase

k-NN: K-nearest neighbors algorithm

CysC: Cystatin C

CatD: Cathepsin D

GFR: glomerular filtration rate

PTI: Protein Thiolation Index

QoL: Quality of Life scores

PM: Precision Medicine

DL: Deep learning

CNN: convolutional neural networks

BMI: Body Mass Index

AKUSSI_jointpain: AKU Severity Score Index joint pain

AKUSSI_spinalpain: AKU Severity Score Index spinal pain

KOOSpain: Knee injury and Osteoarthritis Outcome Score pain

KOOSsymptoms: Knee injury and Osteoarthritis Outcome Score symptoms

KOOSdaily_living: Knee injury and Osteoarthritis Outcome Score daily living

KOOSsport: Knee injury and Osteoarthritis Outcome Score sport

KOOS_QOL: Knee injury and Osteoarthritis Outcome Score Quality of Life

hapVAS: Global pain visual analog scale

HAQ-DI: Health Assessment Questionnaire Disability Index

5. Reference

1. Cicaloni V, Spiga O, Dimitri GM, Maiocchi R, Millucci L, Giustarini D, Bernardini G, Bernini A, Marzocchi B, Braconi D, Santucci A. (2019) "Interactive alkaptonuria database: investigating clinical data to improve patient care in a rare disease." The FASEB Journal. Nov;33(11):12696-703.

2. Spiga O, Cicaloni V., Fiorini C, Trezza A, Visibelli A, Millucci L, et al. (2020) "Machine learning application for development of a data-driven predictive model able to investigate quality of life scores in a rare disease." ORPHANET JOURNAL OF RARE DISEASES, 15(1), 46.

3. A) Spiga O, Cicaloni V, Visibelli A, Davoli A, Paparo M.A, Orlandini M, Vecchi B, Santucci A. (2021) "Towards a Precision Medicine Approach Based on Machine Learning for Tailoring Medical Treatment in Alkaptonuria." Int. J. Mol. Sci. 22, 1187.

4. B) Spiga O, Cicaloni V, Dimitri GM, Pettini F, Braconi D, Bernini A, Santucci A. (2021) "Machine learning application for patient stratification and phenotype/genotype investigation in a rare disease." Briefings in Bioinformatics Volume 22, Issue 5.

5. Ranganath LR, Psarelli EE, Arnoux JB, Braconi D, Briggs M, Bröijersén A, Loftus N, Bygott H, Cox TF, Davison AS, Dillon JP. (2020) "Efficacy and safety of once-daily nitisinone for patients with alkaptonuria (SONIA 2): an international, multicentre, open-label, randomised controlled trial." The Lancet Diabetes & Endocrinology; 8(9):762-72.

6. Rossi A, Giacomini, G, Cicaloni V, Galderisi S, Milella M.S, Bernini A, Millucci L, Spiga O, Bianchini M, Orlandini M, Vecchi B, Santucci A. (2021) "AKUImg: A database of cartilage images of Alkaptonuria patients." Computers in Biology and Medicine.

7. Bereczki D. (2012) "Personalized medicine: a competitor or an upgrade of evidence-based medicine?" Per Med 9(2):211–221.

8. Hafen E, Kossmann D, Brand A. (2014) "Health data cooperatives–citizen empowerment." Methods Inf Med 53(8).

9. Lehrach H. (2015) "Virtual clinical trials, an essential step in increasing the effectiveness of the drug development process. Public Health Genomics." 18(6):366–371.

10. Roden DM. (2015) “Cardiovascular pharmacogenomics: current status and future directions.” J Hum Genet 61(1):79–85.

11. Leyens L, Horgan D, Lal JA, Steinhausen K, Satyamoorthy K, Brand A. (2014) “Working towards personalization in medicine: main obstacles to reaching this vision from today’s perspective.” Per Med 11(7):641–649.

12. Haga SB. (2017) “Precision Medicine and Challenges in Research and Clinical Implementation.” In Principles of Gender-Specific Medicine 2017 Jan 1 (pp. 717-732). Academic Press.

13. Biomarkers Definition Working Group. (2001) “Biomarkers and surrogate endpoints: preferred definitions and conceptual framework.” Clin Pharmacol Therapeutics 69:89–95.

14. Laifenfeld D, Drubin DA, Catlett NL, Park JS, Van Hooser AA, Frushour BP, de Graaf D, Fryburg DA, Deehan R. (2012) “Early patient stratification and predictive biomarkers in drug discovery and development: a case study of ulcerative colitis anti-TNF therapy.” Adv Exp Med Biol 736:645-53.

15. Scheer Genant Halfman S, Mammann L, Leyens L, Reumann M, Brand A. (2017) “Personalized Medicine: What's in it for Rare Diseases?.” Adv Exp Med Biol 1031:387-404.

16. Trusheim MR, Burgess B, Xinghua Hu S, Long T, Averbuch SD, Flynn AA, Lieftucht A, Mazumder A, Milloy A, Shaw AP, Swank D, Wang J, Berndt ER, Goodsaid F, Palmer MC. (2011) “Quantifying factors for the success of stratified medicine. Nature Reviews Drug Discovery” 10:11. 37

17. Ogino S, Fuchs CS, Giovannucci E. (2012) “How many molecular subtypes? implications of the unique tumor principle in personalized medicine.” Expert Rev Med Diagn 12(6):621–628.

18. Aronson SJ, Rehm HL. (2015) “Building the foundation for genomics in precision medicine.” Nature.

19. Ascher, DB et al. (2019) “Homogentisate 1,2-dioxygenase (HGD) gene variants, their analysis and genotype–phenotype correlations in the largest cohort of patients with AKU.” Eur. J. Hum. Genet.

20. La Du BN, Zannoni VG, Laster L, Seegmiller JE (1958) “The nature of the defect in tyrosine metabolism in alcaptonuria.” J. Biol. Chem. 230, 251–260 .

21. Garrod AE (1908) “ON INBORN ERRORS OF METABOLISM.” The Croonian Lectures, Lancet.

22. Phornphutkul C et al. (2002) “Natural history of alkaptonuria.” N. Engl. J. Med.

23. Zatkova A, Ranganath L, Kadasi L. (2020) “Alkaptonuria: Current perspectives. Application of Clinical Genetics”.

24. Titus, GP et al. (2000) “Crystal structure of human homogentisate dioxygenase.” Nat. Struct. Biol.

25. Nemethova M. et al. (2016) “Twelve novel HGD gene variants identified in 99 alkaptonuria patients: Focus on ‘black bone disease’ in Italy.” Eur. J. Hum. Genet.

26. Milch RA. (1961) “Studies of Alcaptonuria: A Genetic Study of 58 Cases Occurring in Eight Generations of Seven Inter-related Dominican Kindreds.” Arthritis Rheum.

27. Bernardini G, Leone G, Millucci L, Consumi M, Braconi D, Spiga O, Galderisi S, Marzocchi B, Viti C, Giorgetti G, Lupetti P. (2019) “Homogentisic acid induces morphological and mechanical alteration of ochronotic cartilage in alkaptonuria. Journal of cellular physiology. 234(5):6696-708.

28. Bernini A, Petricci E, Atrei A, Baratto MC, Manetti F, Santucci A. (2021) “A molecular spectroscopy approach for the investigation of early phase ochronotic pigment development in Alkaptonuria.” Scientific Reports., 11(1):1-4.

29. Braconi D, Millucci L, Spiga O, Santucci A. (2020) “Cell and tissue models of alkaptonuria. “Drug Discovery Today: Disease Models.; 31:3-10.

30. Braconi D, Millucci L, Bernardini G, Santucci A. (2015) “Oxidative stress and mechanisms of ochronosis in alkaptonuria. Free Radical Biology and Medicine”.

31. Braconi, D et al. (2010) “Proteomic and redox-proteomic evaluation of homogentisic acid and ascorbic acid effects on human articular chondrocytes.” J. Cell. Biochem.

32. Braconi D et al. (2011) “Redox-proteomics of the effects of homogentisic acid in an in vitro human serum model of alkaptonuric ochronosis.” J. Inherit. Metab. Dis.

33. Millucci L et al. (2014) “Secondary amyloidosis in an alkaptonuric aortic valve.” Int. J. Cardiol.

34. Millucci L et al. (2012) “Alkaptonuria is a novel human secondary amyloidogenic disease. “Biochim. Biophys. Acta - Mol. Basis Dis.

35. Millucci L et al. (2015) “Amyloidosis in alkaptonuria. Journal of Inherited Metabolic Disease”

36. Gabay C, Kushner I. (1999)“Acute-phase proteins and other systemic responses to inflammation. New England Journal of Medicine” .

37. Braconi D et al. (2016) “Comparative proteomics in alkaptonuria provides insights into inflammation and oxidative stress.” Int. J. Biochem. Cell Biol.

38. Braconi, D et al. (2018) “Inflammatory and oxidative stress biomarkers in alkaptonuria: data from the DevelopAKUre project.” Osteoarthr. Cartil.

39. Cho, S. J., Weiden, M. D. & Lee, C. G. (2014) “Chitotriosidase in the pathogenesis of inflammation, interstitial lung diseases and COPD.” Allergy, Asthma Immunol. Res.

40. Ranganath LR, Cox TF.(2011) “Natural history of alkaptonuria revisited: Analyses based on scoring systems.” Journal of Inherited Metabolic Disease.

41. Vilboux T et al. (2009) “Mutation spectrum of homogentisic acid oxidase (HGD) in alkaptonuria.” Human Mutation.

42. Spiga O et al. (2018) “A new integrated and interactive tool applicable to inborn errors of metabolism: Application to alkaptonuria.” Comput. Biol. Med.

43. Spiga O, Cicaloni V, Bernini A, Zatkova A, Santucci A. (2017) “ApreciseKUre: An approach of Precision Medicine in a Rare Disease.” BMC Med. Inform. Decis. Mak.

44. Cicaloni V. et al. (2016) “Towards an integrated interactive database for the search of stratification biomarkers in Alkaptonuria.” PeerJ Prepr.

45. Croda-Todd MT, Soto-Montano XJ, Hernández-Cancino PA, Juárez-Aguilar E. (2007) “Adult cystatin C reference intervals determined by nephelometric immunoassay,” Clin. Biochem. 40:13–14

46. Randers E, Kristensen JH, Erlandsen EJ, Danielsen H. (1998) “Serum cystatin C as a marker of the renal function.” Scand. J. Clin. Lab. Invest. 25 (7) 585–592.

47. Khalkhali-Ellis Z ,Hendrix Mary JC. (2014) “Two faces of cathepsin D: physiological guardian angel and pathological demon.” Biol. Med. 6:2.

48. Lenarčič B, Krašovec M, Ritonja A, Olafsson I, Turk V. (1991) “Inactivation of human cystatin C and kininogen by human cathepsin D.” FEBS Lett. 280 (2): 211–215.

49. Faria B, Vidinha J, Pêgo C, Correia H, Sousa T. (2012) “Impact of chronic kidney disease on the natural history of alkaptonuria.” Clinical Kidney Journal 5 (4): 352–355.

50. Giustarini D, Galvagni F, Colombo, G., Dalle-Donne, I., Milzani, A., Aloisi,A.M., andRossi,R. (2017) “Determination of protein thiolation index (PTI) as a biomarker of oxidative stress in human serum.” Anal.Biochem. 538, 38–41

51. Roos EM, Lohmander LS (2003) “The Knee injury and Osteoarthritis Outcome Score (KOOS): From joint injury to osteoarthritis.” Health Qual. Life Outcomes, 1, 1–8.

52. Rodríguez JM et al. (2000) “Structural and functional analysis of mutations in alkaptonuria.” Hum. Mol. Genet.

Figure 1.JPEG

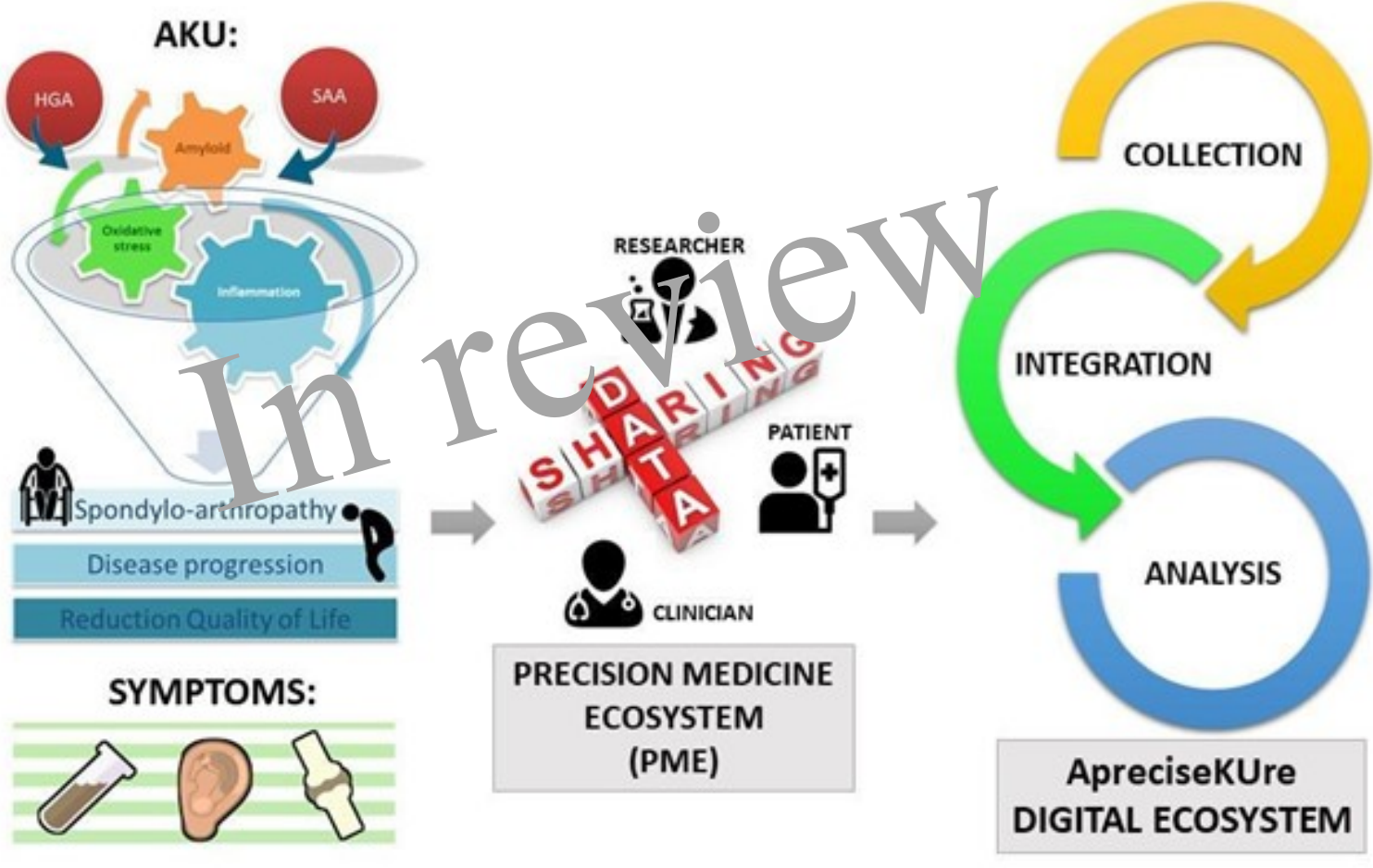


Figure 2.JPEG

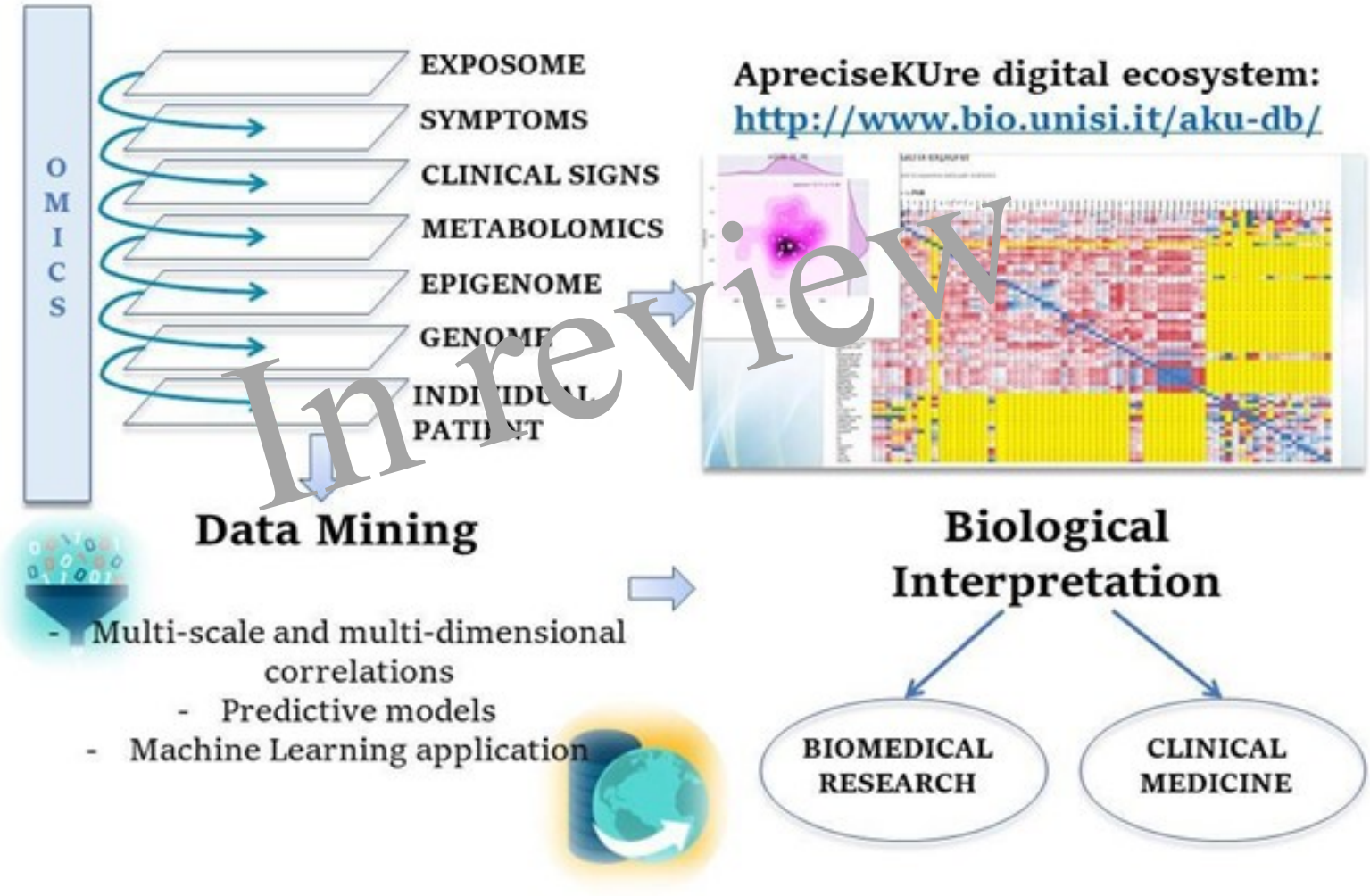
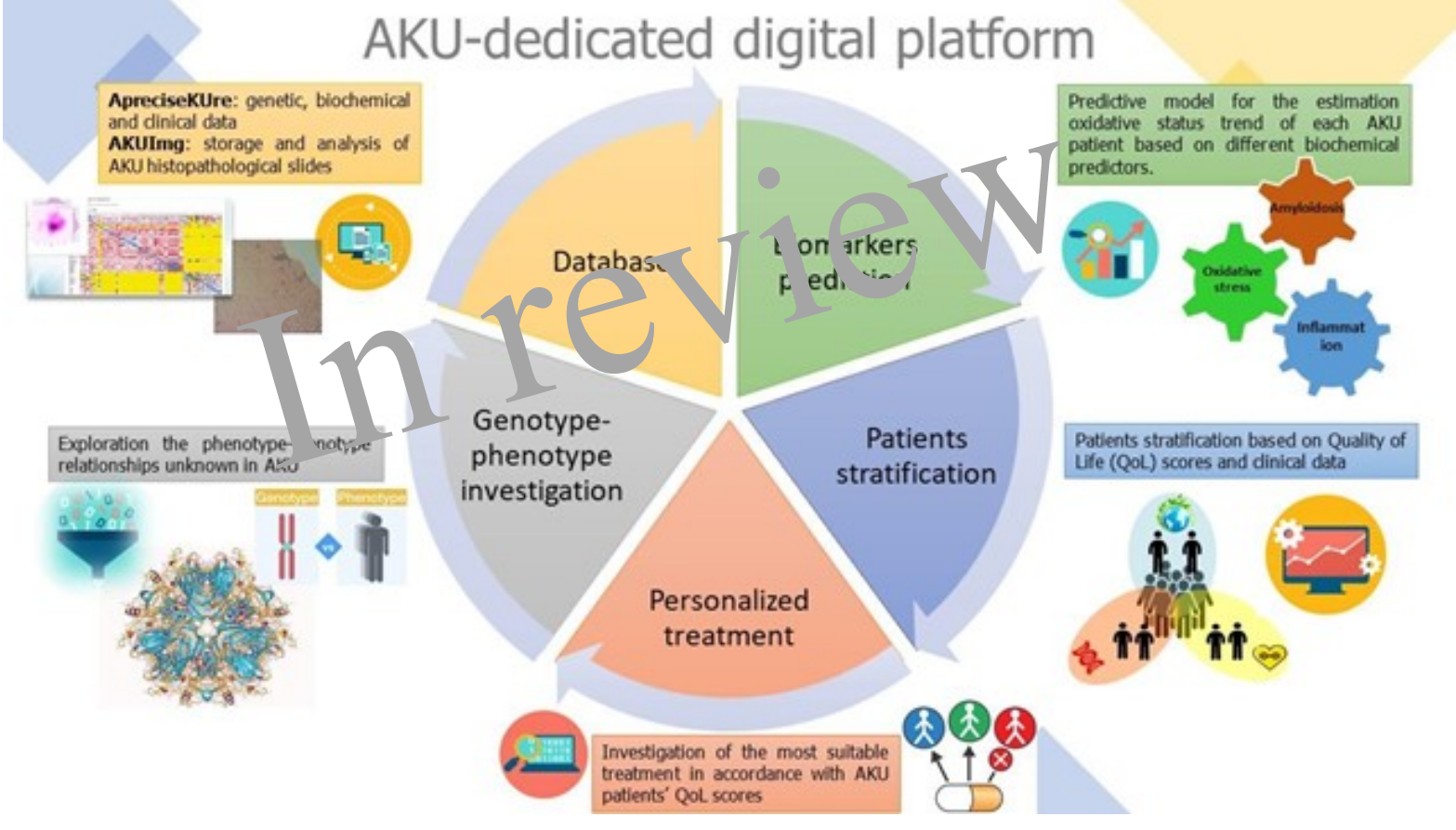


Figure 3.JPEG



Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 192 (2021) 1696–1704

Procedia Computer Science

www.elsevier.com/locate/procedia

25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

CaregiverMatcher: graph neural networks for connecting caregivers of rare disease patients

Filippo Guerranti[a,**], Mirco Mannino[a,*], Federica Baccini[d,e,**], Pietro Bongini[a,b,**], Niccol Pancino[a,b,**], Anna Visibelli[a,c,**], Sara Marziali[a,**]

[a]*Department of Information Engineering and Mathematics, University of Siena, 53100, Siena, Italy*
[b]*Department of Information Engineering, University of Florence, 50139, Florence, Italy*
[c]*Department of Biotechnology, Chemistry and Pharmacy, University of Siena, 53100, Siena, Italy*
[d]*Department of Computer Science, University of Pisa, 56124 Pisa,Italy*
[e]*Institute for Informatics and Telematics, CNR, 56124 Pisa, Italy*

Abstract

Rare diseases affect a growing number of individuals. One key problem for patients and their caregivers is the difficulty in reaching experts and associations competent on a particular disease. As a consequence, caregivers, often family members of the patient, learn much about the disease from their own experience. CaregiverMatcher is a proof of concept providing a smart solution to build a network of caregivers, linked by a matching mechanism based on graph neural networks. The caregivers and their experience with rare diseases are described by node features. Associations and care centers are invited to share their knowledge on the platform.

