Emilio Dorigatti

Cancer immunotherapy design and analysis through discrete optimization, positive-unlabeled learning, and semi-structured regression models

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Eingereicht am 17.07.2023

Emilio Dorigatti

Cancer immunotherapy design and analysis through discrete optimization, positive-unlabeled learning, and semi-structured regression models

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Eingereicht am 17.07.2023

Erster Berichterstatter: Prof. Dr. Bernd Bischl
Zweiter Berichterstatter: Prof. Dr. David Rügamer
Dritter Berichterstatter: Prof. Dr. Oliver Kohlbacher

Tag der Disputation: 28.05.2024

Acknowledgments

I would like to express my sincere gratitude to Prof. Dr. Bernd Bischl, Dr. Benjamin Schubert, Prof. Dr. David Rügamer, and Dr. Mina Rezaei for their invaluable advice, guidance, supervision, mentorship, and inspiration, without which the content of this Ph.D. thesis could have never be conceived, and my formation as a scientist grossly incomplete. I would also like to thank Prof. Dr. Oliver Kohlbacher for their willingness to act as the third reviewer for my Ph.D. thesis, as well as PD Dr. Fabian Scheipl and Prof. Dr. Christian Heumann for their availability to be part of the examination panel at my Ph.D. defense.

I am grateful to all my colleagues and coauthors, in particular Alejandra, Anil, Cornelius, Emy, Faye, Felix, Ingo, Jann, Jonas, Juan, Julian A., Julian R., Katharina, Sabrina, and Yang, for their inspiration and support in both professional and personal matters, which they unknowingly provided in countless occasions and circumstances, showing me new and better ways to handle difficult situations.

I will forever be thankful to my wife Jiawen for her unwavering support during both the darkest and brightest moments of this Ph.D., including a global pandemic, as well as all other members of my family for caring and showing their love when it matters the most.

Summary

From ideation to market availability, developing new drugs and therapies can take more than a billion dollars and a decade of work. Clinical testing in human subjects is a particularly time-consuming phase of the development process, and nine out of ten clinical trials fail to demonstrate safety and/or efficacy of the treatments. This delays the introduction to the market by years, and makes the treatment more expensive for end consumers. The safety and efficacy of any given treatment is determined by characteristics of patients and diseases, but our limited ability to identify such factors inevitably leads to reduced success rates of clinical trials, because of overly broad categorization of diseases and patients. Cancer treatments in particular are plagued by low response rates, with therapies often failing to clear the tumor.

The recent introduction of novel computational and experimental tools in clinical practice, mostly enabled by artificial intelligence techniques, led to the discovery of a large number of previously unknown *biomarkers*, i.e., chemical factors that differentiate sub-populations of patients and sub-types of diseases, leading to an improved understanding of the variables that drive the efficacy of therapies. At the same time, advances in experimental techniques generated an exponential increase in the amount of available data characterizing the molecular landscape of patients, making computational tools a necessity to recognize patterns and identify promising directions to develop new therapies, in an approach known as *precision medicine*.

This thesis contributes to the precision medicine revolution by introducing an expert opinion paper about potential uses of artificial intelligence in this practice, novel computational tools to aid the development of cancer immunotherapies, and methodological advances to confront some challenges arising from the complex data modalities frequently found in this field. From an applied perspective, this thesis introduces two frameworks for cancer vaccine design based on discrete optimization, complemented by a benchmark of machine learning predictors that are used in conjunction with such frameworks. Then, recognizing the frequent absence of negative examples with which to train machine learning models for such biological problems, this thesis introduces two methods to learn from this type of data with a particular focus on imbalanced distributions. Finally, enabling practitioners to interpret the effect of tabular data such as clinical variables of a patient, modeled jointly with non-tabular data including radiology and histopathology images, this thesis presents a method to perform correct statistical inference in semi-structured regression models. One application of such models, predicting the spread of COVID-19 in Germany, highlights the advantage of such hybrid modeling.

Zusammenfassung

Von der Idee bis zur Marktreife kann die Entwicklung neuer Medikamente und Therapien mehr als eine Milliarde Dollar und ein Jahrzehnt Arbeit in Anspruch nehmen. Dabei stellen klinische Studien am Menschen eine besonders zeitaufwändige Phase des Entwicklungsprozesses dar, und in neun von zehn Fällen gelingt es nicht, die Sicherheit und/oder Wirksamkeit der Behandlungen nachzuweisen. Dadurch verzögert sich die Markteinführung um Jahre, und die Behandlung wird für die Endverbraucher teurer. Die Sicherheit und Wirksamkeit einer bestimmten Behandlung hängt von den Charakteristika der Patienten und Krankheiten ab. Aber unsere begrenzte Fähigkeit, solche Faktoren zu identifizieren, führt unweigerlich zu geringeren Erfolgsquoten bei klinischen Versuchen, weil Krankheiten und Patienten zu breit kategorisiert werden. Insbesondere Krebsbehandlungen haben mit niedrigen Ansprechraten zu kämpfen, da die Therapien den Tumor oft nicht beseitigen können.

Die jüngste Einführung neuartiger computergestützter Instrumente in der klinischen Praxis, die größtenteils durch Techniken der künstlichen Intelligenz ermöglicht werden, führte zur Entdeckung einer großen Zahl bisher unbekannter Biomarker.Das sind Faktoren, die Subpopulationen von Patienten und Subtypen von Krankheiten unterscheiden, was zu einem besseren Verständnis der Variablen führt, die die Wirksamkeit von Therapien bestimmen. Gleichzeitig haben Fortschritte bei den experimentellen Techniken zu einem exponentiellen Anstieg der verfügbaren Datenmenge geführt, die die molekulare Landschaft der Patienten charakterisiert. Dies führt dazu, dass computergestützte Werkzeuge eine Notwendigkeit geworden sind, um Muster zu erkennen und vielversprechende Richtungen für die Entwicklung neuer Therapien zu identifizieren.

Diese Arbeit leistet einen Beitrag zur Revolution der Präzisionsmedizin, indem sie ein Expertengutachten über den möglichen Einsatz künstlicher Intelligenz in dieser Praxis vorstellt. Dabei werden auch neuartige computergestützte Werkzeuge zur Unterstützung der Entwicklung von Krebsimmuntherapien und methodische Fortschritte zur Bewältigung einiger Herausforderungen beleuchtet, die sich aus den komplexen Datenmodalitäten ergeben, die in diesem Bereich häufig anzutreffen sind. Desweiteren, werden in der Arbeit aus einer angewandten Perspektive zwei Rahmenwerke für die Entwicklung von Krebsimpfstoffen vorgestellt, die auf diskreter Optimierung beruhen, ergänzt durch einen Benchmark von Prädiktoren für maschinelles Lernen, die in Verbindung mit solchen Rahmenwerken verwendet werden. Ferner, in Anbetracht des häufigen Fehlens von Negativbeispielen, mit denen maschinelle Lernmodelle für solche biologischen Probleme trainiert werden können, werden in dieser Arbeit zwei Methoden zum Lernen aus dieser Art von Daten mit besonderem Schwerpunkt auf unausgewogenen Verteilungen vorgestellt. Schließlich wird eine Methode zur korrekten statistischen Inferenz in semi-strukturierten Regressionsmodellen vorgestellt, die es Praktikern ermöglicht, die Auswirkungen von tabellarischen Daten wie klinischen Variablen eines Patienten zu interpretieren, die gemeinsam mit nicht-tabellarischen Daten wie radiologischen und histopathologischen Bildern modelliert werden. Eine Anwendung solcher Modelle, die Vorhersage der Ausbreitung von COVID-19 in Deutschland, verdeutlicht den Vorteil einer solchen hybriden Modellierung.

Contents

List of Figures

1 Motivation and Outline

1.1 Precision medicine: each patient has a different disease

People have always looked for ways to alleviate pain and suffering. Ancient China developed a holistic system of healing that encompassed various modalities such as herbal remedies, acupuncture, and specific exercises emphasizing the balance and flow of vital energy (Qi) within the body. In parallel, western thought held illness to be a punishment from God because of a person's sins. The most common forms of treatment were therefore aimed at placating divine displeasure, and included prayers, penances, pilgrimages, and even exorcisms, just in case the illness was rather caused by evil spirits possessing one's body. Meanwhile, the Islamic culture made great contributions to the understanding of human anatomy and medical practice, most notably ideating the modern concept of the hospital as a place to care for the sick.

The Renaissance brought foreign scholarship in focus, allowing Europeans to catch up with other cultures with more developed medical knowledge, and to contribute new discoveries during the Enlightenment. Germ theory, according to which disease was caused by certain microorganisms, was one of the most important conceptual paradigm shifts that occurred in this period, and obsoleted most previous theories, in particular the idea that illness was caused by "smelly" air, the so-called "miasma" theory (Rhodes et al., 2023).

Empirical investigators at this time focused on diseases with clear *phenotypic* variations, i.e., visible differences in the physical or chemical characteristics of sick patients compared to healthy individuals. On the basis of these observations, scientists hypothesized that small differences among individuals were caused by the presence or absence of certain personal "factors," and that particular combinations of these factors could lead to diseases. This notion sparked considerable debate in the scientific community regarding the connection between discrete factors and continuous phenotypic traits such as height, and the question was solved by assuming that phenotypic traits are affected in small degree by a large number of distinct factors, thus giving the appearance of continuous variations within a population. Eventually, the modern notion of genes was linked to such factors, leading to the birth of *genomics*, the field that studies genetic mutations and their association with an organism's phenotype.

Initial systematic efforts in drug development in the 19-th century were based on isolating compounds from plants and extracting anti-toxin serum from animals previously exposed to a disease. This costly and time-expensive process was obsoleted by the advent of synthetic organic chemistry, allowing pharmaceutical companies to conduct large-scale screening of synthetic compounds. However, the productivity boost of this technology proved short-lived, as companies were now facing considerable hurdles in testing randomly-generated compounds, most of which did not work. As a result, research focus shifted towards *rational* design, studying the way in which diseases operated to manually identify new potential drug targets. The quest for mechanistic understanding of diseases was the main driver, and benefactor, of advances in genomics technologies (Debouck and Metcalf, 2000; Emilien et al., 2000), reducing sequencing cost by five orders of magnitude in the last twenty years alone (Wetterstrand, 2021).

A typical drug development pipeline is structured as a funnel, where increasing amounts of resources are committed to the investigation of a reducing number of promising compounds. Broadly speaking, such a pipeline is composed of several iterations of the following stages (Hughes et al., 2011):

1. Target identification and validation: A target is a biological entity that is involved in the disease. Drugs are designed to interact with a specific target, alleviating the disease by blocking or enhancing certain processes. Potential targets are identified through basic research and by using data mining techniques on available biomedical datasets (Yang et al., 2009). Because most connections are not causal in nature, however, all potential targets have to be validated through wet-lab experiments (Dugger et al., 2018).

2. Assay development: Molecules that interact with the target are called *hits*. In order to quantify to what extent interactions are successful, specialized assays and wet-lab experimental protocols may have to be developed *de novo*, or adapted from existing ones.

3. Hit identification: The assays developed in the previous step are used to screen compounds and molecules, seeking those that interact with the target. A few of the most promising hits are then selected, often based on heuristics, for further analyses, where functional assays are used to determine whether the hit has any effect on the disease, besides interacting with the target.

4. Lead optimization: Hit molecules that successfully clear all tests become *leads*, whose molecular structures are further refined to improve their interaction with the target, for example by increasing potency, so that smaller doses are required, selectivity, so as to avoid unwanted interactions with other molecules, and other physiochemical properties that ensure the drug is safe and effective.

5. Clinical studies: Finally, leads that could be optimized to satisfactory levels are tested through clinical trials to certify their safety, effectiveness, potential side-effects and their severity, in living organisms.

Despite the efficient allocation of resources through such a funnel, developing a new drug cumulatively costs more than one billion US dollars and takes ten to 15 years on average, with more than nine out of ten compounds failing to clear the clinical trials, despite the lengthy and meticulous process used to derive lead candidates (Hughes et al., 2011).

The wealth of data and string of failures accumulated in drug discovery soon highlighted that different patients responded differently to identical treatment, sometimes even developing worse conditions in response of a certain drug rather than clearing the disease. This further increased interest in finding genetic factors that determine patients' response to interventions, thus inverting the drug discovery pipeline from a “disease-to-drug-to-patient” paradigm into a “patient-to-disease-to-drug” approach. Conveniently for pharmaceutical companies, this approach married the ethical duty of providing appropriate care to all patients with the economic incentive of increasing the success rate of clinical trials. Similar ideas were also applied to preventive care, recognizing that not only the disease itself, but also the modality and timing of its occurrence are very personal and different among patients. The principle of tailoring treatments to individuals, either by improving the allocation of existing therapies or by developing entirely new drugs, is known as *personalized medicine* (Goetz and Schork, 2018). An intermediate milestone towards personalized medicine is *precision medicine*, whereby patients are divided into small cohorts, on which a certain treatment is known to be effective, based on relevant molecular signatures called *biomarkers*.

The field of oncology was one of the first adopters of precision medicine, and personalized medicine then, techniques, since cancer is the quintessential example of a highly patient-specific disease (Zitvogel and Kroemer, 2017). Cancer is the leading cause of death in high-income countries (Dagenais et al., 2020), and is the result of malicious mutations in a cell's genome, common events that randomly occur during cell division as a result of errors while duplicating its genome. The impact of mutation ranges from inconsequential, having no impact on the life of a cell, to catastrophic, inducing cell death shortly thereafter. Most mutations are dealt with internally by DNA repair mechanisms that correct errors and damaged DNA, and mutations that are not detected by this quality control generally cause a cell to malfunction or to behave unnaturally, producing anomalous signals that lead the immune system to dispose of this cell. However, there exist a remote possibility that a sequence of mutations, accumulated over a cell's ancestry, allow it to grow and proliferate uncontrollably at the expense of other healthy cells, while escaping detection or elimination by the immune system, thus forming a *tumor*. Although mutations are infrequent, and each only has a minuscule probability of resulting in a tumor, the size of the human genome, the number of cells in a human body, and the rate at which they divide, make the probability of a tumor occurring during an individual's lifetime noticeable. Moreover, even though most cancers have a few typical "signature" mutations, their randomness makes most mutations in any single tumor highly patient-specific.

1.2 Artificial Intelligence (AI): pattern recognition at scale

The growing amount and complexity of data about patients and their diseases collected in the lab led to the birth of *bioinformatics* in the 1960s (Gauthier et al., 2019). Experimental limitations of the most common sequencing method used at the time (Edman et al., 1949) prevented scientists from directly finding the amino acid sequence of proteins longer than 50 or 60 amino acids. To do this, it was necessary to fragment these long proteins into small chunks, determine the sequence of each chunk separately, and reconstruct the sequence of the whole protein by inspecting the overlaps among chunks. The tedious process of reassembling these short reads was the first of many tasks to be automated with a computer program, and remains a fundamental step in modern bioinformatics pipelines. Other foundational issues were approached shortly thereafter. For example, comparing sequences of proteins with the same function, but found in organisms of different species, made it possible to relate all known living organisms in the "tree of life." The evolutionary history of life can be reconstructed by comparing the sequence of common proteins found in different organisms, based on the rationale that species that diverged a long time ago from a common ancestors had more time to accumulate mutations compared to species that diverged more recently. This means that the larger the number of differences, the further in the past two species diverged. This same principle was used to trace the spread of SARS-CoV-2 variants in the recent COVID-19 pandemic (Forster et al., 2020; Li et al., 2020a). In the past sixty years, sequencing technology and bioinformatics mutually reinforced each other's exponential growth, such that it is simply unimaginable, nowadays, not to use computers to analyze biological data.

Bioinformatics is only one of the countless fields that were created by the advent of computers and automated information processing. Artificial Intelligence (AI) started as an endeavor to make computers think and act like humans (Russell, 2010), and the question of whether a mechanical machine can become "intelligent" was raised much earlier. Initial computerized approaches were based on manipulating discrete units of information called *symbols*, for example by applying the rules of logic to reach certain conclusions from a set of known facts. Early successes in the 1950s were followed by a great deal of enthusiasm and grandiose promises on the capabilities that AI systems would reach in the next years. Instead, bitter disappointment followed, as it was soon apparent that these systems failed to scale beyond toy examples due to the combinatorial explosion of trivial or irrelevant conclusions that could be generated by blindly applying logical inference rules. Research therefore shifted to domain-specific *expert systems*, that promised to replace actual human experts in complex fields such as healthcare, business and engineering. Initial encouraging results of the 1970s generated, again, a wave of enthusiasm and subsequent disappointment, as codifying experts' knowledge into a set of comprehensive and well-defined rules and facts turned out to be much harder than anticipated. Later approaches to AI, therefore, narrowly focused on specific sub-problems of intelligence and cognition, such as the ability to learn from experience, the ability to communicate with others, and the ability to plan a sequence of actions to achieve a certain outcome, with each field following separate approaches. Eventually, it was noticed that methods based on learning were far more effective than methods involving manual knowledge engineering, and in the last decades the growing availability of data and computational power made the field of *machine learning* rise to prominence as the most effective approach to AI, encompassing the majority of modern applications. The realization that general-purpose learning methods, supported by extensive data and computing power, could significantly surpass intricate algorithms leveraging expert knowledge that AI researchers have been striving to develop over the past six decades came to be known as *the bitter lesson* (Sutton, 2019).

Today, the unique ability of AI systems to learn and detect new patterns in large amounts of data far surpasses that of human experts, and it is thus no surprise that many extremely diverse fields of research have greatly benefited from this technology. For example, AI is being used to improve the design of fusion reactors by modeling plasma (Kates-Harbeck et al., 2019), to provide more accurate weather forecasts by modeling atmospheric events (Schultz et al., 2021), to estimate pollution and land cover from satellite data (Rezaee et al., 2018), and many other applications throughout most modern scientific endeavors (Jordan and Mitchell, 2015). The health sciences, in particular, present a plethora of challenges amenable to AI, including biomedical image analysis, prognosis, patient care, and clinical decision support (Ravì et al., 2016; Miotto et al., 2018; Norgeot et al., 2019; Esteva et al., 2019).

1.3 AI as a fundamental driver of precision medicine

It is now a commonly held belief that precision medicine will immensely benefit from AI techniques, increasing the accuracy by which treatments are tailored to individual patients, as well as helping develop new treatments that are more effective (Boniolo et al. , 2021). With AI, it will be possible to identify increasingly complex associations between diseases and specific genetic and molecular factors to enable even more accurate stratification of patients into cohorts that are likely to respond to certain treatments, and uncover disease subtypes requiring different drugs (Figure 1.1) AI can help medical researchers to explore a wider landscape of possible treatments by reducing required

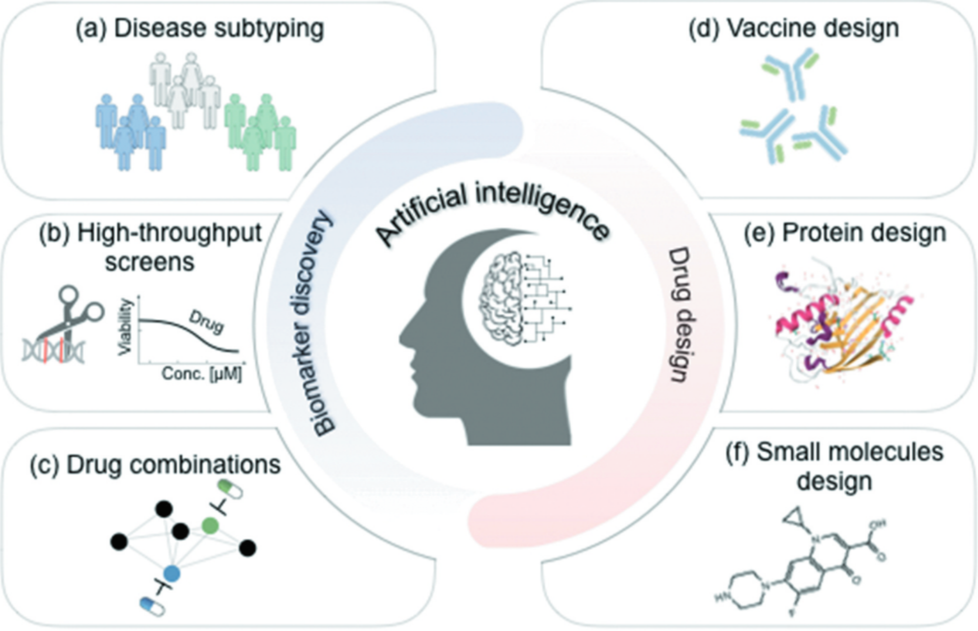

Figure 1.1: Artificial intelligence can support precision medicine and early drug discovery by discovering more reliable biomarkers (left), uncovering different disease subtypes (a), quickly screening thousands of drug compounds (b), and suggesting novel combinations of synergistic drugs (c). It can also improve the effectiveness of drugs (right) by designing personalized vaccines (d) and drugs based on large proteins (e) and small-molecule (f) that are better suited to specific patient cohorts. Figure credit: Boniolo et al. (2021)

experimental efforts and increasing the efficiency of high-throughput screening procedures. AI is also poised to revolutionize the drug discovery pipeline (Gawehn et al., 2016; Chen et al., 2018; Paul et al., 2020), making it more efficient by predicting various properties of molecules, such as stability, functionality, toxicity, solubility, etc., as well as the interaction strength between a drug and its target, optimizing molecular structures accordingly with little human input.