Keywords: Cross-platform application; Caregivers; Rare diseases; Graph Neural Networks; Deep Learning.

1. Introduction

Any disorder which has a low prevalence in the target population, typically chronic and potentially life-threatening, is known as *rare disease*. In the United States, a disease is defined as *rare* when it affects less than 200,000 people in the US population. This definition was introduced in the Orphan Drug Act of 1983 with the aim of regulating the production of drugs for the treatment of such diseases. In the European Union a rare disease is defined as a disorder "with an incidence of less than 1 per 2000 people". This was first established in the EU legislation in Regulation (EC) 141/2000 of 16 December 1999.

* Corresponding author
** Equal contributions
E-mail address: mirco.mannino@student.unisi.it

1877-0509
10.1016/j.procs.2021.08.174

According to Eurordis [1], the estimated number of rare diseases is higher than 6,000 and, depending on the local definitions of *rarity*, the prevalence of people suffering from them varies between 3.5% and 5.9%. This results in 263-446 millions of affected people worldwide [2]. Therefore, although rare disorders have a relatively low prevalence, the number of patients is still very large.

An important role in the support of people affected by rare diseases is that of the *caregiver*. Caregivers provide daily assistance to people with impairments caused by ageing, chronic diseases, infirmities, etc. They can be either members of the patient's family, or people hired for providing help. An ISTAT report of 2015 [3] states that, in the European Union, 15 out of 100 people provide assistance to individuals with impairments at least once a week. Regarding care for family members, this number slightly decreases to 13. This indicates that the role of the caregiver is mostly occupied by family members, who rarely have got an adequate education on how to deal with people affected by some impairments. In addition, the constant attention to the patient's needs, and the social isolation that the role of being caregivers entails are at the basis of the obstacles they have to deal with in the daily assistance [4]. This aspect becomes even more relevant when the assisted patient is affected by a rare disease. The diagnosis is often a slow and difficult process [5] which can lead to sudden changes in the life of a patient. Consequently, it is often challenging for a caregiver, be it a family member or not, to give immediately the appropriate support to the patient.

A further obstacle is represented by the fact that rare diseases dedicated associations are generally dispersed around the world. This makes it difficult for caregivers and their patients to communicate with specialized centers, resulting in the lack of psychological and practical support. In order to cope with the issues of isolation and poor communication with healthcare professionals, a network of caregivers is extremely valuable [6].

In this work, a proposal for the design of a cross–platform application in support of the caregiver's experience is presented. The proposal is called *CaregiverMatcher*, and its aim would be to create a network of caregivers assisting people affected by rare diseases. Technically, CaregiverMatcher would exploit graph neural networks (GNNs [7]) to perform a matching (an association) between caregivers, based on information about the assisted patients. Consequently, CaregiverMatcher would give the caregiver the opportunity to establish a direct contact with other caregivers that face similar issues in daily assistance. Moreover, CaregiverMatcher would make some informative sections available, which would be dedicated to improve the knowledge about rare diseases. These sections would be curated by doctors and associations joining the platform, and they would also include contacts to medical centers and associations. In summary, besides offering the opportunity of a direct contact between caregivers to promote the sharing of experiences, CaregiverMatcher attempts to facilitate the exchange of information between associations, doctors and caregivers in the field of rare diseases.

The paper is organized as follows: in Section 2, some related works are presented.Section 3 explains the architectural structure of CaregiverMatcher. Section 4 highlights the strengths and limitations of the proposed application. Finally, Section 5 presents the conclusions of the paper.

2. Related works

Several technologies have been proposed with the aim of improving the health and well-being of caregivers, by enabling them to communicate with other caregivers [8, 9, 10]. These approaches can reduce difficulties related with the access to health care providers and resources to give an appropriate support to the assisted patient [11]. Other applications focus instead on the creation of a "health team" community, through which caregivers and patients can request information and receive feedback [12]. Finally, other support technologies provide an emergency channel through which the patient can immediately be put in contact with the caregiver [13].

However, to the best of our knowledge, existing applications do not exploit Machine Learning techniques as a mean of facilitating communications between caregivers, associations and doctors. The proposal represents an innovation in this direction, because CaregiverMatcher would select groups of caregivers facing similar daily issues by exploiting graph neural networks (GNNs).

GNNs, first introduced in [7] and [14], are deep neural networks designed to process graphs, which have been proven to be universal approximators on graph-structured inputs under certain conditions [15]. This family of models includes several types of architectures, that differ for structure and performed tasks (for an exhaustive taxonomy of existing GNN models see [16] and [17]). The first models to be introduced [7, 14] exploit recurrent neural networks to learn a node's representation in a graph. Later, convolutional GNNs (GCNs) were introduced [18, 19, 20, 21]. In

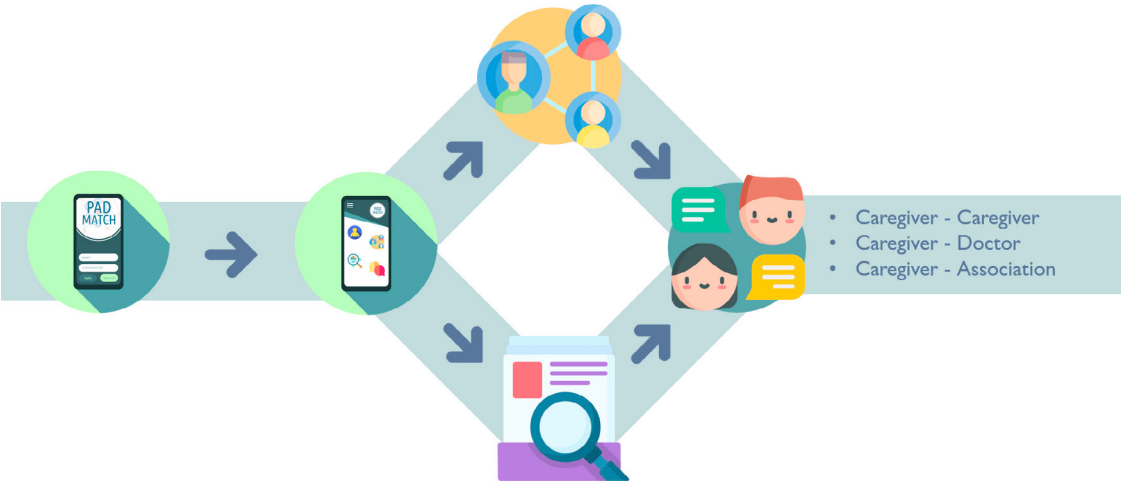


Fig. 1. General architecture of CaregiverMatcher mobile app. From the left: to access the platform, caregivers log in with username and password. Four sections are available in the home page: *Profile*, to manage personal and patient data; *Chat*, where all messages and chat conversations are stored; *Get Informed* to retrieve rare diseases information as well as associations or doctors contacts; *Match* to start the matching process. As a result, caregivers can then connect with patient associations, specialized clinicians and other caregivers.

order to learn node representations, GCNs exploit convolutional operations to aggregate the information contained in a node's neighbourhood. Moreover, the type of convolution applied to the graph has led to the definition of many different architectures, e.g. Graph Isomorphism Networks [22].

3. Materials and Methods

CaregiverMatcher main purpose is to allow caregivers to share their experience with other people, and to spread knowledge gained in the course of their assistance to patients affected by a rare disease. CaregiverMatcher can be described as a free and easy–to–use multi–platform network application, where the user can both provide and request psychological or technical support, which comes in the form of a simple chat conversation with other caregivers, patient associations and specialized doctors. This could lead to benefits in many aspects of daily life, including, for example, psychological health. In addition, more expert caregivers can spread their experience, in order to make "newcomers" benefit of the advises which have been shared on the platform.

Finally, the feeling of abandonment a caregiver can experience [23], may be mitigated by the possibility offered by CaregiverMatcher to promote the communication with associations and doctors.

3.1. Proposed architecture

During the registration on the platform, the caregiver creates a personal profile and provides some information about the patient. Information may include personal data such as age, gender, height and weight, blood group, as well as health condition, symptoms and characteristics of the specific disease affecting the patient. For example, information regarding organs involved in the disease, condition of reduced mobility, blindness, difficulties in breathing and communicating may be included. It is worth noting that these data are anonymous and intentionally generic, in order to protect the patients identity. Personal information cannot be accessed by other users: patients data are exclusively used in the matching process by the machine learning model.

Moreover, patients and caregivers data can be updated in the dedicated *Profile* section, in order to keep track of the course of the disease, as well as of the experience of the caregiver. Multiple patient records are available in the profile section to include caregivers providing assistance to more than one patient at the same time. If for some reason a new

patient replaces one of those already present in the caregiver's profile, the experience gained with the previous one is still used by CaregiverMatcher as useful information for the matching process.

In addition to the modifiable *Profile* section, CaregiverMatcher makes other three sections accessible to the user:

- *Connecting* section, where the user can discover the usernames of the matched caregivers. The association is performed by using the Deep Learning techniques described in Section 3.2, by exploiting information on similar assisted patients and similar life condition of the caregivers.
- *Chat* section, where the caregiver can communicate in real time with other people after the matching process and where all chat sessions are stored.
- *Get Informed* section, a specific area where doctors and associations provide easily understandable documentation about several rare diseases and useful links to external websites. Moreover, direct links to associations and/or doctors are available, so as to offer other communication channels to the user.

3.2. Deep learning-based matching process

The core of CaregiverMatcher is its ability to connect caregivers by means of deep learning techniques. From a practical point of view, the application checks the compatibility between caregivers based on both patient personal information and health condition. In particular, a graph neural network is exploited by the application to perform a matching between caregivers living in similar conditions. The input to the model is constituted by a vector of information regarding both the caregiver and the assisted patient health conditions. As several information could be collected to describe a caregiver in the network, it could happen that the input to the machine learning model is high dimensional. Therefore, in order to facilitate the data processing, some specific architectures, such as autoencoders, could be used to obtain a compressed representation of the input data.

The mathematical structure CaregiverMatcher is built on is constituted by graphs.

A graph is a common data structure composed of two basic elements: a finite set of nodes (vertices) and a set of arcs (edges) connecting them. In this context, nodes represent entities such as patients, while edges stand for relationships between entities, such as being affected by the same disease. Every node is then associated to a label which includes a compressed representation of patient personal information as well as health condition or symptoms and characteristics of the corresponding disease.

In order to add the caregiver information as well, another type of node, labeled with caregiver personal data, is included in the graph and connected to the patient nodes through an assistive–type relational edge. As a consequence, a heterogeneous graph is obtained. This kind of graph has been commonly used to abstract and model complex systems, in which entities of different types interact. The resulting graph is then composed of two types of nodes and edges: the former represent both patient and caregiver and the latter represents both patient-patient and caregiver-patient relationships.

Neural Networks and Deep Learning have been shown to be efficient in processing graph structured data, both in the homogeneous [16] and heterogeneous [24] domains. In particular, CaregiverMatcher matching function is based on graph neural networks (GNNs), a special class of deep learning models capable of correctly processing data in the graph domain [7], leveraging on node features and on relationships between nodes. GNNs can also include heterogeneous structural information, i.e. different types of nodes and edges, and heterogeneous features associated with each node type.

The GNN model exploited in the present application is asked to predict whether an edge exists between each pair of caregiver nodes. The predicted presence or absence of an edge represents the existence of a caregiver–caregiver relationship, and it is weighted according to a real–valued similarity score describing how compatible their profiles are: the higher the score, the higher the compatibility between the connected users.

Eventually, once the matching process has been completed, the user is returned a list of similar caregivers, filtered as needed by setting some parameters in a dedicated section: by default, the first five compatible caregivers are shown, in decreasing order with respect to the similarity score.

3.3. Graph neural network model

Graph neural networks (GNNs) are deep neural network models capable of processing graph-structured data [7]. The model input is defined as a graph $G = (V, E)$, where V is the set of nodes, $E \subseteq V \times V$ is the set of edges, and every node $v_i \in V$ is labeled with a feature vector l_i. The neighborhood of a node is defined as a function $Ne(v_i) = \{v_j : (v_j, v_i) \in E\}$ assigning a set of neighbors $Ne(v_i)$ to each node $v_i \in V$. A GNN implements two functions: a state updating function $f()$ that allows the network to define and update a state s_i for each node v_i, and an output function $g()$ that calculates the output based on the node states. Depending on the problem at hand, the output function can be defined on a set of nodes $V_{out} \subseteq V$ (see Figure 2), a set of edges $E_{out} \subseteq E$, or the whole graph.

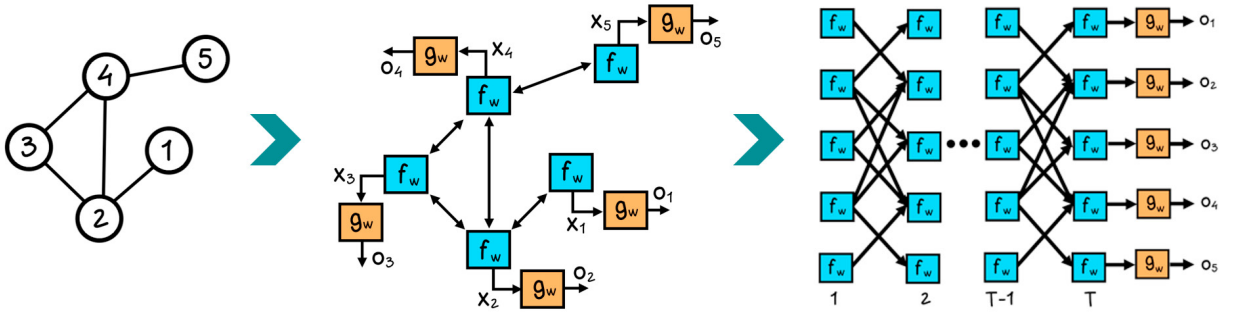


Fig. 2. Graph neural network model for general node-focused application. GNNs create an encoding network, an architecture which replicates the input graph structure by using two MLPs as building blocks. An MLP implements a state transition function f on each node; the other one implements an output function g on each node or edge (or on a subset of them). The network unfolds itself in time and space, respectively, by replicating the MLP units on each node of the input graph, and by iterating the state computations until a stable point or a maximum number of iterations is reached. In the resulting feedforward network, called *unfolding network*, each level corresponds to a time instant, and contains a copy of all the elements of the encoding network, which determines connections between the various layers.

Since the objective is to match a new (caregiver) node with the most similar (caregivers) nodes in the network, by calculating a matching score, the new node is connected to all the existing caregiver nodes. Then, an edge regression task is assigned to the GNN, where E_{out} will correspond to the subset of edges connected to the new node. After the scores have been predicted, all the edges connecting the new node to the rest of the graph, but the top-5 matching nodes, will be erased from the graph. The state updating function is defined in Eq. (1), while the edge-based output function is defined in Eq. (2)

$$s_i^k = f(s_i^{k-1}, \phi(\{s_j^{k-1} j : v_j \in Ne(v_i)\})) \tag{1}$$

$$y_{i,j} = g(s_i^m, s_j^m) \tag{2}$$

In particular, after initializing the state of each node with its feature vector $s_i^0 = l_i$, the state calculation is iterated m times (with m being set as a hyperparameter). The state at each $1 \leq k \leq m$ iteration is calculated as in Eq. (1), based on the node state, and on the states of its neighbours at iteration $k-1$, which are grouped by an aggregation function ϕ (sum, average, or another custom aggregation method). Therefore, nodes exchange information by sending their state vectors through the outcoming edges, and by receiving the states of their neighbours through the incoming edges. This process is called *message passing*, and allows to exploit the relationships (edges) between the nodes without breaking the graph structure. This represents an advantage in the use of graph neural networks with respect to other methods (e.g. random walks) which encode the graph into a vector before processing the graph information, often leading to a loss of structural information. Each function is implemented with a Multi-Layer Perceptron (MLP) module [25]. The state network is replicated on each node of the input graph. The output network is placed on the entities (nodes or edges) for which an output is required, depending on the type of problem. In the case of CaregiverMatcher, the output network is located on each edge linking two caregivers. Finally, all the replicas of the same MLP share their parameters [26].

3.4. Autoencoders

As many different rare diseases and patient health conditions exist, the input features of the nodes labels may be represented in a high-dimensional space. This may result in high computational costs if caregivers and patients are described by large vectors, since the model tends to work on all the available data.

In order to facilitate data processing, compressed representations of the nodes features can be obtained by using appropriate techniques. In particular, autoencoders are unsupervised machine learning models which are often used to learn compact representations of high–dimensional vectors [27, 28]. Theyre usually designed as feedforward neural networks composed of two sub–units: an encoder and a decoder. The former has the role of compressing the input data, while the latter learns to reconstruct the original input starting from the compressed version provided by the encoder. The last layer of the encoder could be thought of as a bottleneck which forces a compressed representation of the original input. This allows to use the encoder as a data preprocessing tool to perform feature dimensionality reduction on raw data. This is obtained, in practice, by first training the autoencoder, and then by exploiting the hidden layer output to train the machine learning model. In particular, the dimensionality of nodes features can be reduced with an autoencoder, and the resulting representation used as node labels in the GNN learning procedure.

It is worth noting that this is a special kind of unsupervised task, called self-supervised, since there is no need for a supervisor to give the correct answer to the network, as its target is the input itself.

4. Discussion: strengths and limitations

CaregiverMatcher is a cross-platform application designed to facilitate the communication between caregivers, by offering them the possibility to interact with specialists and rare disease associations. The interaction with other caregivers, with doctors and with associations are thought, in particular, for caregivers assisting patients affected by rare diseases.

Firstly, CaregiverMatcher has an inutitive and easy–to–use interface. Caregivers only have to register to the application and to select the desired page (*Chat*, *Get Informed*, ...). A GNN model will automatically look for a correspondence between a caregiver and other users with similar needs, based on the features of the assisted patients.

Secondly, the language used in CaregiverMatcher is easy to comprehend. This assumes a particular importance as caregivers can have different levels of education. A further advantage brought by CaregiverMatcher is the low cost of its usage, design, and implementation. Indeed, the user only needs to have a mobile phone or a PC to have access to the application. Moreover, the GNN exploited by CaregiverMatcher would perform the matching operation with a high level of accuracy in a reasonable computational time. Nevertheless, it has to be pointed out that a high level of efficiency of the application can be reached only after a variable (yet not quantifiable) amount of time. Actually, the GNN model will require a consistent amount of data (a consistent number of registered users) to efficiently perform a matching between the nodes of the network (patients and caregivers). However, even if an accurate matching is not immediately available to the user, the simplified informative pages on rare diseases, and the related useful links to web pages on related topics could be exploited as soon as CaregiverMatcher is launched.

Another crucial point in the development of CaregiverMatcher is that it is thought not only for supportive care, but also for sharing knowledge and experiences among caregivers. In the past, there have been attempts to improve support to caregivers (see, for example, the COPE project [29, 30]). However, these attempts were mostly focused on giving supportive care, rather than on offering a concrete possibility of sharing experiences. In contrast, CaregiverMatcher is designed to offer a multidisciplinary psychological support to caregivers. The opportunity to virtually talk with other caregivers with similar experiences results in an occasion for giving/receiving help in daily problems. It is well established that support groups for caregivers lead to improvements in psychological well-being, caregiver burden, and social consequences [31]. Indeed a lot of caregivers have limited access to information and resources that exist in their communities, and often report feelings of isolation and inadequate social support [32]. The chat page of CaregiverMatcher gives the caregiver a real–time opportunity to express private feelings, by establishing a connection with caregivers assisting patients with similar diseases.

In particular, the writing process immediately after an emotionally charged event has already been described as therapeutic [33]. Interventions comprising provision of information, psycho–educational and supportive interventions offered by professionals and associations have the aim of improving the well-being of the caregiver [34]. Associa-

tions are responsible for producing reliable educational material, information on diseases, available treatments and the location of knowledgeable clinicians. Finally, CaregiverMatcher provides a solution for practical needs as well, especially during the spread of the COVID-19 pandemic. The system puts the caregivers in contact with medical specialists, when possible, avoiding unnecessary visits to care centres and reducing the risks and difficulties connected to traveling, especially in the case of patients suffering from rare diseases [35].

5. Conclusions

The purpose of this paper is to make a project proposal for a machine learning based application which supports caregivers in daily life. The proposed solution, CaregiverMatcher, consists in a free and easy–to–use multi–platform application which facilitates communication between caregivers, patients associations and specialists. A direct communication channel between caregivers is realized by means of a graph neural network, which performs a matching between similar caregivers based on information regarding both the assisted patient and the living conditions of the caregiver. The use of graph neural networks is the most innovative aspect of this proposal, as GNNs allow to efficiently process input data in the graph domain by exploiting both features describing the graph nodes (in this case, caregivers and patients), and the edges (relations) between them. Moreover, as the potentially high dimensionality of the input features describing the nodes of the graph may affect the performance of the GNN, an option to enrich the application would be to use an autoencoder to obtain a compressed representation of the input data. A further advantage in this proposal is that CaregiverMatcher would perform the matching with a heterogeneous graph as input. This allows to take into account relevant information about the assisted patient, as well as problems encountered by the caregiver in the daily assistance.

Overall, CaregiverMatcher aims at improving the caregivers knowledge not only by providing direct contact with other caregivers, but also by offering a section of easily understandable information material on rare diseases, provided by associations, doctors and health professionals, with useful links to get in touch with doctors or associations, as well as to external websites or to additional material.

In conclusion, CaregiverMatcher may result in benefits in many aspects of caregivers life, including mental health, by providing psychological and practical support, along with the possibility to easily access reliable educational material offered by professionals and associations.

Acknowledgements

We want to thank Professor Monica Bianchini for supporting us in this work. This project and all the ideas described so far are based on the *PADMatch* finalist project proposed by the authors of this paper for the italian *Rare Disease Hackathon 2020 (digital edition)*, an online event held by *Forum Sistema Salute* with the aim of promoting technological innovation applied to rare diseases, so as to facilitate the lives of patients suffering from them. In particular, PADMatch offered a possible solution for the challenge *The caregiver: which needs? Competencies and supporting technologies*, entering the final round of the event.

References

[1] Eurordis. About rare diseases. URL https://www.eurordis.org/about-rare-diseases.

[2] Stéphanie Nguengang Wakap, Deborah M Lambert, Annie Olry, Charlotte Rodwell, Charlotte Gueydan, Valérie Lanneau, Daniel Murphy, Yann Le Cam, and Ana Rath. Estimating cumulative point prevalence of rare diseases: analysis of the orphanet database. *European Journal of Human Genetics*, 28(2):165–173, 2020. URL https://doi.org/10.1038/s41431-019-0508-0.

[3] ISTAT. Condizioni di salute e ricorso ai servizi sanitari in Italia e nell'Unione europea - Indagine Ehis 2015, 2017. URL https://www.istat.it/it/archivio/204655.

[4] Maryam Navaie-Waliser, Penny H. Feldman, David A. Gould, Carol Levine, Alexis N. Kuerbis, and Karen Donelan. When the Caregiver Needs Care: The Plight of Vulnerable Caregivers. *American Journal of Public Health*, 92(3):409–413, mar 2002. ISSN 0090-0036, 1541-0048. URL https://doi.org/10.2105/AJPH.92.3.409.

[5] J Thevenon, Y Duffourd, A Masurel-Paulet, M Lefebvre, F Feillet, S El Chehadeh-Djebbar, J St-Onge, A Steinmetz, F Huet, M Chouchane, V Darmency-Stamboul, P Callier, C Thauvin-Robinet, L Faivre, and J.B. Rivire. Diagnostic odyssey in severe neurodevelopmental disorders: toward clinical whole-exome sequencing as a first-line diagnostic test. *Clinical Genetics*, 89(6), 2016. ISSN 00099163. doi: 10.1111/cge.12732. URL https://doi.org/10.1111/cge.12732.

[6] Eylin Palamaro Munsell, Ryan P. Kilmer, James R. Cook, and Charlie L. Reeve. The effects of caregiver social connections on caregiver, child, and family wellbeing. *American Journal of Orthopsychiatry*, 82(1):137–145, 2012. ISSN 1939-0025, 0002-9432. URL https://doi.org/10.1111/j.1939-0025.2011.01129.x.

[7] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009. URL https://doi.org/10.1109/TNN.2008.2005605.

[8] Ianacare, support family caregivers and the people in their lives who want to help. URL https://www.ianacare.com/how-it-works.

[9] CaringBridge, keep your family and friends updated during a difficult time., . URL https://www.caringbridge.org/how-it-works/.

[10] Caring Village, free help for caregivers, . URL https://www.caringvillage.com/.

[11] Med Guide : Medication Optimization supported via Telehealth. URL https://illuminate.health/home-2/med-guide/.

[12] Matheus Costa Stutzel, Michel Pedro Filippo, Alexandre Sztajnberg, Rosa Maria E.M. da Costa, André da Silva Brites, Luciana Branco da Motta, and Célia Pereira Caldas. Multi-part quality evaluation of a customized mobile application for monitoring elderly patients with functional loss and helping caregivers. *BMC Medical Informatics and Decision Making*, 19(1):140, 2019. ISSN 1472-6947. URL https://doi.org/10.1186/s12911-019-0839-3.

[13] Fbio Ferreira, Flvio Dias, Jo£o Braz, Ricardo Santos, Roberto Nascimento, Carlos Ferreira, and Ricardo Martinho. Protege: A mobile health application for the elder-caregiver monitoring paradigm. *Procedia Technology*, 9:1361–1371, 2013. ISSN 2212-0173. URL https://doi.org/10.1016/j.protcy.2013.12.153.

[14] Alessandro Sperduti and Antonina Starita. Supervised neural networks for the classification of structures. *IEEE Transactions on Neural Networks*, 8(3):714–735, 1997.

[15] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. Computational capabilities of graph neural networks. *IEEE Transactions on Neural Networks*, 20(1):81–102, 2008. URL https://doi.org/10.1109/TNN.2008.2005141.

[16] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, 2021. URL https://doi.org/10.1109/TNNLS.2020.2978386.

[17] Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks, 2018.

[18] Thomas N Kipf and Max Welling. Semi–supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

[19] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*, pages 3844–3852, 2016. URL https://doi.org/10.5555/3157382.3157527.

[20] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and deep locally connected networks on graphs. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014. URL https://doi.org/10.1109/TNN.2008.2005141.

[21] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in neural information processing systems*, pages 1024–1034, 2017.

[22] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[23] Paola Cardinali, Laura Migliorini, and Nadia Rania. The caregiving experiences of fathers and mothers of children with rare diseases in italy: Challenges and social support perceptions. *Frontiers in Psychology*, 10:1780, 2019. ISSN 1664-1078. URL https://doi.org/10.3389/fpsyg.2019.01780.

[24] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous graph transformer. In *Proceedings of The Web Conference 2020*, WWW '20, page 27042710, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370233. URL https://doi.org/10.1145/3366423.3380027.

[25] Alberto Rossi, Matteo Tiezzi, Giovanna Maria Dimitri, Monica Bianchini, Marco Maggini, and Franco Scarselli. Inductive–transductive learning with graph neural networks. In Trentin E. Pancioni L., Schwenker F., editor, *Artificial Neural Networks in Pattern Recognition. ANNPR 2018, Lecture Notes in Computer Science, vol 11081*, pages 201–212. Springer, Cham, 2018. URL https://doi.org/10.1007/978-3-319-99978-4_16.

[26] Niccol Pancino, Alberto Rossi, Giorgio Ciano, Giorgia Giacomini, Simone Bonechi, Paolo Andreini, Franco Scarselli, Monica Bianchini, and Pietro Bongini. Graph neural networks for the prediction of proteinprotein interfaces. In *28th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (online event)*, pages 127–132, 2020. URL http://hdl.handle.net/11365/1117820.

[27] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006. URL https://doi.org/10.1126/science.1127647.

[28] Laurens Van Der Maaten, Eric Postma, and Jaap Van den Herik. Dimensionality reduction: a comparative review. *J Mach Learn Res*, 10 (66-71):13, 2009.

[29] Mike Nolan and Ian Philip. Cope: towards a comprehensive assessment of caregiver need. *British Journal of Nursing*, 8(20):1364–1372, 1999. URL https://doi.org/10.12968/bjon.1999.8.20.1364.

[30] KJ McKee, I Philp, G Lamura, C Prouskas, Birgitta Öberg, Barbro Krevers, L Spazzafumo, B Bien, C Parker, MR Nolan, et al. The cope index–a first stage assessment of negative impact, positive value and quality of support of caregiving in informal carers of older people. *Aging & mental health*, 7(1):39–52, 2003. URL https://doi.org/10.1080/1360786021000006956.

[31] Ling-Yu Chien, Hsin Chu, Jong-Long Guo, Yuan-Mei Liao, Lu-I Chang, Chiung-Hua Chen, and Kuei-Ru Chou. Caregiver support groups in patients with dementia: A meta-analysis. *International journal of geriatric psychiatry*, 26:1089–98, 10 2011. URL https://doi.org/10.1002/gps.2660.

[32] Kenneth E. Miller, Heba Ghalayini, Maguy Arnous, Fadila Tossyeh, Alexandra Chen, Myrthe van den Broek, Gabriela V. Koppenol-Gonzalez, Joy Saade, and Mark J.D. Jordans. Strengthening parenting in conflict-affected communities: development of the caregiver support intervention. *Global Mental Health*, 7:e14, 2020. doi: 10.1017/gmh.2020.8. URL https://doi.org/10.1017/gmh.2020.8.

[33] Brian Perron. Online support for caregivers of people with a mental illness. *Psychiatric Rehabilitation Journal*, 26(1):70, 2002. URL https://doi.org/10.2975/26.2002.70.77.

[34] Sharon Nelis, Catherine Quinn, and Linda Clare. Information and support interventions for informal caregivers of people with dementia. *Cochrane Database of Systematic Reviews*, 2007. URL https://doi.org/10.1002/14651858.CD006440.

[35] Sridhar Vaitheswaran, Monisha Lakshminarayanan, Vaishnavi Ramanujam, Subashini Sargunan, and Shreenila Venkatesan. Experiences and needs of caregivers of persons with dementia in india during the covid-19 pandemica qualitative study. *The American Journal of Geriatric Psychiatry*, 28(11):1185–1194, 2020. ISSN 1064-7481. URL https://doi.org/10.1016/j.jagp.2020.06.026.

International Journal of
Molecular Sciences

MDPI

Review

Multi-Omics Model Applied to Cancer Genetics

Francesco Pettini [1,*,†], Anna Visibelli [2,†], Vittoria Cicaloni [3], Daniele Iovinelli [2] and Ottavia Spiga [2]

[1] Department of Medical Biotechnology, University of Siena, Via M. Bracci 2, 53100 Siena, Italy
[2] Department of Biotechnology, Chemistry and Pharmacy, University of Siena, Via A. Moro 2, 53100 Siena, Italy; anna.visibelli@student.unisi.it (A.V.); daniele.iovinelli@student.unisi.it (D.I.); ottavia.spiga@unisi.it (O.S.)
[3] Toscana Life Sciences Foundation, Via Fiorentina 1, 53100 Siena, Italy; v.cicaloni@toscanalifesciences.org
* Correspondence: francesco.pettini@dbm.unisi.it; Tel.: +39-3755461426
† These authors contributed equally to this work.

Abstract: In this review, we focus on bioinformatic oncology as an integrative discipline that incorporates knowledge from the mathematical, physical, and computational fields to further the biomedical understanding of cancer. Before providing a deeper insight into the bioinformatics approach and utilities involved in oncology, we must understand what is a system biology framework and the genetic connection, because of the high heterogenicity of the backgrounds of people approaching precision medicine. In fact, it is essential to providing general theoretical information on genomics, epigenomics, and transcriptomics to understand the phases of multi-omics approach. We consider how to create a multi-omics model. In the last section, we describe the new frontiers and future perspectives of this field.

Keywords: data analysis; artificial intelligence; precision medicine; machine learning models; computational oncology; cancer disease; omics tools

Citation: Pettini, F.; Visibelli, A.; Cicaloni, V.; Iovinelli, D.; Spiga, O. Multi-Omics Model Applied to Cancer Genetics. *Int. J. Mol. Sci.* **2021**, *22*, 5751. https://doi.org/10.3390/ijms22115751

Academic Editors: Claudiu T. Supuran, Paola Gratteri and Silvia Selleri

Received: 19 April 2021
Accepted: 26 May 2021
Published: 27 May 2021

1. Introduction

Last fact sheets from World Health Organization (WHO), updated to March 2021, reports cancer is the second leading cause of death worldwide, accounting for nearly 10 million deaths in 2020. Approximately 70% of the deaths from cancer occur in low- and middle-income countries. Breast, lung, colorectal, and prostate cancers are the most common [1].

A correct cancer diagnosis is essential for adequate and effective treatment because every tumor is involved in interactions with non-cancer elements such as gene-environment interactions (GxE), micro-environmental interactions, and those with the immune system; intercellular interactions within the tumor environment; and intracellular interactions, such as transcriptional regulation and gene co-expression, signaling and metabolic pathways, as well as protein interactions (Figure 1) [2].

This is the reason why only an integrating framework among different omics layers can gather and organize the knowledge gained with each experimental approach into mechanistic or semi-mechanistic descriptions of the biological phenomenon [3].

Multi-omics model is defined as a biological approach that, by using one or more current high-throughput experimental techniques, can investigate physiological or pathological phenomena and characterize biomolecular systems at different levels. As a matter of fact, each omics contributes on a specific fashion to shape the actual biological phenotype of interest.

Thus, a comprehensive recognition of molecular networks based on multi-omics data has an important scientific role to understand the molecular mechanisms of cancer, but this is possible only because of bioinformatics application [4]. Computational oncology can be defined as an integrative discipline incorporating scientific backgrounds from the mathematical, physical, and computational fields to get a deeper understanding on malignancies [2].

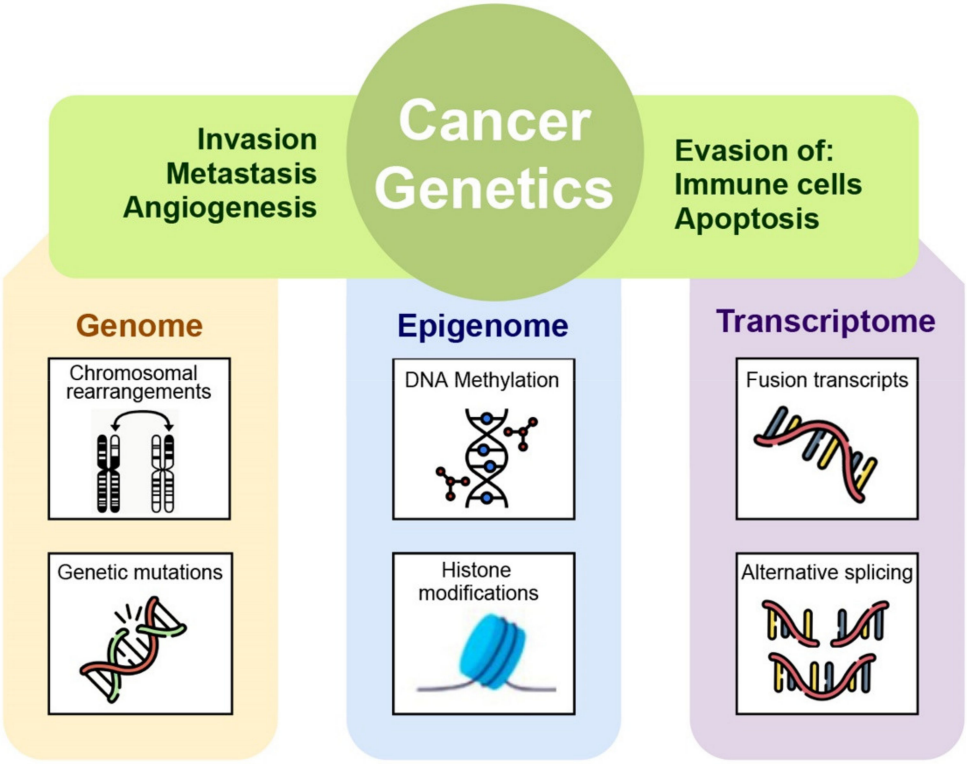


Figure 1. The many levels of interactions found in a cancer system, that can be measured via the different omics technologies, such as genomics, epigenomics, transcriptomic, and proteomic.

In the coming age of omics technologies, next gen sequencing, proteomics, metabolomics, and other high throughput techniques will become the usual tools in biomedical cancer research. However, their integrative approach is not trivial due to the broad diversity of data types, dynamic ranges and sources of experimental and analytical errors characteristic of each omics [2]. The multi-omics systematic study of cancer found many different factors involved in the development/maintenance of the malignant state such as genetic aberrations, epigenetic alterations, changes in the response to signaling pathways, metabolic alterations, and many others [5]. The advent of high-throughput technologies has permitted the development of systems biology. The system biology paradigm tries to analyze cancer as a complex and intricate pathology and to gain insight into its molecular origin by taking into account the different contributions like DNA mutations, deregulation of the gene expression, metabolic abnormalities, and aberrant pathway signaling [2].

The essential basis of systems biology is to consider a biological phenomenon as a system of interconnected elements such as many complex molecular and environmental components interacting with each other at different levels. For example, tumor behavior is determined by a combination of changes in genomic information possibly associated with abnormal gene expression, protein profiles, and different cellular pathways. In this scenario, the complex interaction of DNA and proteins in replication, transcription, metabolic, and signaling networks are considered the decisive causes for cancer cells dis-functioning [2]. The integration of multi-omics data provides a platform to connect the genomic or epigenomic alterations to transcriptome, proteome, and metabolome networks underling the cellular response to a perturbation. Powerful and sophisticated computational tools can identify the interconnection between genomic aberrations with differentially expressed mRNAs, proteins, and metabolites associated with cancer-driven cellular perturbation [6]. If on the one hand this aspect provides an opportunity to better study the cellular response, on the other hand it poses a challenge for systems biology-driven modelling. Therefore, the next step of systems biology approach focuses on dynamic models that can deal with thousands of mRNA, protein, and metabolite changes developing effective strategies to

administer personalized cancer therapy [7]. Summarizing, the main goal of the systems biology research driven by multi-omics data is to develop predictive models that are refined by experimental validations in order to select patients based on personalized multi-omics data and stratifying them to determine who are most likely to benefit from targeted therapies [6,8].

Definition and detection of cancer-distinctive features allow the investigation of the transition process of a normal cell to malignancy. Generally, the hallmarks involve phenotypic and molecular changes in several metabolic pathways such as uncontrolled proliferation by blocking growth suppressors, reprogramming of energy metabolism, evading immune destruction, resisting cell death, angiogenesis, and metastasis [9]. These variations in cellular machinery are driven by molecular aberration in several omics layers such as genome, epigenome, transcriptome, proteome, and metabolome within cancer cells. Specifically, by applying next generation sequencing to cancer cell genomes, it is possible to reveal how mutations in proliferative genes like B-raf drives the activation of mitogen-activated protein- (MAP-) kinase signaling pathway underlying an uncontrolled cell proliferation [10]. Molecular aberrations leading to cancer are involved not only in genomic mutational events but also in the epigenome. In particular, aberrant epigenetic mechanisms can be responsible for silencing of certain cancer suppressor genes [11]. The multistep processes of invasion and metastasis require a transition of epithelial cell toward mesenchymal phenotype to colonize distant sites. Recent studies have revealed that epithelial-mesenchymal transition is induced by specific transcription factors that coordinate the invasion and metastasis processes [9]. By applying transcriptomics techniques it is possible to investigate the transcription factors involved in transcription regulatory networks assumed to be activated in malignancy. Moreover, manifestations of cancer hallmarks also affected cellular metabolism, in fact tumor cells can reprogram glucose metabolism and energy production pathways detectable with a metabolomics approach [6].

2. Genomics and Molecular Processes

2.1. Cancer Gene Types

In general, cancer disrupts cellular relations and results in the dysfunction of vital genes. This disturbance is affective in the cell cycle and enhances abnormal proliferation [12,13]. There are three main types of cancer genes that control cell growth and can cause cancer to develop:

1. Oncogenes. These, when mutated, actively promote cell proliferation. They are formed when proto-oncogenes that promote cell division are improperly activated, so they are not known to be inherited. They may lead to increased/dysregulated expression of the gene in a new location or to production of fusion proteins with new functions [14]. Two common oncogenes are HER2 and RAS.
2. Gatekeeper genes. These are protective genes, also known as tumor suppressor genes. Normally, they negatively control cell growth by monitoring and controlling the cell phases or repairing mismatched DNA. Autosomal recessive mutations in tumor suppressor gene cause loss of function effect at the cellular level, inducing cells to grow uncontrollably, which may eventually form a tumor. Examples of tumor-suppressor genes include BRCA1, BRCA2, and p53 or TP53. Germline mutations in BRCA1 or BRCA2 genes increase a woman's risk of developing hereditary breast or ovarian cancers and a man's risk of developing hereditary prostate or breast cancers. They also increase the risk of pancreatic cancer and melanoma in women and men [15]. The most mutated gene in people with cancer is p53 or TP53. More than 50% of cancers involve a missing or damaged p53 gene. Most p53 gene mutations are acquired. Germline p53 mutations are rare, but patients who carry them are at a higher risk of developing many different types of cancer [15].
3. Carekeeper genes. These fix the mistakes made when DNA is copied. Many of them function as tumor suppressor genes. BRCA1, BRCA2, and p53 are all DNA repair genes. If a person has an error in a DNA repair gene, mistakes remain uncorrected.

Then, the mistakes become mutations. These mutations may eventually lead to cancer, particularly mutations in tumor suppressor genes or oncogenes. Mutations in DNA repair genes may be inherited or acquired. Lynch syndrome is an example of the inherited kind. BRCA1, BRCA2, and p53 mutations and their associated syndromes are also inherited [14].

As said before genetic changes that promote cancer can be inherited from our parents if the changes are present in germ cells, which are the reproductive cells of the body (eggs and sperm). Such changes, called germline changes, are found in every cell of the offspring. Cancer-causing genetic changes can also be acquired during one's lifetime and are called somatic (or acquired) changes [14]. Next, we will take these aspects into consideration.

2.2. Genomic Instability

Somatic mutations, based on their function, involves driver mutations, conferring growth advantage to the cancer cells. Otherwise, acquired mutations do not confer any growth advantage to the cancer cells nor contribute to cancer development [16]. Chromosomal changes are highly variable, they can be grouped into two general categories [17]:

- Balanced structural changes; the genetic material is equally exchanged, even if genetic information was rearranged into an abnormal gene;
- Unbalanced or nonreciprocal structural changes; the exchange is not equally distributed, and genetic material is added or lost. This can range from the loss or gain of a single base pair to the loss or gain of the entire chromosomes.

From Knudson's two hits hypothesis [18] studies to the present, scientists suggested that the primary pathogenetic changes in cancer result from balanced rearrangements, while the secondary hits that occur during cancer progression are from unbalanced changes (see Table A1).

In Wilms' tumor and retinoblastoma, gene deletions or inactivation are responsible for cancer development [17]. Deletions, inversions, and translocations are commonly detected in the Philadelphia chromosome, the first balanced chromosomal mutation described in cancer cells; it is the result of a reciprocal translocation between chromosomes 9 and 22 with breakpoints in the c-abl gene on chromosome 9 and the c-bcr gene on chromosome 22. The fusion gene created by this rearrangement encodes a tyrosine kinase that promotes cancer in white blood cells (chronic myeloid leukemia) [19]. Burkitt's lymphoma is another type of cancer associated with reciprocal translocations involving chromosome 8 and a chromosome carrying an immunoglobulin gene (2, 14, or 22). The translocations juxtapose c-myc to the genes for the immunoglobulin genes, causing overexpression of c-myc in B cells. The c-myc gene encodes a transcription factor that activates genes for cell division [20]. Large portions of chromosomes can also be lost, as occurred on chromosomes 1p and 16q in solid tumor cells [17]. Gene duplications and increases in gene copy numbers can also contribute to cancer and can be detected, for example, in many sarcomas. The chromosomal region 12q13-q14 encodes a binding protein called MDM2, which is known to bind to a tumor suppressor called p53. Amplification of MDM2 prevents p53 from regulating cell growth, which can result in tumor formation. Also, mutations in carekeeper genes can additionally lead to rearrangements and duplications [17]. Most of the cancers harbor more than one driver gene mutation. Breast, colorectal, and prostate cancers require from five to seven driver mutations for cancer initiation and progression, while hematological malignancies may require fewer. TP53, RB1, EGFR, and KRAS, are widely known mutated genes in various cancer types, whereas others are rare and/or restricted to one cancer [16].