The pattern recognition capabilities of AI are also essential to deal with the extreme variability of cancer. Cancer immunotherapies rely on instructing the patient's immune system to attack mutation products that are specifically associated with cancer cells and differentiate them from healthy ones. A comparison of tumor specimens with healthy tissue samples collected from the patient can reveal hundreds or thousands of differences. Most of them, however, cannot be used for immunotherapy, as they are not recognized by the immune system. Moreover, the antigens that can be recognized by the immune system of a patient are not necessarily recognized by the immune system of another patient, due to inherent genetic variability built into certain components of the immune system. All in all, this means that it is not feasible to produce generic, off-the-shelf cancer treatments (Shetty and Ott, 2021). AI approaches are thus essential, and increasingly applied, to screen cancer neoantigens in order to determine the optimal subset that should be included in an immunotherapy for a given patient (Shetty and Ott, 2021). Vaccines personalized in such a way are one of several treatment options for cancer, whose benefits include increased flexibility and specificity, ease of manufacturing, long-term protection against relapse (Blass and Ott, 2021), as well as opportunities for preventive treatment (Finn, 2018). Indeed, several clinical trials already demonstrated safety and effectiveness of cancer vaccines (Abd-Aziz and Poh, 2022).

1.4 Outline

Drug discovery and precision medicine thus present vast challenges and opportunities to be tackled with AI. This thesis contributes several AI techniques in this regard, with a particular focus on cancer vaccines and related challenges.

Chapter 2 builds an understanding of relevant domain knowledge in biology with the end goal of understanding cancer vaccines (Section 2.3.3). The immune system is introduced in Section 2.1, describing first innate immunity mechanisms that target a broad range of generic disease patterns (Section 2.1.1), followed by a description of how adaptive immunity against a specific antigen is formed (Section 2.1.2). As cancer immunotherapies aim at stimulating adaptive immunity against neoantigens generated by mutations, a more detailed discussion on how neoantigens are recognized follows in Section 2.2. Finally, Section 2.3 describes the emergence of cancer, why it is so difficult to cure, and the different treatment options that are available (Section 2.3.2), concluding with a focus on cancer vaccines (Section 2.3.3).

Chapter 3 introduces AI and the main techniques used in the contributions of this thesis. Section 3.1 describes how machines can learn from data by adjusting a suitable model of the world to match available examples. Deep neural networks, a specific type of extremely flexible models, especially suited for unstructured data such as images and text, are introduced in Section 3.2, while Section 3.3 describes methods to learn from data that does not contain negative examples, as is often the case in applications that deal with antigen processing (Section 2.2). As modern clinical and medical applications frequently include both tabular and non-tabular data such as images, Section 3.4 explores how deep neural network can be combined with models that are more suited to handle tabular data, so as to harness the benefits of both approaches. Finally, Section 3.5 discusses computational approaches to design a specific type of cancer vaccines upon which some contributions of this thesis are centered.

The contributions of this thesis start from applications on cancer immunotherapy design in Chapter 4. First, an expert opinion article outlines the many ways in which AI will boost precision medicine (Section 4.1), from biomarker discovery to drug design, including vaccines. The second (Section 4.2) and third (Section 4.3) contributions introduce two vaccine design frameworks based on mixed-integer linear programming that improve over previous methods by jointly formulating and solving two problems that were approached separately by previous methods (Section 3.5). Such frameworks are crucially dependent on accurate predictions of an important biological event, therefore in Section 4.4 we present a survey and benchmark of the current landscape of computational predictors, arguing about possible future directions of the field.

Recognizing the frequent lack or scarcity of negative examples that can be used to learn predictors for many biological events, Chapter 5 introduces two contributions that improve machine learning methods for this scenario, known as positive-unlabeled learning (Section 3.3). Section 5.1 introduces a generic method, based on model-agnostic semi-supervised learning techniques, to improve performance of positive-unlabeled learning models on imbalanced datasets, where the majority of events to classify are actually negative. A method specialized to deal with image data is then proposed in Section 5.2, where the latest advances in representation learning are applied to positive-unlabeled learning, thus obtaining considerably higher performance compared to alternative approaches.

In Chapter 6 we present two contributions to semi-structured regression models, hybrid approaches that combine deep learning with statistical regression techniques to learn from both tabular and non-tabular data modalities at the same time, while providing interpretable effects for the tabular part. First, in Section 6.1 we use such approach to predict the number of COVID-19 cases in each district in Germany, by combining geographical data with information about the population of each district. Next, in Section 6.2 we then show that traditional inference methods for semi-structured regression models result in increased false-positive rates, and propose an alternative that ameliorates the issue.

Lastly, Chapter 7 contains some concluding remarks, summarizing the novelty of the presented contributions (Section 7.1), critically reviewing the research methodology followed (Section 7.2), and discussing exciting new possibilities for future work (Section 7.3).

2 Biological Background

This chapter establishes epitope vaccines as one of the promising treatment modalities for cancer (Section 2.3), after presenting a basic introduction to the immune system (Section 2.1), and to the main events that occur when the body processes a vaccine (Section 2.2).

2.1 Basic concepts in immunology

This section presents an introduction to immunology based on the first few chapters of the excellent Murphy and Weaver (2016). The immune system comprises all of the tissues and mechanisms that the body employs to defend itself from *pathogens*, i.e., harmful organisms that cause disease. The most notable characteristic of the immune system is its ability to adapt over the lifetime of an individual, evolving to resist to threats it had never seen before. Indeed, it has been known since at least ancient Greece that surviving a disease gives greater protection towards it later in life, and this principle was already exploited in the 1400s by the Chinese and Middle Eastern civilizations to create the first vaccines. A more systematic study of this phenomenon occurred during 1800s, when investigations on the serum of animals immune to a certain disease led to the discovery of *antibodies*. Antibodies are small proteins that confer immunity to a disease by interacting with *antigens*, specific parts of the microorganism that caused the disease, called *pathogen*.[1] Besides antibodies, the immune system is composed of many types of cells called *leukocytes*, or white blood cells, as well as different types of molecules and proteins, that together cooperate to eliminate threats. Immune cells either permanently reside in peripheral tissues and organs, or circulate in the bloodstream, or circulate in the lymphatic system, a system of vessels that drains and cleanses extracellular fluid, including the damage caused by ongoing infections.

Pathogens are broadly divided, based on their size and *modus operandi*, into viruses, bacteria, fungi, and parasites, each requiring different methods to be opposed. Not all foreign organisms inside the human body are threats to be eliminated, however. In fact, many tissues rely on the work of *commensal* organisms to operate properly, living in symbiosis with the human body for mutual advantage. The immune systems deals with pathogens using three basic strategies: avoidance, resistance, and tolerance. Avoidance mechanisms entail all those countermeasures that prevent pathogens from entering the body in the first place, including physical and chemical barriers such as the skin and the mucosal coating of the nose walls. Resistance strategies entail all those countermeasures that reduce and eliminate pathogens, and are usually enacted by a variety of molecular and cellular functions collectively called *effector mechanisms*. Finally, tolerance mechanisms improve the ability of tissues to withstand the damage caused by pathogens, and are mostly found in plants.

[1] For example, COVID-19 vaccines induce the generation of antibodies targeting the spike protein (the antigen) of the SARS-CoV-2 virus (the pathogen).

2.1.1 Innate immunity

Innate immunity represents the first line of defense against pathogens, composed of relatively simple and generic, non-adaptable countermeasures. Most anatomical barriers on the interface between the body and the external world, for example, employ simple chemical substances and a variety of antimicrobial proteins to impede pathogens. Among these, the *complement* is a group of proteins that act together and, possibly, in conjunction with antibodies to destroy, or facilitate destruction, of foreign organisms, for example through *lysis*, i.e., the destruction of their external membrane. Surviving pathogens encounter more elaborate cellular defenses initiated by *sensor cells*. Sensor cells employ receptors on their surface to detect a variety of anomalous molecules and substances that are not usually found in the extracellular environment. The existence of these substances suggests the presence of infiltrating pathogens, or cellular damage they caused, and thus indicates an ongoing infection. Sensor cells detect substances that are essential cellular components, without which pathogens would not be able to function properly, making them invariant to evolutionary pressure and thus excellent targets for recognition. In response to such anomalous patterns, sensor cells either try to destroy the pathogen, or involve other immune cells by using the appropriate mediator. These mediators are small molecules that convey important signals to cells that bear the appropriate receptor on their surface, which respond appropriately once they detect these mediators. More than a hundred are known, some being widely recognized by immune cells, and some being very specific. Two broad categories exist of mediators exist: (1) *chemochines* attract immune cells from the bloodstream into the infected tissue, and trigger other symptoms, collectively known as *inflammation*, that improve the efficacy of the immune response, while (2) *cytokines* enable and amplify certain functionalities of the target cell. Inflammation makes nearby blood vessels larger and more permeable, allowing white blood cells that circulate in the blood to reach the infection site in the surrounding tissues.[2] It also increases the local body temperature to aid in killing the pathogens, and it increases the flow of lymph, so as to drain the waste substances that result from fighting against the pathogen. The lymph transports antigens to nearby lymphoid tissues, where the adaptive immune response is initiated.

2.1.2 Adaptive immunity

Adaptive immunity requires more time to be initiated compared to the innate immune response, because it involves antigen-specific cells that have to be formed in response to the antigen itself. Antigen-specificity refers to the fact that lymphocytes, i.e., the cells in the adaptive immune system, have a receptor that only recognizes a single, specific antigen, unlike cells in the innate immune system, whose receptor can recognize and deal with a wide range of antigens. This specificity makes the adaptive immune response extremely effective at dealing with any sort of threat, at the cost of a slower response. Normally, lymphocytes are inert, or *naive*, and continuously circulate between lymphoid organs such as lymph nodes and the spleen through the lymphatic system, until they encounter their cognate antigen, i.e., the antigen that is recognized by their receptor. Lymphoid organs favor the encounter between naive lymphocytes and antigens coming from active infection sites. These antigens are usually carried by specialized cells called *antigen presenting cells*, since their role is to present antigens to naive lymphocytes, seeking the one that

[2] Some blood cells leak as well, making the skin appear redder.

has a matching receptor. Certain inflammatory inducers supercharge this mechanism, increasing the flow of lymph and attracting antigen presenting cells and naive lymphocytes to nearby lymphoid organs, increasing the probability of finding matching pairs of antigen and receptor.

When a match occurs, the lymphocyte *proliferate* into tens of thousands of clones that all carry exactly the same receptor, in a process called *clonal expansion.* Expanding lymphocytes *differentiate* into one of several sub-types, gaining full functionality to oppose the pathogen through certain specific effector mechanism. Some of these clones, instead of directly fighting the pathogen, become *memory cells*, and are responsible for the increased immunity towards the same antigen later in life by enabling a faster adaptive response. Two major types of lymphocytes exist, with widely different *effector* functionality and development: B cells and T cells, named after their place of origin, respectively the bone marrow and the thymus. Activated B cells, also known as plasma cells, secrete antibodies with the same antigen specificity as the plasma cell's receptor. Antibodies coat pathogens, making it hard for them to damage tissues and, at the same time, making it easier for other immune cells to deal with the pathogen by interacting with bound antibodies. Activated T cells differentiate into four classes. *Cytotoxic* T cells kill other cells, such as those infected by viruses or certain single-cellular pathogens, while *helper* and *regulatory* T cells provide specific signals to orchestrate the immune response by controlling other immune cells, respectively by boosting and inhibiting certain types of immune responses.

The receptors of both T and B cells are composed of a *constant region*, that is the same for every receptor and keeps it attached to the parent cell, and a *variable region* that recognizes the antigen. The variable region of the receptor recognizes a very specific part of the antigen called *epitope*, having a molecular structure that is complementary to that of the receptor's variable region, similarly to how a key fits in a lock. Antibodies function in a similar manner, however a few types exist that differ in their constant region. The constant region sticks out after antibodies are bound to the antigen, and is recognized by other immune cells that react differently depending on the type of antibody. While antibodies and B cell receptors can bind to almost any chemical structure that is found in the extracellular space, T cell receptors only recognize epitopes that are bound to specialized protein complexes, called Major Histocompatibility Complexes, or *MHC molecules*, found on the surface of other cells. MHC molecules present epitopes that result from proteins produced inside the cell, thereby enabling the immune system to know whether the cell is working properly or not. Both cancer and viral infections cause cells to malfunction and produce abnormal proteins, some fragments of which are presented on the surface by the MHC. When a cytotoxic T cell recognizes a MHC-bound epitope, it destroys the cell by releasing toxins.

There exists an astronomical number of possible antigenic sequences, each of which requires a different receptor in order to be recognized, yet, the human genome only contains a few hundreds of genetic segments that can be used to construct the variable region of a receptor. During lymphocyte development, a subset of these segments is randomly selected, mutated, and permanently joined together to form the variable region of the lymphocyte's receptor. It is the combinatorial diversity resulting from this *genetic recombination* process that allows a few hundreds basic components to result in more than 10^{20} possible receptors, of which at least 10^7 different ones are present on lymphocytes circulating in the human body at any given moment. A *positive selection* process leads to the survival of receptors that are useful to fight pathogens: If a naive lymphocyte does not recognize its cognate antigen within a certain amount of time it simply dies off, leading to a constant renewal of the receptor pool in circulation, while upon antigen recognition lymphocytes proliferate and differentiate, leading to the accumulation over time of memory cells harboring useful receptors. A similar process of *negative selection* eliminates developing lymphocytes whose receptor is activated by *self-antigens*, benign parts of the individual host that should not be attacked, thus giving rise to *immunological tolerance* towards friendly tissues, or autoimmune diseases when something in this process goes wrong. This is a biological solution to the multi-armed bandit problem!

2.2 Antigen processing and presentation

Cytotoxic T cells, also called CD8 T cells because of a particular surface receptor they harbor, are the main cancer-killing cells, and rely on epitope presentation by MHC molecules to recognize and eliminate tumors. Two types of MHC molecules exist: MHC Class I, or MHC-I for short, presents epitopes to cytotoxic T cells, while MHC Class II, or MHC-II, presents epitopes to helper T cells, that coordinate the immune response rather than killing cancer cells directly. This difference is necessary because MHC-I molecules are used by all nucleated cells to present intracellular peptides, thus allowing CD8 T cells to eliminate other cells that produce abnormal proteins due to, for example, a viral infection or cancer, while MHC-II molecules are only found on immune cells, and other specific types of cells when stimulated by appropriate cytokines, thus involving helper T cells during specific phases of the immune response to pathogens.

Naive CD8 T cells must be activated by *dendritic cells*, a type of antigen presenting cells carrying the appropriate epitope on their MHC-I molecules, together with a number of biological signals to activate the T cell. When this happens, the T cell acquires effector cytotoxic capabilities, and is able to kill any other non-immune cell presenting the same epitope. Normally, MHC-I molecules only present peptides from endogenous proteins, however dendritic cells are also able to present peptides derived from exogenous proteins through the *cross-presentation* pathway, which is essential to generate anti-tumor responses.

Before being presented by the MHC, epitopes undergo a sequence of steps collectively called *antigen processing pathway* (Figure 2.1). Having a firm understanding of these processes is essential, as cancer cells disrupt them several ways to avoid detection and elimination by the immune system. Both cytotoxic and helper T cells should be involved for an effective anti-cancer response, however, as the main goal of cancer immunotherapies is to restore the functionality of cytotoxic T cells, the following presentation is focused on the MHC-I pathway, synthesizing concepts from Maupin-Furlow (2012) and Blum et al. (2013).

2.2.1 Proteasomal cleavage

The epitopes presented on the MHC-I originate from proteins that were be fragmented into pieces that are small enough to fit on *peptide binding groove* of the molecule. Such fragments are typically between eight and ten amino acids long, while the originating antigen can include thousands or even more. This degradation process is performed by a protein complex called *proteasome*, a tubular-shaped construct that cleaves the antigen into peptides of approximately the right length. Most proteins targeted by the proteasome are tagged by specific amino acid sequences called *degrons*, which in eukaryotes correspond to ubiquitins. The entry point of the proteasome is guarded by caps that consume adenosine triphosphate (ATP), the energy store of cells, to unfold and straighten the protein to be degraded while removing degrons, followed by an anti-chamber

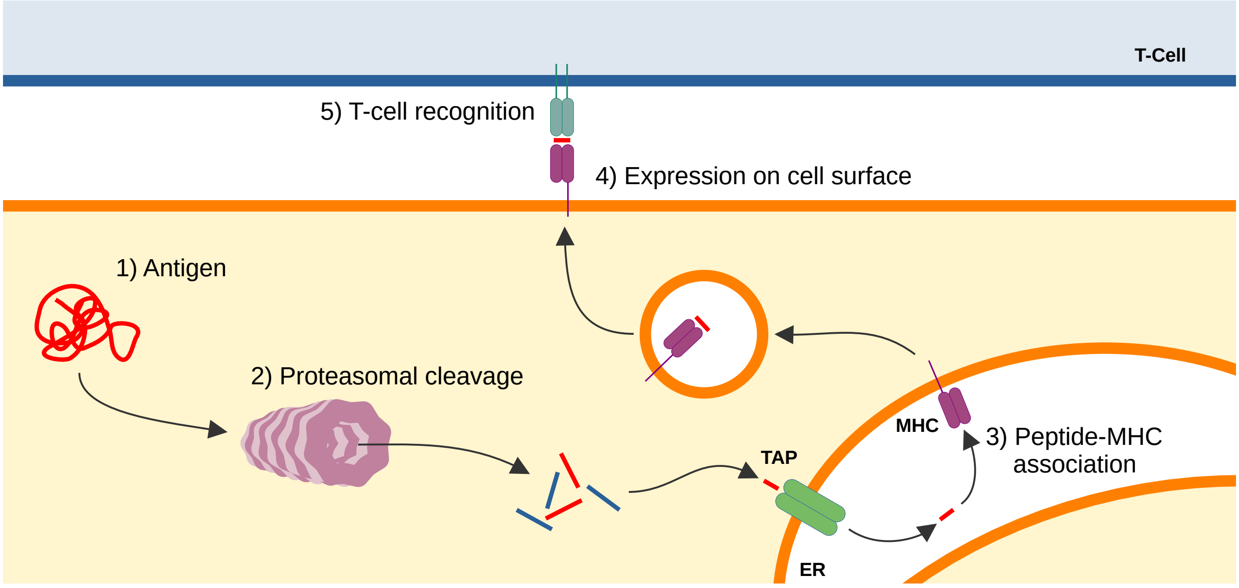

Figure 2.1: The major events in the MHC-I antigen processing pathway. Antigens (1) are cleaved in short fragments by the proteasome (2). Some of these peptides are then transported into the endoplasmic reticulum (ER) through the Transporter associated with Antigen Processing (TAP). A fraction of these peptides bind to the Major Histocompatibility Complex (MHC, 3) and the resulting construct is then expressed on the cell surface (4), where they can be inspected by passerby T-cells and possibly trigger an appropriate immune response (5). Figure inspired from Dorigatti et al. (2022a)

serving as a buffer zone to equalize the rate of proteolysis with the rate of protein ingestion. Besides helping with immunosurveillance, proteasomes also contribute to protein quality control by degrading damaged proteins, and regulate important cellular processes by removing certain regulatory proteins at key locations and time-points.

Proteasomes can be differentiated into *constitutive* proteasome and *immunoproteasome*. While the main function of the constitutive proteasome is related to normal cellular processes necessary for cell growth survival, the immunoproteasome is responsible for generating most of the MHC-I-bound peptides, although the constitutive proteasome can also contribute to peptide presentation. The immunoproteasome has a slightly different composition than the constitutive proteasome, and thus different cleavage specificity, that allows it to produce peptides that are specifically tuned for MHC-I presentation, and its presence and activity is greatly increased in response to interferon-gamma (IFN-γ), a cytokine that signals inflammation.

The proteasome and the proteins it digests are all located in the *cytosol*, the main workspace where most intracellular processes occur, including the translation of mRNA into new proteins by *ribosomes*. Viruses hijack rybosomes, and lead cells to produce new copies of the virus, rather than the proteins the cells needs to function properly. and some of these viral proteins, instead of being assembled into a functional virus, are presented on MHC-I molecules, revealing the cell to be infected. Cancerous cells, similarly, produce mutated proteins, some pieces of which end up on the MHC-I, and some types of vaccines, including mRNA vaccines, also exploit this process within dendritic cells, inducing them to present epitopes from the vaccine antigen on the MHC-I.