2.3. Epigenomic Instability

It is largely proved that genomic instability is a reductive model; studies demonstrated epigenetic errors resulting in aberrant gene silencing/activation [21]. According to the definition, epigenetics is a dynamic situation in the study of cell fate, that alter the structure of DNA without directly affecting and mutating its sequence [22]. In fact, mutations occurred in the elements that regulate the expression or repression of the genome, such

as transcription factors and noncoding RNAs, with a consistent effect on the coordination of multiple biological processes. These elements can be divided into three roles: "writers" and "erasers" refer to enzymes that transfer or remove chemical groups to or from DNA or histones, respectively; "readers" are proteins that can recognize the modified DNA or histones [23].

In tumor tissues, different tumor cells show various patterns of histone modification, genome-wide or in individual genes, demonstrating that epigenetic heterogeneity exists at a cellular level and suggesting that tumorigenesis is the consequence of the combined action of multiple epigenetic events [24]. For example, the repression of gatekeeper genes is usually caused by DNA modification in the methylation of CpG islands together with hypoacetylated and hypermethylated histones [25]. Gene silencing experiments identified several hallmarks of epigenetic events, including histone H3 and H4 hypoacetylation, histone H3K9 methylation, and cytosine methylation [26]. Major epigenetic modifications are classified as DNA modifications, histone modifications, effects of non-coding RNA (Figure 2).

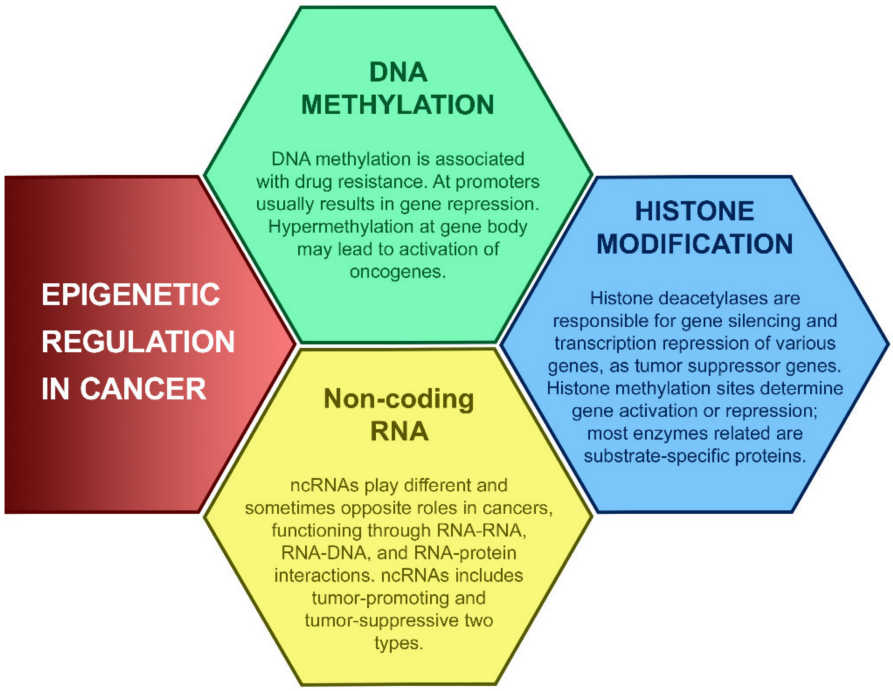


Figure 2. Epigenetic regulations in cancer. Alterations in epigenetic modifications in cancer regulate various cellular responses, including cell proliferation, apoptosis, invasion, and senescence. Through DNA methylation, histone modification, and noncoding RNA regulation, epigenetics play an important role in tumorigenesis. These main aspects of epigenetics present reversible effects on gene silencing and activation via epigenetic enzymes and related proteins.

DNA methylation typically occurs at CpG sites (cytosine-phosphate-guanine) sites. This methylation results in the conversion of the cytosine to 5-methylcytosine. The formation of Me-CpG is catalyzed by enzymes called DNA methyltransferases (DNMTs). This modification is common in body cells; in tumors, we can observe an hypomethylation of the genome [27] that results in initiate and propagate oncogenesis, by inducing chromosome instabilities and transcriptional activation of oncogenes and pro-metastatic genes, such as *r-ras* [28]. This state is accompanied by a region- and gene-specific hypermethylation of multiple CpG islands [29,30]. Hypermethylation of CpG islands in the promoter region of a tumor suppressor or otherwise cancer-related gene is often associated with transcriptional silencing of the related gene. Numerous genes associated to various pathways are known and rapidly identified; actually, genes involved in signal transduction (*APC*), DNA repair (*MGMT*, *MLH1*, *BRCA1*), detoxification (*GSTP1*), cell cycle regulation (*p15*, *p16*, *RB*),

differentiation (*MYOD1*), angiogenesis (*THBS1*, *VHL*), and apoptosis (*Caspases*, *p14*, *DAPK*) are reported and largely studied [31] (for a complete overview on key regulatory factors of DNA methylation in cancer we suggest to see Table 1 of [23]). DNA methylation can act as one hit having the same functional effect as a genetic point mutation, as proven by numerous experiments in which re-establishing expression of tumor suppressor genes could be reached through drugs inducing demethylation. Epimutations can inactivate one of the two alleles, while the other is lost through genetic mechanisms or silence both alleles [32]. For example, a study conducted on 50 RB patients (45 unilateral and 5 bilateral), selected from an initial cohort of 476 RB cases diagnosed over a period of 17 years at the Retinoblastoma Referral Centre of Siena (Ophthalmology Department, AOUS), provided evidence supporting the identification of a constitutional epimutation acting as the first "hit" in the Knudson model of RB development and suggests that epimutations do not represent a frequent cause of RB predisposition but this is an understudied etiological phenomenon and, besides promoter methylation, other untested epigenetic events may reduce gene expression, phenocopying RB onset [33]. Epigenetic changes occur at higher frequency with respect to genetic changes and might be especially important in the first phase of human neoplasia; aberrant promoter methylation is initiated at ~1% of all CpG islands and as much as 10% become methylated during the multistep process of tumorigenesis [34]. As a stable nucleic-acid-based modification with limited dynamic range that is technically easy to handle, DNA methylation is a promising biomarker for non-invasive detection of different tumor types [35–38]. Besides early detection, the methylation status of CpG islands can be used to characterize and classify cancers. While for example, breast, or testicular tumors show global low levels of methylation, some other tumor types such as colon tumors, acute myeloid leukemias, or gliomas are characterized by high levels of methylation, although some heterogeneity is observed in almost all tumor types. So, methylation patterns can be an important hallmark to identify and classify the different types of human cancers [34,39]. DNA methylation profile can be used also to predict and monitor the response to anti-neoplastic treatment [39,40].

Histones are made up of amino acids, like all other proteins. Amino acids located in the tail of them are targets for enzymes that attach or remove chemical markers, therefore the potential main site of histone modifications, in particular, lysine (Lys) and serine (Ser) are common targets. Histone modification is a relatively complicated process compared to DNA methylation that involves only two types of enzymes, which can add or remove only methyl groups to cytosine. When histone modification patterns are altered, it can lead to unregulated activity or silencing of genes related to cancer onset [24,41].

All families of protein involved in chromatin remodeling pathways are associated with cancer, although in most cases, the molecular mechanisms underlying their functions remain unknown [42]. Overall reduction of mono acetylated H4K16 forms the majority of histone modifications in cancer cells [43]. Other modifications, as histone H3 acetylation and methylations, interfere with the chromatin remodeling status, leading to repression or activation of transcription [44]. In contrast ubiquitination is a larger covalent modification, commonly related to the H2B. H2BK123ub1 modification involves the addition of ubiquitin chain to histone H2B and this modification results in regulating transcriptional initiation and elongation, while H2AK119ub1 is involved in gene silencing [45]. Similarly, phosphorylated forms of histones, H3S10ph and H2BS32ph, are implicated in the expression of proto-oncogenes, such as MYC, JUN, and FOS [46] (for a more detailed overview on important enzymes or proteins that regulate histone modification we suggest to see Tables 2 and 3 in reference [23]).

Epigenetic-related noncoding RNAs (ncRNAs) include microRNAs (miRNAs), small interfering RNA (siRNAs), Piwi-interacting RNA (piRNAs), and long noncoding RNAs (lncRNAs). MiRNAs, one of the most studied ncRNAs, are small RNAs (from 19 to 22 nucleotides in length) known to influence gene expression by way of targeting messenger RNA (mRNA) [24]. Generally, they can be classified into tumor-promoting and tumor-suppressing miRNAs. In fact, during tumorigenesis we can observe that oncogenic

miRNAs such as miR-155, miR-21, and miR-17-92 are usually overexpressed, while miRNAs such as miR-15-16 are downregulated. There is another type of miRNA, cellular context-dependent miRNAs, working in tumorigenesis. For example, miR-146 has been shown to be overexpressed in multiple cancers, whereas a study of Garcia et al. has proven that miR-146 can reduce the expression of BRCA1. At the same time, the expression of proteins and enzymes is also regulated by certain miRNAs. The miR-101 reduces EZH2 expression, and abnormal downregulation of miR-101 has been observed in several types of cancers [24]. As miRNAs play critical roles in regulating functions of the cells, disruption in their structure and turnover can also cause diseases [47]. CLL is the first human disease that is associated with miRNA disorders [48]. miRNAs can be used as cancer diagnosis biomarker as determinant of cancer prognosis and patient overall survival. MiRNAs can be used to classify myeloid malignancies. For example, in a study of meta-analysis performed by Erdogan et al. [49], they identified 13 miRNAs of interest from a total of 42 MDS samples and 45 controls studied, 8 of which proved statistically significant on real-time polymerase chain reaction verification. LncRNAs are another diffused group of ncRNAs that can play an important role in tumorigenesis. Some of them are cancer type-specific, such as PCGEM1 in prostate cancer and HEIH in hepatocellular carcinoma. Many aberrant lncRNAs have been discovered in various cancers; for example, dysregulation of HOTAIR has been found in lung, pancreatic, and colorectal cancer [24]. ncRNAs can either be directly involved in tumorigenesis or indirectly affect tumor development by participating in other epigenetic events [24].

3. Roles of Computational Approach in Multi-Omics Era

Computational approach plays central roles not only in the analysis of high-throughput experiments, but also in data acquisition, in processing of raw file derived from several instruments, in storage and management of large streams of omics information and in the data model integration. Bioinformatics workflow management systems can be used in developing and in application of a certain pipeline. Examples of such systems include Galaxy [50], Snakemake [51], Nextflow [52], and the general-purpose Common Workflow Language [53]. Several tools for omics data studies are available in Bioconductor project as packages for the R language [54] and in Biopython project [55].

3.1. Data Acquisition

All the omics technologies have a specific role to figure out the complex phenotype of cells especially in complex diseases like cancer. Knowledge of the biological molecular basis of different cellular signaling pathways does not involve only genes and transcripts, in fact, proteins and metabolites are particularly important to predict the phenotypic alterations for diagnosis and prognosis of cancer, and for this reason, in this chapter, we will spend some words about them. Table 1 represents a summary of the applications of different NGS-based and mass spectrometry-based techniques which are at the basis of different omics data acquisition approaches.

3.1.1. Genomics

To date, genomics approach has highly sustained the finding and investigation of variations at both the germline and somatic levels thanks to many progresses in genome-exome sequencing techniques, for instance from the Sanger sequencing-based approaches to the NGS-based sequencing. Bioinformatics has always had a central role in the analysis of downstream genetic data. For example, in the multiscale scale project "The Cancer Genome Atlas" (TCGA), researchers used NGS sequencing associated to bioinformatics tools with the aim to discover somatic mutational landscape across thousands of tumor samples and to understand the complexity underlying different cancer types [56,57]. For the analysis of NGS data a sequence aligner tool is used on the sequence data (stored in FASTQ format). Some popular aligners are the stand-alone BWA [58], Bowtie [59],

Bowtie2 [60], and SNAP [61], with aligned sequences being stored in SAM (Sequence Alignment Map, text-based) or BAM (Binary Alignment Map) files.

Table 1. Summary of the applications of different techniques for sequencing, which are at the basis of different omics data acquisition approaches. Genomics, epigenomics, and transcriptomics are based on NGS techniques, whereas proteomics and metabolomics are driven by mass-spectrometric (LC-MS/MS) method. The main goal of genomics, epigenomics, and transcriptomics is the screening of genome-wide mutations, the identification of altered epigenomic modifications, and exploring differential RNA expression, while for proteomics and metabolomics is the identification of differentially regulated proteins and metabolites (reprinted from reference [6]).

OMICS	TYPE	PRINCIPLE	APPLICATION	BIOINFORMATICS TOOLS
GENOMICS	Whole exome sequencing	NGS	Exome-wide mutational/analysis	BWA Bowtie Bowtie2 SNAP SAM BAM
	Whole genome sequencing	NGS	Genome-wide mutational/analysis	
	Targeted gene/exome sequencing	Sanger sequencing	Mutational analysis in targeted gene/exon	
EPIGENOMICS	Methylomics	Whole genome bisulfite sequencing	Genome-wide mapping of DNA methylation pattern	Methylation-Array-Analysis SICER2 PeakRanger GEM MUSIC PePr DFilter MACS
	ChIP-sequencing	NGS	Genome-wide mapping of epigenetic marks	
TRANSCRIPTOMICS	RNA-sequencing	NGS	Genome-wide differential gene expression analysis	Bowtie STAR kallisto Salmon
	Microarray	Hybridization	Differential gene expression analysis	
PROTEOMICS	Deep-proteomics	Mass-spectrometry	Differential protein expression analysis	MaxQuant Perseus
METABOLOMICS	Deep-metabolomics	Mass-spectrometry	Differential metabolite expression analysis	Metab metaRbolomics Lipidr

3.1.2. Epigenomics

Epigenomics is concerned with the genome-wide identification of chemical modifications (i.e., methylation and acetylation of DNA) which are involved in regulatory mechanisms controlling gene expression and cellular phenotypes [62]. Chromatin immunoprecipitation (ChIP) assays-coupled NGS (ChIP-seq) and methylation analysis through whole-genome bisulfite sequencing (WGBS) or bisulfite sequencing (BSSeq) are the most widely used methods in epigenomics analysis [6]. By exploiting the advances in NGS field, it is now possible to analyze genome-wide methylome patterns at a single nucleotide resolution and to detect the methylated cytosine bases in genomic DNA. Data from array-based techniques can be analyzed using dedicated packages such as *methylationArrayAnalysis* [63], whereas for ChIP-seq data processing tools like SICER2 [64], PeakRanger [65], GEM [66], MUSIC [67], PePr [68], DFilter [69], and MACS [70] are used.

3.1.3. Transcriptomics

The detection and quantification of RNA transcripts (mRNA, noncoding RNA and microRNAs) is possible owing to the employment of several transcriptomics techniques. Differently from the static nature of genome, transcriptome dynamically changes as consequence of temporal cellular and extracellular stimuli. Microarray was the technique of choice to detect alterations in cellular mRNA levels in a high-throughput manner owing to its ability to quantify the relative abundance of mRNAs for thousands of genes at the same time. Microarrays are widely used to facilitate the identification of genes with differential expression between normal and cancer conditions. With the advent of NGS, the identification of the presence and the abundance of RNA transcripts in genome-wide manner became possible. In contrast to microarrays technique, RNA-seq does not depend on the transcript-specific probes and thus can effectively perform an unbiased detection of novel transcripts, also the less abundant, with high specificity and sensitivity. Starting points for RNA-seq bioinformatics analysis include alignment-based methods, such as Bowtie [59], and STAR [71], or alignment-free methods, such as kallisto [72] and Salmon [73]. Cancer-related omics experiments often rely on specific, tailor-made analytic pipeline. TCGA and other repositories give the great opportunity to analyze the omics data by a pan-cancer approach where different types of cancers can be compared in terms of genomic and transcriptomic landscapes [74].

3.1.4. Proteomics and Metabolomics

Given the high complexity and dynamic range of proteins, their identification and quantification in large scale are significantly challenging. Proteomic analyses are applied to identify and quantify the set of proteins present within a biological system of interest. Progressions of the tandem mass-spectrometry (LC-MS/MS) techniques in terms of resolution, accuracy, quantitation, and data analysis have made it a solid instrument for both the identification and quantification of cells proteome [75]. Recently, the advent of cutting edge high-resolution "Orbitrap" mass-spectrometer instruments associated with powerful computational tools (i.e., MaxQuant [76] and Perseus [77]) simplified the genome-wide detection of all expressed proteins in human cells and tissues paving the way for a first draft of the human proteome [78,79]. MS-based proteomics techniques have been extensively applied also to investigate the proteome alteration in several human cancer tissues [80]. In particular, the study of cancer proteomes is a promising path for biomarkers and therapeutic targets identification because proteins are the molecular unit from which cellular structure and function arise [81].

The application of MS techniques is not restricted to proteomics but rather can be extended to smaller molecules such as metabolites. Metabolomics is characterized by the quantifications of metabolites that are synthesized as products of cellular metabolic activities, such as amino acids, fatty acids, carbohydrates, and lipids. Their levels can be dynamically altered in disease states reflecting aberrant metabolic functions in complex disorders like cancer. Indeed, metabolic variations are significant contributors to cancer development [82]. This is the reason why cancer metabolomics has become an important research topic in oncology [83], with the aim to get new insights on cancer progression and potential therapeutic targets. Lipidomics is a subset of metabolomics [84], specifically cancer lipidomics has recently led to the identification of novel biomarkers in cancer progression and diagnosis [85]. Metabolomics is still an ongoing field with the potential to be highly effective in the discovery of biomarkers, especially in cancer. This is possible due to the support of bioinformatics tools like metab package [86], which provides an analysis pipeline for metabolomics derived from gas chromatography-MS data, or metaRbolomics package [87], which is a general toolbox that goes from data processing to functional analysis. Similarly, the lipidr package [88] is an analogous framework focused on lipidomics data processing.

3.2. Data Management

The huge amount of data deriving from different omics analyses need to be adequately collected and stored. Challenges of data management include defining the type of data to be stored and how to store it, the policies for data access, sharing, use, and finally, long-term archiving procedures [89]. One of the most successful repositories regarding application of multi-omics approach in cancer is NIHs Genome Data Commons (GDC) [90] containing all data generated by the Cancer Genome Atlas (TCGA) project [74]. TCGA project has performed integrative analysis of more than 30 human cancer types with the aim to create a publicly available comprehensive platform for collecting the molecular alterations in the cancer cells at the forefront of multi-omics research [74]. Information about aberrations in the DNA and chromatin of the cancer-genomes from thousands of tumors have been catalogued by matching with the normal genomes and linking these aberrations to RNA and proteins levels. Moreover, it provides data for method development and validation usable in many current projects. In 2020, the collaboration of an international team has completed the most comprehensive study of whole cancer genomes, significantly improving the fundamental understanding of cancer, and indicating new directions for developing diagnostics and treatments. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Project (PCAWG, or the Pan-Cancer Project) involved more than 1300 scientists and clinicians from 37 countries, analyzed more than 2600 whole genomes of 38 different tumor types. Commenting this aspect, Rameen Beroukhim, an associate member of the Broad Institute, said: "It was heartening that this very large group was able to bring together disparate resources and work to come up with some groundbreaking findings". Additionally, Gad Getz, an institute member and the director of the Cancer Genome Computational Analysis Group at the Broad Institute, director of bioinformatics at the Massachusetts General Hospital's (MGH) Cancer Center and professor of pathology at Harvard Medical School, said: "This large international effort shows the breadth of the types of research and new biological insight that are possible using whole cancer genome data". He continued: "By analyzing the largest collection of whole cancer genomes studied thus far, we created the most comprehensive catalog of mutational signatures to date, this catalog can be used to understand the mechanisms that generate mutations and drive cancer in each patient" [91]. The Pan-Cancer Project improved and developed new methods for exploring not only exome, that represent the 1 percent of the genome, but, also, the remaining 99 percent of the genome, which includes regions that regulate the activity of genes.

With the genomics, epigenomics, and transcriptomics data from over 11,000 tumors representing 33 of the most prevalent forms of cancer, the Pan-Cancer Atlas represents an exceptional chance for a comprehensive and integrated analysis to extend our current knowledge of how a normal cell achieves cancer hallmarks. The pan-cancer analysis involving multi-omics data in combination with structured bioinformatics and statistical instruments provides an effective platform to recognize common molecular signatures for the stratification of patients affected by different cancer types and uncover shared molecular pathology of different cancer types for designing tailored therapies. Investigation of the massive amount of cancer-specific data deposited in TCGA requires special bioinformatics methods to mine biologically meaningful information. Several analytic and visualization platforms have been already developed to support the rapid analysis of TCGA data. For instance, cBioPortal provides the opportunity to visualize, analyze, and download large-scale cancer genomics data sets [92]. The impulse for open data in the field of biomedical genomics is important to make data available in public repositories for improving and accelerating scientific discovery, although there are ethical and technological challenges to be overcome.

3.3. Data Integration

The need to integrate multi-omics data has led to the development of new theoretical algorithms and methods that are able to extract biologically significant information of clinical relevance.

Unsupervised data integration refers to the cluster of methods that draw an inference from of an unlabeled input dataset. Learning consists in detecting intrinsic regularities and relationships between the data, without any prior knowledge about the data itself. Examples of unsupervised techniques are matrix factorization methods, Bayesian methods, network-based methods, and multi-step analysis. CNAmet is a powerful multi-step integration tool for CNV, DNA methylation, and gene expression data [93]. The identification of genes which are synergistically regulated by methylation and CNV data, allow the understanding of biological process behind cancer progression.

Supervised methods involve the use of a dataset for which the phenotype label is known. In this way, when the system has learned a given task, it will be able to generalize, or to use the experience gained to solve problems that provide the same basic knowledge. Supervised data integration methods are built via information of available known labels from the training omics data. The most common supervised techniques are Network-based methods, Multiple Kernel Learning methods, and multi-step analysis. For example, Feature Selection Multiple Kernel Learning (FSMKL) is a method which uses the statistical score for feature selection per data type per pathway, improving the prediction accuracy for cancer detection.

Semi-supervised integration methods, lies between supervised and unsupervised methods, takes both labeled and unlabeled samples to develop learning algorithm. It is particularly useful in cases where we have a partial knowledge about the data, or if the collection and sampling phase of labeled data is too expensive to be carried out exhaustively. Semi-supervised data integration methods are usually graph-based. Graph-based semi-supervised learning (SSL) methods have been applied to cancer diagnosis and prognosis predictions.

The combination of different biological layers, with the aim to discover a coherent biological signature, remain a challenging process. Furthermore, multi-omics combinations are not necessarily capable to achieve better diagnostic results. Selecting an optimal omics combination is not trivial, since there are economic and technical constraints in the clinical setting in which such diagnostic tools are to be deployed [94]. Machine Learning Bioinformatic approaches play an important role in the design of such studies.