2.2.2 MHC binding

Peptides produced by the immunoproteasome are bound to MHC-I molecules in the endoplasmic reticulum (ER), a membrane system that forms a series of sacs (empty pockets) inside eukaryotic cells that helps with protein synthesis and transportation. Peptides enter the ER through the transporter associated with antigen processing (TAP), essentially a gate into the pocket formed by the ER. Before binding, the N-terminals of peptides in the ER are further trimmed by the endoplasmic reticulum aminopeptidase-1 (ERAP1) to the optimal length preferred by the MHC-I, i.e., eight to ten amino acids.

Attached to the TAP is the *peptide loading complex* (PLC), a number of proteins that fix free-floating peptides upon empty MHC molecules before they can be transported to the cell surface. Similarly to T and B cell receptors, MHC molecules are also composed of a constant domain and a highly-polymorphic variable domain, allowing them to bind to arbitrary antigens. Empty MHC molecules are held in place adjacent to the TAP by *tapasin*, such that a free-floating peptide can be fixed onto the groove formed by variable region of the MHC molecule by two additional proteins called *chaperone calreticulin* (CRT) and *ERp57*.

If the binding affinity (strength) between the peptide and the MHC is strong enough, a conformation change in the MHC molecule detaches it from the PLC, and frees it to migrate outside of the ER and onto the cell surface. If the binding affinity is not high enough, however, the MHC will not be able to dissociate, and the loaded peptide will be removed to make place for a more suitable peptide. This process of peptide *editing* ensures that only high-affinity peptides, so-called *immunodominant*, are presented to T cells. The disruption of this delicate choreography in cancer cell is a major reason why tumors are so difficult to eradicate.

2.3 Oncoimmunology

Cancer is the second leading cause of death worldwide, after cardiovascular diseases, and the first leading cause in high-income countries (Dagenais et al., 2020). Fundamentally, cancer is unrestricted and undesired growth of a group of cells at the expense of surrounding cells, and, eventually, the host itself. Initial cancer treatments were aimed at removing the mass of abnormal cells through chemical or mechanical processes, but the severity of the resulting side-effects encouraged research into alternative forms of treatment, starting from understanding of the biological processes involved in cancer. Zitvogel and Kroemer (2017) present a comprehensive summary of the major concepts of this topic, of which a brief summary follows.

2.3.1 Carcinogenesis

Cancer is marked by a constant struggle between the tumor and the immune system. This struggle, according to the *3E hypothesis*, develops through three separate stages: initial Elimination of abnormal cells by the immune system, followed by an Equilibrium between the cancer and the immune system, and a final Escape where cancer cells overpower the immune system and grow unchecked.

The Elimination phase is centered on several *immunosurveillance* mechanisms, whereby immune cells quickly eliminate nascent tumors before they become harmful, using many of the same strategies that used for external pathogens (Section 2.1.2). Tumors originate by a gradual accumulation of malignant genetic abnormalities over time caused by random mutations. Mutations arise normally upon cell division, and are facilitated by factors such as UV light, pollution, viruses, chronic inflammation, hereditary diseases, etc. Harmless mutations can accumulate over the progeny of a cell, and normal cells have a variety of safeguards that detect damaged DNA and attempt to repair it, or trigger cell-death when not possible. Most mutations that somehow elude these mechanisms are nonetheless detected by the immune system and the surrounding intracellular environment, leading again to the elimination of the cell. In the end, however, this continuous and persistent emergence of random mutations results in an appreciable probability that a specific combination enables a cell to escape all immunosurveillance mechanisms enacted by the immune system. This cell is now able to proliferate, and all daughter cells inherit the malignant mutations that allow them to escape immunosurveillance, plus additional new mutations generated upon division.

At this point, there is a constant struggle by the immune system to prevent the cancer from growing uncontrolled, leading to a situation of temporary Equilibrium. This phase is marked by the creation of a *tumor microenvironment*, a dynamic "battleground" that surrounds the tumor and includes a variety of immune cells. These immune cells form tertiary lymphoid structures, highly organized lumps of immune cells that allow a faster *in situ* adaptive response, and a stroma (barrier) of cells to contain tumor growth while allowing infiltration from external immune cells. The tumor microenvironment is a highly dynamic place, whose composition continually evolves over time in response to cancer growth, and is correlates with prognosis both before and after immunotherapy. For example, high densities of cytotoxic T cells are correlated to increased overall survival rates in most, but not all, tumors, while increased presence of regulatory T cells is associated with poor prognosis. The amount of CD8 T cells and regulatory T cells are only two of many known *biomarkers* that help predicting the appropriate treatment.

During the equilibrium phase, eliminated cancer cells are continuously replaced by new variants harboring different sets of mutations. This process, akin to natural selection, is called *immunoediting*, and eventually leads to the formation of mutations that make cancer cells invisible to and untouchable from the immune system. These mutations enable many *immunosuppression* mechanisms that prevent the immune system from working properly within the tumor microenvironment, thus preventing CD8 T cells from killing cancerous cells. These mechanisms involve, for example, down-regulating components of the antigen processing pathway such as proteasomal cleavage and MHC production (Section 2.2), increasing the presence of regulatory T cells preventing CD8 T cells from operating properly, enacting measures to mediate the exhaustion of CD8 T cells, and so on. Tumor escape may also be followed by *metastasis*, whereby some cancerous cells migrate away from the main originating tumor and establish themselves in other organs and tissues, originating new cancer sites that will also grow. Based on these insights, the latest consensus on cancer treatment is to operate within the tumor microenvironment, trying to reinstate the mechanisms that allow the immune system to fight cancer on its own, while still administering treatments that kill cancer cells directly.

2.3.2 Cancer immunotherapies

The first treatments for cancer aimed at killing the tumor cells through exogenous mechanisms, for example special drugs (*chemotherapy*) or high-energy X-ray or proton beams (*radiotherapy*). Although effective, such therapies are not applicable after metastasis, and often present severe side-effects, including collateral damage of healthy tissues. Cancer immunotherapies, instead, try to exploit the immune system of the patient and reinstate the mechanisms that enable CD8 T cells to kill cancer cells (Koury et al., 2018). Generally speaking, there are three main obstacles that must be overcome to produce an effective immunotherapy (Mellman et al., 2011): first, it must stimulate antigen presentation by dendritic cells; second, it must generate protective T cell responses; and third, it must reverse immunosuppression mechanisms in the tumor microenvironment.

Adoptive T cell therapies involve extracting T cells from a patient, allowing a suitable subset to grown and expand *in vitro* (i.e., in the lab), and re-administering the new T cells to the patient together with some immunostimulants (Zitvogel and Kroemer, 2017). This type of therapy has relatively high response rates and results in durable immunity, as it tends to generate T cells targeting several tumor antigens at the same time, however it is not very scalable since cell growth in the lab is a laborious process (Koury et al., 2018). Chimeric antigen receptor (CAR) T cells are an alternative to autologous T cells that are genetically engineered to target a specific tumor antigen, and have also proven a successful and flexible treatment modality.

Immune checkpoints are mechanisms of the immune system that modulate the immune response to prevent unwanted damage to healthy tissues, and one of the main mechanisms of cancer immunosuppression is the abuse of checkpoints to prevent the immune system to eliminate the tumor. Inhibitory checkpoints reduce the functionality of effector lymphocytes, for example to wind down the immune response after the threat has been eliminated, while stimulatory checkpoints enhance their functionality, for example at the beginning of an infection. By up-regulating inhibitory checkpoints, and down-regulating stimulatory checkpoints, cancer cells are able to disable immune cells in their surroundings. Treatments that prevent tumors to manipulate such checkpoints are among the most effective options available, although their response rates are still low due to the complexity of the biological processes involved (He and Xu, 2020). Just like all biological signaling mechanisms, immune checkpoints involve the interaction between receptors and ligands, and treatments prevent such interactions by using antibodies specifically designed to “clog” the receptors or “trap” the ligand, thus preventing them from interacting and delivering the respective message.

Similarly to immune checkpoint inhibitors, antibodies can also be used to target certain surface receptors that are over-expressed by cancer cells, thus reducing proliferation and increasing cell death. Moreover, the constant region of such antibodies, after they are bound to the surface of cancer cells, can trigger other components of the innate immune system (Zitvogel and Kroemer, 2017). Unlike normal antibodies, bispecific antibodies are able to bind to two different antigens, and are used to direct immune cells towards cancer cells, for example by attaching a peptide-MHC complex to cancer cells that down-regulated MHC expression, thus making the cell recognizable to cytotoxic T cells again (Koury et al., 2018; Dahlén et al., 2018).

2.3.3 Cancer vaccines

Vaccines stimulate the immune system to produce protective responses against a specific antigen. Prophylactic, or preventive, vaccines are administered before patients catch a disease, and are aimed at forming a population of memory T and B cells that ensures a faster and stronger immune reaction when the subject comes in contact with the antigen again. Effective prophylactic cancer vaccines already exist for cancers of viral origin, such as hepatitis B and human papillomavirus (HPV). Therapeutic vaccines are instead administered to treat an existing disease, and in the case of cancer are targeted to tumor-specific antigens resulting from mutations in cancer cells, also known as *neoantigens*. As most of these mutations are highly patient-specific, therapeutic cancer vaccines are an example of personalized medicine (Sahin and Türeci, 2018).

Identifying such antigens is a long process, requiring extensive experimental procedures that leverage next-generation sequencing to analyze the genome of cancer cells and compare it with healthy cells from the same patient, identifying, ranking, and selecting the most promising antigens through computational techniques (Gopanenko et al., 2020). These computational tools are not yet entirely reliable and produce a large number of false positives, i.e., suggest neoantigens that do not evoke an immune response, therefore careful experimental validation is still required (Shetty and Ott, 2021). As identifying and validating neoantigens is such a lengthy process, an alternative approach is to target public neoantigens occurring relatively frequently across different types of cancer, so-called *driver* mutations because they are essential for the tumor (Pearlman et al., 2021).

The chosen set of neoantigens has to be delivered to the patient so as to induce a strong and effective immune response, most importantly stimulating dendritic cells to pick up the neoantigens encoded by the vaccine and presenting them to T cells to enable their effector mechanism. One possible way of doing so is to directly load the neoantigens onto dendritic cells *ex vivo*, i.e., in the wet-lab, and administering these dendritic cells to the patient (Saxena and Bhardwaj, 2018). Another option is to stimulate dendritic cells *in vivo* by delivering the neoantigens in such a way that they appear to be originating from a pathogen, thus undergoing the same antigen processing pathway (Section 2.2). The most direct way of doing so is to genetically engineer a virus to express the neoantigens when they infect a healthy cell, instead of generating new copies of the virus itself (Humphreys and Sebastian, 2018). Another option is to encode the antigens into DNA (Gary and Weiner, 2020) or mRNA (Pardi et al., 2018) strands, and stimulating dendritic cells to pick up and express these neoantigens.

All these possibilities are continually being tested and refined through a number of clinical trials, and based on their results the consensus is that cancer vaccines are a safe and feasible treatment modality that can produce immune responses in most patients when paired with checkpoint inhibitors, although a lot of work still needs to be done before they can become a routine option for cancer treatment (Shetty and Ott, 2021; Blass and Ott, 2021; Abd-Aziz and Poh, 2022).

2.3.4 Connection to precision medicine and contributions

As a highly patient-specific disease, cancer treatment was one of the first adopters and motivators for precision medicine techniques. After offering an overview of how AI can advance the effectiveness of precision medicine techniques (Section 4.1), this thesis proposes two frameworks to design cancer vaccines (Sections 4.2 and 4.3) that differ in how the vaccine is formulated. The design of these vaccines requires the prediction of the outcome of several events in the antigen processing pathway, motivating a survey and benchmark we conducted (Section 4.4), as well as the development of novel methods to handle the datasets frequently found in the field (Sections 5.1 and 6.2).

3 Methodological Background

In its infancy, AI was partitioned into planning and reasoning, natural language processing, and machine learning (Russell, 2010). In the last decades, however, machine learning approaches, propelled by artificial neural networks, have been increasingly applied to the other two fields and now constitute the dominant approach to AI.

3.1 Fundamental concepts in Machine Learning

Machine learning enables a computer to learn a specific behavior from a set of provided examples of that behavior, and to apply the same behavior to new situations. From a high-level perspective, machine learning approaches are differentiated based on the type and amount of available examples. In supervised learning, each example is a pair of “situation” and “expected behavior,” while in unsupervised learning only the situations are given, but not the desired behavior. Semi-supervised learning is a mix of both, where certain situations have an associated behavior and other situations do not. In reinforcement learning, no examples are given, but the machine is able to interact with an environment and observe how the environment changes as a result of the actions it took. A reward signal is then given to the machine, whose goal is to identify the sequence of actions that lead to the maximum reward. As the contributions of this thesis are focused on supervised and semi-supervised learning, the concepts introduced in this section are focused on these settings, adapting content from excellent books on the subject (Bishop and Nasrabadi, 2006; Murphy, 2012; MacKay, 2003; Mohri et al., 2018).

3.1.1 The machine learning blueprint

Regardless of the specific learning setting, there are three basic ingredients for any machine learning method: the data, the model, and the optimization procedure.

Data

“Data” refers to the examples that are available to the machine to learn the desired behavior. We denote this set with $\mathcal{D}_l = \{(x^{(i)}, y^{(i)})\}_{i=1}^{n_l}$, $n_l > 0$, where $x^{(i)} \in \mathcal{X}$ is a situation, and $y^{(i)} \in \mathcal{Y}$ the associated response. The dataset $\mathcal{D}_l$ is commonly seen as an independent and identically distributed (i.i.d.) sample from an unknown probability distribution $p_{\mathcal{X}\mathcal{Y}}$ defined on $\mathcal{X} \times \mathcal{Y}$. In semi-supervised learning, the machine can also access an additional dataset $\mathcal{D}_u = \{x^{(i)}\}_{i=1}^{n_u}$, $n_u > 0$, of situations without an associated expected response, coming from a probability distribution $p_{\mathcal{X}}$ which is usually, but not necessarily, the marginal of $p_{\mathcal{X}\mathcal{Y}}$. Hereafter, we denote with $\mathcal{D}$ the entire set of available examples, with $\mathcal{D} := \mathcal{D}_l$ in supervised learning applications, and $\mathcal{D} := \mathcal{D}_l \cup \mathcal{D}_u$

in semi-supervised learning, and the context will make it clear which alternative we are referring to, if a it makes a difference.

For reasons of clarity, in the remainder of the thesis scalar values will be denoted with italic symbols (e.g., $a \in \mathbb{R}$), vectors with lowercase bold letters (e.g, $\boldsymbol{x} \in \mathbb{R}^d$), higher-order tensors, including matrices, with uppercase bold letters (e.g., $\boldsymbol{X} \in \mathbb{R}^{n \times d}$). Moreover, as in the majority of machine learning applications, including those mentioned in this introduction, $\mathcal{X} \subseteq \mathbb{R}^d$ for some $d > 1$ and $\mathcal{Y} \subseteq \mathbb{R}$, we follow the conventions above and always denote the examples as $\boldsymbol{x}^{(i)}$ and $y^{(i)}$. Finally, to keep the notation uncluttered, the samples in $\mathcal{D}$ will not be indexed explicitly unless necessary.

Model

We assume that there exists a (possibly random) function $c : \mathcal{X} \to \mathcal{Y}$, called a *concept*, that connects each situation to its desired response induced by $p_{\mathcal{X}\mathcal{Y}}$. The goal of the machine is to learn this function in the most accurate way possible based on the dataset $\mathcal{D}$. In doing so, we restrict the machine to only consider possible functions, or *hypotheses*, from a set $\mathcal{H}$ that is decided beforehand by practitioners. Note that $\mathcal{H}$ often does not contain the true concept c that connects $\mathcal{X}$ to $\mathcal{Y}$, thus the problem is, in general, how to find the "best" member of $\mathcal{H}$ to approximate c.

For example, in binary classification it is assumed that the situations in $\mathcal{X}$ belong to one of two types or *classes*, i.e., $\mathcal{Y} = \{-1, 1\}$, and the machine should *classify* each situation in $\mathcal{X}$ into the correct class. Logistic regression classifiers decide which class an example $\boldsymbol{x}$ belongs to based on which side of a separating hyperplane it falls to. Logistic regression thus corresponds to the following hypothesis class:

$$\mathcal{H} = \left\{ \boldsymbol{x} \mapsto \text{sign} \left(\boldsymbol{x}^\top \boldsymbol{w} + b \right) : \boldsymbol{x}, \boldsymbol{w} \in \mathbb{R}^d, b \in \mathbb{R} \right\} \tag{3.1}$$

Frequently, the hypotheses in $\mathcal{H}$ have the same functional form, and are parameterized by an item θ of a set Θ. Usually, $\Theta \subseteq \mathbb{R}^{d'}$ for some d' that need not be the same as the dimensionality of $\mathcal{X}$, therefore we will follow the convention of denoting the parameters $\boldsymbol{\theta}$ as vectors. For example, the parameter vector of a logistic regression classifiers in Equation (3.1) is $\boldsymbol{\theta} := |\boldsymbol{w}^\top \quad b|^\top$.

A large part of machine learning research is focused on understanding the properties of known hypothesis classes, and creating new and more powerful ones. Section 3.1 will introduce specific classes that are particularly relevant for this thesis.

Optimization

In order to find the hypothesis in $\mathcal{H}$ that best matches the concept c originating the dataset $\mathcal{D}$, it is necessary to define a means of comparison between different hypotheses. This can be done by, for example, comparing the responses predicted by the hypothesis under consideration with the expected responses in the dataset, and seeking a hypothesis which makes few mistakes. For this, we define a *loss function* $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ that provides a "penalty" for incorrect predictions, together with an *optimization procedure* that scans $\mathcal{H}$ and seeks the hypothesis f^* with the lowest loss:

$$f^* = \arg\min_{f \in \mathcal{H}} \mathcal{R}(f) = \arg\min_{f \in \mathcal{H}} \mathbb{E}_{(\boldsymbol{x},y) \sim p_{\mathcal{X}\mathcal{Y}}} \left[\ell(f(\boldsymbol{x}), y) \right] \tag{3.2}$$

Where the expectation is called the risk $\mathcal{R}(f)$ of f. In some cases, the loss ℓ that we are interested in minimizing cannot be used directly, e.g. because it would make the optimization too hard to solve, therefore an appropriate *surrogate loss* $\mathcal{L}$ is used instead. In case of classification, as is the logistic regression example in Equation (3.1), a commonly used risk function is the classification error $\ell(\hat{y}, y) := \mathbb{1}[\hat{y} \neq y]$, becoming the probability of mis-classification under the expectation of Equation (3.2), and a common surrogate loss is the cross-entropy, which can be justified by a probabilistic argument presented below in Section 3.1.3.

Since $p_{\mathcal{X}\mathcal{Y}}$ is unknown, the expectation in Equation (3.2) cannot be computed directly, therefore in practice we resort to estimating it using the dataset $\mathcal{D}$, which was assumed to be sampled from that distribution. The hypothesis $\hat{f}$ that minimizes the risk averaged on the dataset is thus defined as:

$$\hat{f} = \arg\min_{f\in\mathcal{H}} \hat{\mathcal{R}}(f) = \arg\min_{f\in\mathcal{H}} \frac{1}{n_l} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_l} \ell(f(\boldsymbol{x}), y) \tag{3.3}$$

where the average loss on the dataset is called the *empirical risk* $\hat{\mathcal{R}}(f)$ of f. This approach is known as *empirical risk minimization*, and many algorithms to solve this problem exist, their applicability depending on characteristics of the hypothesis class, the dataset, etc. Later, two examples will be presented: gradient descent (Section 3.2.1), and iterated least squares (Section 3.4.1).

3.1.2 Generalization

A crucial question is how to relate $\hat{\mathcal{R}}(\hat{f})$ to $\mathcal{R}(\hat{f})$ and $\mathcal{R}(f^*)$. In other words, understanding how much worse the hypothesis $\hat{f}$ that was found using a surrogate loss $\mathcal{L}$ on the dataset $\mathcal{D}$ is, compared to the best possible hypothesis f^* that minimizes the desired loss ℓ on the data distribution $p_{\mathcal{X}\mathcal{Y}}$. Theoretical results in the field are usually expressed with "probably approximately correct" statements that bound the difference in risks with high probability (Mohri et al., 2018):

$$p\left(\left|\mathcal{R}(\hat{f}) - \mathcal{R}(f^*)\right| \leq \epsilon\right) \geq 1 - \delta \tag{3.4}$$

where δ depends on ϵ, the size of the dataset, the "complexity" of $\mathcal{H}$, and whether a surrogate loss was used (Bartlett et al., 2006). Intuitively, complex hypothesis classes are able to accurately model a large number of datasets from many different distributions, and indeed one of the most common measures of complexity quantifies the ability of a hypothesis class to achieve low loss on datasets with random labels (Koltchinskii and Panchenko, 2000).