3.3.1. Multi-Omics Datasets

Selecting an appropriate dataset that allows for easy manipulation and data calculations could affect the performance of a computational model and reduce the main obstacles to multi-omics data analysis by improving data science applications of multiple omics datasets:

- The MultiAssayExperiment Bioconductor database [95] contains the information of different multi-omics experiments, linking features, patients, and experiments;
- The STATegRa dataset [96] has the advantage of allowing the sharing of design principles, increasing their interoperability;
- MOSim tool [97] provides methods for the generation of synthetic multi-omics datasets.

3.3.2. The Problem of Missing Data

Integrating large amounts of heterogeneous data is currently one of the major challenges in systems biology, due to the increase in available data information [98]. The problem of missing and mislabeled samples, is a common problem in large-scale multi-omics studies [99]. It is common for datasets to have missing data related to some individuals. This often happens in clinical studies, where patients can forget to fill out a form. In other cases, it is possible that the acquisition of data reveals to be too expensive, need much time to be obtained or it is difficult to measure. Missing row values for a table are difficult to

manage because most statistical methods cannot be applied directly to incomplete datasets. In recent years, several approaches have already been proposed to address missing row values [100]. The missRow package combines multiple imputation with multiple factor analysis to deal with missing data [99]. The omicsPrint method detects data linkage errors and family relations in large-scale multiple omics studies [101].

3.3.3. Exploratory Data Analysis

Understanding the nature of the data is a critical step in omics analysis [102]. For this purpose, it is possible to use exploratory data analysis (EDA) techniques which allow better assessments at a further modeling step. The main techniques for EDA include cluster analysis and dimension reduction, both widely applied to transcriptomics data analysis [103]. While *cluster analysis* consists of a set of methods for grouping objects into homogeneous classes, based on measures related to the similarity between the elements, *dimension reduction* is the process of reducing the number of variables, obtaining a set of variables called "*principal*." Both cluster analysis [104] and dimension reduction [105] are applied to cancer studies, as shown in Table 2.

Table 2. Main cluster analysis and dimension reduction package tools applied to cancer studies.

Package Tools	Description
OMICsPCA	Omics-oriented tools for PCA analysis [106]
CancerSubtypes	Contains clustering methods for the identification of cancer subpopulations from multi-omics data [107]
Omicade4	Implementation of multiple co-inertia analysis (MCIA) [108]
Biocancer	Interactive multi-omics data exploratory instrument [109]
iClusterPlus	Integrative cluster analysis combining different types of genomic data [110]

Together with dimensionality reduction and data clustering, data visualization is also an important part of EDA [2]. The combinations of these three factors make it possible to identify complex patterns, subpopulations within a dataset, and understand the variability within a phenomenon. Even if the scatter plot is the most common method for data visualization, there are other visualization tools available. Hexbins [111] can be used to explore sc-RNAseq data, while Circos diagram [112] can be used for the detailed representation of multi-omic data and their position in specific genomic regions.

Recently it is stated that mapping omics data to pathway networks could provide an opportunity to biologically contextualize the data. A network representation of multi-omics data can enhance every aspect of the multi-omics analysis because the functional level of biological description is fundamentally composed of molecular interactions [2]. The main tools for a network representation of multi-omics data are Pathview [113] and Graphite [114].

3.3.4. Machine Learning Models

In recent years, machine learning has been proved to be capable of solving many biomedical problems. These mathematical models can represent the relationships between observed variables and provide a useful description of biological phenomena. A ML tool can perform several tasks, including classification task in which the input data are divided into two or more classes and the learning system produces a model capable of assigning one class among those available to each input. These models have important biomedical applications [94], because they are capable of discriminating between health and disease, or between different diseases outcomes [2]. In a regression task instead, the output belongs to a continuous rather than discrete domain. These models provide insights into the molecular mechanisms driving physiological states, reveal interactions between different omics, and have been used in prognostic tools [115]. In this context, due to the large amounts of heterogeneous data, the removal of non-informative characteristics which

simplifies the model, increases its performance, and makes it less expensive to measure, reveals to be a crucial process [2]. Feature selection algorithm is a process which selects the variables that contribute most to the prediction, removing the irrelevant or less important features that can negatively contribute to the performance of the model. Both classification and regression ML techniques combined with feature selection algorithms have been widely used for cancer prognosis and prediction [2]. Moreover, many packages, which combine exploratory, supervised, and unsupervised tools, have been recently implemented in oncology. Table 3 provides a list of some of these new tools.

Table 3. Main packages tools implemented in oncology for machine learning.

Package Tools	Description
mixOmics	R package for the multivariate analysis of biological datasets with a specific focus on data exploration, dimension reduction, and visualization [116].
DIABLO	Package for the identification of multi-omic biomarker panels capable of discriminating between multiple phenotypic groups. It can be used to understand the molecular mechanisms that guide a disease [117].
MOFA	Package for discovering the principal sources of variation in multi-omics data sets [118].
Biosigner	Package for the identification of molecular signatures from large omics datasets in the process of developing new diagnostics [119].
omicRexposome	Package that uses high-dimensional exposome data in disease association studies, including its integration with a variety of high-performance data types [120].
OmicsLonDA	Package that identifies the time intervals in which omics functions are significantly different between groups [121].
Micrographite	Package that provides a method to integrate micro-RNA and mRNA data through their association to canonical pathways [122].
pwOmics	Package for integrating multi-omics data, adapted for the study of time series analyses [123].

3.3.5. Functional Enrichment Approaches

The interpretation of a ML model results could be a difficult task. A strategy that can provide readily interpretable results consist in mapping omic data on functional characteristics, in order to make them more informative and to associate them with a wider body of biomedical knowledge [2]. Some functional enrichment approaches are listed below:

- Over-Representation Analysis (ORA) [124];
- Gene-Set Enrichment Analysis (GSEA) [125];
- Multi-Omics Gene-Set Analysis (MOGSA) [126];
- Massive Integrative Gene Set Analysis (MIGSA) [127];
- Exploratory Data Analysis (PCA) [128];
- Divergence Analysis [129].

The first two enrichment approaches, ORA and GSEA, are feature extraction methods generally employed as dimensionality reduction methods. The output of these methods could be the starting points for more complex models such as interactions among functions. In particular, ORA method is based on a statistical evaluation of the fraction of pathway components found among a user-selected list of biological components. This input list fulfils the specific criteria (i.e., log fold change, statistical significance, and cutting-off the majority of components from the input list such as all the genes of a microarray experiment). GoMiner [130] is one of the most popular examples of ORA method. It was developed for gene-expression analysis of microarray data. It takes as input a set of over-/under-expressed genes plus the complete set list of the microarray, then it calculates

over-/under-representation for Gene Ontology categories by means of Fisher's exact test. Similarly, GSEA was developed for gene expression analysis from microarray data. The input is a list of ranked genes in accordance with their differential gene expression between two phenotypic classes. For each set of genes, an enrichment score (ES) is calculated based on a Kolmogorov–Smirnov pathway-level statistic. Multiple hypothesis testing is applied for the evaluation of ES significance. In the study of [131], the GSEA methodology was used to validate the proliferative role of growth-supporting genes involved in cancer treatment [132]. Multi-omics gene-set analysis (MOGSA) is an enrichment approach that uses multivariate analysis, which consists in integrating multiple experimental and molecular data types measured on the same data set. The method projects the features across multiple omics data sets to reduce dimensional spaces and calculates a gene set score with the most significant features. MOGSA's multi-omics approach compensates for missing information in each single data type to find sets of genes not obtainable from the analysis of single omics data. A different approach is the massive integrative gene set analysis (MIGSA). It allows to compare large collections of datasets from different sources and create independent functional associations for each omic layer. The utility of MIGSA was demonstrated in [133] by applying the multi-omics perspective method to functionally characterize the molecular subtypes of breast cancer. There are enrichment approaches, such as pathwayPCA and divergence analysis methods, which use functional aggregation as support for other data analysis studies. In pathwayPCA, exploratory data analysis is performed using statistical methodologies to analyze the functional enrichment of each omics set and aggregating them via consensus. pathwayPCA overcomes alternative methods for identifying disease-associated pathways in integrative analysis. Among various case studies, the model was applied for the identification of sex-specific pathway effects in kidney cancer for the construction of integrative models for the prediction of the patient's prognosis and for the study of heterogeneity in an ovarian cancer dataset. Divergence analysis method instead, is an enrichment approach that uses functional aggregation to classify large amounts of omics data. The omic profile is reduced to a digital representation based on that of a set of samples taken from a baseline population. The state of a subprofile that is not within the basic distribution is interpreted as "divergent." In [134] an application of the divergence analysis within the study of metabolic differences among the interpersonal heterogeneous cancer phenotypes has been described.

4. Novelty, Challenges, and Future Perspective

The computational approach plays a central role in improving our current cancer diagnostic capabilities [135]. The understanding of the cancer progression, the new therapeutic interventions, and the discovery of novel cancer biomarkers need to adopt and integrate different omics strategies at multiple levels. To achieve this aim, as suggested in the work of [136] there are five essential challenges in the omics integration workflow: (1) experimental challenges, (2) individual omics datasets, (3) integration issues, (4) data issues, and (5) biological knowledge.

1. Experimental challenges: an accurate sample preparation in a multi-omics perspective becomes one of the major experimental challenges, with the aim to achieve a universal sample collection and preparation protocol for generating multiple omics datasets.
2. Individual omics datasets: data preprocessing is also another significant challenge. This process can be performed on each omic dataset independently before merging significant results or after the production of a unique merged dataset. Moreover, the information included in each individual omic dataset requires very different standardization and scaling approaches, operating in different numerical and time scales.
3. Integration issues: data integration issues increases the difficulty of accounting for false positives in merged datasets. Additional problems include the management of rigorous approaches based on statistical models with respect to less rigorous approaches that include a biological interpretation. In comparison to a single omics study, a multi-omics approach has the benefit to allow a deeper understanding of

how the tumoral transformation is affecting the flow of information from different omics levels resulting in a bridge between cancerous genotype and the phenotype.

4. Data issues: the storage of omics data is very important for reproducibility. To this end, new omic platforms are being developed to provide essential clinical data for insights into the prognosis and diagnosis of diseases.
5. Biological knowledge: the interpretation of the outputs of computational models requires a deep knowledge of the biological system under study, in order to discriminate results that are not biologically relevant.

Despite these challenges the application of bioinformatics data integration and analysis, as well as the use of molecular modeling algorithms, allow to formulate many predictions of drug–target interactions to greatly facilitate guided drug development and guided drug resistance prevention [137]. Artificial intelligence (AI) approaches act on many aspects related to cancer therapy, including drug discovery and development and how these drugs are clinically validated and ultimately administered to patients [138]. The convergence of AI and cancer therapy has led to multiple benefits in terms of cost and time reduction. AI methods, ranging from regression models to neural networks can accelerate drug discovery, harness biomarkers to accurately match patients to clinical trials, and truly customize cancer therapy using only patients' own data.

In conclusion, the design and development of methods that integrate different multi-omic computational approaches in order to create robust and reliable models can lead to enormous advances in understanding the biology of cancer. As bioinformatics tools evolve, they must become user-friendly, interconnected, interoperable, and powerful for intensive analyses. In this context, integrated omics is not just an ensemble of computational tools, but a cohesive paradigm for deeper biological interpretation of multi-omics datasets that will potentially reveal novel details into cancer investigation. Although this field is still under development, many advances are constantly being made, with the development of new updated algorithmic approaches.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Examples of main chromosomal translocations associated to different cancers.

Translocation	Associated Diseases	Fused Genes/Proteins	
		First	Second
t(8;14)(q24;q32)	Burkitt's lymphoma	c-myc on chromosome 8	IGH@ (immunoglobulin heavy locus) on chromosome 14
		gives the fusion protein lymphocyte-proliferative ability	induces massive transcription of fusion protein
t(11;14)(q13;q32)	Mantle cell lymphoma	cyclin D1 on chromosome 11	IGH@ (immunoglobulin heavy locus) on chromosome 14
		gives fusion protein cell-proliferative ability	induces massive transcription of fusion protein
t(14;18)(q32;q21)	Follicular lymphoma (~90% of cases)	IGH@ (immunoglobulin heavy locus) on chromosome 14	Bcl-2 on chromosome 18
		induces massive transcription of fusion protein	gives fusion protein anti-apoptotic abilities

Table A1. *Cont.*

Translocation	Associated Diseases	Fused Genes/Proteins	
		First	Second
t(10;(various))(q11;(various))	Papillary thyroid cancer	RET proto-oncogene on chromosome 10	PTC (papillary thyroid cancer)—Placeholder for any of several other genes/proteins
t(2;3)(q13;p25)	Follicular thyroid cancer	PAX8—paired box gene 8 on chromosome 2	PPARγ1 (peroxisome proliferator-activated receptor γ 1) on chromosome 3
t(8;21)(q22;q22)	Acute myeloblastic leukemia with maturation	ETO on chromosome 8	AML1 on chromosome 21 found in ~7% of new cases of AML, carries a favorable prognosis and predicts good response to cytosine arabinoside therapy
t(9;22)(q34;q11) Philadelphia chromosome	Chronic myelogenous leukemia (CML), acute lymphoblastic leukemia (ALL)	Abl1 gene on chromosome 9	BCR ("breakpoint cluster region" on chromosome 22
t(15;17)(q22;q21)	Acute promyelocytic leukemia	PML protein on chromosome 15	RAR-α on chromosome 17 persistent laboratory detection of the PML-RARA transcript is strong predictor of relapse
t(12;15)(p13;q25)	Acute myeloid leukemia, congenital fibrosarcoma, secretory breast carcinoma, mammary analogue secretory carcinoma of salivary glands, cellular variant of mesoblastic nephroma	TEL on chromosome 12	TrkC receptor on chromosome 15
t(9;12)(p24;p13)	CML, ALL	JAK on chromosome 9	TEL on chromosome 12
t(12;16)(q13;p11)	Myxoid liposarcoma	DDIT3 (formerly CHOP) on chromosome 12	FUS gene on chromosome 16
t(12;21)(p12;q22)	ALL	TEL on chromosome 12	AML1 on chromosome 21
t(11;18)(q21;q21)	MALT lymphoma	BIRC3 (API-2)	MLT
t(1;11)(q42.1;q14.3)	Schizophrenia		
t(2;5)(p23;q35)	Anaplastic large cell lymphoma	ALK	NPM1
t(11;22)(q24;q11.2–12)	Ewing's sarcoma	FLI1	EWS
t(17;22)	DFSP	Collagen I on chromosome 17	Platelet derived growth factor B on chromosome 22
t(1;12)(q21;p13)	Acute myelogenous leukemia		
t(X;18)(p11.2;q11.2)	Synovial sarcoma		
t(1;19)(q10;p10)	Oligodendroglioma and oligoastrocytoma		
t(17;19)(q22;p13)	ALL		
t(7,16) (q32–34;p11) or t(11,16) (p11;p11)	Low-grade fibromyxoid sarcoma	FUS	CREB3L2 or CREB3L1

References

1. Ferlay, J.; Ervik, M.L.F.; Colombet, M.; Mery, L.; Piñeros, M. *Global Cancer Observatory: Cancer Today*; International Agency for Research on Cancer: Lyon, France, 2021.
2. De Anda-Jáuregui, G.; Hernández-Lemus, E. Computational Oncology in the Multi-Omics Era: State of the Art. *Front. Oncol.* **2020**, *10*. [CrossRef] [PubMed]
3. Hernandez-Lemus, E.; Reyes-Gopar, H.; Espinal-Enriquez, J.; Ochoa, S. The Many Faces of Gene Regulation in Cancer: A Computational Oncogenomics Outlook. *Genes* **2019**, *10*, 865. [CrossRef]
4. Long, Y.; Lu, M.; Cheng, T.; Zhan, X.; Zhan, X. Multiomics-Based Signaling Pathway Network Alterations in Human Non-functional Pituitary Adenomas. *Front. Endocrinol.* **2019**, *10*. [CrossRef]

5. Du, W.; Elemento, O. Cancer systems biology: Embracing complexity to develop better anticancer therapeutic strategies. *Oncogene* **2015**, *34*, 3215–3225. [CrossRef]
6. Chakraborty, S.; Hosen, M.I.; Ahmed, M.; Shekhar, H.U. Onco-Multi-OMICS Approach: A New Frontier in Cancer Research. *Biomed. Res. Int.* **2018**, *2018*, 9836256. [CrossRef]
7. Werner, H.M.J.; Mills, G.B.; Ram, P.T. Cancer Systems Biology: A peek into the future of patient care? *Nat. Rev. Clin. Oncol.* **2014**, *11*, 167–176. [CrossRef]
8. GuhaThakurta, D.; Sheikh, N.A.; Meagher, T.C.; Letarte, S.; Trager, J.B. Applications of systems biology in cancer immunotherapy: From target discovery to biomarkers of clinical outcome. *Expert Rev. Clin. Pharmacol.* **2013**, *6*, 387–401. [CrossRef] [PubMed]
9. Hanahan, D.; Weinberg, R.A. Hallmarks of Cancer: The Next Generation. *Cell* **2011**, *144*, 646–674. [CrossRef]
10. Davies, M.A.; Samuels, Y. Analysis of the genome to personalize therapy for melanoma. *Oncogene* **2010**, *29*, 5545–5555. [CrossRef] [PubMed]
11. Berdasco, M.; Esteller, M. Aberrant epigenetic landscape in cancer: How cellular identity goes awry. *Dev. Cell* **2010**, *19*, 698–711. [CrossRef] [PubMed]
12. Seto, M.; Honma, K.; Nakagawa, M. Diversity of genome profiles in malignant lymphoma. *Cancer Sci.* **2010**, *101*, 573–578. [CrossRef]
13. Cigudosa, J.C.; Parsa, N.Z.; Louie, D.C.; Filippa, D.A.; Jhanwar, S.C.; Johansson, B.; Mitelman, F.; Chaganti, R.S. Cytogenetic analysis of 363 consecutively ascertained diffuse large B-cell lymphomas. *Genes Chromosomes Cancer* **1999**, *25*, 123–133. [CrossRef]
14. Society, C.C. Genetic Changes and Cancer Risk. Available online: https://www.cancer.ca/en/cancer-information/cancer-101/what-is-cancer/genes-and-cancer/genetic-changes-and-cancer-risk/?region=on (accessed on 30 May 2020).
15. (NIH). The Genetics of Cancer. Available online: https://www.cancer.gov/about-cancer/causes-prevention/genetics (accessed on 30 May 2020).
16. Chakravarthi, B.V.; Nepal, S.; Varambally, S. Genomic and Epigenomic Alterations in Cancer. *Am. J. Pathol.* **2016**, *186*, 1724–1735. [CrossRef] [PubMed]
17. Lobo, I. Chromosome Abnormalities and Cancer Cytogenetics. *Nat. Educ.* **2008**, *1*, 25–44.
18. Knudson, A.G. Two genetic hits (more or less) to cancer. *Nat. Rev. Cancer* **2001**, *1*, 157–162. [CrossRef] [PubMed]
19. Kang, Z.-J.; Liu, Y.-F.; Xu, L.-Z.; Long, Z.-J.; Huang, D.; Yang, Y.; Liu, B.; Feng, J.-X.; Pan, Y.-J.; Yan, J.-S.; et al. The Philadelphia chromosome in leukemogenesis. *Chin. J. Cancer* **2016**, *35*, 48. [CrossRef]
20. Cowling, V.H.; Turner, S.A.; Cole, M.D. Burkitt's lymphoma-associated c-Myc mutations converge on a dramatically altered target gene response and implicate Nol5a/Nop56 in oncogenesis. *Oncogene* **2014**, *33*, 3519–3527. [CrossRef] [PubMed]
21. Hesson, L.B.; Hitchins, M.P.; Ward, R.L. Epimutations and cancer predisposition: Importance and mechanisms. *Curr. Opin. Genet. Dev.* **2010**, *20*, 290–298. [CrossRef]
22. Hassanpour, S.H.; Dehghani, M. Review of cancer from perspective of molecular. *J. Cancer Res. Pract.* **2017**, *4*, 127–129. [CrossRef]
23. Cheng, Y.; He, C.; Wang, M.; Ma, X.; Mo, F.; Yang, S.; Han, J.; Wei, X. Targeting epigenetic regulators for cancer therapy: Mechanisms and advances in clinical trials. *Signal. Transduct Target.* **2019**, *4*, 62. [CrossRef]
24. Seligson, D.B.; Horvath, S.; Shi, T.; Yu, H.; Tze, S.; Grunstein, M.; Kurdistani, S.K. Global histone modification patterns predict risk of prostate cancer recurrence. *Nature* **2005**, *435*, 1262–1266. [CrossRef] [PubMed]
25. Fahrner, J.A.; Eguchi, S.; Herman, J.G.; Baylin, S.B. Dependence of histone modifications and gene expression on DNA hypermethylation in cancer. *Cancer Res.* **2002**, *62*, 7213–7218. [PubMed]
26. Ben-Porath, I.; Cedar, H. Epigenetic crosstalk. *Mol. Cell* **2001**, *8*, 933–935. [CrossRef]
27. Feinberg, A.P.; Vogelstein, B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature* **1983**, *301*, 89–92. [CrossRef] [PubMed]
28. Ehrlich, M. DNA methylation in cancer: Too much, but also too little. *Oncogene* **2002**, *21*, 5400–5413. [CrossRef] [PubMed]
29. Laird, P.W. Cancer epigenetics. *Hum. Mol. Genet.* **2005**, *14*, R65–R76. [CrossRef]
30. Jones, P.A.; Baylin, S.B. The epigenomics of cancer. *Cell* **2007**, *128*, 683–692. [CrossRef]
31. Tost, J. DNA Methylation: An Introduction to the Biology and the Disease-Associated Changes of a Promising Biomarker. *Mol. Biotechnol.* **2010**, *44*, 71–81. [CrossRef]
32. Balmain, A.; Gray, J.; Ponder, B. The genetics and genomics of cancer. *Nat. Genet.* **2003**, *33*, 238–244. [CrossRef]
33. Gelli, E.; Pinto, A.M.; Somma, S.; Imperatore, V.; Cannone, M.G.; Hadjistilianou, T.; De Francesco, S.; Galimberti, D.; Curro, A.; Bruttini, M.; et al. Evidence of predisposing epimutation in retinoblastoma. *Hum. Mutat* **2019**, *40*, 201–206. [CrossRef]
34. Costello, J.F.; Frühwald, M.C.; Smiraglia, D.J.; Rush, L.J.; Robertson, G.P.; Gao, X.; Wright, F.A.; Feramisco, J.D.; Peltomäki, P.; Lang, J.C.; et al. Aberrant CpG-island methylation has non-random and tumour-type–specific patterns. *Nat. Genet.* **2000**, *24*, 132–138. [CrossRef]
35. Duruisseaux, M.; Martínez-Cardús, A.; Calleja-Cervantes, M.E.; Moran, S.; Castro de Moura, M.; Davalos, V.; Piñeyro, D.; Sanchez-Cespedes, M.; Girard, N.; Brevet, M.; et al. Epigenetic prediction of response to anti-PD-1 treatment in non-small-cell lung cancer: A multicentre, retrospective analysis. *Lancet Respir. Med.* **2018**, *6*, 771–781. [CrossRef]
36. Duruisseaux, M.; Esteller, M. Lung cancer epigenetics: From knowledge to applications. *Semin. Cancer Biol.* **2018**, *51*, 116–128. [CrossRef] [PubMed]
37. Kim, J.Y.; Choi, J.K.; Jung, H. Genome-wide methylation patterns predict clinical benefit of immunotherapy in lung cancer. *Clin. Epigenetics* **2020**, *12*, 119. [CrossRef] [PubMed]

38. Goltz, D.; Gevensleben, H.; Vogt, T.J.; Dietrich, J.; Golletz, C.; Bootz, F.; Kristiansen, G.; Landsberg, J.; Dietrich, D. CTLA4 methylation predicts response to anti-PD-1 and anti-CTLA-4 immunotherapy in melanoma patients. *JCI Insight* **2018**, *3*. [CrossRef]
39. Sigin, V.O.; Kalinkin, A.I.; Kuznetsova, E.B.; Simonova, O.A.; Chesnokova, G.G.; Litviakov, N.V.; Slonimskaya, E.M.; Tsyganov, M.M.; Ibragimova, M.K.; Volodin, I.V.; et al. DNA methylation markers panel can improve prediction of response to neoadjuvant chemotherapy in luminal B breast cancer. *Sci. Rep.* **2020**, *10*, 9239. [CrossRef]
40. Yu, D.-H.; Waterland, R.A.; Zhang, P.; Schady, D.; Chen, M.-H.; Guan, Y.; Gadkari, M.; Shen, L. Targeted p16(Ink4a) epimutation causes tumorigenesis and reduces survival in mice. *J. Clin. Investig.* **2014**, *124*, 3708–3712. [CrossRef]
41. Majumder, A.; Syed, K.M.; Mukherjee, A.; Lankadasari, M.B.; Azeez, J.M.; Sreeja, S.; Harikumar, K.B.; Pillai, M.R.; Dutta, D. Enhanced expression of histone chaperone APLF associate with breast cancer. *Mol. Cancer* **2018**, *17*, 76. [CrossRef] [PubMed]
42. Sharma, S.; Kelly, T.K.; Jones, P.A. Epigenetics in cancer. *Carcinogenesis* **2009**, *31*, 27–36. [CrossRef]
43. Fraga, M.F.; Ballestar, E.; Villar-Garea, A.; Boix-Chornet, M.; Espada, J.; Schotta, G.; Bonaldi, T.; Haydon, C.; Ropero, S.; Petrie, K.; et al. Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. *Nat. Genet.* **2005**, *37*, 391–400. [CrossRef] [PubMed]
44. Fullgrabe, J.; Kavanagh, E.; Joseph, B. Histone onco-modifications. *Oncogene* **2011**, *30*, 3391–3403. [CrossRef]
45. Bannister, A.J.; Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res.* **2011**, *21*, 381–395. [CrossRef]
46. Rossetto, D.; Avvakumov, N.; Cote, J. Histone phosphorylation: A chromatin modification involved in diverse nuclear events. *Epigenetics* **2012**, *7*, 1098–1108. [CrossRef]
47. Nam, S.; Li, M.; Choi, K.; Balch, C.; Kim, S.; Nephew, K.P. MicroRNA and mRNA integrated analysis (MMIA): A web tool for examining biological functions of microRNA expression. *Nucleic Acids Res.* **2009**, *37*, W356–W362. [CrossRef]
48. Musilova, K.; Mraz, M. MicroRNAs in B-cell lymphomas: How a complex biology gets more complex. *Leukemia* **2015**, *29*, 1004–1017. [CrossRef]
49. Erdogan, B.; Facey, C.; Qualtieri, J.; Tedesco, J.; Rinker, E.; Isett, R.B.; Tobias, J.; Baldwin, D.A.; Thompson, J.E.; Carroll, M.; et al. Diagnostic microRNAs in myelodysplastic syndrome. *Exp. Hematol.* **2011**, *39*, 915–926.e2. [CrossRef] [PubMed]
50. Goecks, J.; Nekrutenko, A.; Taylor, J.; Galaxy, T. Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **2010**, *11*, R86. [CrossRef]
51. Koster, J.; Rahmann, S. Snakemake—A scalable bioinformatics workflow engine. *Bioinformatics* **2012**, *28*, 2520–2522. [CrossRef] [PubMed]
52. Di Tommaso, P.; Chatzou, M.; Floden, E.W.; Barja, P.P.; Palumbo, E.; Notredame, C. Nextflow enables reproducible computational workflows. *Nat. Biotechnol* **2017**, *35*, 316–319. [CrossRef]
53. Peter, A.; Michael, R.C.; Nebojša, T.; Brad, C.; John, C.; Michael, H.; Andrey, K.; Dan, L.; Hervé, M.; Maya, N.; et al. Common Workflow Language, v1.0. 2016. Available online: https://escholarship.org/uc/item/25z538jj (accessed on 30 May 2020).
54. Huber, W.; Carey, V.J.; Gentleman, R.; Anders, S.; Carlson, M.; Carvalho, B.S.; Bravo, H.C.; Davis, S.; Gatto, L.; Girke, T.; et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **2015**, *12*, 115–121. [CrossRef] [PubMed]
55. Cock, P.J.; Antao, T.; Chang, J.T.; Chapman, B.A.; Cox, C.J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; et al. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423. [CrossRef]
56. Kandoth, C.; McLellan, M.D.; Vandin, F.; Ye, K.; Niu, B.; Lu, C.; Xie, M.; Zhang, Q.; McMichael, J.F.; Wyczalkowski, M.A.; et al. Mutational landscape and significance across 12 major cancer types. *Nature* **2013**, *502*, 333–339. [CrossRef] [PubMed]
57. Lawrence, M.S.; Stojanov, P.; Polak, P.; Kryukov, G.V.; Cibulskis, K.; Sivachenko, A.; Carter, S.L.; Stewart, C.; Mermel, C.H.; Roberts, S.A.; et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **2013**, *499*, 214–218. [CrossRef]
58. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef]
59. Langmead, B.; Trapnell, C.; Pop, M.; Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **2009**, *10*, R25. [CrossRef] [PubMed]
60. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [CrossRef] [PubMed]
61. Zaharia, M.; Bolosky, W.J.; Curtis, K.; Fox, A.; Patterson, D.; Shenker, S.; Stoica, I.; Karp, R.M.; Sittler, T. Faster and More Accurate Sequence Alignment with SNAP. *arXiv* **2011**, arXiv:1111.5572.
62. Piunti, A.; Shilatifard, A. Epigenetic balance of gene expression by Polycomb and COMPASS families. *Science* **2016**, *352*, aad9780. [CrossRef] [PubMed]
63. Maksimovic, J.; Phipson, B.; Oshlack, A. A cross-package Bioconductor workflow for analysing methylation array data. *F1000Res* **2016**, *5*, 1281. [CrossRef]
64. Zang, C.; Schones, D.E.; Zeng, C.; Cui, K.; Zhao, K.; Peng, W. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* **2009**, *25*, 1952–1958. [CrossRef]
65. Feng, X.; Grossman, R.; Stein, L. PeakRanger: A cloud-enabled peak caller for ChIP-seq data. *BMC Bioinform.* **2011**, *12*, 139. [CrossRef]
66. Guo, Y.; Mahony, S.; Gifford, D.K. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput. Biol.* **2012**, *8*, e1002638. [CrossRef]

67. Harmanci, A.; Rozowsky, J.; Gerstein, M. MUSIC: Identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework. *Genome Biol.* **2014**, *15*, 474. [CrossRef] [PubMed]
68. Zhang, Y.; Lin, Y.H.; Johnson, T.D.; Rozek, L.S.; Sartor, M.A. PePr: A peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-Seq data. *Bioinformatics* **2014**, *30*, 2568–2575. [CrossRef] [PubMed]
69. Kumar, V.; Muratani, M.; Rayan, N.A.; Kraus, P.; Lufkin, T.; Ng, H.H.; Prabhakar, S. Uniform, optimal signal processing of mapped deep-sequencing data. *Nat. Biotechnol.* **2013**, *31*, 615–622. [CrossRef] [PubMed]
70. Zhang, Y.; Liu, T.; Meyer, C.A.; Eeckhoute, J.; Johnson, D.S.; Bernstein, B.E.; Nusbaum, C.; Myers, R.M.; Brown, M.; Li, W.; et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **2008**, *9*, R137. [CrossRef]
71. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **2013**, *29*, 15–21. [CrossRef] [PubMed]
72. Bray, N.L.; Pimentel, H.; Melsted, P.; Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **2016**, *34*, 525–527. [CrossRef]
73. Patro, R.; Duggal, G.; Love, M.I.; Irizarry, R.A.; Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **2017**, *14*, 417–419. [CrossRef] [PubMed]
74. Cancer Genome Atlas Research Network; Weinstein, J.N.; Collisson, E.A.; Mills, G.B.; Shaw, K.R.; Ozenberger, B.A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J.M. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **2013**, *45*, 1113–1120. [CrossRef]
75. Iwamoto, N.; Shimada, T. Recent advances in mass spectrometry-based approaches for proteomics and biologics: Great contribution for developing therapeutic antibodies. *Pharmacol. Ther.* **2018**, *185*, 147–154. [CrossRef]
76. Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26*, 1367–1372. [CrossRef]
77. Tyanova, S.; Temu, T.; Sinitcyn, P.; Carlson, A.; Hein, M.Y.; Geiger, T.; Mann, M.; Cox, J. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **2016**, *13*, 731–740. [CrossRef]
78. Kim, M.S.; Pinto, S.M.; Getnet, D.; Nirujogi, R.S.; Manda, S.S.; Chaerkady, R.; Madugundu, A.K.; Kelkar, D.S.; Isserlin, R.; Jain, S.; et al. A draft map of the human proteome. *Nature* **2014**, *509*, 575–581. [CrossRef]
79. Wilhelm, M.; Schlegl, J.; Hahne, H.; Gholami, A.M.; Lieberenz, M.; Savitski, M.M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; et al. Mass-spectrometry-based draft of the human proteome. *Nature* **2014**, *509*, 582–587. [CrossRef]
80. Shruthi, B.S.; Vinodhkumar, P.; Selvamani. Proteomics: A new perspective for cancer. *Adv. Biomed. Res.* **2016**, *5*, 67. [CrossRef] [PubMed]
81. Yakkioui, Y.; Temel, Y.; Chevet, E.; Negroni, L. Integrated and Quantitative Proteomics of Human Tumors. *Methods Enzym.* **2017**, *586*, 229–246. [CrossRef]
82. Vazquez, A.; Kamphorst, J.J.; Markert, E.K.; Schug, Z.T.; Tardito, S.; Gottlieb, E. Cancer metabolism at a glance. *J. Cell Sci.* **2016**, *129*, 3367–3373. [CrossRef] [PubMed]
83. Armitage, E.G.; Ciborowski, M. Applications of Metabolomics in Cancer Studies. *Adv. Exp. Med. Biol* **2017**, *965*, 209–234. [CrossRef] [PubMed]
84. Yang, K.; Han, X. Lipidomics: Techniques, Applications, and Outcomes Related to Biomedical Sciences. *Trends Biochem. Sci.* **2016**, *41*, 954–969. [CrossRef] [PubMed]
85. Perrotti, F.; Rosa, C.; Cicalini, I.; Sacchetta, P.; Del Boccio, P.; Genovesi, D.; Pieragostino, D. Advances in Lipidomics for Cancer Biomarkers Discovery. *Int. J. Mol. Sci.* **2016**, *17*. [CrossRef] [PubMed]
86. Aggio, R.; Villas-Boas, S.G.; Ruggiero, K. Metab: An R package for high-throughput analysis of metabolomics data generated by GC-MS. *Bioinformatics* **2011**, *27*, 2316–2318. [CrossRef] [PubMed]
87. Stanstrup, J.; Broeckling, C.D.; Helmus, R.; Hoffmann, N.; Mathe, E.; Naake, T.; Nicolotti, L.; Peters, K.; Rainer, J.; Salek, R.M.; et al. The metaRbolomics Toolbox in Bioconductor and beyond. *Metabolites* **2019**, *9*. [CrossRef] [PubMed]
88. Mohamed, A.; Molendijk, J.; Hill, M.M. lipidr: A Software Tool for Data Mining and Analysis of Lipidomics Datasets. *J. Proteome Res.* **2020**. [CrossRef] [PubMed]
89. Jansen, P.; van den Berg, L.; van Overveld, P.; Boiten, J.W. Research Data Stewardship for Healthcare Professionals. *Fundam. Clin. Data Sci.* **2019**, 37–53. [CrossRef]
90. Grossman, R.L.; Heath, A.P.; Ferretti, V.; Varmus, H.E.; Lowy, D.R.; Kibbe, W.A.; Staudt, L.M. Toward a Shared Vision for Cancer Genomic Data. *N. Engl. J. Med.* **2016**, *375*, 1109–1112. [CrossRef] [PubMed]
91. McPherson, S. Collaboration Generates Most Complete Cancer Genome Map. 2020. Available online: https://news.harvard.edu/gazette/story/2020/02/big-step-toward-identifying-all-cancer-causing-genetic-mutations/ (accessed on 30 May 2020).
92. Cerami, E.; Gao, J.; Dogrusoz, U.; Gross, B.E.; Sumer, S.O.; Aksoy, B.A.; Jacobsen, A.; Byrne, C.J.; Heuer, M.L.; Larsson, E.; et al. The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2012**, *2*, 401–404. [CrossRef] [PubMed]
93. Louhimo, R.; Hautaniemi, S. CNAmet: An R package for integrating copy number, methylation and expression data. *Bioinformatics* **2011**, *27*, 887–888. [CrossRef]
94. Yoo, B.C.; Kim, K.H.; Woo, S.M.; Myung, J.K. Clinical multi-omics strategies for the effective cancer management. *J. Proteom.* **2018**, *188*, 97–106. [CrossRef]

95. Ramos, M.; Schiffer, L.; Re, A.; Azhar, R.; Basunia, A.; Rodriguez, C.; Chan, T.; Chapman, P.; Davis, S.R.; Gomez-Cabrero, D.; et al. Software for the Integration of Multiomics Experiments in Bioconductor. *Cancer Res.* **2017**, *77*, e39–e42. [CrossRef] [PubMed]
96. Consortia. STATegRa: Classes and methods for multi-omics data integration. *R Package Version 1.24.0*. 2020. Available online: https://www.bioconductor.org/packages/release/bioc/html/STATegRa.html (accessed on 30 May 2020).
97. Martínez-Mira, C.; Conesa, A.; Tarazona, S. MOSim: Multi-Omics Simulation in R. *bioRxiv* **2018**, 421834. [CrossRef]
98. Gomez-Cabrero, D.; Abugessaisa, I.; Maier, D.; Teschendorff, A.; Merkenschlager, M.; Gisel, A.; Ballestar, E.; Bongcam-Rudloff, E.; Conesa, A.; Tegner, J. Data integration in the era of omics: Current and future challenges. *BMC Syst. Biol.* **2014**, *8* (Suppl. 2), 1. [CrossRef] [PubMed]
99. Voillet, V.; Besse, P.; Liaubet, L.; San Cristobal, M.; Gonzalez, I. Handling missing rows in multi-omics data integration: Multiple imputation in multiple factor analysis framework. *BMC Bioinform.* **2016**, *17*, 402. [CrossRef]
100. Pigott, T.D. A Review of Methods for Missing Data. *Educ. Res. Eval.* **2001**, *7*, 353–383. [CrossRef]
101. Van Iterson, M.; Cats, D.; Hop, P.; BIOS Consortium; Heijmans, B.T. omicsPrint: Detection of data linkage errors in multiple omics studies. *Bioinformatics* **2018**, *34*, 2142–2143. [CrossRef]
102. Meng, C.; Zeleznik, O.A.; Thallinger, G.G.; Kuster, B.; Gholami, A.M.; Culhane, A.C. Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief. Bioinform* **2016**, *17*, 628–641. [CrossRef]
103. Brazma, A.; Culhane, A.C. Algorithms for gene expression analysis. In *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2005. [CrossRef]
104. Streicher, K.L.; Zhu, W.; Lehmann, K.P.; Georgantas, R.W.; Morehouse, C.A.; Brohawn, P.; Carrasco, R.A.; Xiao, Z.; Tice, D.A.; Higgs, B.W.; et al. A novel oncogenic role for the miRNA-506-514 cluster in initiating melanocyte transformation and promoting melanoma growth. *Oncogene* **2012**, *31*, 1558–1570. [CrossRef]
105. Biton, A.; Bernard-Pierrot, I.; Lou, Y.; Krucker, C.; Chapeaublanc, E.; Rubio-Pérez, C.; López-Bigas, N.; Kamoun, A.; Neuzillet, Y.; Gestraud, P.; et al. Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. *Cell Rep.* **2014**, *9*, 1235–1245. [CrossRef] [PubMed]
106. Das, S.; Tripathy, D.S. OMICsPCA: An. R Package for Quantitative Integration and Analysis of Multiple Omics Assays from Heterogeneous Samples. R Package Version 1.5.0. 2019. Available online: https://www.bioconductor.org/packages/release/bioc/html/OMICsPCA.html (accessed on 30 May 2020).
107. Xu, T.; Le, T.D.; Liu, L.; Su, N.; Wang, R.; Sun, B.; Colaprico, A.; Bontempi, G.; Li, J. CancerSubtypes: An R/Bioconductor package for molecular cancer subtype identification, validation and visualization. *Bioinformatics* **2017**, *33*, 3131–3133. [CrossRef] [PubMed]
108. Meng, C.; Kuster, B.; Culhane, A.C.; Gholami, A.M. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinform.* **2014**, *15*, 162. [CrossRef]
109. Mezhoud, K. bioCancer: Interactive Multi-Omics Cancers Data Visualization and Analysis. R package version 1.16.0. 2020. Available online: http://kmezhoud.github.io/bioCancer (accessed on 30 May 2020).
110. Shen, R.; Olshen, A.B.; Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **2009**, *25*, 2906–2912. [CrossRef] [PubMed]
111. Freytag, S. schex: Hexbin plots for single cell omics data. R package version 1.2.0. Available online: https://github.com/SaskiaFreytag/schex (accessed on 30 May 2020).
112. Krzywinski, M.; Schein, J.; Birol, I.; Connors, J.; Gascoyne, R.; Horsman, D.; Jones, S.J.; Marra, M.A. Circos: An information aesthetic for comparative genomics. *Genome Res.* **2009**, *19*, 1639–1645. [CrossRef] [PubMed]
113. Luo, W.; Brouwer, C. Pathview: An R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* **2013**, *29*, 1830–1831. [CrossRef] [PubMed]
114. Sales, G.; Calura, E.; Cavalieri, D.; Romualdi, C. Graphite—A Bioconductor package to convert pathway topology to gene network. *BMC Bioinform.* **2012**, *13*, 20. [CrossRef]
115. Syed-Abdul, S.; Iqbal, U.; Jack Li, Y.C. Predictive Analytics through Machine Learning in the clinical settings. *Comput Methods Programs Biomed.* **2017**, *144*, A1–A2. [CrossRef]
116. Rohart, F.; Gautier, B.; Singh, A.; Le Cao, K.A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* **2017**, *13*, e1005752. [CrossRef]
117. Singh, A.; Shannon, C.P.; Gautier, B.; Rohart, F.; Vacher, M.; Tebbutt, S.J.; Le Cao, K.A. DIABLO: An integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* **2019**, *35*, 3055–3062. [CrossRef]
118. Argelaguet, R.; Velten, B.; Arnol, D.; Dietrich, S.; Zenz, T.; Marioni, J.C.; Buettner, F.; Huber, W.; Stegle, O. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **2018**, *14*, e8124. [CrossRef]
119. Rinaudo, P.; Boudah, S.; Junot, C.; Thevenot, E.A. biosigner: A New Method for the Discovery of Significant Molecular Signatures from Omics Data. *Front. Mol. Biosci.* **2016**, *3*, 26. [CrossRef]
120. Hernandez-Ferrer, C.; Wellenius, G.A.; Tamayo, I.; Basagana, X.; Sunyer, J.; Vrijheid, M.; Gonzalez, J.R. Comprehensive study of the exposome and omic data using rexposome Bioconductor Packages. *Bioinformatics* **2019**, *35*, 5344–5345. [CrossRef]
121. Metwally, A.A.; Zhang, T.; Snyder, M. OmicsLonDA: Omics Longitudinal Differential Analysis. R package version 1.4.0. 2020. Available online: https://github.com/aametwally/OmicsLonDA (accessed on 30 May 2020).
122. Calura, E.; Martini, P.; Sales, G.; Beltrame, L.; Chiorino, G.; D'Incalci, M.; Marchini, S.; Romualdi, C. Wiring miRNAs to pathways: A topological approach to integrate miRNA and mRNA expression profiles. *Nucleic Acids Res.* **2014**, *42*, e96. [CrossRef]