While complex hypothesis classes enable practitioners to explain a wide range of different phenomena, there are often several different hypotheses that explain a given dataset equally well, as measured by the loss function, especially when the dataset does not contain many samples in relation the the complexity of the hypothesis class. Moreover, it is often not desirable to select a hypothesis that explains the dataset "too well," because the observations may be corrupted by a certain amount of noise that should not be explained by the hypothesis. In fact, even though the optimization problem is always cast in terms of the empirical risk (Equation (3.3)) the actual goal of optimization is to minimize the true risk (Equation (3.2)), i.e., to find a hypothesis that *generalizes* well to new samples from the underlying data distribution (and, in advanced applications, to samples from other distributions as well). To achieve this, the principle of Occam's razor is applied, which suggests to choose the simpler hypothesis among a set of competing ones.

This is most commonly done via *regularization*, also called *structural risk minimization*, i.e., enforcing an upper limit on the complexity $\mathcal{C}$ of the hypothesis found through the use of a Lagrange multiplier:

$$\hat{f} = \arg\min_{f\in\mathcal{H}} \left[\hat{\mathcal{R}}(f) + \lambda \cdot \mathcal{C}(f)\right] \tag{3.5}$$

The complexity measure depends on $\mathcal{H}$, and is different from the complexity measure used for risk bounds (Equation (3.4)) because that does not result in actionable minimization problems. A common measure employed for suitable models is the L_p norm of the parameter vector, i.e., $\mathcal{C}(f_{\boldsymbol{\theta}}) = ||\boldsymbol{\theta}||_p^p$, with different values of p resulting in hypotheses with different properties. The best value for λ should guarantee good generalization on unseen data, therefore it is commonly chosen by resampling procedures, which repeatedly partition $\mathcal{D}$ into "training" and "validation" subsets, and use the former to solve Equation (3.5), and the latter to assess the performance of $\hat{f}$ on unseen data, eventually selecting the λ that resulted in the best average validation performance.

3.1.3 Probabilistic models

An important special case of the optimization procedure mentioned above is *maximum likelihood estimation*, which is applied when building probabilistic models of the data. This type of models assumes a parametric form for the (unknown) distribution of $y|\boldsymbol{x}$, and optimizes its parameters to minimize the difference, as quantified by the Kullbach-Leibler divergence, between the observed and assumed distributions. Let $c(y|\boldsymbol{x})$ denote this conditional distribution induced by the concept under consideration, and $p_{\boldsymbol{\theta}}(y|\boldsymbol{x})$ the distribution assumed by the model parameterized by $\boldsymbol{\theta}$. The optimal parameters $\boldsymbol{\theta}^*$ are, then, those that minimze the distance between the two distributions:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\Theta} \mathrm{D}_{\mathrm{KL}}(c(y|\boldsymbol{x})||p_{\boldsymbol{\theta}}(y|\boldsymbol{x})] = \arg\min_{\boldsymbol{\theta}\in\Theta} -\int_{\mathcal{X}\times\mathcal{Y}} \ln\frac{p_{\boldsymbol{\theta}}(y|\boldsymbol{x})}{c(y|\boldsymbol{x})} c(y|\boldsymbol{x})\mathrm{d}p_{\mathcal{X}\mathcal{Y}}(\boldsymbol{x},y) \tag{3.6}$$

Assuming that the dataset contains i.i.d. samples, and estimating the integral as an empirical average over the dataset, while discarding $c(y|\boldsymbol{x})$ in the denominator as it is constant given the dataset, results in:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \sum_{(\boldsymbol{x},y)\in\mathcal{D}} \ln p_{\boldsymbol{\theta}}(y|\boldsymbol{x}) = \arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}(\boldsymbol{\theta}) \tag{3.7}$$

where the summation is called the *likelihood* of the parameters $\boldsymbol{\theta}$, and this approach is thus known as *maximum likelihood*. Maximum likelihood estimation fits the empirical risk minimization framework in Equation (3.3) by using the negative log-likelihood as loss function, as shown in the rightmost term of Equation (3.7). Assuming that the model $p_{\boldsymbol{\theta}}(y|\boldsymbol{x})$ corresponds to the true conditional $c(y|\boldsymbol{x})$, the parameters $\hat{\boldsymbol{\theta}}$ estimated via Equation (3.7) converge in distribution to a multivariate Normal with expectation the true parameters $\boldsymbol{\theta}^*$:

$$\hat{\boldsymbol{\theta}} \sim \mathcal{N}\left(\boldsymbol{\theta}^*, -\mathbb{E}\left[\nabla^2_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right]\right) \tag{3.8}$$

where $\nabla^2_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$ is the Hessian of the negative log likelihood computed at $\boldsymbol{\theta}^*$ This result is extremely important to derive the uncertainty of the components in $\hat{\boldsymbol{\theta}}$, thus guiding practitioners in interpreting their models.

A probabilistic interpretation of logistic regression presented in Equation (3.1), for example, is to assume that the responses are generated from a Bernoulli distribution $\mathrm{Ber}(\cdot)$ whose parameter is proportional to the distance from the hyperplane defined by the model, i.e., $p_{\boldsymbol{\theta}}(y|\boldsymbol{x}) = \mathrm{Ber}(f_{\boldsymbol{\theta}}(\boldsymbol{x}))$, with $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sigma(\boldsymbol{x}^\top \boldsymbol{w} + b)$ and $\sigma(x) = (1 + e^{-x})^{-1}$ the logistic sigmoid function, used to squash this distance in the interval $(0, 1)$. The log-likelihood of this model is:

$$\begin{aligned} \log p_{\boldsymbol{\theta}}(y|\boldsymbol{x}) &= \begin{cases} \log(f_{\boldsymbol{\theta}}(\boldsymbol{x})) & \text{if } y = 1 \\ \log(1 - f_{\boldsymbol{\theta}}(\boldsymbol{x}])) & \text{if } y = 0 \end{cases} \\ &= y \log(f_{\boldsymbol{\theta}}(\boldsymbol{x})) + (1 - y) \log(1 - f_{\boldsymbol{\theta}}(\boldsymbol{x})) \end{aligned} \tag{3.9}$$

The negative log-likelihood is then used as loss function into Equation (3.3) or, more frequently, into Equation (3.5) with an L_1 or L_2 norm as penalty. Equation (3.9) is recognizable as the cross-entropy between the predictions and the true labels, meaning that empirical risk minimization with this loss is equivalent to seeking a decoder $f_{\boldsymbol{\theta}}$ that produces the shortest possible message length for events in the distribution $y|\boldsymbol{x}$ (MacKay, 2003).

3.2 Fundamental concepts in deep learning

The discussion until now was very general, and arguably most of the exciting things in machine learning occur when focusing on the hypothesis class $\mathcal{H}$. We already introduced logistic regression as an example hypothesis class that, despite its simplicity, is still very commonly used. In this section, we are going to introduce deep neural networks, particularly powerful hypothesis classes that are the drivers of a present scientific and economic revolution, with their performance raising important questions about the nature of human intelligence. The set of techniques that enable deep neural networks to learn so well is called deep learning, whose fundamental concepts are introduced in Goodfellow et al. (2016).

3.2.1 The deep learning blueprint

As a particular machine learning technique, deep learning can be defined in terms of the same blueprint we used in Section 3.1.1, starting from the data, to the model, to the optimization techniques.

Data

Deep neural networks excel at learning from unstructured data sources, such as images, audio, text, graphs, etc. These types of data posed considerable challenges to traditional machine learning methods, because they required practitioners to spend considerable efforts in developing *ad hoc feature engineering* methods that extract more meaningful signals from the data before a traditional machine learning method could be used. Deep neural networks are instead able to automatically identify meaningful features in the data through a hierarchy of flexible transformations that extracts more and more abstract and general patterns, until a final decision can be made. These pattern extractors can be formulated so that they are able to deal with any domain that presents some set of pre-determined symmetries (Bronstein et al., 2016), including images, graphs and point clouds. Graph data, for example, is frequently found in bioinformatics applications (Zhang et al., 2021) and plenty of other domains (Zhou et al., 2020), while point clouds are produced by LIDAR sensors in self-driving cars (Li et al., 2020b) and robotics applications (Liu et al., 2019; Guo et al., 2020).

Model

Neural networks were introduced in the 1950s as a model of neurons in the brain (Rosenblatt, 1958). A *perceptron*, like a real neuron, receives signals from d different input sources, and "fires" when the combined signal is larger than a threshold $-b$:

$$f(\boldsymbol{x}) = \mathbb{1}\left[\sum_{i=1}^{d} x_i w_i + b > 0\right] \tag{3.10}$$

Each source x_i is associated with a weight w_i that specifies how important the signal from that source is for the output of the neuron, and the neuron can learn by adjusting the weights to produce the desired output signal. While perceptrons were originally believed to enable machines to "walk, talk, see, write, reproduce itself and be conscious of its existence" (Olazaran, 1996), their inability to learn simple functions such as the exclusive-or (XOR) was quickly pointed out (Minsky and Papert, 1969) and led to their temporary demise. Soon thereafter, however, researchers realized that multiple "layers" of perceptrons, each receiving the output of perceptrons in the previous layer, and sending its output to perceptrons of the next layer, could model the XOR function, and considerably more complicated signals as well. In fact, it was theoretically proven that two layers of neurons are enough to learn any function subject to certain technical constraints (Hornik et al., 1989). Only recently, however, neural networks such as those became popular, propelled by computational advances that allowed them to learn from tens of thousands of examples, vastly outperforming competing approaches (Krizhevsky et al., 2017). The field assumed its popular name of *deep learning* as "deep" models with a large number of "narrow" layers were found to outperform "shallow" models with "wide" layers, after a series of tricks were introduced (He et al., 2016; Glorot et al., 2011; He et al., 2015; Kingma and Ba, 2014). This is however far from an universal principle, as excessively deep models with thousands of layers fell out of fashion after some years, in favor of models that carefully balance width and depth depending on the number of training samples (Tan and Le, 2019).

Multi-layer perceptrons are the simplest example of *feedforward* neural networks. This type of neural network uses multiple layers of neurons to progressively learn more and more abstract concepts, until a final output prediction is provided. A feedforward neural network can be formalized as the composition of L functions, each corresponding to a layer of the network:

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \left(f^{(L)}_{\boldsymbol{\theta}_L} \circ \ldots \circ f^{(1)}_{\boldsymbol{\theta}_1}\right)(\boldsymbol{x}) \tag{3.11}$$

Each of these layers has its own vector of parameters, collectively grouped into $\boldsymbol{\theta}$, and the transformation they apply need not be the same, although a few common variations are used in the majority of models. Each layer operates on tensors, and can alter the dimensionality of its input, until the final layer produces a vector of the same dimensionality as the items in $\mathcal{Y}$. Multilayer perceptrons are composed of dense, or fully-connected, layers:

$$f^{\text{dense}}(\boldsymbol{x}) := \phi\left(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}\right) \tag{3.12}$$

which apply an affine transformation to $\boldsymbol{x} \in \mathbb{R}^d$ using a matrix $\boldsymbol{W} \in \mathbb{R}^{d' \times d}$ and a bias vector $\boldsymbol{b} \in \mathbb{R}^{d'}$, together comprising the parameters $\boldsymbol{\theta}$, followed by a non-linearity, ϕ, also called *activation function*. Typical choices for ϕ include the sigmoid, the hyperbolic tangent, the ReLU (Glorot et al., 2011) or its variations (Maas et al., 2013; Hendrycks and Gimpel, 2016), and many others.

Another common transformation is the discrete convolution of the input $\boldsymbol{x}$ with a set of J (one-dimensional, in this example) filters of size K:

$$f^{\text{conv}}(\boldsymbol{x})_{ij} := \phi\left(\sum_{k=1}^{K} x_{i+k-1}w_{jk} + b_j\right) \quad \forall 1 \leq j \leq J, 1 \leq i \leq d-K+1 \tag{3.13}$$

Convolutional layers (LeCun et al., 1989) are used to identify a certain pattern regardless of its position in the input sequence, and are especially useful in computer vision, audio processing, and sequence modeling both for natural language and genomics, although they are being superseded by more powerful alternatives such as vision transformers (Dosovitskiy et al., 2021). Convolutional layers are usually interleaved with *pooling* layers, that reduce the dimensionality of their inputs by aggregating features in the same neighborhood. For example, a (one-dimensional) max pooling layer of size K and stride $s \geq 1$ maps each "window" of K elements to its maximum, considering windows separated by s elements:

$$f^{\text{maxpool}}(\boldsymbol{x})_i := \max\left\{x_{(i-1)s+1}, \ldots, x_{\max\{d,(i-1)s+K\}}\right\} \quad \forall 1 \leq i \leq \lceil d/s \rceil \tag{3.14}$$

Pooling layers do not have parameters, but are still very useful to infuse a degree of location invariance to neural networks using convolutional layers. One last example is the graph convolution (Kipf and Welling, 2017), that operates on graphs by aggregating the feature vectors corresponding to adjacent nodes:

$$f^{\text{GCN}}(\boldsymbol{X}) := \phi\left(\boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{X}^\top + \boldsymbol{b}\boldsymbol{1}^\top\right) \tag{3.15}$$

where $\boldsymbol{A} \in \mathbb{R}^{m\times m}$ is the adjacency matrix of the graph, having m vertices and self-connections (i.e., $A_{ii} = 1$ for every i), $\boldsymbol{D} \in \mathbb{R}^{m\times m}$ is a diagonal matrix of node degrees with $D_{ii} = \sum_j A_{ij}$, and $\boldsymbol{W} \in \mathbb{R}^{m\times d}$ and $\boldsymbol{b} \in \mathbb{R}^m$ the usual layer weights and biases. Unlike the previous examples, which operated on column vectors, the input $\boldsymbol{X} \in \mathbb{R}^{m\times d}$ in Equation (3.15) is a matrix whose rows contain feature vectors for each node, and the transformation can be seen as a specific type of *message passing* layer (Gilmer et al., 2017), where the features of each node are transformed by aggregating the feature vectors of its neighbors.

Many other types of layer exist, and deep learning researchers continuously come up with new transformations and ways of combining them. Notable mentions are long-short term memory networks (Hochreiter and Schmidhuber, 1997) and Transformers (Vaswani et al., 2017) to learn from sequential data, ResNets (He et al., 2016) for vision tasks, U-Nets (Ronneberger et al., 2015) for image segmentation tasks, Mask-R CNN for object detection (He et al., 2017), and countless others.

Optimization

Following the framework introduced in Section 3.1, fixing the number and type of layers in Equation (3.11) results in a well-defined hypothesis class $\mathcal{H}$; the remaining question is, then, how to solve the empirical risk minimization problem of Equation (3.3). Most of the algorithms commonly used are variations of a simple iterative procedure that, at each iteration t, modifies the current parameters $\boldsymbol{\theta}^{(t)}$ by finding the change that results in the largest decrease in loss. Such changes, when small enough, correspond to the negative gradient of the loss function, which is usually approximated using only a small *batch* of examples $\mathcal{B}^{(t)} \subset \mathcal{D}$. Besides reducing the required computational resources, using small batches aids generalization (Keskar et al., 2016). All in all, this procedure is called *gradient descent* and is summarized in the following relation:

$$\boldsymbol{\theta}_l^{(t+1)} := \boldsymbol{\theta}_l^{(t)} - \eta \nabla_{\boldsymbol{\theta}_l} \left[\sum_{(\boldsymbol{x},y)\in\mathcal{B}^{(t)}} \ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y) \right] \qquad \forall 1 \leq l \leq L \tag{3.16}$$

The procedure is initialized from a randomly chosen initial vector of parameters $\boldsymbol{\theta}_l^{(0)}$, and the *learning rate* $\eta > 0$ controls how quickly the weights are adapted. Note that Equation (3.16) is applied to the parameters of each layer of the network separately, whose gradients are computed using a recursive procedure called *backpropagation* (Rumelhart et al., 1985). A vast amount of research is done to improve the stability and convergence speed of Equation (3.16), including better initialization for $\boldsymbol{\theta}^{(0)}$ (He et al., 2015), dynamic learning rate that adapts over time (Kingma and Ba, 2014), choosing batches with the most informative examples (Hoffer et al., 2020), efficiently including curvature information (Martens and Grosse, 2015), etc., as well as developing theoretical foundations and understanding (Sun, 2019).

3.2.2 Deep uncertainty quantification

Recall from Section 3.1 that it is critical to understand how well the empirical risk minimizer generalizes to unseen examples, and how much worse that is compared to the best estimator in $\mathcal{H}$. While the field of statistical learning theory provides high probability, worst-case bounds, as in Equation (3.4), statistical inference is concerned with finding a range of likely hypotheses given the observed dataset. Since the dataset is a noisy sample from the data distribution, there is some unavoidable uncertainty in any hypothesis found using exclusively this dataset. Many sources of uncertainty affect the quality of predictive models, ranging from data issues such as noise and non-representativeness, to inference issues such as training and hyperparameter tuning, among others. Nonetheless, it is necessary to accurately quantify the uncertainty of predictive models in order to ensure that their predictions are fair, can be trusted, and are safe to use – especially in domains such as medicine and autonomous driving, where real harm could be done if real-world decisions are based on wrong, uncertain predictions (see, e.g., Begoli et al., 2019; Michelmore et al., 2020; Verma and Rubin, 2018).

Broadly speaking, two types of uncertainty are considered: *aleatoric* uncertainty is caused by noise in the data collection procedures, while *epistemic* uncertainty stems from the modeling approach that is used (Hüllermeier and Waegeman, 2021). While epistemic uncertainty can, in principle, be eliminated by collecting enough data, aleatoric uncertainty cannot be reduced unless cleaner data is collected. The asymptotic result of Equation (3.8) is an example of epistemic uncertainty, and for simple models it is possible to use this result to derive useful expressions for both aleatoric and epistemic uncertainty, as described later in Section 3.4.1. The large number of parameters in deep neural networks, however, makes them considerably harder to study, preventing, for example, direct computation of the Hessian required in Equation (3.8). Model-agnostic numerical methods obtain samples from the parameter's distributions (Neal, 2011; Hoffman et al., 2014), as opposed to its analytical form, are also hampered by excessive computational complexity, and are essentially inapplicable to real-world deep learning models (Izmailov et al., 2021b).

For these reasons, a wide range of approximate methods for uncertainty quantification have been proposed recently. From a high level perspective, these can be partitioned into those focusing on *weight-space* uncertainty and those focusing on *functional* uncertainty: while methods in the former class (e.g., Blundell et al., 2015; Welling and Teh, 2011; Maddox et al., 2019; Daxberger et al., 2021; Lakshminarayanan et al., 2017) try to derive a distribution for the model's parameters, methods in the latter class (e.g. Gal and Ghahramani, 2015; Wilson et al., 2015; Alaa and van der Schaar, 2020) only focus on deriving a distribution for the model's predictions. Weight-space uncertainty methods, in order to quantify the uncertainty in the predictions, usually require repeatedly sampling from the posterior distribution of the weights, and running a separate forward pass for each sample, aggregating the model's predictions into their mean and variance for a Gaussian approximation (for example). Another distinction is between *ad hoc* and *post hoc* methods, where the former type of method estimates uncertainty with specific procedures as the network is being trained, and the latter derives the uncertainty after the training process is completed.

Given all these competing approaches (Abdar et al., 2021; Gawlikowski et al., 2021), it is not trivial to understand when they work well and when they do not, with each method having their own strength and weaknesses and considerable effort spent on benchmarks and comparisons (Wilson and Izmailov, 2020; Izmailov et al., 2021a; Abdar et al., 2021; Wang and Yeung, 2016; Gawlikowski et al., 2021), as well as the development of theoretical foundations of deep learning (Roberts et al., 2022; Bartlett et al., 2021).

3.2.3 Semi-supervised deep learning

Up until now, the discussion always assumed the presence of labeled examples with which to compute the loss, comparing the model's predictions with the response expected for the given input pattern. In many practical applications, however, labeled data is scarce, and most examples in the dataset do not have an associated label. Because of the practical relevance of this situation, many methods were developed to improve a model leveraging the signal in the unlabeled examples, leading to what is called *semi-supervised learning*.

All semi-supervised learning methods rely on at least one of three assumptions regarding the underlying data distribution $p_{\mathcal{X}\mathcal{Y}}$ (Van Engelen and Hoos, 2020): (1) the smoothness assumption states that two “similar” samples $\boldsymbol{x}^{(i)}$ and $\boldsymbol{x}^{(y)}$ should have “similar” labels $y^{(i)}$ and $y^{(j)}$, (2) the low-density assumption, according to which, in classification problems, relatively few samples lie close to the separating boundary between different classes in $\mathcal{Y}$, and (3) the manifold assumption states that samples from $p_{\mathcal{X}\mathcal{Y}}$ lie on or close to a manifold with lower dimensionality compared to that of $\mathcal{X}$. By using these assumptions, semi-supervised learning methods can use unlabeled samples from the marginal data distribution on $\mathcal{X}$ to infer information about their labels following, broadly speaking, three different directions (Chapelle et al., 2006). Generative models learn a distribution $p(\boldsymbol{x}|y)$ conditioned on the response, which is marginalized in order to estimate the density $p(\boldsymbol{x})$ at each unlabeled sample. The simplest example of this kind of methods is a mixture of Gaussians fit through the expectation-maximization algorithm (Dempster et al., 1977). A second class of models explicitly leverages the low-density assumption to push the decision function away from labeled and unlabeled samples, for example by extending the maximum-margin mechanism of support vector machines to unlabeled data (Collobert et al., 2006). The last class of methods represents the data through a graph whose nodes are labeled and unlabeled examples in the dataset, and edges are weighted by the distance of the samples they connect. By computing the distance of two arbitrary data points as the smallest total edge weight of all paths connecting the two data points, these methods are implicitly based on the manifold assumption and leverage its structure to propagate information to nearby points (Xiaojin and Zoubin, 2002).

An alternative taxonomy (Van Engelen and Hoos, 2020) divides semi-supervised learning approaches into three categories named intrinsically semi-supervised methods, unsupervised preprocessing methods, and wrapper methods. Intrinsically semi-supervised methods extend the loss function to explicitly handle unlabeled data points. Wrapper methods, most commonly called *pseudo-labeling* methods, leverage a traditional supervised learning method trained on the original labeled data to generate additional "pseudo-labeled" data from the unlabeled portion of the dataset, so that a better classifier can be trained. Unsupervised preprocessing methods transform, aggregate, or summarize the unlabeled data in some way to improve a traditional supervised learning method that is applied in a second step to the labeled portion of the dataset. For example, the unlabeled data can be used to jointly reduce the dimensionality of the labeled data, cluster it, extract useful features, etc. The contribution of this thesis leverage methods from the two latter categories, which are thus covered in more detail in the remaining of this section.

Pseudo-labeling

Pseudo-labeling (Yarowsky, 1995) is an iterative process composed of two stages: an initial training stage, where a fully supervised model is learned on all the available labeled data, and a second pseudo-labeling stage, where the model's predictions are used to generate new *pseudo-labels* for a subset of the unlabeled data. In the next iteration, a fully supervised model is trained jointly on the original labeled data together with the pseudo-labeled data, assuming that the pseudo-labels are correct. This procedure is repeated several times, until either the entire unlabeled dataset is labeled, or some other stopping condition is reached. If the pseudo-labels are indeed correct, the performance of the supervised model increases at every round, while incorrect pseudo-labels gradually degrade the performance of the classifier due to a sort of *confirmation bias*. To avoid this, the examples to pseudo-label are determined based on on some form of confidence measure. Different choices on the way the examples to be pseudo-labeled are selected, how they are used in later iterations, and the stopping criteria exist (Triguero et al., 2015).

Pseudo-labeling was adapted to deep neural networks by Lee et al. (2013), who assigned a progressively increasing weight to the loss of pseudo-labeled samples to reflect the increasing confidence of the classifier as training proceeds. The issue of confirmation bias in deep semi-supervised learning was studied in detail by Arazo et al. (2020), where several techniques to reduce the typical overconfidence of deep neural networks were applied (Zhang et al., 2018; Tanaka et al., 2018; Grandvalet and Bengio, 2004), and distilled into a working formula by Rizve et al. (2021) through the use of explicit uncertainty quantification via Monte Carlo dropout (Gal and Ghahramani, 2016).

Contrastive learning

Contrastive learning is an example of the "unsupervised preprocessing" type of semi-supervised learning algorithms. It leverages unlabeled data to train a neural network *encoder* that is able to compress samples into good *representations*, which can be subsequently used to learn a predictive model using the labeled data (Jaiswal et al., 2020). Contrastive learning is thus composed of an initial phase where representations are learned using a *pretext* task, and a second phase where these representations are used to learn solve the actual *downstream* task of interest. Ideally, the representations are (1) expressive and of low dimensionality, (2) only capture abstract and high-level concepts that are useful for the downstream task and invariant to certain changes in the input data, and (3) disentangle latent factors of variations to promote their re-use and interpretability (Le-Khac et al., 2020). Contrastive learning was originally introduced as a more biologically-plausible alternative (Becker and Hinton, 1992) to the backpropagation algorithm (Section 3.2.1), and similar concepts leveraged shortly thereafter for unsupervised signature identification (Bromley et al., 1993).

Contrastive learning was popularized by the formula of Chen et al. (2020b), which proposed a general structure for pretext tasks composed of three steps: (1) creating one or more augmentations of each input sample, appearing different but being semantically identical, (2) using the encoder to derive their representations, and (3) using a *contrastive loss* to push the representations of a sample and its augmentations to be similar among each other and dissimilar from the augmentations of other samples. The "similarity" of two representations is generally quantified as the cosine similarity, while the type of augmentations that should be applied depends on the data and downstream task. A commonly used contrastive loss is the infoNCE (Oord et al., 2018), which, given a batch $\mathcal{B} \subseteq \mathcal{D}_u$ of unlabeled examples, a random augmentation $t : \mathcal{X} \rightarrow \mathcal{X}$ sampled from a suitable distribution, an encoder $f : \mathcal{X} \rightarrow \mathbb{R}^k$, and a similarity function $s : \mathbb{R}^k \times \mathbb{R}^k \rightarrow \mathbb{R}^+$, is computed as:

$$\mathcal{L}_{\text{infoNCE}} = -\sum_{\boldsymbol{x}^{(i)} \in \mathcal{B}} \log \frac{e^{s\left(f(\boldsymbol{x}^{(i)}), f(t(\boldsymbol{x}^{(i)}))\right)}}{e^{s\left(f(\boldsymbol{x}^{(i)}), f(t(\boldsymbol{x}^{(i)}))\right)} + \sum_{\boldsymbol{x}^{(j)} \in \mathcal{B} : \boldsymbol{x}^{(j)} \neq \boldsymbol{x}^{(i)}} e^{s\left(f(\boldsymbol{x}^{(i)}), f(\boldsymbol{x}^{(j)})\right)}} \tag{3.17}$$

This loss compares the representations of $\boldsymbol{x}^{(i)}$ and $t(\boldsymbol{x}^{(i)})$ against the representations of every other sample in the batch, which are assumed to belong to different classes compared to $\boldsymbol{x}^{(i)}$ and thus require dissimilar representations. However, it is possible that two samples actually belong to the same class, thus the second term of the denominator in Equation (3.17) introduces a certain degree of *sampling bias*. Based on this observation, Chuang et al. (2020a) proposed a correction that removes this bias, given the probability π that two unlabeled samples belong to the same class. Given M random augmentations $t_1, \dots, t_M$, and using $N = |\mathcal{B}| - 1$, the biased term in Equation (3.17) should be replaced by:

$$N \max\left\{\frac{1}{e}, \frac{1}{1-\pi}\left(\frac{1}{N}\sum_{\boldsymbol{x}^{(j)} \in \mathcal{B} : \boldsymbol{x}^{(j)} \neq \boldsymbol{x}^{(i)}} e^{s\left(f(\boldsymbol{x}^{(i)}), f(\boldsymbol{x}^{(j)})\right)} - \pi \frac{1}{M}\sum_{k=1}^{M} e^{s\left(f(\boldsymbol{x}^{(i)}), f(t_k(\boldsymbol{x}^{(i)}))\right)}\right)\right\} \tag{3.18}$$

Current research in contrastive learning focuses on finding good augmentation strategies for vision (Misra and Maaten, 2020; Qian et al., 2021), audio (Oord et al., 2018), and language (Devlin et al., 2018), improving training techniques (Chuang et al., 2020b), and understanding its theoretical properties (Garrido et al., 2022; Saunshi et al., 2019; Arora et al., 2019). The theoretical analysis of Chuang et al. (2020a), for example, provides generalization bounds for downstream classification tasks when the infoNCE loss modified with Equation (3.18) is used to learn representations.

3.3 Positive-unlabeled learning

A challenging setting in semi-supervised learning deals with the complete absence of negatively-labeled examples, which tends to arise in several applications of great practical interest, from bioinformatics (Li et al., 2021) to business and security (Jaskie and Spanias, 2019). In these scenarios, constraints on data collection or labeling make gathering negative examples very difficult, or even impossible. However, it is still relatively easy collect positive examples and large number of unlabeled samples, which contain both positives and negatives examples without any label to differentiate them. This scenario is known in the machine learning literature as positive-unlabeled learning, and differs from one-class classification (Ruff et al., 2018; Li et al., 2010) as negative examples are also available, albeit without label. Even when negative examples are available, if their number is too limited, it is advisable to preserve them exclusively for evaluating the model's performance, rather than for training (Oliver et al., 2018), as evaluation with positive-unlabeled data poses some challenges.

3.3.1 Data

Let $\mathcal{Y} = \{-1, 1\}$, and the data distribution be factorized as follows:

$$p_{\mathcal{X}\mathcal{Y}}(\boldsymbol{x}, y) = \pi \cdot p(\boldsymbol{x}|y = 1) + (1 - \pi) \cdot p(\boldsymbol{x}|y = -1) \tag{3.19}$$

where $\pi := p(y = 1)$ is the *class prior*, i.e., the prior probability of an example belonging to the positive class. Positive-unlabeled learning is, then, a semi-supervised binary classification task with the complication that $y = 1$ for all $\boldsymbol{x} \in \mathcal{D}_l$, and the unlabeled dataset $\mathcal{D}_u$ contains samples from the marginal $p_{\mathcal{X}}(\boldsymbol{x}) = p_{\mathcal{X}\mathcal{Y}}(\boldsymbol{x}, y = -1) + p_{\mathcal{X}\mathcal{Y}}(\boldsymbol{x}, y = 1)$.

There are two similar, but different, settings under which the positive examples in $\mathcal{D}_u$ are assumed to be generated. In the *single-training set* scenario, a single dataset is sampled from $p_{\mathcal{X}\mathcal{Y}}$, and a fraction of the positive examples is unlabeled according to some mechanism. In the *case-control* scenario, instead, the positives and unlabeled are assumed to come from separate sources. In both cases, the positives in $\mathcal{D}_l$ are sampled from $p(\boldsymbol{x}|y = 1)$, and the unlabeled in $\mathcal{D}_u$ from $p_{\mathcal{X}\mathcal{Y}}$, but whereas in the case-control scenario it is possible to control the number of available positives separately by gathering more samples from their respective source, in the single-training set scenario gathering more positives also results in more unlabeled samples. The following discussion assumes the single-training set scenario, as it has received considerably more attention in the literature (Bekker and Davis, 2020).

Unlabeled examples are either truly negative, or positives that were not labeled. Learning from this type of data, therefore, requires further assumptions on the labeling mechanism. Under the *selected completely at random* (SCAR) assumption, the probability that a positive sample is labeled is a constant, while under the *selected at random* assumption the probability that a positive is labeled depends on the value of its features $\boldsymbol{x}$. The SCAR assumption enables traditional classification methods to be applied to the positive-unlabeled learning setting by predicting whether a sample is labeled or unlabeled, as the probability that a sample is positive is just a constant factor from the probability that it is labeled (Elkan and Noto, 2008).

3.3.2 Model

The majority of positive-unlabeled learning methods make use of the SCAR assumption. In this case, assuming that the class prior π is known, it is possible to directly derive a risk estimator to be minimized (du Plessis et al., 2014). From Equation (3.19), we obtain an expression for the risk of a classifier f factorized on the positive and negative classes:

$$\mathcal{R}(f) = \pi\mathbb{E}_{\boldsymbol{x}|y=1}[\ell(f(\boldsymbol{x}), 1)] + (1-\pi)\mathbb{E}_{\boldsymbol{x}|y=-1}[\ell(f(\boldsymbol{x}), -1)] \tag{3.20}$$

The problem in positive-unlabeled learning is that there are no negative example to estimate the negative risk, i.e., the second term on the right-hand side of Equation (3.20). The distribution of negatives can be obtained from Equation (3.19) as $(1-\pi)p(\boldsymbol{x}|y=-1) = p_{\mathcal{X}\mathcal{Y}}(\boldsymbol{x}, y) - \pi p(\boldsymbol{x}|y=1)$, and with this the risk on negative data is:

$$\begin{align}
&(1-\pi)\mathbb{E}_{\boldsymbol{x}|y=-1}[\ell(f(\boldsymbol{x}), -1)] \tag{3.21}\\
&\quad= \int_{\mathcal{X}} \ell(f(\boldsymbol{x}), -1)(1-\pi)p(\boldsymbol{x}|y=-1)\mathrm{d}\boldsymbol{x} \tag{3.22}\\
&\quad= \int_{\mathcal{X}} \ell(f(\boldsymbol{x}), -1)\left[p(\boldsymbol{x}) - \pi p(\boldsymbol{x}|y=1)\right]\mathrm{d}\boldsymbol{x} \tag{3.23}\\
&\quad= \left[\int_{\mathcal{X}} \ell(f(\boldsymbol{x}), -1)p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right] - \left[\pi\int_{\mathcal{X}} \ell(f(\boldsymbol{x}), -1)p(\boldsymbol{x}|y=+1)\mathrm{d}\boldsymbol{x}\right] \tag{3.24}\\
&\quad= \mathbb{E}_{\boldsymbol{x}}[\ell(f(\boldsymbol{x}), -1)] - \pi\mathbb{E}_{\boldsymbol{x}|y=1}[\ell(f(\boldsymbol{x}), -1)] \tag{3.25}
\end{align}$$

Notably, Equation (3.25) can be estimated from positive (second term) and unlabeled (first term) data only. Plugging back into Equation (3.20), we finally get the risk for positive-unlabeled data:

$$\begin{aligned}
\mathcal{R}_{\mathrm{PU}}(f) &= \pi\mathbb{E}_{\boldsymbol{x}|y=1}[\ell(f(\boldsymbol{x}), 1)] + \mathbb{E}_{\boldsymbol{x}}[\ell(f(\boldsymbol{x}), -1)] - \pi\mathbb{E}_{\boldsymbol{x}|y=1}[\ell(f(\boldsymbol{x}), -1)]\\
&= \pi\mathcal{R}_p^1(f) + \mathcal{R}_u^{-1}(f) - \pi\mathcal{R}_p^{-1}(f)
\end{aligned} \tag{3.26}$$

du Plessis et al. (2014) showed that Equation (3.26) is a consistent and unbiased estimator of the true risk in Equation (3.20), and can thus used to train a binary classifier using positive-unlabeled data only, as long as the loss function ℓ satisfies the following *symmetry condition*:

$$\ell(\hat{y}, 1) + \ell(\hat{y}, -1) = 1 \tag{3.27}$$

A common loss satisfying this condition is the sigmoid loss $\ell_\sigma(\hat{y}, y) := \sigma(-y\hat{y})$.

Kiryo et al. (2017) noted that highly flexible hypothesis classes such as deep neural networks are in practice able to "overfit" the risk in Equation (3.25) and make it negative, even though it is theoretically bound to be non-negative. Their solution simply forces this term to be non-negative, giving rise to the nnPU loss that forms the basis of many a positive-unlabeled learning algorithms (Kiryo et al., 2017):

$$\mathcal{R}_{\mathrm{nnPU}}(f) = \pi\mathcal{R}_p^1(f) + \max\left\{0, \mathcal{R}_u^{-1}(f) - \pi\mathcal{R}_p^{-1}(f)\right\} \tag{3.28}$$

The careful reader surely did not miss the similarity between the second term of Equation (3.28) and the denominator of the debiased contrastive loss in Equation (3.18); this is not coincidental, as both take into account an unknown portion of π samples as positive, or similar, among all the unlabeled. Estimating such risks requires, in practice, knowledge of π, for which several estimation methods were proposed (Christoffel et al., 2016; Elkan and Noto, 2008; Zeiberg et al., 2020; Bekker and Davis, 2018; Ramaswamy et al., 2016; Garg et al., 2021). While some present works still assume a given class prior (Chen et al., 2020c; Zhao et al., 2022; Hammoudeh and Lowd, 2020; Acharya et al., 2022), an emerging research stream tries to avoid this two-step estimation procedure, and to develop methods that do not rely on its estimation (Chen et al., 2020a; Hu et al., 2021; Gong et al., 2021). Another research direction tries to handle biases in the data selection (Kato et al., 2019; Hsieh et al., 2019), and particularly relevant for the contributions of this thesis is the work of Su et al. (2021), that introduced a modified risk estimator more suited to imbalanced distributions with very low π. This risk estimator is based on re-weighting the positives such that their cumulative loss equals a portion π', usually set to 1/2, of the total loss:

$$\mathcal{R}_{\text{imbnnPU}} = \pi'\mathcal{R}_p^1(f) + \max\left\{0, \frac{1-\pi'}{1-\pi}\mathcal{R}_u^{-1}(f) - \frac{(1-\pi')\pi}{1-\pi}\mathcal{R}_p^{-1}(f)\right\} \tag{3.29}$$

Regardless of the chosen risk estimator, models for positive-unlabeled data are learned using the same algorithms as any other deep learning model (Section 3.2.1).

3.3.3 Evaluation

The lack of negative data complicates not only learning, but also evaluation of classifiers learned on positive-unlabeled data. For example, while recall can be estimated from positive samples only, by computing the proportion of positive predictions in this dataset, precision cannot be estimated directly, as it is impossible to know the fraction of truly positive samples among those that are predicted so. Lee and Liu (2003) proposed a metric based on positive-unlabeled data that behaves similarly to the F-score, in the sense that it is large when both precision and recall, evaluated on the true labels, are large, and is small when either one is small. Menon et al. (2015) showed that the balanced error rate and the area under the ROC curve (AUC) are the only two performance measures that are immune to a particular type of label corruption that includes positive-unlabeled learning, and can be optimized even when the corruption parameter (π, in this case) is not known. Jain et al. (2017) focused on AUC estimation, and showed that the AUC on positive-unlabeled data is a lower bound for the AUC on clean data, and that the latter can be derived from the former with knowledge of the true class prior or an estimation thereof. In a similar vein, Ramola et al. (2018) showed how to use the true class prior to correct the (balanced) accuracy, F-score and Matthew's correlation coefficient.

The practical implication of these results is that, when the desired performance metric is either the AUC or the balanced error rate, it is possible to optimize that metric treating the unlabeled samples as negatives, and the resulting classifier is also optimal under the same metric computed using the true positive and negative labels from the given dataset. Importantly, this procedure enables practitioners to treat π as a hyperparameter, and optimize it by choosing the value that leads to the highest AUC or balanced error rate.

3.3.4 Connection to precision medicine and contributions

A large part of artificial intelligence applications to cancer immunotherapies consist in learning to predict certain events of the antigen processing pathway (Section 2.2), most importantly proteasomal cleavage, TAP transport, MHC binding and presentation, and T cell activation. Each of these events can be investigated in isolation by collecting data through *in vitro* experiments performed in the wet-lab. Such experiments are, however, expensive, time-consuming, and result in low quantities of data that do not necessarily reflect the entire spectrum of biological mechanisms occurring *in vivo* sufficiently accurately. On the other end of the spectrum, recent advances in high-throughput technologies such as liquid chromatography tandem mass spectrometry (LC-MS/MS) made it possible to collect large numbers of MHC-presented peptides that underwent the entire antigen processing pathway (Caron et al., 2015). Such technologies only detect a small portions of MHC ligands presented on the cells' surface, making positive-unlabeled learning the natural formalization of learning from this type of data. In fact, the most widely used MHC binding predictor (Reynisson et al., 2020) follows the case-control data generating process by creating artificial negative examples from the entire human genome, and leverages pseudo-labeling (Section 3.2.3) to learn from ambiguous data points (Alvarez et al., 2019). Proteasomal cleavage predictors are similarly trained with a mixture of positive examples derived from MHC ligands and synthetically-generated negatives (Keşmir et al., 2002).

In this thesis, two contributions to positive-unlabeled learning are presented. First, in Section 5.1, we propose a model-agnostic positive-unlabeled learning framework based on pseudo-labeling (Section 3.2.3), and using epistemic uncertainty (Section 3.2.2) for the selection process, showing its benefits on imbalanced datasets, and applying it to the problem of proteasomal cleavage prediction (Section 4.4). Second, in Section 5.2, we combine positive-unlabeled learning with contrastive representation learning (Section 3.2.3), showing that such representations are highly beneficial for the downstream classification task.

3.4 Semi-structured regression

Deep neural networks (Section 3.2) can be described as "unstructured" models, since they excel at handling unstructured, non-tabular data types such as images, audio, text, etc., and struggle dealing with tabular data (Shwartz-Ziv and Armon, 2022), where earlier techniques still excel.