123. Wachter, A.; Beissbarth, T. pwOmics: An R package for pathway-based integration of time-series omics data using public database knowledge. *Bioinformatics* **2015**, *31*, 3072–3074. [CrossRef] [PubMed]
124. Boyle, E.I.; Weng, S.; Gollub, J.; Jin, H.; Botstein, D.; Cherry, J.M.; Sherlock, G. GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **2004**, *20*, 3710–3715. [CrossRef]
125. Subramanian, A.; Tamayo, P.; Mootha, V.K.; Mukherjee, S.; Ebert, B.L.; Gillette, M.A.; Paulovich, A.; Pomeroy, S.L.; Golub, T.R.; Lander, E.S.; et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15545–15550. [CrossRef] [PubMed]
126. Meng, C.; Basunia, A.; Peters, B.; Gholami, A.M.; Kuster, B.; Culhane, A.C. MOGSA: Integrative Single Sample Gene-set Analysis of Multiple Omics Data. *Mol. Cell Proteom.* **2019**, *18*, S153–S168. [CrossRef] [PubMed]
127. Rodriguez, J.C.; Merino, G.A.; Llera, A.S.; Fernandez, E.A. Massive integrative gene set analysis enables functional characterization of breast cancer subtypes. *J. Biomed. Inf.* **2019**, *93*, 103157. [CrossRef] [PubMed]
128. Odom, G.J.; Ban, Y.; Colaprico, A.; Liu, L.; Silva, T.C.; Sun, X.; Pico, A.R.; Zhang, B.; Wang, L.; Chen, X. PathwayPCA: An R/Bioconductor Package for Pathway Based Integrative Analysis of Multi-Omics Data. *Proteomics* **2020**, e1900409. [CrossRef]
129. Dinalankara, W.; Ke, Q.; Xu, Y.; Ji, L.; Pagane, N.; Lien, A.; Matam, T.; Fertig, E.J.; Price, N.D.; Younes, L.; et al. Digitizing omics profiles by divergence from a baseline. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 4545–4552. [CrossRef]
130. Zeeberg, B.R.; Feng, W.; Wang, G.; Wang, M.D.; Fojo, A.T.; Sunshine, M.; Narasimhan, S.; Kane, D.W.; Reinhold, W.C.; Lababidi, S.; et al. GoMiner: A resource for biological interpretation of genomic and proteomic data. *Genome Biol.* **2003**, *4*, R28. [CrossRef] [PubMed]
131. Folger, O.; Jerby, L.; Frezza, C.; Gottlieb, E.; Ruppin, E.; Shlomi, T. Predicting selective drug targets in cancer through metabolic networks. *Mol. Syst. Biol.* **2011**, *7*, 501. [CrossRef]
132. García-Campos, M.A.; Espinal-Enríquez, J.; Hernández-Lemus, E. Pathway Analysis: State of the Art. *Front. Physiol.* **2015**, *6*, 383. [CrossRef]
133. Rocha, D.; García, I.A.; González Montoro, A.; Llera, A.; Prato, L.; Girotti, M.R.; Soria, G.; Fernández, E.A. Pan-Cancer Molecular Patterns and Biological Implications Associated with a Tumor-Specific Molecular Signature. *Cells* **2020**, *10*, 45. [CrossRef]
134. Baloni, P.; Dinalankara, W.; Earls, J.C.; Knijnenburg, T.A.; Geman, D.; Marchionni, L.; Price, N.D. Identifying Personalized Metabolic Signatures in Breast Cancer. *Metabolites* **2020**, *11*, 20. [CrossRef]
135. Parkinson, D.R.; Johnson, B.E.; Sledge, G.W. Making personalized cancer medicine a reality: Challenges and opportunities in the development of biomarkers and companion diagnostics. *Clin. Cancer Res.* **2012**, *18*, 619–624. [CrossRef]
136. Misra, B.B.; Langefeld, C.D.; Olivier, M.; Cox, L.A. Integrated Omics: Tools, Advances, and Future Approaches. *J. Mol. Endocrinol.* **2018**. [CrossRef]
137. Tolios, A.; De Las Rivas, J.; Hovig, E.; Trouillas, P.; Scorilas, A.; Mohr, T. Computational approaches in cancer multidrug resistance research: Identification of potential biomarkers, drug targets and drug-target interactions. *Drug Resist. Updat.* **2020**, *48*, 100662. [CrossRef]
138. Agrawal, P. Artificial Intelligence in Drug Discovery and Development. *J. Pharmacovigil.* **2018**, *6*, 80–93. [CrossRef]

Journal of Bioinformatics and Computational Biology
(2020) 2050028 (17 pages)

DOI: 10.1142/S0219720020500286

A deep attention network for predicting amino acid signals in the formation of α-helices

A. Visibelli[*,§], P. Bongini[†,‡], A. Rossi[†,‡], N. Niccolai[*] and M. Bianchini[†]

[*]*Department of Biotechnology, Chemistry and Pharmacy*
University of Siena, 53100, Siena, Italy

[†]*Department of Information Engineering and Mathematics*
University of Siena, 53100, Siena, Italy

[‡]*Department of Information Engineering*
University of Florence, 50139, Florence, Italy
[§]*anna.visibelli@student.unisi.it*

Received 5 June 2020
Accepted 10 June 2020
Published 6 August 2020

The secondary and tertiary structure of a protein has a primary role in determining its function. Even though many folding prediction algorithms have been developed in the past decades — mainly based on the assumption that folding instructions are encoded within the protein sequence — experimental techniques remain the most reliable to establish protein structures. In this paper, we searched for signals related to the formation of α-helices. We carried out a statistical analysis on a large dataset of experimentally characterized secondary structure elements to find over- or under-occurrences of specific amino acids defining the boundaries of helical moieties. To validate our hypothesis, we trained various Machine Learning models, each equipped with an attention mechanism, to predict the occurrence of α-helices. The attention mechanism allows to interpret the model's decision, weighing the importance the predictor gives to each part of the input. The experimental results show that different models focus on the same subsequences, which can be seen as codes driving the secondary structure formation.

Keywords: α-Helices; proteins; machine learning; attention mechanism.

1. Introduction

Knowledge of the secondary and tertiary structure of a protein is fundamental in understanding its function and its structure-function relationships. It is now well established that protein structures are mainly determined by their amino acid sequences.[1] Protein folding prediction techniques have been based on this hypothesis for decades, but they have not yet reached an accuracy comparable to that of the traditional methods, like NMR spectroscopy, X-ray crystallography, or cryo-electron tomography. These processes, though, are expensive and overly time consuming.

The rapid progresses in genomics have led to the discovery of millions of protein sequences, less than 0.2% of which were resolved with the traditional methods. Advanced computational approaches for the prediction of secondary and tertiary structures would solve this problem allowing to process large amounts of sequences in a reasonable time. Nevertheless, although gradual developments have been made in the prediction methods based on local information on the amino acid sequence, their results have a lower quality compared to those of the experimental techniques, which can also take into account the protein evolutionary information and constitute the ground–truth for the performance evaluation of predictive methods.

Efficiently predicting the occurrence of secondary structure motifs in proteins can represent a solid basis towards accurate predictions of 3D native structures and, actually, the search for a performing method for secondary structure prediction, over the last five decades, produced a wide variety of different statistical approaches. One of the first solutions was SIMPA,[2] a nearest neighbor classifier based on a sequence similarity matrix. In BSPSS,[3] a probabilistic model was developed, formulating the secondary structure prediction as a general Bayesian inference problem. Another example of a static inference approach is SOPM,[4] which makes use of a sequence similarity score in order to predict secondary structures. Finally, IPSSP[5] extends the hidden semi-Markov model (HSMM) used in BSPSS, taking residue dependencies into account, in order to make a better prediction.

Recently, also neural networks have been applied to the secondary structure prediction task, showing promising results. In Ref. 6, a hybrid modular architecture was proposed based on a combination of BSPSS with a neural network. Differently, an ensemble model, composed by a Multi–Class Support Vector Machine (M–SVM) and a neural network, was described in Ref. 7. PSIpred[8] is a server, hosted at UCI, which implements a simple and accurate secondary structure prediction method, incorporating two feed-forward neural networks. For each amino acid in the sequence, the first neural network is fed with a window of 15 amino acids. A second neural network is then used to refine the structure predicted by the first network. Since their first appearance (almost 20 years ago) Long Short-Term Memories (LSTMs), a special kind of Recurrent Neural Network, have produced the state-of-the-art results in many Machine Learning (ML) tasks involving sequential data, such as in speech recognition.[9] and in machine translation.[10] LSTMs are tailored to process sequences, since they can focus on both local and global contextual features. LSTMs have already been applied to predict secondary structures in proteins,[11,12] obtaining very good results.

In this paper, we faced the problem of finding helical moieties in proteins from a somewhat different perspective, searching for small conserved amino acid signals, that delimitate α-helices — or, in other words, define their boundaries — and are used by Nature to drive their formation. To this aim, we first carried out a statistical analysis and then implemented and compared three different ML approaches to identify such signals. In particular, in Sec. 2, the concentration of each amino acid is calculated considering three positions outside and four positions inside known

helices, as reported in CATH database.[13] Actually, the obtained results show how some amino acids are both hyper- or under-represented in biologically explainable ways. In Sec. 3, three ML models, specifically built for the task of amino acid signal identification, are described. These ML architectures are trained to discriminate sequences which contain an α-helix from those which do not. Each of them is equipped with an attention module, which measures the importance given by the model to each feature in each sequence position, allowing an interpretation of its behavior. The experimental results reported in Sec. 5 show how all the models focus on the most important information, suggesting that the amino acids located at the sequence boundaries are fundamental in determining the occurrence of α-helices. Finally, Sec. 7 collects some conclusions and suggests future perspectives.

2. Amino Acid Concentration at the Helix Boundaries

The most common type of secondary structure in proteins is the α-helix. Linus Pauling was the first to predict the existence of α-helices,[14] which was confirmed a few years later by X-ray crystallographic determinations of myoglobin and haemoglobin structures, respectively resolved by Max Perutz[15] and John Kendrew.[16]

The idea that Nature interprets some signals to locally fold a protein into an α-helix is not completely new, even if, in the past, it was declined based on different assumptions, for instance searching for combinations of unusual codons.[17] Our point of view, instead, is that of searching for abnormal concentrations of some amino acids, in particular positions located at the helix boundaries.

In the following, we first describe how the helix dataset was gathered, the preprocessing phase on the collected data and their statistical analysis, together with a biological interpretation of the obtained results.

2.1. *Dataset collection*

A set of proteins, divided in three main classes (mainly-α, mainly-β, and α-β proteins) was downloaded from the CATH database (`http://www.cathdb.info/browse/tree`). All the sequences were then analyzed in order to find as many helices as possible. Sequences and secondary structure information were extracted from the PDB entries, using the Kabsch and Sander DSSP algorithm,[18] which exploits backbone dihedral angles and hydrogen bonds, to assign each amino acid to a secondary structure.

2.2. *Preprocessing*

In an α-helix, each turn is composed by an average number of 3.6 residues. Therefore, to ensure that each helix includes at least two turns, helices shorter than eight residues were discarded. Since signals that trigger the helix formation can also be located outside the helix sequence itself, we analyzed the complete sequence, taking into account also two or three amino acids before and after each helix. We labeled the

Table 1. Helices obtained from DSSP files.

	Alpha domains	Beta domains	Alpha Beta domains	Tot
Total helices	11592	1864	25776	39347
>8,(1)	8343	631	16846	25873
>8,(2)	3818	328	8945	13094
>8,(3)	1220	121	2906	4249

sequences with two or three external residues with the suffix 2H or 3H, respectively. Table 1 shows the distribution of sequences in each category.

In Table 1, the first row shows the total number of helices in each class, without taking into account their length. The other rows convey the number of helices longer than eight residues, which also include one (1), two (2) and three (3) amino acids before and after the helix. Due to the small number of helices in the mainly-β domains, this class was not used in the following analysis. Moreover, to obtain a non-redundant dataset, the sequences were scanned with CATH S100, with the purpose of removing sequences with a 100% identity. All the α-helices were then grouped according to their size, as shown in Fig. 1.

The lengths of the collected α-helices vary between 8 and 40 residues, with a mean helical length of approximately 10 residues. There is a gradual decrease in the helix population as the helical length increases beyond 14 residues. Moreover, helices longer than 40 residues are rarely found in proteins, whereas the longest helix in our dataset is composed by 60 residues.

2.3. *Statistical analysis*

In order to establish if the biological signals actually exist which can delimit the α-helix, we first evaluate the *residue propensity value* for each amino acid in the

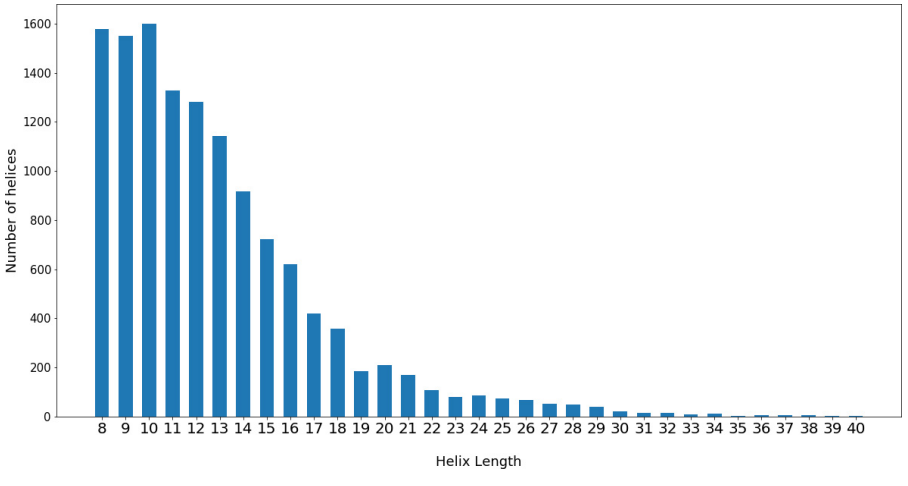


Fig. 1. Length distribution of 2H helices.

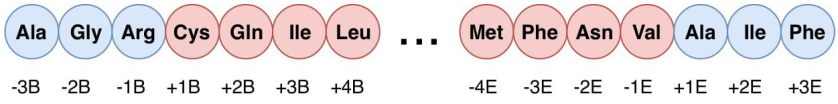


Fig. 2. Position coding along helix sequences.

positions shown in Fig. 2, namely three positions outside and four positions inside the helix.

The residue propensity value (ϵ_a) describes in which percentage a particular amino acid occupies a given position, and can be calculated based on the following equation:

$$\epsilon_a = \frac{a_s}{n_s} \times 100,$$

where a_s and n_s are the number of residues of type a in position s and the total number of residues, respectively. Propensity values can then be compared with the percentage concentration of each amino acid in the entire dataset. The results obtained are reported in Tables 2 and 3. Colored values represent residue frequencies that deviate more than 4% from the standard concentration value, being negative and positive deviations highlighted in blue and red. The second column of both tables shows the frequency of each amino acid within the dataset.

Tables 2 and 3 report very similar propensity values. The case of Glycine is particularly interesting as it has a very high frequency of occurrence in the terminal

Table 2. Residue propensity values in 2H helices.

Code	%	−2B	−1B	+1B	+2B	+3B	+4B	−4E	−3E	−2E	−1E	+1E	+2E
Ala	8,7	8,1	4,8	9,	9,6	10,3	14,4	12,8	16,6	12	13,5	12,1	4,5
Cys	1,2	1	1,2	1,1	0,6	0,9	1,3	1,1	2	1	1	1,4	0,8
Asp	5,8	7,9	10,6	4,2	7,5	7,3	2,7	3,7	3,4	3,1	4,6	3,5	5,3
Glu	7,1	8,8	5,5	6,5	12	12,2	4,3	7,9	6,8	8,2	10	7,2	5,4
Phe	4,1	2,9	3,3	4,7	3	3,5	5,1	4,4	4,3	3,5	3,6	3,5	2
Gly	7	8,9	15,7	7,3	7,1	7,2	3,3	2,7	3	1,9	2,5	9,2	29,7
His	2,4	2,4	2,4	1,8	2,1	2,4	1,9	2,2	1,6	2,6	2,1	3,4	2,6
Ile	6	3,6	2,2	5,8	4,8	4,3	8,9	7,3	5,4	7,8	4,2	2,6	1,9
Lys	5,8	6,2	3,7	6,6	5,3	4,7	4,8	6	5,6	8,6	8,5	7,6	7,5
Leu	10,1	5,5	5,1	10,5	7	7,2	15,1	13,9	19,2	14,6	14,1	11,1	5,4
Met	1,7	1,4	1,4	1,9	1,3	1,5	2,7	2,5	2,8	2,3	2,5	2,1	1
Asn	4	4,8	8,2	2,1	3,3	2,8	2,3	2,9	2,9	2,7	3,7	6,6	6,8
Pro	4,4	7,7	2,6	5,8	8,2	4,8	0	0,5	0,5	0,1	0	0	4,2
Gln	3,8	3,3	3	3,4	4	5,2	3,9	4,6	3,9	4,9	4,9	5,2	4,6
Arg	5,3	4,9	3,6	5,5	4,9	4,3	6	7	5,5	7,6	6,8	6,5	5,1
Ser	5,8	9,3	11,5	5,1	5,6	5,9	3,7	3,9	4,2	3,7	6	7,2	5,3
Thr	5,1	5,9	8,2	4,6	4,8	5,4	4,6	3,7	2,7	3,3	3,7	3,6	3,2
Val	7	4,1	2,6	7	5,1	5,5	9,5	7	5,2	7,5	4,3	3,1	2,4
Trp	1,3	0,9	1,2	1,8	1,2	1,1	1,7	1,9	1,1	1,4	0,9	0,7	0,4
Tyr	3,4	2,5	3,1	4,2	2,7	3,5	3,9	4,1	3,2	3,2	3,1	3,4	1,9

Table 3. Residue propensity values in 3H helices.

Code	%	−3B	−2B	−1B	+1B	+2B	+3B	+4B	−4E	−3E	−2E	−1E	+1E	+2E	+3E
Ala	8,7	9,7	7,7	5,2	9,8	9,6	10,4	12,9	11,9	14,6	11,9	12,2	10,6	6,3	5,6
Cys	1,2	1,4	1,2	1,4	0,8	0,6	1	1,3	1,3	1	1,3	1,3	1,6	0,9	1,1
Asp	5,8	4,9	8,7	9,4	4	7,1	6,5	1,9	3,4	4,3	2,9	3,7	3,8	6,1	6
Glu	7,1	7,2	9,7	5,4	5,9	11,9	11,5	4,5	7,5	9,4	9	7,7	6,6	6,8	6,1
Phe	4,1	4,8	3,1	3,9	4,3	3,4	4,1	5,3	5,1	3	3,6	4,7	4	2,5	4,1
Gly	7	11,7	8,9	13,7	7	8,4	5,9	4	2,4	3,2	1,9	2,4	10,5	14,7	15,9
His	2,4	2,4	2,5	2,4	1,9	1,7	2,5	1,8	1,9	1,7	2	2,5	3,1	3,1	3,5
Ile	6	3,7	3,3	2,8	6,3	4	4,3	9,5	8,1	4,8	8,1	4,9	3,5	2,3	4,4
Lys	5,8	5,6	5,7	4,3	7,2	5,2	4,7	4,6	6,1	7	7,3	7,6	7,3	8,4	7,1
Leu	10,1	6,6	5,6	6	10,4	6,4	7,4	16	14,9	13,9	15,8	16,8	9,5	6,5	8,3
Met	1,7	1,7	1,6	1,4	2	1,3	1,5	3	2,8	2,6	2,3	3	1,9	1,2	1,2
Asn	4	3	5,5	7,4	2	3,7	3	2,7	3,2	4	2,8	3,6	6,3	6,5	5,6
Pro	4,4	7,5	5,4	2,6	5,5	8,4	4,2	0	0,5	0,7	0,1	0	0	8,1	3,4
Gln	3,8	3,3	3,4	3,4	3,2	3,9	5,3	3,7	4,5	5,1	5	3,8	4,2	5,2	3,6
Arg	5,3	4,3	4,6	3,3	5,8	4,7	3,8	4,8	6,2	6,6	6,3	5,8	6	5,3	5,4
Ser	5,8	7,5	9,4	12,2	5,6	6,4	6,1	3,6	3,9	4,6	3,9	5,9	9,2	6,4	4,8
Thr	5,1	5	6,2	7,8	4,8	4,9	6,1	4,4	3,1	3,8	3,4	3,9	4,1	4,2	4
Val	7	5	4	3	6,7	4,5	5,6	10,2	7,2	5	7,4	5,6	3,5	2,9	5
Trp	1,3	1,1	1	1,2	1,8	1,4	1,1	1,8	2,1	1,2	1,4	0,9	0,8	0,6	0,9
Tyr	3,4	3,4	2,6	3,2	4,9	2,5	4,7	4	4,1	3,4	3,6	3,8	3,7	2,3	4

regions of helices (in position +2E and +3E). Especially, for 2H helices, the frequency of Gly increases from an average value of 7 to 29.7.

Known for not being strong helix conformers, Glu, Lys and Arg residues have a low frequency of occurrence in the examined positions. This is due to their charge that makes them repel each other by preventing the formation of α-helices. It is also known that negatively charged amino acids are often found near the amino-terminus of the helical segment, where they have a stabilizing interaction with the positive charge of the helix dipole; a positively charged amino acid at the amino-terminal end is destabilizing. The opposite is true at the carboxyl-terminal of the helical segment. Consequently, negatively charged amino acids, as Glu and Asp, increase their frequency at the beginning of the helices. For the same reason, positively charged amino acids, as Lys, Arg and His, increase their frequency, often slightly, at the end of the helix. On the one hand, the frequencies of Ala and Leu in position +4B, −4E, −3E, −2E, −1E, are higher than their propensity values inside helices (respectively, 8.7 and 10.1). On the other hand, as expected, Pro and Gly have low frequencies in the central positions, whereas they have high concentrations at the ends of the helix. Indeed, Pro has a very large side chain, and it is well known that Prolines get in the way of α-helix formation. Pro either breaks or twists a helix because it cannot donate an amide hydrogen bond (having no amide hydrogen), and also because its side chain interferes sterically with the backbone of the preceding turn inside a helix, which forces a bend of about 30° in the helix axis. However, Pro is often seen as the first residue of a helix (with a propensity value of 5.5–5.8 in our test, which is greater than the average concentration, 4.4), probably due to its structural rigidity. At the other

extreme, Gly also tends to disrupt helices because its high conformational flexibility makes it entropically expensive to adopt the relatively constrained α-helical structure. Finally, Ala, Leu, Glu and Met have an especially high propensity to belong to the inner part of helices.

3. Predictive Models

The statistical investigation carried out in Sec. 2 shows that informative patterns can be evidenced at the beginning and at the end of amino acid sequences representing α-helices. In order to validate this assumption, we compare three ML approaches, which rely only on sequence data, equipped with attention modules, to decide if a short fixed–length amino acid sequence represents or not an α-helix. The ML models used for the classification of sequence data are a Random Forest Classifier (RFC), a MultiLayer Perceptron (MLP), and a Long-Short Term Memory (LSTM) recurrent architecture. RFC is an additive model which makes predictions by combining decisions from a set of base models, like decision tree classifiers. MLPs can be used for this task, since we consider fixed-length sequences of 14 amino acids (see Fig. 2), while LSTMs are employed supposing that they can better capture the very nature of protein data, even for short sequences. Parametric details of the compared architectures are shown in Table 4.

3.1. *Input sequence representation*

Predictive models are trained on amino acid sequences without further information. To represent each amino acid, the simplest solution is one-hot encoding, i.e. each residue is represented by a binary vector of dimension 20, in which only the element corresponding to a particular amino acid gets a value of 1, while all the other entries are set to 0. Anyway, since one-hot encoding is verbouse and sparse, we also realized a dense representation, exploiting Word2Vec,[19] a technique widely used in natural language processing. Taking a corpus of text in input, this method builds a vocabulary of the words in the corpus, and learns a representation for each word, based on the corpus semantics, as shown in Fig. 3. In our case, each amino acid corresponds to a word, and its dense output vector representation has length 5.

While MLPs and RFCs models deal with a 2D input (a matrix containing patterns represented as vectors), the LSTM input can be seen as a 3D tensor [samples; timesteps; features], where "samples" account for the cardinality of the learning set,

Table 4. Models' hyperparameters.