While there is no consensus on the definitions of tabular and non-tabular data, a possible heuristic criterion is related to the "shape" of the data, in the sense that tabular data is often represented by vectors and non-tabular data by tensors, however this need not be the case in general. Another intuitive distinction of the two relies on the type of modeling approaches that make sense: linear models can be used for tabular data, but make no sense on non-tabular data, where neural networks are the model of choice (nowadays). While practitioners certainly have an intuitive notion of what model makes sense for a given datasets, a formal distinction could be based on the underlying geometry of the data, or absence thereof: non-tabular data could be characterized as possessing certain intrinsic symmetries that affect the input features without affecting the response, while tabular data does not have any symmetry. Formally, a decision rule $f : \mathcal{X} \rightarrow \mathcal{Y}$ is invariant to a transformation $g : \mathcal{X} \rightarrow \mathcal{X}$ if we have $f(x) = f(g(x))$ for every x. For tabular data, no such transformation exists, while for non-tabular data there is one or more transformations that do not affect the response. A typical example is convolutional neural networks learning from image data: the sliding window design of convolutions in Equation (3.13) applies the same filter to all parts of the image, allowing the network to detect the same patterns regardless of where they occur in the image, thus making the network invariant to spatial shifts such as $g(x_{ij}) = x_{i-1,j-1}$. In tabular data, each covariate is associated to a well-specified feature of the input sample, for example, age or sex of a person, and any sort of modification to this covariate results in a new, different sample, with a potentially different response.

In certain domains, however, it may be desirable or required to model the data such that certain modifications do not influence the response: for example, when designing a system to perform automated decisions related to people, it is often the case that such decisions should not depend on protected attributes such as age, sex, race, etc. In these cases, the hypothesis class under consideration should be restricted to decision rules that are invariant to variations to one or more of these attributes, even if the dataset at hand does not exhibit such property. Geometric deep learning (Bronstein et al., 2016) studies how to design neural networks to be invariant (i.e., $f(g(x)) = f(x)$) or equivariant (i.e., $f(g(x)) = g(f(x))$) to transformations such as rotations and translations, while the field of fairness is concerned with alternative definitions of invariance that are more suited to handle societal issues (Mehrabi et al., 2021; Corbett-Davies and Goel, 2018). Therefore, we posit that the distinction between tabular and non-tabular data depends on the input space itself as well as the concept under consideration.

3.4.1 Structured models

Learning from tabular data is very commonly approached through linear models, as they are flexible enough to model complex responses and include non-linear covariate effects, while being easily interpretable thanks to distributional statements on their parameters. Importantly, they make it possible to infer the distribution of the estimated parameters, enabling practitioners to understand in a principled manner whether and how strongly each covariate influences the response, unlike more flexible class of models. Wood (2017) presents a comprehensive treatment of the concepts summarized in this section.

Data

Generalized additive models (GAMs; Hastie and Tibshirani, 1990) are one of the most powerful methods to model tabular data. They extend ordinary linear models by allowing responses to follow a *general* distribution in the exponential family, and by modeling smooth, non-linear covariate effects through *additive* functions. Further extensions (GAMLSS; Rigby and Stasinopoulos, 2005) remove the exponential family restriction and allow responses to follow distributions with an arbitrary number of parameters.

The exponential family includes many popular distributions, including Gaussian, Poisson, Binomial, Gamma, etc., that are suitable to model responses that occur in many practical applications of interest such as binary data, proportions, and counts. For example, ecology and evolution studies include breeding success, infection status, mortality rates, number of offsprings, etc. (Bolker et al., 2009), engineering applications include modeling failure and wear rates (Myers et al., 2012), and computational biology datasets frequently contain counts (Luecken and Theis, 2019; Hu et al., 2012).

A distribution $\mathcal{D}$ belongs to the exponential family if its probability density function $p_\mathcal{D}$ can be written as

$$p_\mathcal{D}(y) = \exp\left[(y\tau - b(\tau))/a(\phi) + c(y,\phi)\right] \tag{3.30}$$

where a, b and c are arbitrary functions with the reals as both domain and codomain, and ϕ and τ are real constants respectively called the *scale* and *canonical*[1] parameters of $\mathcal{D}$. Note that $\mathbb{E}[y] = b'(\tau) =: \mu$. To ease the notation in the results presented later, we also define a function $V(\mu) = b''(\tau)/\omega$ such that the variance of $y \sim \mathcal{D}$ can be expressed as $\mathbb{V}[y] = V(\mu)/\phi$, with ω an arbitrary constant to parameterize $a(\phi) = \phi/\omega$.

Model

Formally, a GAM has the following form:

$$y \sim \mathcal{D}(\mu, \phi) \quad , \quad \mu = g^{-1}\left(\boldsymbol{\gamma}^\top \boldsymbol{x} + \sum_{j=1}^{J} f_j(\boldsymbol{x})\right) \tag{3.31}$$

where $\mathcal{D}(\mu, \phi)$ is a distribution in the exponential family with mean μ and scale parameter ϕ, $g : \mathbb{R} \to \mathbb{R}$ is an invertible *link function* that connects the linear predictor to the mean of the distribution $\mathcal{D}$, $\boldsymbol{\gamma} \in \mathbb{R}^d$ is a vector of parameters, and the collection of $f_j : \mathbb{R}^d \to \mathbb{R}$ are parameterized smooth, non-linear functions of certain covariates. For convenience, the link function g is usually chosen so that the linear predictor for the mean equals the canonical parameter $\tau^{(i)}$ of $\mathcal{D}$.

Each function f_j models non-linearity as the weighted sum of independent contributions from K_j different *basis* functions $b_{jk} : \mathbb{R}^d \to \mathbb{R}$:

$$f_j(\boldsymbol{x}) = \sum_{k=1}^{K_j} \beta_{jk} b_{jk}(\boldsymbol{x}) \tag{3.32}$$

with all weights $\beta_{jk} \in \mathbb{R}$. Different types of basis functions result in different types of modeled functions; for example, univariate piecewise linear functions of the l-th covariate can be modeled by basis functions of the form $b_{jk}(\boldsymbol{x}) = \max\{\boldsymbol{x}_l - z_{jk}, 0\}$, where $z_{jk} \in \mathbb{R}$ is a *node* that indicates when the k-th basis becomes active. Cubic splines are one of the most commonly used smooths, as they represent the "smoothest" function that interpolates a set of data points, where smoothness is defined in terms of its integrated second derivative (Green and Silverman, 1993). Smoothers of the form Equation (3.32) can also be designed for multivariate inputs, such as geographic coordinates (Wood, 2003, 2006).

The terms $b_{jk}(\boldsymbol{x})$ in Equation (3.32) can be pre-computed and gathered in a single vector called the *basis expansion* of $\boldsymbol{x}$ for the j-th smooth. For multivariate inputs, several smooths can be used to transform different covariates, and their respective basis expansions concatenated into a single *design* vector, such that, in the end, the predictor in Equation (3.31) has form $\mu = g^{-1}(\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}})$ where $\tilde{\boldsymbol{x}}$ contains the basis expansions of all smooths, and $\boldsymbol{\theta}$ includes $\boldsymbol{\gamma}$ and all parameters β_{ij} of all smooths. Additionally, smoothness penalties for f_j, if necessary, are designed so as to be easily computable as quadratic forms $\lambda_j \boldsymbol{\beta}_j^\top \boldsymbol{S}_j \boldsymbol{\beta}_j$ for a suitable matrix $\boldsymbol{S}_j$ that depends on the functional form of the basis functions b_{jk}, and where $\boldsymbol{\beta}_j$ contains all parameters β_{jk} of the j-th smooth and λ_j is the strength of the penalization.

[1] Traditionally, the canonical parameter is indicated with θ, but here we use τ instead to avoid clashes with the model's parameter vector $\boldsymbol{\theta}$.

Optimization

The optimization of GAMs proceeds via maximum likelihood estimation (Section 3.1.3). For fixed values of $\lambda_1, \dots, \lambda_J$, denoting all of the smoothness penalties, the parameter vector $\boldsymbol{\beta}$ is estimated via the penalized iteratively re-weighted least squares algorithm (PIRLS, Wood, 2017), while the smooth penalties can be estimated via (generalized) cross-validation (Stone, 1977; Craven and Wahba, 1978), restricted maximum likelihood (REML; Anderssen and Bloomfield, 1974), or several other methods (Wood, 2017).

3.4.2 Semi-structured models

While GAMs can be used for tabular data, and deep neural networks for non-tabular data, several applications of practical interest include both tabular and non-tabular data at the same time. This situation is especially common in medicine, where most studies include non-tabular data such as medical images, including CT scans, fMRI and histopathology images, genomics sequence data, etc., paired with tabular data about the patient and their condition, such as age, sex, body-mass index, medication, etc. (Huang et al., 2022; Isobe et al., 2022; Zheng et al., 2020). One possibility for jointly modeling both data modalities would be to use a deep neural network of appropriate architecture with two branches, one for each modality. In this model, tabular data would be modeled by a sequence of one or more fully-connected layers, until a final layer that merges the effect of both data types (Wolf et al., 2022). However, using a single layer for the tabular branch prevents the model from learning any non-linear and interaction effects, and using multiple layers would result in a model whose predictive power could be inferior to that of a GAM (Shwartz-Ziv and Armon, 2022), and harder to interpret and explain (Molnar et al., 2022).

Semi-structured models are an alternative approach to jointly modeling tabular and non-tabular data that does not force practitioners to choose between flexibility and interpretability. Semi-structured models combine a structured predictor, such as a GAM, for the tabular data, with an unstructured predictor, i.e., a deep neural network, for the non-tabular data (Kopper et al., 2021; Baumann et al., 2021; Kook et al., 2022). Semi-structured models can be thought of as GAMs with an additional observation-specific offset corresponding to the predictions of the deep neural network used to process the non-tabular data.

Formally, definining $\boldsymbol{x} \in \mathcal{X}$ the tabular features of a sample, $\boldsymbol{z} \in \mathcal{Z}$ its non-tabular features, and $h : \mathcal{Z} \to \mathbb{R}$ a deep neural network of appropriate architecture, a semi-structured model has the following form:

$$\mu = g^{-1}\left(\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}} + h(\boldsymbol{z})\right) \tag{3.33}$$

where μ is understood in the context of Equation (3.31) and $\tilde{\boldsymbol{x}}$ as an appropriate basis expansion of $\boldsymbol{x}$. Semi-structured models are also fitted through maximum likelihood, however the presence of the deep neural network considerably complicates this procedure, especially when h also takes $\boldsymbol{x}$ as input (Rügamer et al., 2023). In the end, semi-structured models retain the distributional results for $\boldsymbol{\beta}$ that maximum likelihood affords, thus remaining as interpretable as GAMs are, for the tabular features, and still allowing alternative interpretability methods (Tjoa and Guan, 2020; Linardatos et al., 2020) for the deep neural network as well as the entire model, if necessary.

3.4.3 Connection to precision medicine and contributions

Semi-structured models can be used in clinical and medical settings to model the joint effect of patient metadata, as tabular features, and medical images produced by devices such as CT scans, fMRI and histopathology images. Such models could be used to make the development of new drugs and treatments cheaper and faster by enabling more accurate and efficient clinical trials, for example by identifying subtle but important relationships between different variables that might not be detectable from unstructured data alone, helping to build trust in the model by clinicians and medical professionals thanks to their transparency and interpretability, and identifying patient subgroups that are likely to respond well to particular drugs or treatments by jointly analyzing patient metadata and unstructured information.

This thesis contributes an application of semi-structured models to predict the future number of COVID-19 cases in each German district and for each sex/age stratum (Section 6.1). A second contribution (Section 6.2) shows that previous inference methods for semi-structured models, by ignoring the uncertainty of the deep neural network (Section 3.2.2), provided excessively narrow confidence intervals for the coefficients of the structured model, resulting in inflated false-positive rates. This contribution proposes to incorporate the uncertainty of the deep neural network into Equation (3.33) by treating its predictions as a random offset, and propagating its variance to the estimated structured coefficients.

3.5 Discrete optimization for vaccine design

As described in Section 2.3.3, epitope vaccines (EV) can be used for cancer treatment by incorporating a subset of the neoantigens in the tumor. Even though EV design does not directly make use of machine learning techniques, the data-model-optimization framework introduced in Section 3.1 is still applicable, and we will thus follow it again. An EV contains short epitopes that can trigger an immune response, joined together into a longer polypeptide for ease of delivery (Figure 3.1).

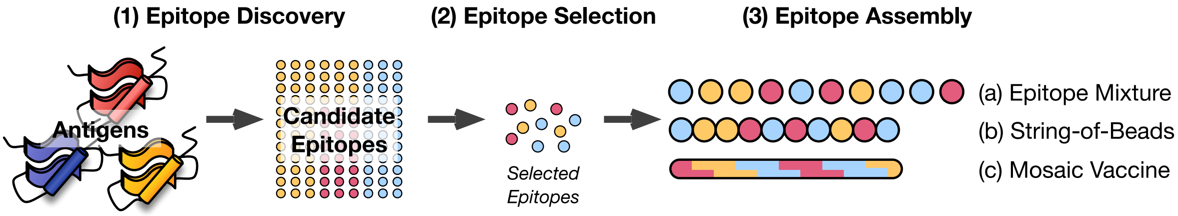

Figure 3.1: The major steps in an epitope vaccine (EV) design pipeline (Figure credit: Dorigatti and Schubert (2020a)). First, epitopes are discovered from antigen sequences through computational tools that predict the immune response. Second, a subset of such epitopes are selected to be included in the EV. Third, the epitopes are delivered separately (3a), or assembled into a longer polypeptide by concatenating them (3b), or by leveraging overlaps (3c).

3.5.1 Data

The EV design process starts from identifying the target epitopes from the antigens that the vaccine should protect against. As detailed in Chapter 2, the main factor to identify epitopes is immunogenicity, i.e., their potential to activate T cells when being presented on MHC-I surface molecules. Due to the large number of potential epitopes, and the cumbersome experimental procedures required to determine whether a candidate epitope does or does not activate T cells, there is great interest in computationally predicting T cell activation without resorting to the wet-lab (Peters et al., 2020). However, computational prediction T cell activation for generic epitopes is still extremely difficult to perform, owing to the extreme variety in the epitopes and receptors themselves and the relative scarcity of relevant data. Therefore, binding affinity to the MHC is used instead as a proxy for T cell activation (Peters et al., 2020). Predicting MHC-I binding is a relatively easier task, for which methods of increasing complexity were developed. While earlier methods were based on kernels (Pfeifer and Kohlbacher, 2008; Ren et al., 2011) or linear models (Racle et al., 2019; Bassani-Sternberg et al., 2017), the latest and most effective models are all based on deep learning (Reynisson et al., 2020; O'Donnell et al., 2020; Shao et al., 2020). Our understanding of the factors that determine immune recognition of epitopes is continuously expanding (Lang et al., 2022), and to accommodate the progression of knowledge, all EV design frameworks assume that each epitope is associated to a single immunogenicity score that summarizes all that is known about its potential to trigger an immune response.

Formally, let $\mathcal{E}$ denote the set of candidate epitopes that were identified, and $I : \mathcal{E} \to \mathbb{R}$ the immunogenicity of an epitope $e \in \mathcal{E}$. Furthermore, let $\mathcal{A}$ be the set of MHC alleles in the target population, with $p(a)$ the probability that an allele $a \in \mathcal{A}$ is found in an individual. Commonly, the immunogenicity of an epitope is computed based on the interactions of the epitope with each allele, without considering interaction effects between different alleles. Moreover, such interactions are quantified from the binding strength $s(e,a) : \mathcal{E} \times \mathcal{A} \to \mathbb{R}^+$ between e and a, measured as the half-maximal inhibitory concentration (IC_{50}) that indicates how weakly e interacts with a. The IC_{50} is measured in nanomolars (nM), and in general terms an IC_{50} of less than 50 is considered as strong binding, less than 500 is moderate binding, less than 5,000 denotes weak binding, and more than 50,000 is undetectable, although different alleles have different sensitivities (Sette et al., 1994). Normalizing the binding strength between 0 and 1 through a log transformation gives the common definition of immunogenicity:

$$I(e) = \sum_{a \in \mathcal{A}} p(a) \left[1 - \log_{50{,}000}(s(e,a)) \right] \tag{3.34}$$

As discussed above, $s(e,a)$ can be predicted by several tools (Nielsen et al., 2007; Reynisson et al., 2020; O'Donnell et al., 2020), and is used as a proxy for T cell activation (Peters et al., 2020). We also define an indicator $c(e,a) := \mathbb{1}[s(e,a) \le \tau_a]$ to identify whether an epitope binds to an allele or not, based on an affinity threshold τ_a. Finally, let $\mathcal{P}$ denote the sequences of antigens that the vaccine should protect against. In a hypothetical COVID-19 EV, for example, $\mathcal{P}$ would contain all of the available sequences of the spike protein, while vaccines for HIV (Ng'uni et al., 2020) and influenza (Wei et al., 2020) target structural proteins that form the outer shell of the virus, and cancer vaccines would target the neoepitopes generated by mutations. With a slight abuse of notation, we also introduce the indicator $c(e,p) := \mathbb{1}[e \in p]$ indicating whether the epitope e is found in the antigen p.

3.5.2 Model

At the core of EV design frameworks is the selection, among all possible candidates, of the set $\mathcal{V} \subseteq \mathcal{E}$ of epitopes with the largest possible immunogenicity. Other considerations also play a role in the selection, depending on the desired use of the vaccine. For example, vaccines that target a specific virus should prioritize conserved epitopes, i.e., epitopes that are rarely mutated, thus appearing in the majority of sequences in $\mathcal{P}$. The lack of mutations in an epitope suggests that it is functionally very important, so that any alteration would lead to a fatal loss of function for the virus, making it an excellent vaccination target. Moreover, the EV should contain epitopes that are recognized by a large number of MHC alleles to ensure that the vaccine is effective for the world's population. This effect is partially encoded into the immunogenicity in Equation (3.34), but it is appropriate to explicitly force this diversity in order to produce a fair EV that is effective for all patient sub-populations. Personalized cancer vaccines do not have such stringent requirements if they are targeted towards a specific individual. Choosing epitopes to maximize the immunogenicity in Equation (3.34) subject to certain constraints suggests to approach EV design as a discrete optimization problem.

Epitope selection

Let us associate a binary decision variable $x_e \in \{0,1\}$ for each epitope $e \in \mathcal{E}$, such that its value equals one if and only if $e \in \mathcal{V}$. The goal of the optimization problem is to find the value of each decision variable such that the total immunogenicity is maximized, while selecting at most $N_e > 0$ epitopes. Similarly, we introduce one decision variable $y_a \in \{0,1\}$ for each allele $a \in \mathcal{A}$, and one $z_p \in \{0,1\}$ for each antigen $p \in \mathcal{P}$, indicating whether the respective allele or antigen is covered by the vaccine, and respective minimum number $N_a > 0$ and $N_p > 0$ of alleles and antigens that should be covered. The maximum number of epitopes N_e is based on pharmacological considerations such as the ease of producing the vaccine, the efficiency by which it is delivered and consumed by the body, etc., while determining N_a and N_p presents a clear trade-off between the overall immunogenicity of the vaccine and its general applicability. Altogether, a simple model for EV design, loosely inspired from the one proposed by Toussaint et al. (2008), could be:

$$
\begin{aligned}
\text{Maximize} \quad & \sum_{e\in\mathcal{E}} I(e)x_e \\
\text{Subject to} \quad & \sum_{e\in\mathcal{E}} x_e \leq N_e \\
& \sum_{a\in\mathcal{A}} y_a \geq N_a \\
& \sum_{p\in\mathcal{P}} z_p \geq N_p \\
& \sum_{e\in\mathcal{E}} x_e c(e,a) \geq y_a \qquad \forall a \in \mathcal{A} \\
& \sum_{e\in\mathcal{E}} x_e c(e,p) \geq z_p \qquad \forall p \in \mathcal{P} \\
& x_e, y_a, z_p \in \{0,1\} \quad \forall e \in \mathcal{E}, a \in \mathcal{A}, p \in \mathcal{P}
\end{aligned}
\tag{3.35}
$$

One of the first approaches to this discrete optimization problem (Vider-Shalit et al., 2007) used genetic algorithms to select epitopes, developing a fitness function to evaluate the vaccine sequence in terms of population coverage and immunogenicity, instead of using explicit constraints as in the formulation above. Lundegaard et al. (2010) used instead a greedy epitope selection method, whereby the set of epitopes selected for the vaccine is iteratively expanded by selecting the highest-ranking epitope according to a measure that balances their immunogenicity with the diversity of the set of already selected epitopes. Toussaint et al. (2008) started the tradition of approaching the EV design problem through mixed-integer linear programming, provably maximizing the immunogenicity subject to antigen and population coverage and epitope conservation, and open-sourcing their algorithm through a web-server (Toussaint and Kohlbacher, 2009).