Model	MLP I	MLP II	LSTM I	LSTM II	LSTM III	LSTM IV
Encoding	One-Hot	Word2Vec	One-Hot	Word2Vec	One-Hot	Word2Vec
LSTM units			30	33	48	54
Dense units	135	135	10	10	20	20
Attention	Before	Before	After	After	Before	Before

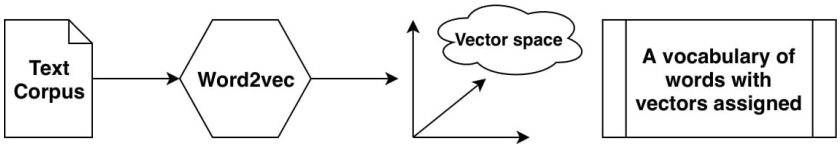


Fig. 3. Representation of the Word2Vec tool.

"timesteps" corresponds to the length of each amino acid sequence, and "features" is the dimensionality of the residue representation (5). The structure of the input tensor is illustrated in Fig. 4.

3.2. *Attention mechanism*

Attention is one of the most important cognitive processes in human beings. When dealing with any problem, instead of processing the whole bulk of information at their disposal, humans focus only on the details which are important for understanding and solving the problem itself. A very similar approach can be attached to the ML models. Even though attention, for instance in neural networks, is very loosely related to the visual attention mechanism found in humans, it has been applied to a wide variety of applications, from text summarization.[20] to image description.[21]

An attention module takes n arguments $y_1, \ldots, y_n$ and a context c in input, and returns a vector z, which is the summary of the y_i focused on the information related to the context c. More precisely, the attention module calculates a weighted arithmetic mean of the y_i, where the weights are chosen according to the relevance of each y_i in the given context c.

In the MLP model, the attention mechanism is applied directly on the input, as reported in Fig. 5.

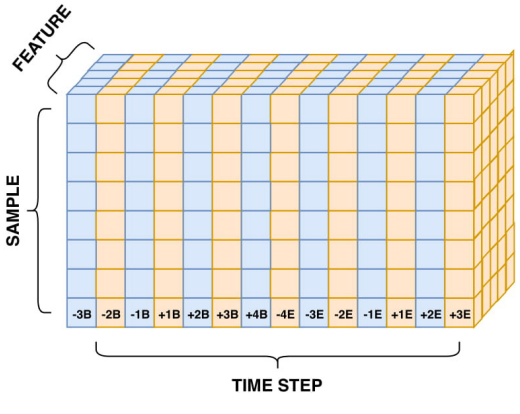


Fig. 4. 3D input tensor for LSTMs.

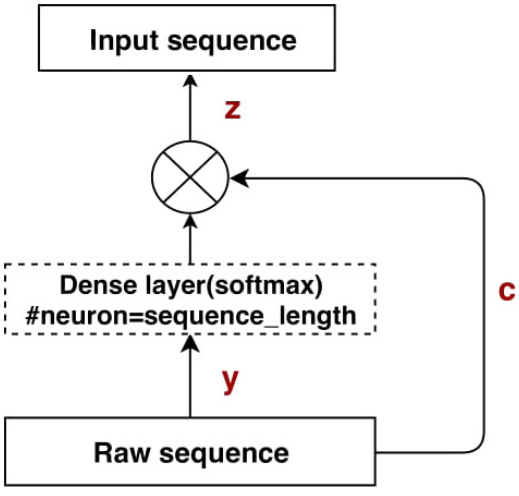


Fig. 5. MLP attention.

Concerning the LSTM model, instead, we applied the attention module in two different positions within the network, respectively, after and before the LSTM layer, as shown in Figs. 6 and 7.

This is done by transposing the input in the 3D tensor [samples; features; timesteps], feeding it to a softmax which estimates the weight distributions, that are then combined with the input sequence. Although both methods are valid, the disadvantage of applying the attention module after the LSTM is that the high-dimensional space spanned by the LSTM might be trickier to interpret. The permutation in the attention mechanism (Figs. 6 and 7) allows us to move the focus on

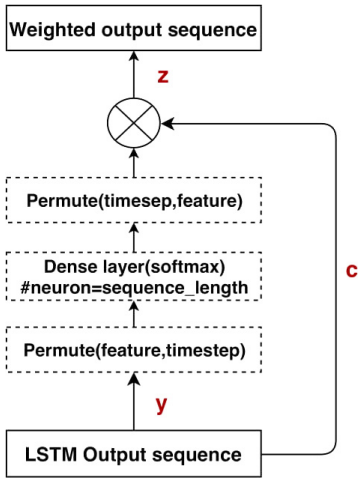


Fig. 6. Attention after LSTM.

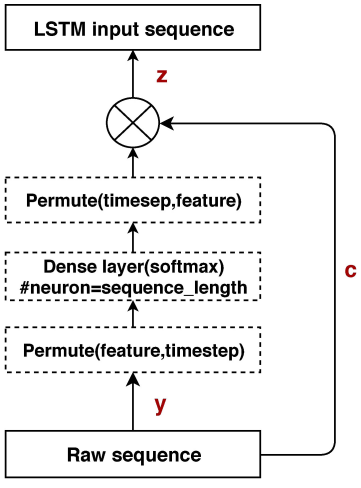


Fig. 7. Attention before LSTM.

each timestep rather than on each feature. Our primary interest is, in fact, to measure the importance of each sequence position in determining the occurrence of α-helices. Secondarily, and only in the one-hot case, the focus on the single feature in each timestep allows us to evaluate the importance of the presence of a particular amino acid in a given position.

In the RFC model, the attention mechanism is implemented in terms of the importance of each sequence position with respect to the model decision that can be evaluated as an average across the base decision tree classifiers.

In order to visualize the focus of attention throughout the sequence positions, each network was trained and tested on 20 different 10-fold cross-validation runs, each using a different dataset split. The attention vectors were averaged over the 10 folds in every test.

4. Experimental Setup

Our dataset is composed of 4127 3H helices and 7767 non-helix sequences, collected from the three main protein classes, described in Table 1.

Six different Neural Networks, all built with roughly the same number of parameters, are used in the experiments. Two of them share the same MLP architecture, a dense SeLU layer followed by a two-unit softmax output layer. The other four models include a single LSTM layer followed by a dense SeLU hidden layer and by a two-unit softmax output layer. Details on architectural hyperparameters can be found in Table 4.

The hyperparameters, shown in Table 4, were selected after a grid search, which consists in trying every possible configuration in order to find the parameter set which guarantees the highest accuracy. In our experiments, each model was tested on

the validation set, to avoid an excessive adaptation of the hyperparameters on the training data. The model is trained with the Adam optimizer, minimizing the categorical cross-entropy loss function.

With respect to RFCs, the parameter values were determined after a grid search. We implemented a forest with 1000 trees with a maximum depth of 50 for the RFC using One-Hot Encoding (RFC I) and 21 for the RFC using Word2Vec (RFC II). We used the default values for the minimum number of samples per leaf (1) and the minimum number of samples required to split an internal node (2).

5. Results

In order to visualize the focus of attention throughout the sequence positions, a set of experiments was performed. Each network was trained and tested on 20 different 10-fold cross-validation runs, each using a different dataset split. The attention vectors were averaged over the 10 folds in every test. Even if optimizing the classification accuracy is out of the scope of this paper, actually obtaining performing models means that we can be confident in their results or, in other words, that they have been able to select the correct information inside the data to solve the classification problem. Therefore, for the sake of completeness, we evaluated the performance for all the ML approaches, almost always obtaining values greater than 80%, as shown in Table 5.

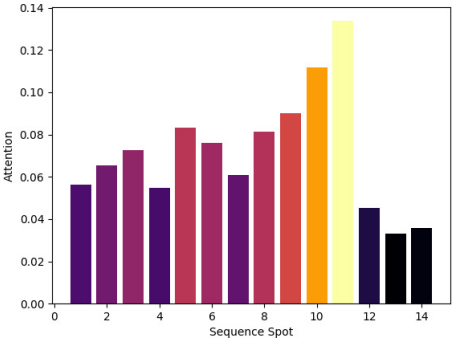
The best architecture, namely the LSTM II network, scored more than 84%, based on Word2Vec embedding and with the attention module posed after the LSTM layer. Nevertheless, networks whose attention mechanism is located after the LSTM layer return attention vectors which slightly depend on the split. On the contrary, attention modules placed before the LSTM layer produce very similar vectors across the runs, showing almost no dependence on the split. This independence is also found in the MLP and RFC models, which produce very similar attention vectors. Therefore, one example, representative of each model, can be considered to illustrate the related attention vectors, as shown in Fig. 8.

Regardless of the ML model and the encoding used, the attention focuses principally on the last position upstream with respect to the 5′ end of the helix, and on the three last positions inside the helix.

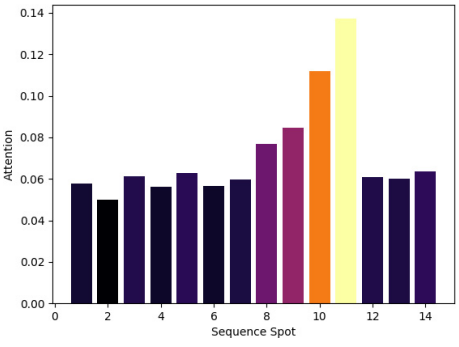
From the bar charts, it emerges that most of the information which defines the presence of an α-helix is contained in the helix itself, except position -1B. Furthermore, the information at the end of the motif looks more relevant with respect to what is present in the other considered position both inside and outside the helix.

Table 5. Accuracy (Acc %) and Standard Deviation (SD) of the six models.

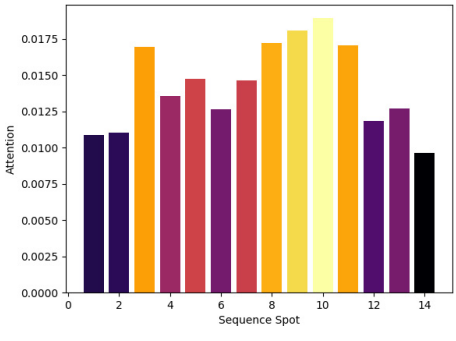
Model	MLP I	MLP II	LSTM I	LSTM II	LSTM III	LSTM IV	RFC I	RFC II
Acc %	82.00	78.10	82.74	84.08	81.84	83.35	82.30	82.02
SD	0.005	0.006	0.009	0.011	0.007	0.014	0.007	0.007



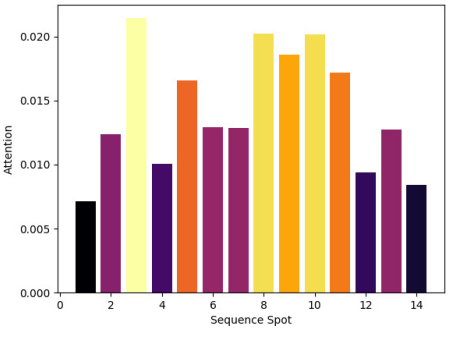
(a) Average attention LSTM III (one-hot).



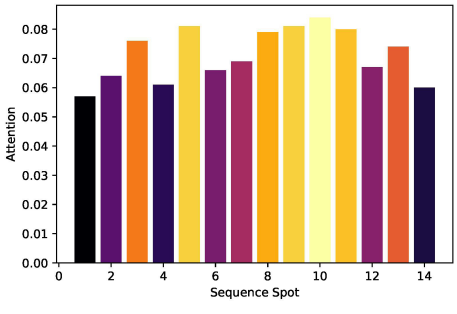
(b) Average attention LSTM IV (Word2Vec).



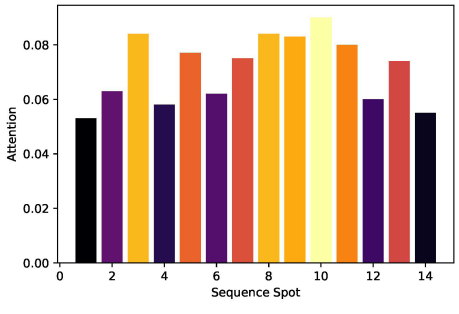
(c) Average attention MLP I (one-hot).



(d) Average attention MLP II (Word2Vec).



(e) Average attention RFC (one-hot).



(f) Average attention RFC (Word2Vec).

Fig. 8. Models average attention.

One of the advantages of one-hot encoding with respect to Word2Vec is the possibility to quantify the importance of the single amino acids in each position of the attention vector. This can be visualized with the help of the heatmap in Fig. 9, calculated for LSTM III. Each row of the heatmap corresponds to a sequence position (from top to bottom), while each column corresponds to one amino acid (indicated by its one letter code).

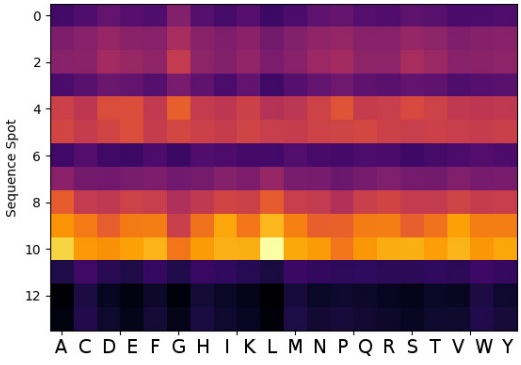


Fig. 9. Attention heatmap.

In Fig. 9, the focus of attention on the last three positions of the helix is highlighted. In particular, the LSTM III network concentrates on the occurrence of the amino acids which are particularly abundant in those positions (see Table 3): Leucine (15,8% in position −2E), Alanine (14,6% in position −3E) and Valine (7,4% in position −2E). On the contrary, the less common residues, such as Glycine (1,9% in position −2E) and Proline (0,1% in position −2E), seem not so relevant for the classification, according to the attention mechanism. The correlation between the heatmap in Fig. 9 (calculated considering 3H helices) and the residue propensity values in Table 3 is evident and suggests that the network attention principally focuses on highly concentrated amino acids (f.i., Glutamic acid at the beginning of the helix, Alanine and Leucine at the end), while it does not take into account down-concentrated amino acids (f.i., Proline and Glycine at the end of the helix, which are represented by dark blue cells).

Figure 9 also underlines the fundamental role of Leucine in α-helix stabilization, as it appears to be the most abundant amino acid at the helix carboxy-termini. The fact that Leucine is the least affected by translation errors, due to its six different codons, seems to make it more preferable than other strong α-helix formers, such as Glutamic acid, Alanine and Methionine, in the position where helices must collapse.

6. Availability and implementation

All the code are implemented in Python, and the experiments were performed on a single GPU Nvidia 1080 TI. For the sake of reproducibility, data, source codes and models evaluated in this paper were made freely available at `https://github.com/annavisibelli/DL4Helices`.

7. Conclusions

Given a protein sequence (its primary structure), the first step towards the prediction of the three-dimensional native configuration consists in determining its

secondary structures. This means telling which backbone regions are likely to form helices, strands, and β-turns, U-bent structures obtained when a β-strand reverses its direction in an antiparallel β-sheet.

Secondary structure prediction algorithms employ a variety of computational techniques, including neural networks, finite state automata, hidden Markov models, clustering techniques and genetic algorithms.

Based on the intuition that signals should exist, in the form of particular amino acid concentrations, which determine the formation of secondary structures and define their extension, in this paper, we carried out a statistical analysis of the amino acid concentrations in the vicinity of α-helices. We then compared ML approaches to predict their formation and reveal the fundamental role some positions play in protein folding. In particular, three different ML methods were used, equipped with an attention module, to predict the presence of amino acid signals for the occurence of α-helices. The attention mechanism, integrated in our prediction methods, can actually derive useful information on protein sequence profiles.

The obtained experimental results demonstrate the power of ML techniques in extracting information from protein data to make predictions on the protein structural features, based only on the amino acid sequence. Both MLPs, RFCs and LSTMs can interpret the nature of protein data and focus on the long-conserved pieces of information which are fundamental in the formation of secondary structures. Moreover, having demonstrated that both the statistical an the ML approaches focus on the same positions to ascertain the presence of an α-helix has a twofold impact. On the one hand, it reinforces the biological intuition of the presence of amino acid signals delimiting helical moieties; on the other hand, it ensures the interpretability of the results produced by ML approaches, showing how what we repute biologically significant is also important for the nework decision.

It is a matter of future research to extend the proposed approach to the prediction of signals defining other common secondary structures, namely β-sheets and U-turns. Actually, such a local information can be potentially applied to support 3D protein structure prediction. Indeed, the combination of the existing sophisticated deep learning techniques with a deeper knowledge of the primary structure information content — in the form of amino acid signals which regulate the formation of secondary structures or of protein residue — residue contacts,[22] etc. will play soon a significant role, at least for a rough but rapid prediction of the structure of new proteins.

Indeed, our results are very encouraging, and suggest to continue using ML approaches for the secondary structure prediction. Nevertheless, many challenges remain open, requiring the development of alternative strategies to complement/improve existing techniques.

Acknowledgments

Bongini and Rossi that contributed equally to this work.

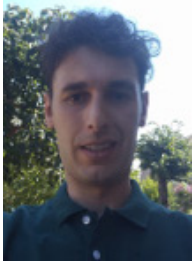
References

1. Anfinsen CB, Principles that govern the folding of protein chains, *Science* **181**:223–230, 1973.
2. Levin JM, Exploring the limits of nearest neighbour secondary structure prediction, *Protein Eng Des Selection* **10**:771–776, 1997.
3. Schmidler SC, Liu JS, Brutlag DL, Bayesian Segmentation of Protein Secondary Structure, *J Comput Biol* **7**:233–248, 2000.
4. Geourjon C, Deléage G, SOPM: A self–optimized method for protein secondary structure prediction, *Protein Eng Des Selection* **7**:157–164, 1994.
5. Aydin Z, Altunbasak Y, Borodovsky M, Protein secondary structure prediction for a single-sequence using hidden semi-Markov model, *BMC Bioinf* **7**:178, 2006.
6. Bidargaddi N, Chetty M, Kamruzzaman J, Combining segmental semi-Markov models with neural networks for protein secondary structure prediction, *Neurocomput* **72**:3943–3950, 2009.
7. Bouziane H, Messabih B, Chouarfia A, Effect of simple ensemble methods on protein secondary structure prediction, *Soft Comput* **19**:1663–1678, 2014.
8. Jones DT, Protein secondary structure prediction based on position-specific scoring matrices, *J Mol Biol* **292**(2):195–202, 1999.
9. Bourlard H, Morgan N, *Connectionist Speech Recognition: A Hybrid Approach*, Kluwer Academic Publishers, 1994.
10. Kalchbrenner N, Blunsom P, Recurrent continuous translation models, in *Proc ACL Conf Empirical Methods in Natural Language Processing*, pp. 1700–1709, 2013.
11. Li Z, Yu Y, Protein secondary structure prediction using cascaded convolutional and recurrent neural networks, *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 2560–2567, 2016.
12. Heffernan R, Yang Y, Paliwal K, Zhou Y, Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility, *Bioinf* **33**(18):2842–2849, 2017.
13. Bennett CF, Bray JE, Harrison AP, Martin N, Shepherd A, Sillitoe I, Thornton J, Orengo CA, Pearl FMG, The CATH database: An extended protein family resource for structural and functional genomics, *Nucl Acids Res* **31**(1):425–455, 2003.
14. Pauling L, Corey RB, Branson HR, The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain, *Proc Natl Acad Sci* **37**(4):205–211, 1951.
15. Perutz MF, Rossmann MG, Cullis AF, Muirhead H, Will G, North ACT, Structure of Hæmoglobin: A Three-Dimensional Fourier Synthesis at 5.5–Å. Resolution, Obtained by X-Ray Analysis, *Nature*, **185**(4711):416–422, 1960.
16. Kendrew JC, Dickerson RE, Strandberg BE, Hart RG, Davies DR, Phillips DC, Shore VC, Structure of Myoglobin: A three-dimensional fourier synthesis at 2 Å. resolution, *Nature*, **185**(4711):422–427, 1960.
17. Brunak S, Engelbrecht J, Protein structure and the sequential structure of mRNA: α-Helix and β-sheet signals at the nucleotide level, *Proteins: Structure, Function Bioinf* **25**(2):237–252, 1996.
18. Kabsch W, Sander C, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers* **22**(12):2577–2637, 1983.
19. Mikolov T, Corrado G, Chen K, Dean J, Efficient estimation of word representations in vector space, in *Proc Int Conf Learning Representation*, pp. 1–12, 2013.
20. Rush AM, Chopra S, Weston J, A neural attention model for abstractive sentence summarization, in *Proc EMNLP 2015*, pp. 379–389, 2015.

21. Xu K, Ba JL, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel RS, and Bengio Y, Show, Attend and Tell: Neural image caption generation with visual attention, in *Proc ICML 2015*, Vol. 37, 2048–2057, 2015.
22. Eickholt J, Cheng J, Predicting protein residue–residue contacts using deep networks and boosting, *Bioinf* **28**:3066–3072, 2012.

Anna Visibelli is Ph.D. student in Biochemistry and Molecular Biology at the University of Siena. She obtained the Bachelor Degree in Mathematics and the Master Degree in Applied Mathematics with honors at the University of Siena in October 2018. Her current research is focused on applying mathematical techniques to biology.

Pietro Bongini is a Ph.D. student in Smart Computing at the University of Florence. He is affiliated with the Department of Information Engineering and Mathematics, at the University of Siena. In July 2015, he obtained a bachelor degree in Information Engineering, followed in July 2018 by a Master Degree with honors in Computer and Automation Engineering, both at the University of Siena. His main research interests are bioinformatics and machine learning on structured data.



Alberto Rossi Received the B.E. degree in information engineering from the University of Siena, Siena, Italy, in 2013, and the M.S. degree in computer and automation engineering from the University of Siena, Italy, in 2017, with a score of 110 cum laude. He is working toward the Ph.D. degree in smart computing at the University of Florence, Florence, Italy. His research interests include deep learning, data mining and computer vision.

Monica Bianchini received the Laurea degree cum laude in Applied Mathematics in 1989 and the Ph.D. degree in Computer Science and Control Systems in 1995 from the University of Florence, Italy. She is currently an Associate Professor at the Department of Information Engineering and Mathematics of the University of Siena. Her main research interests are in the field of machine learning, with emphasis on neural networks for structured data and deep learning, approximation theory, bioinformatics, and image processing. She served/serves as an Associate Editor for IEEE Transactions on Neural Networks, Neurocomputing, In. J. Knowledge-Based and Intelligent Engineering Systems, Int. J. Computers in Healthcare, and has been the editor of numerous books and special issue in international journals on neural networks/ structural pattern recognition. She is a permanent member of the editorial board of IJCNN, ICANN, ICPR, ESANN, ANNPR, and KES.

Neri Niccolai is a Full professor in Biochemistry at the University of Siena, he received the Laurea in Chemistry from the University of Florence, Italy, in 1971. Post-doctoral fellow at the University of Wisconsin-Madison in 1977–1979 and visiting professor at the School of Pharmacy of the University of London in 1985–1986. He is currently Senior Professor at the Department of Biochemistry, Chemistry and Biotechnology of the University of Siena. As a Structural Biologist, his research activity spans from NMR investigation on protein surface accessibility to bioinformatic analyses of protein-ligand interfaces. Niccolai's awards and honors include the Gold Medal from the Italian Society of Chemistry in 2014 for his relevant contribution in the field of biological NMR. Currently, SCOPUS assigns to him 130 papers in ISI journals.