Epitope assembly

The EV problem in Equation (3.35) only concerns epitope selection, i.e., the determination of which epitopes should be included in the vaccine, however it was experimentally observed that delivering each epitope separately does not result in sufficient immune responses (Cornet et al., 2006; Livingston et al., 2001). Instead, the epitopes should be assembled into a single, longer polypeptide sequence, for example by concatenating them, in order to favors uptake by the immune system. The same studies highlighted that the effectiveness of this kind of polypeptide vaccines is highly dependent upon the specific ordering of the epitopes, thus opening another problem to be solved for designing effective EV vaccines.

Toussaint et al. (2011) tackled this issue by casting it as a traveling salesperson problem, with each epitope corresponding to a city, and the distance between the cities, in this case non-symmetric, being inversely proportional to the proteasomal cleavage efficiency (Section 2.2.1) at the junction between the two epitopes. Considering proteasomal cleavage is essential to ensure that the polypeptide is fragmented in such a way that the original epitopes are recovered, without producing extraneous, and potentially dangerous, fragments. To formalize this problem, consider a fully-connected graph whose set of nodes numbered from 1 to n correspond to the epitopes. Each edge (i,j) is associated to a non-negative weight $w_{ij}\in\mathbb{R}^+$ and a binary decision variable $x_{ij}\in\{0,1\}$ that indicates whether the j-th epitope follows the i-th in the final EV. The basic epitope assembly problem then corresponds to finding the tour in the graph with the minimum cost, which can be formulated as a mixed-integer linear program as shown by Miller et al. (1960):

$$
\begin{aligned}
\text{Maximize} \quad & \sum_{0\le i\neq j\le n}\sum w_{ij}x_{ij} & & \\
\text{Subject to} \quad & \sum_{\substack{i=0\\ i\neq j}}^{n} x_{ij} = 1 & & 1\le j\le n \\
& \sum_{\substack{j=0\\ j\neq i}}^{n} x_{ij} = 1 & & 1\le i\le n \\
& u_i - u_j + n\cdot x_{ij} \le n-1 & & 1\le i\neq j\le n \\
& x_{ij}\in\{0,1\} & & 1\le i,j\le n
\end{aligned}
\tag{3.36}
$$

This formulation leverages a dummy node with index 0 from where the tour starts and ends, and is connected to all other nodes by edges of zero weight. It additionally uses a “node potential” variable u_i for each node other than the dummy, forcing any solution to only visit nodes ordered by increased potential, thus excluding all solutions that contain multiple disjoint tours. Specifically, the constraint on node potentials forces the potential of each node in the tour to be greater than the potential of the node immediately preceding, and smaller than the potential of the node immediately following it. The only tours that satisfy this constraint visit the dummy node, as it is the only way to "reset" the increasing potential before starting a second loop. This constraint, together with the remaining constraints forcing a single tour to pass from each vertex, including the dummy, exclude solutions that containing disjoint tours. This particular constraint is known as the "MTZ" subtour elimination, from the names of the authors Miller, Tucker, and Zemlin, but other forms exist too, chiefly the "DFJ constraint", also named after its authors Dantzig, Fulkerson, and Johnson (Dantzig et al., 1954).

The formulation of Toussaint et al. (2011) extended Equation (3.36) in two ways: (1) by optimizing two-residue linker sequences between adjacent epitopes, in order to further increase proteasomal cleavage efficiency, and (2) by minimizing the immunogenicity of unwanted epitopes resulting from incorrect cleavage of the vaccine sequence. Schubert and Kohlbacher (2016) further improved this formulation to allow for variable-length linker sequences to be designed during the optimization.

3.5.3 Optimization

The latest EV design frameworks (Toussaint et al., 2008, 2011; Schubert and Kohlbacher, 2016), including those contributed by this thesis, are based on mixed-integer linear programming. For a detailed treatment of linear programming, the reader is referred to Bertsimas and Tsitsiklis (1997). Formally, a linear program is a constrained optimization problem of the form:

$$\text{Maximize} \quad \boldsymbol{c}^\top \boldsymbol{x} \tag{3.37}$$

$$\text{Subject to} \quad \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \tag{3.38}$$

where $\boldsymbol{A} \in \mathbb{R}^{n_c \times n_v}$, $\boldsymbol{b}, \boldsymbol{c}, \boldsymbol{x} \in \mathbb{R}^{n_v}$ encode the specifics of the problem being solved having n_c constraints and n_v variables, and the optimization is done with respect to $\boldsymbol{x}$. Linear programs either have (1) no solutions, when the constraints in Equation (3.38) cannot be all satisfied at the same time (infeasible problem), or when the objective in Equation (3.37) can be made arbitrarily large (unbounded problem), or (2) a single solution, or (3) an unlimited number of solutions, all having the same objective value. When all of the decision variables $\boldsymbol{x}$ are real numbers, the solution, or its absence, can be found by the simplex algorithm or interior-point methods, while for discrete or binary variables a divide-and-conquer strategy called branch-and-bound is usually applied.

3.5.4 Connection to precision medicine and contributions

As extensively discussed in Section 2.3.3, EVs are a promising platform for cancer treatment, as well as viral infections such as SARS-CoV-2 (Kar et al., 2020), HIV (Ng'uni et al., 2020), and influenza (Wei et al., 2020).

In this thesis, we propose a general mathematical formulation of the EV design problem that unifies the epitope selection and assembly problems in Equations (3.35) and (3.36), together with additional constraints on vaccine coverage and conservation (Section 4.2). We then present a specialization of this framework in Section 4.3 that is focused on a specific type of EV design, and simplifies the specification of constraints relating to the assembly problem (Section 3.5) and introduces a more accurate computational evaluation of the designed vaccines by incorporating a proteasomal cleavage (Section 2.2.1) predictor into the linear program.

4 Contributions on Immunotherapy Design

4.1 Artificial intelligence in early drug discovery enabling precision medicine

In this review, we discuss the current state of drug discovery in precision medicine (Section 1.1), and present our vision of how artificial intelligence will impact biomarker discovery and drug design. Current precision medicine applications in early drug discovery are still based on a handful of biomarkers, while recent technologies allow for more and better characterization of patients and their diseases. Artificial intelligence approaches to analyze such data will be fundamental to enable truly personalized approaches to drug design and impact clinical practice (Section 1.3).

Contributing article: Boniolo, F.* and Dorigatti, E.* and Ohnmacht, A. J.* and Saur, D. and Schubert B. and Menden, M. P. Artificial intelligence in early drug discovery enabling precision medicine. *Expert Opinion on Drug Discovery 16 (9), 991-1007 https://doi.org/10.1080/17460441.2021.1918096*. (* share first authorship).

Copyright information:

Author contributions: Emilio Dorigatti wrote the introduction on artificial intelligence and Section 3 about drug design, including vaccine design, protein design, and small-molecule design. Fabio Boniolo and Alexander Ohnmacht wrote the remaining parts of the manuscript. Dieter Saur, Benjamin Schubert Michael Menden provided advice and guidance, and proof-read the manuscript. All authors wrote the last section on the expert opinion.

4.2 Graph-theoretical formulation of the generalized epitope-based vaccine design problem

Epitope-based vaccines are a flexible vaccine platform that can be used to treat cancer (Section 2.3.3) and other diseases. Due to their complex nature, bioinformatics plays a pivotal role in their development, however, existing algorithms address only specific parts of the design process, or are unable to provide formal guarantees on the quality of the solution (Section 3.5). In this article, we unify the two problems of epitope selection and epitope assembly, introducing a single mathematical formalism that encompasses all prevalent design principles. We then formulate this problem through a mixed-integer linear program, thus guaranteeing optimality of the solution.

Contributing article: Dorigatti, E. and Schubert, B. (2020) Graph-theoretical formulation of the generalized epitope-based vaccine design problem. *PLoS computational biology 16 (10), e1008237. https://doi.org/10.1371/journal.pcbi.1008237.*

Author contributions: Emilio Dorigatti conceived and developed the method and the experimental design, performed the experiments, interpreted the results. Benjamin Schubert advised in all phases of the project. All authors wrote the manuscript.

4.3 Joint epitope selection and spacer design for string-of-beads vaccines

This article introduces another vaccine design framework that, while focused on a specific type of design formulation, simplifies the specification of constraints relating to the epitope assembly problem (Section 3.5), and introduces a more accurate computational evaluation of the designed vaccines. Both innovations stem from incorporating a proteasomal cleavage predictor (Section 2.2) into the linear program, thus enabling users to specify novel constraints and objectives that directly relate to cleavage, and enabling the linear program to design optimal spacers as part of the solution process. This relieves users from doing this as a separate, and costly, pre-processing step, as required by our previous vaccine design framework (Section 4.2).

Contributing article: Dorigatti, E. and Schubert, B. (2020) Joint epitope selection and spacer design for string-of-beads vaccines. *Bioinformatics, Volume 36, Issue Supplement_2, December 2020, Pages i643–i650. (Proceedings of the 19th European Conference on Computational Biology). https://doi.org/10.1093/bioinformatics/btaa790.*

Author contributions: Emilio Dorigatti conceived and developed the method and the experimental design, performed the experiments, interpreted the results. Benjamin Schubert advised in all phases of the project. All authors wrote the manuscript.

4.4 Proteasomal cleavage prediction: state-of-the-art and future directions

Recognizing the importance of accurate proteasomal cleavage (Section 2.2) to design epitope vaccines (Sections 4.2 and 4.3), in this article we reviewed current available predictors, and performed a benchmark of such tools together with a wide range of deep learning architectures and training regimes. We found that the amount of training data is the single most important determinant of predictive performance, as most methods performed very similarly, and thus argue that data quality is more important than methodological advances to move the field forward.

Contributing article: Ziegler, I. and Ma, B. and Bischl, B. and Dorigatti, E.* and Schubert, B.* (2023) Proteasomal cleavage prediction: state-of-the-art and future directions. *https://doi.org/10.1101/2023.07.17.549305*. (* share last authorship).

Author contributions: Emilio Dorigatti conceived and directed the benchmark, developed the experimental design, performed the literature review, implemented and executed the SVM, logistic regression, and PCM baselines. Ingo Ziegler and Bolei Ma implemented and executed all experiments involving deep architectures. Emilio Dorigatti, Ingo Ziegler, and Bolei Ma wrote the manuscript. Benjamin Schubert advised in all phases of the project. Bernd Bischl advised on the experimental design and interpretation of the results.

5 Contributions on Positive Unlabeled Learning

5.1 Positive-Unlabeled Learning with Uncertainty-aware Pseudo-label Selection

In this article, we tackle the problem of positive-unlabeled learning (Section 3.3) by using pseudo-labeling (Section 3.2.3), selected based on the epistemic uncertainty of an ensemble of deep neural networks (Section 3.2.2). Through a series of benchmarks we show that our method achieves competitive performance, especially when the dataset is highly imbalanced, i.e., with low positive prior π, and is applicable to any data type, unlike most other methods that are specialized to image data. We also apply our method to the practical problem of proteasomal cleavage prediction (Section 2.2.1), achieving considerably higher performance.

Contributing article: Dorigatti, E. and Goschenhofer, J. and Schubert, B. and Rezaei M., and Bischl B. (2022) Positive-Unlabeled Learning with Uncertainty-aware Pseudo-label Selection. *arXiv preprint arXiv:2201.13192* *https://arxiv.org/abs/2201.13192*.

Copyright information: This article is licensed under a Creative Commons Attribution, Non Commercial, Share Alike 4.0 International Public License (CC BY-NC-SA 4.0, https://creativecommons.org/licenses/by-nc-sa/4.0/).

Author contributions: Emilio Dorigatti ideated the method, implemented the main algorithm and some of the supporting code, and performed all experiments involving the proposed method. Jann Goschenhofer helped in refining the method, developed most of the supporting code, and performed all experiments with the external baselines. Emilio Dorigatti and Jann Goschenhofer wrote the manuscript together and interpreted the results with the support of Mina Rezaei, Benjamin Schubert and Bernd Bischl. Mina Rezaei, Benjamin Schubert and Bernd Bischl advised on the experimental design, and the interpretation of the results.

5.2 Robust and Efficient Imbalanced Positive-Unlabeled Learning with Self-supervision

In this article we leverage contrastive learning (Section 3.2.3) to learn good image representations, which we use to train a positive-unlabeled learning classifier (Section 3.3). Through a suite of experiments we show that such representations considerably improve predictive performance both in the balanced and imbalanced settings.

Contributing article: Dorigatti, E.* and Schweisthal, J.* and Bischl, B. and Rezaei, M. (2022) Robust and Efficient Imbalanced Positive-Unlabeled Learning with Self-supervision. *arXiv preprint arXiv:2209.02459* *https://arxiv.org/abs/2209.02459*. (* share first authrship).

Copyright information: This article is licensed under a Creative Commons Attribution 4.0 International Public License (CC BY 4.0, https://creativecommons.org/licenses/by/4.0/).

Author contributions: Emilio Dorigatti developed the theoretical section of the paper, helped in the experimental design and interpretation of the results. The method was conceived and developed by Jonas Schweisthal, who also performed the experiments. Mina Rezaei proposed and led the project, assisted in the experimental design, and interpretation of the results. All authors wrote the manuscript.

6 Contributions on Semi-Structured Regression

6.1 Combining graph neural networks and spatio-temporal disease models to improve the prediction of weekly COVID-19 cases in Germany

Section 3.4 introduces semi-structured regression as a technique to combine structured models (i.e., generalized additive models, GAMs) for tabular data and unstructured models (i.e., deep neural networks, DNNs) for non-tabular data. In this paper, we used a SSR model to predict the number of future COVID-19 cases in each German district, combining tabular information about the population composition of the district, such as density and number of people for each age and gender group, with geographical information about districts and their relationships, such as the number of people traveling from one to another, modeled through a graph neural network. We demonstrate that our semi-structured GAM with a zero-inflated Poisson outperforms other models including deep neural networks, GAMs without the non-tabular data, graph networks without tabular data, and gradient boosted trees.

Contributing article: Fritz, C.* and Dorigatti, E.* and Rügamer, D. (2022). Combining graph neural networks and spatio-temporal disease models to improve the prediction of weekly COVID-19 cases in Germany. *Sci Rep 12, 3930. https://doi.org/10.1038/s41598-022-07757-5.* (* share first authrship).

Copyright information: This article is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0, http://creativecommons.org/licenses/by/4.0/).

Author contributions: Emilio Dorigatti implemented the graph neural network and the semi-structured regression framework using PyTorch and ran the experiments with the inputs of the other authors. Cornelius Fritz conceived the study, led the project, prepared the data. David Rügamer implemented and executed the baselines. All authors participated in designing the experiments, interpreting the results, and writing the manuscript.

6.2 Frequentist Uncertainty Quantification in Semi-Structured Neural Networks

Commonly, semi-structured regression models (Section 3.4) are approached by treating the output of the DNN as an observation-specific offset to be included into a GAM. In this paper, motivated by our earlier investigation in Section 6.1, we argue that this approach ignores the uncertainty of the DNN (Section 3.2.2), and thus results in confidence intervals for the coefficients of the structured model that are too narrow, thus leading to an increased rate of false discoveries. We present simulation studies to highlight the problem, and propose to solve the problem by treating the DNN as a random offset with known variance, given by some deep uncertainty quantification method such as deep ensembles or Monte Carlo dropout. Simulations show that this method results in confidence intervals that better reflect the model's uncertainty, and a practical application on a medical dataset shows that the predictions are more accurate.

Contributing article: Dorigatti, E. and Schubert, B. and Bischl, B. and Rügamer, D. (2023) Frequentist Uncertainty Quantification in Semi-Structured Neural Networks. *Proceedings of the 26th International conference on AI and Statistics, PMLR.* *https://proceedings.mlr.press/v206/dorigatti23a.html*

Copyright information: This article is licensed under a Creative Commons Attribution 4.0 International license (CC BY 4.0, https://creativecommons.org/licenses/by/4.0/).

Author contributions: Emilio Dorigatti ideated and developed the method, performed all the experiments, developed the theory, and wrote the manuscript. David Rügamer advised in all phases of the project. Benjamin Schubert and Bernd Bischl advised on the experimental design and interpretation of the results, and proof-read and approved the manuscript.

7 Concluding remarks

7.1 Summary

Rapid advances in experimental technologies created a wealth of new clinical data describing patients and their diseases. At the same time, contemporary advances in computational and artificial intelligence technologies facilitated the association between these characteristics and the outcome of treatments. The patterns discovered from this data made it clear that disease heterogeneity among patients requires a similar variety of treatments, each one tailored to the specific disease affecting a given patient, and giving rise to a new approach in medical practice called *precision medicine*. In precision medicine, patients and diseases are grouped according to relevant molecular signatures, or biomarkers, instead of superficial phenotypic factors such as frequency, duration, or strength of symptoms. In the extreme case, the treatment is truly tailored to an individual rather than a small cohort of patients, using techniques that are collectively known as *personalized medicine* and include, among others, quickly acquiring relevant data for decision making, designing, producing and administering a suitable drug, and monitoring the treatment outcome. This approach to treatment has been pioneered to cure cancer, one of the leading causes of death worldwide which is particularly unresponsive to traditional medical techniques, as it is caused by randomly-occurring malignant mutations in the cellular DNA.

Starting from this motivation, the contributions presented in this thesis seek to advance the state of the art in cancer treatment, with a particular focus on cancer vaccines. The contributions can be roughly divided in three different areas: applied contributions to vaccine design, methodological contributions to positive-unlabeled learning, and methodological contributions to semi-structured regression models. Concerning the first area, we argued, through a review and expert opinion paper, about the many ways in which artificial intelligence can support and drive precision medicine (Section 4.1), from biomarker discovery to drug design, with a particular focus on epitope vaccine design. Epitope vaccines, introduced in detail in Section 2.3.3, are designed to contain the specific antigens that determine the patient's disease, as discussed in section 3.5, and thus represent one of the most promising precision and personalized treatments. Consequently, this thesis introduced a novel mathematical formulation of the epitope vaccine design problem (Section 4.2), that improved previous approaches by unifying the two sub-problems that need to be solved to design a vaccine: the selection of epitopes to include in the vaccine, and their assembly into a single polypeptide, employing a single mathematical formalism to describe three different types of assembly modalities. The third contribution of this thesis (Section 4.3) improved this epitope vaccine design framework by reformulating the problem in a more user-friendly format, and proposing more accurate *in silico* evaluation scheme of the results. Both of these frameworks are focused on a specific type of epitope vaccine, whose efficacy crucially depends on the quality of the final epitope assembly. Epitopes are assembled so as to elicit proteasomal cleavage (Section 2.2.1) at their junction site, thus motivating the fourth contribution of this thesis: a literature survey and benchmark of current cleavage prediction techniques, concluded by guidelines for possible further developments of the field (Section 4.4).

Besides proteasomal cleavage, another important quantity to be predicted in order to design effective cancer vaccines is the strength of the immune response that can be elicited against an epitope (Section 3.5.1), frequently quantified as the strength with which it binds to the MHC (Section 2.2.2). High-throughput experimental techniques can be used to collect positive examples of MHC binding to train machine learning predictors, however collecting negative examples is significantly more costly and time-consuming (Section 3.3.4). Recognizing this challenge, the second part of this thesis focuses on positive-unlabeled learning (Section 3.3), a branch of semi-supervised learning (Section 3.2.3) that enables binary classifiers to be learned without negatively-labeled examples. First, a generic method based on pseudo-labeling (Section 3.2.3) was proposed, using epistemic uncertainty quantified through deep ensembles (Section 3.2.2) as a selection criterion for pseudo-labels (Section 5.1). Second, recent advances in contrastive learning (Section 3.2.3) were exploited to propose a positive-unlabeled learning method specialized on imaging data (Section 5.2). Experimental results confirm that both of these approaches particularly improve predictive performance on imbalanced data distributions, which are particularly common in precision medicine applications.

While the most recent advances in artificial intelligence (Section 3.2) excel at learning from unstructured, non-tabular data types such as images, audio, and natural language, their predictions on tabular data are not as good as those produced by traditional statistical regression approaches (Section 3.4.1). Commonly, in modern clinical practice, tabular metadata about patients is paired to non-tabular data produced by medical devices, such as microscopes and MRI machines, therefore the development of methods that are able to learn from this kind of multi-modal data can further drive the adoption of precision medicine in the clinic (Section 3.4.3). By this motivation, the third set of contributions of this thesis focuses on semi-structured regression models, combining deep neural networks with traditional statistical regression techniques to enable model learning and interpretation using both data types (Section 3.4). The first contribution in this respect (Section 6.1) demonstrates the advantages of such models in an epidemiological modeling task, by unifying graph neural networks and generalized additive models to predict the number of COVID-19 cases in each district in Germany, for each age and sex stratum, based on mobility patterns and sociological data. A second contribution in the topic (Section 6.2) raises an issue with previous methods used to train semi-structured regression models, namely the fact that they ignored the uncertainty of the neural network when deriving confidence intervals for the coefficients of the structured model on the tabular data, leading to inflated false-positive rates compared to the desired nominal level. A solution for this problem is also proposed, by treating the predictions of the deep neural network as a random offset with given variance, demonstrating improved coverage on simulated data, and higher predictive performance on a glaucoma prediction challenge comprising dermoscopic images and patient metadata.

7.2 Discussion

Research can feel like a random walk, with each idea leading to another, another, and another. In hindsight, the order by which ideas are visited rarely is optimal: if an idea i leads to another idea j, the opposite need not be true, for j in itself may already suggest that i was not so good after all. At the same time, j could have never be found, was it not for i coming before it. For example, although we knew of the importance of proteasomal cleavage prediction for epitope vaccine design, we developed the two frameworks (Sections 4.2 and 4.3) assuming that the predictions were available and reliable. Only later did we perform a systematic review of the state of proteasomal cleavage predictors (Section 4.4), and realized that their predictive performance, especially for N-terminal cleavage, did not seem to be as high as we hoped for. At the same time, were it not for the usefulness of proteasomal cleavage prediction in our vaccine design frameworks, such a review and benchmark would have probably never be performed in the first place, and the community could not have benefited from our survey. The same phenomenon could be observed when we tried to quantify uncertainty for our predictor of COVID-19 cases (Section 6.1), finding available methods lacking and leading to the development of a new inference paradigm (Section 6.2).

The vaccine design frameworks introduced in this thesis stem from a highly interdisciplinary, and continuously advancing, mixture of biological understanding and computational methods, and thus their thorough evaluation inevitably requires careful lab testing (Vitiello and Zanetti, 2017). Most research efforts in cancer vaccines seem to have been mostly focusing on long peptides (Slingluff, 2011; Melief et al., 2015), as opposed to the short peptides used in the vaccine design frameworks proposed in this thesis. Long peptides, by virtue of their length, can be delivered individually, and still elicit immune responses, unlike the short peptides considered by the design frameworks introduced here, that require *ad hoc* assembly procedures. Short peptides for vaccination seem to be exploited by considerably fewer works, and most of them used fixed spacers, rather than spacers optimized *ad hoc* as proposed in this thesis; this could possibly be cause by a number of additional challenges that need to be overcome when working with short peptides (Zhang, 2018). Furthermore, virtually all short-peptide vaccines still do not make use of any spacer optimization frameworks, but opt for using fixed, pre-made linkers, as evidenced by a recent review on the topic (Parvizpour et al., 2020) that all but skirted the issue, despite the community being well-aware of the importance of proper linker design (Shamriz et al., 2016). Whether this happened because the community is not aware of these new design tools, because the tools are not accessible to people who lack computational skills, or because the tools simply do not work, it is unclear to the author of this thesis. It is however likely that further advances in this field, bringing *in silico* optimized cancer vaccines to routine clinical practice, necessitate stronger and deeper collaboration between statisticians and medical doctors, with considerably increased knowledge transfer between fields (Vitiello and Zanetti, 2017), to the extent that it could not be achieved by the author during the few years working on these projects.

7.3 Outlook

Beyond experimental testing and validation, which will certainly suggest further directions for improvement, the works introduced in this thesis offer plenty of opportunities for future methodological research. The vaccine design frameworks (Dorigatti and Schubert, 2020a,b) could be extended to consider the uncertainty in the predictions of immunogenicity and cleavage in the optimization process, by leveraging stochastic programming techniques (Birge and Louveaux, 2011). This would make it possible to optimize the median or worst-case immunogenicity instead of its expectation, thus ensuring that the designed vaccine is effective for all individuals in the target population, as well as alerting practitioners in case the given set of inputs and constraints makes the vaccine unacceptably risky or of dubious efficacy. This would require extending current cleavage, MHC and TCR binding to reliably provide such quantities, which could be done through *post hoc* methods at first, so that current predictors could already be incorporated in such a design framework, and later through the use of predictors trained specifically with uncertainty quantification in mind, as detailed in Section 3.2.2.

Aside from uncertainty quantification, such predictors could be further improved through the positive-unlabeled learning techniques for imbalanced data introduced here. However, especially for MHC binding, the unlabeled data is imbalanced to such a degree that traditional binary classification methods seem to be as successful, if not more successful, than positive-unlabeled approaches. This finding, backed by informal tests and anecdotal evidence from the author and some external collaborators, could be explained by the relative robustness towards noisy labels of the squared error loss (Manwani and Sastry, 2013), used by the most popular MHC binding predictor (Reynisson et al., 2020). Therefore, it seems that, beyond a certain degree of imbalancedness, positive-unlabeled learning problems are better approached through the lens of noisy classification with appropriate loss functions; ideally, the best learning approach for a given problem could be chosen algorithmically in a data-driven fashion. Moving beyond the straightforward option of comparing these two approaches based on their performance on a held-out dataset, it seems likely that there exist a soft transition between the realms of positive-unlabeled learning and noisy classification. A novel loss function, smoothly interpolating between the two options, would therefore be an interesting research avenue.

Another issue with computational epitope prediction is that several experimental studies have found larger amounts of false-positives than expected when using predicted epitopes for immunization (Shetty and Ott, 2021), despite the fact that these tools routinely achieve AUCs of 95% or more in *in silico* evaluations (Reynisson et al., 2020). Part of the issue is certainly because MHC presentation of epitopes ensures neither recognition by nor activation of T cells, which is the end goal of vaccination, especially for cancer neoepitopes. This latter aspect is considerably more complicated to predict computationally due to the vast amount of receptors and sparsity of data measuring such interactions, and considerable research efforts are likely needed in this area, both to gather new data, and to develop predictors using what little data is available at the moment.

With respect to semi-structured regression models, our work on incorporating the uncertainty of the neural network into the confidence intervals of the coefficients of the structured model (Section 6.2) crucially depends on the correctness of the network uncertainty, in the sense that its inaccurate estimation directly translates to inaccurate estimation of the coefficient intervals. A systematic benchmark of uncertainty quantification methods for the neural network is therefore needed to guide practitioners in making the right choices. Furthermore, by relaxing the assumptions on the network uncertainty, it may be possible to provide theoretical guarantees that are more reliable in practice. Moreover, the fitting procedure of such semi-structured regression models is still somewhat cumbersome. The presence of smooth terms or random effects, in particular, requires separate tuning of their penalties via such methods as grid or random searches, as well as iterative fitting of the additive model in a later stage, since most modern estimation software is unable to accept random effects with given, fixed variance. We further elaborated on these points in Appendix D and F of our paper (Section 6.2), but the process could certainly be simplified by developing appropriate inference routines. Such implementation ought to be generic enough to accommodate future advances in deep uncertainty quantification, thus ensuring that the framework remains relevant.

7.4 Conclusion

The quest to form a healthy society has highlighted the staggering complexity of the biological systems governing human life, emphasizing the necessity for increasingly sophisticated tools to understand and combat diseases. Advances in experimental techniques generated immense amount of data, whose analysis required equally significant innovations in computational methods. It is in such realm that the contributions presented in this doctoral thesis are located. By leveraging discrete optimization and machine learning, it is the author's hope that the methods hereby presented will help designing better cancer treatments, thus advancing the fight against the second leading cause of death worldwide.

List of Contributing Publications

Boniolo, F.* and Dorigatti, E.* and Ohnmacht, A. J.* and Saur, D. and Schubert B. and Menden, M. P. Artificial intelligence in early drug discovery enabling precision medicine. *Expert Opinion on Drug Discovery 16 (9), 991-1007* https://doi.org/10.1080/17460441.2021.1918096. (* share first authorship).

Dorigatti, E. and Schubert, B. (2020) Graph-theoretical formulation of the generalized epitope-based vaccine design problem. *PLoS computational biology 16 (10), e1008237.* https://doi.org/10.1371/journal.pcbi.1008237.

Dorigatti, E. and Schubert, B. (2020) Joint epitope selection and spacer design for string-of-beads vaccines. *Bioinformatics, Volume 36, Issue Supplement_2, December 2020, Pages i643–i650. (Proceedings of the 19th European Conference on Computational Biology).* https://doi.org/10.1093/bioinformatics/btaa790.

Ziegler, I. and Ma, B. and Bischl, B. and Dorigatti, E.* and Schubert, B.* (2023) Proteasomal cleavage prediction: state-of-the-art and future directions. https://doi.org/10.1101/2023.07.17.549305. (* share last authorship).

Dorigatti, E. and Goschenhofer, J. and Schubert, B. and Rezaei M., and Bischl B. (2022) Positive-Unlabeled Learning with Uncertainty-aware Pseudo-label Selection. *arXiv preprint arXiv:2201.13192* https://arxiv.org/abs/2201.13192.

Dorigatti, E.* and Schweisthal, J.* and Bischl, B. and Rezaei, M. (2022) Robust and Efficient Imbalanced Positive-Unlabeled Learning with Self-supervision. *arXiv preprint arXiv:2209.02459* https://arxiv.org/abs/2209.02459. (* share first authrship).

Fritz, C.* and Dorigatti, E.* and Rügamer, D. (2022). Combining graph neural networks and spatio-temporal disease models to improve the prediction of weekly COVID-19 cases in Germany. *Sci Rep 12, 3930.* https://doi.org/10.1038/s41598-022-07757-5. (* share first authrship).

Dorigatti, E. and Schubert, B. and Bischl, B. and Rügamer, D. (2023) Frequentist Uncertainty Quantification in Semi-Structured Neural Networks. *Proceedings of the 26th International conference on AI and Statistics, PMLR.* https://proceedings.mlr.press/v206/dorigatti23a.html

Further References

Noraini Abd-Aziz and Chit Laa Poh. 2022. Development of peptide-based vaccines for cancer. *Journal of Oncology*, 2022.

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul W. Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Inf. Fusion*, 76:243–297.

Anish Acharya, Sujay Sanghavi, Li Jing, Bhargav Bhushanam, Dhruv Choudhary, Michael G. Rabbat, and Inderjit S. Dhillon. 2022. Positive unlabeled contrastive learning. *ArXiv*, abs/2206.01206.

Ahmed M. Alaa and Mihaela van der Schaar. 2020. Discriminative jackknife: Quantifying uncertainty in deep learning via higher-order influence functions. *ArXiv*, abs/2007.13481.

Bruno Alvarez, Birkir Reynisson, Carolina Barra, Søren Buus, Nicola Ternette, Tim Connelley, Massimo Andreatta, and Morten Nielsen. 2019. Nnalign_ma; mhc peptidome deconvolution for accurate mhc binding motif characterization and improved t-cell epitope predictions. *Molecular & Cellular Proteomics*, 18(12):2459–2477.

RS Anderssen and Peter Bloomfield. 1974. A time series approach to numerical differentiation. *Technometrics*, 16(1):69–75.

Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. 2020. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. 2019. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. 2006. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156.

Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. 2021. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201.

Michal Bassani-Sternberg, Chloé Chong, Philippe Guillaume, Marthe Solleder, HuiSong Pak, Philippe O Gannon, Lana E Kandalaft, George Coukos, and David Gfeller. 2017. Deciphering hla-i motifs across hla peptidomes improves neo-antigen predictions and identifies allostery regulating hla specificity. *PLoS computational biology*, 13(8):e1005725.

Philipp F. M. Baumann, Torsten Hothorn, and David Rügamer. 2021. Deep conditional transformation models. In *Machine Learning and Knowledge Discovery in Databases. Research Track*, pages 3–18, Cham. Springer International Publishing.

Suzanna Becker and Geoffrey E Hinton. 1992. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163.

Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. 2019. The need for uncertainty quantification in machine-assisted medical decision making. *Nat Mach Intell*, 1(1):20–23.

Jessa Bekker and Jesse Davis. 2018. Estimating the class prior in positive and unlabeled data through decision tree induction. In *AAAI Conference on Artificial Intelligence*.

Jessa Bekker and Jesse Davis. 2020. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4):719–760.

Dimitris Bertsimas and John N Tsitsiklis. 1997. *Introduction to linear optimization*, volume 6. Athena scientific Belmont, MA.

John R Birge and Francois Louveaux. 2011. *Introduction to stochastic programming*. Springer Science & Business Media.

Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*. 4. Springer.

Eryn Blass and Patrick A Ott. 2021. Advances in the development of personalized neoantigen-based therapeutic cancer vaccines. *Nature Reviews Clinical Oncology*, 18(4):215–229.

Janice S Blum, Pamela A Wearsch, and Peter Cresswell. 2013. Pathways of antigen processing. *Annual review of immunology*, 31:443–473.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural network. In *ICML*.

Benjamin M Bolker, Mollie E Brooks, Connie J Clark, Shane W Geange, John R Poulsen, M Henry H Stevens, and Jada-Simone S White. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution*, 24(3):127–135.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6.

Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur D. Szlam, and Pierre Vandergheynst. 2016. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34:18–42.

Etienne Caron, DanielJ Kowalewski, Ching Chiek Koh, Theo Sturm, Heiko Schuster, and Ruedi Aebersold. 2015. Analysis of major histocompatibility complex (mhc) immunopeptidomes using mass spectrometry. *Molecular & Cellular Proteomics*, 14(12):3105–3117.

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien, editors. 2006. *Semi-Supervised Learning*. The MIT Press.

Hongming Chen, Ola Engkvist, Yinhai Wang, Marcus Olivecrona, and Thomas Blaschke. 2018. The rise of deep learning in drug discovery. *Drug discovery today*, 23(6):1241–1250.

Hui Chen, Fangqing Liu, Yin Wang, Liyue Zhao, and Hao Wu. 2020a. A variational approach for learning from positive and unlabeled data. In *Advances in Neural Information Processing Systems*, pages 14844–14854.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020b. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Xuxi Chen, Wuyang Chen, Tianlong Chen, Ye Yuan, Chen Gong, Kewei Chen, and Zhangyang Wang. 2020c. Self-pu: Self boosted and calibrated positive-unlabeled training. In *International Conference on Machine Learning*, pages 1510–1519. PMLR.

Marthinus Christoffel, Gang Niu, and Masashi Sugiyama. 2016. Class-prior estimation for learning from positive and unlabeled data. In *Asian Conference on Machine Learning*, volume 45 of *Proceedings of Machine Learning Research*, pages 221–236.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. 2020a. Debiased contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 8765–8775.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. 2020b. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775.

Ronan Collobert, Fabian Sinz, Jason Weston, Léon Bottou, and Thorsten Joachims. 2006. Large scale transductive svms. *Journal of Machine Learning Research*, 7(8).

Further References

Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023.*

Sébastien Cornet, Isabelle Miconnet, Jeanne Menez, François Lemonnier, and Kostas Kosmatopoulos. 2006. Optimal organization of a polypeptide-based candidate cancer vaccine composed of cryptic tumor peptides with enhanced immunogenicity. *Vaccine*, 24(12):2102–2109.

Peter Craven and Grace Wahba. 1978. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische mathematik*, 31(4):377–403.

Gilles R Dagenais, Darryl P Leong, Sumathy Rangarajan, Fernando Lanas, Patricio Lopez-Jaramillo, Rajeev Gupta, Rafael Diaz, Alvaro Avezum, Gustavo BF Oliveira, Andreas Wielgosz, et al. 2020. Variations in common diseases, hospital admissions, and deaths in middle-aged adults in 21 countries from five continents (pure): a prospective cohort study. *The Lancet*, 395(10226):785–794.

Eva Dahlén, Niina Veitonmäki, and Per Norlén. 2018. Bispecific antibodies in cancer immunotherapy. *Therapeutic advances in vaccines and immunotherapy*, 6(1):3–17.

George Dantzig, Ray Fulkerson, and Selmer Johnson. 1954. Solution of a large-scale traveling-salesman problem. *Journal of the operations research society of America*, 2(4):393–410.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. 2021. Laplace redux-effortless bayesian deep learning. *Advances in Neural Information Processing Systems*, 34:20089–20103.

C Debouck and B Metcalf. 2000. The impact of genomics on drug discovery. *Annual Review of Pharmacology and Toxicology*, 40(1):193–208.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

Emilio Dorigatti and Benjamin Schubert. 2020a. Graph-theoretical formulation of the generalized epitope-based vaccine design problem. *PLOS Computational Biology*, 16(10):e1008237.

Emilio Dorigatti and Benjamin Schubert. 2020b. Joint epitope selection and spacer design for string-of-beads vaccines. *Bioinformatics*, 36(Supplement_2):i643–i650.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations.*

Sarah A Dugger, Adam Platt, and David B Goldstein. 2018. Drug development in the era of precision medicine. *Nature reviews Drug discovery*, 17(3):183–196.

Pehr Edman et al. 1949. A method for the determination of the amino acid sequence in peptides. *Arch. Biochem.*, 22:475–476.

Charles Peter Elkan and Keith Noto. 2008. Learning classifiers from only positive and unlabeled data. In *Knowledge Discovery and Data Mining.*

G Emilien, M Ponchon, C Caldas, O Isacson, and J-M Maloteaux. 2000. Impact of genomics on drug discovery and clinical medicine. *Qjm*, 93(7):391–423.

Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. 2019. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29.

Olivera J Finn. 2018. The dawn of vaccines for cancer prevention. *Nature Reviews Immunology*, 18(3):183–194.

Peter Forster, Lucy Forster, Colin Renfrew, and Michael Forster. 2020. Phylogenetic network analysis of sars-cov-2 genomes. *Proceedings of the National Academy of Sciences*, 117(17):9241–9243.

Yarin Gal and Zoubin Ghahramani. 2015. Bayesian convolutional neural networks with bernoulli approximate variational inference. *ArXiv*, abs/1506.02158.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Saurabh Garg, Yifan Wu, Alexander J Smola, Sivaraman Balakrishnan, and Zachary Lipton. 2021. Mixture proportion estimation and pu learning:a modern approach. In *Advances in Neural Information Processing Systems*, volume 34, pages 8532–8544.

Quentin Garrido, Yubei Chen, Adrien Bardes, Laurent Najman, and Yann Lecun. 2022. On the duality between contrastive and non-contrastive self-supervised learning. *arXiv preprint arXiv:2206.02574*.

Ebony N Gary and David B Weiner. 2020. Dna vaccines: prime time is now. *Current Opinion in Immunology*, 65:21–27.

Jeff Gauthier, Antony T Vincent, Steve J Charette, and Nicolas Derome. 2019. A brief history of bioinformatics. *Briefings in bioinformatics*, 20(6):1981–1996.

Erik Gawehn, Jan A Hiss, and Gisbert Schneider. 2016. Deep learning in drug discovery. *Molecular informatics*, 35(1):3–14.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseo Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, R. Triebel, P. Jung, R. Roscher, M. Shahzad, Wen Yang, R. Bamler, and Xiaoxiang Zhu. 2021. A survey of uncertainty in deep neural networks. *ArXiv*, abs/2107.03342.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2017. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings.

Laura H Goetz and Nicholas J Schork. 2018. Personalized medicine: motivation, challenges, and progress. *Fertility and sterility*, 109(6):952–963.

Chen Gong, Qizhou Wang, Tongliang Liu, Bo Han, Jane Jia You, Jian Yang, and Dacheng Tao. 2021. Instance-dependent positive and unlabeled learning with labeling bias estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:4163–4177.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.

Alexander V Gopanenko, Ekaterina N Kosobokova, and Vyacheslav S Kosorukov. 2020. Main strategies for the identification of neoantigens. *Cancers*, 12(10):2879.

Yves Grandvalet and Yoshua Bengio. 2004. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17.

Peter J Green and Bernard W Silverman. 1993. *Nonparametric regression and generalized linear models: a roughness penalty approach*. Crc Press.

Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. 2020. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(12):4338–4364.

Zayd Hammoudeh and Daniel Lowd. 2020. Learning from positive and unlabeled data with arbitrary positive shift. In *Advances in Neural Information Processing Systems*, volume 33, pages 13088–13099.

T.J. Hastie and R.J. Tibshirani. 1990. Generalized additive models. *Monographs on statistics and applied probability. Chapman & Hall*, 43:335.

Further References

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Xing He and Chenqi Xu. 2020. Immune checkpoint signaling and cancer immunotherapy. *Cell research*, 30(8):660–669.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. 2020. Augment your batch: Improving generalization through instance repetition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8129–8138.

Matthew D Hoffman, Andrew Gelman, et al. 2014. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623.

Kurt Hornik, Maxwell B. Stinchcombe, and Halbert L. White. 1989. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366.

Yu-Guan Hsieh, Gang Niu, and Masashi Sugiyama. 2019. Classification from positive, unlabeled and biased negative data. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2820–2829.

Ming Hu, Ke Deng, Siddarth Selvaraj, Zhaohui Qin, Bing Ren, and Jun S Liu. 2012. Hicnorm: removing biases in hi-c data via poisson regression. *Bioinformatics*, 28(23):3131–3133.

Wenpeng Hu, Ran Le, Bing Liu, Feng Ji, Jinwen Ma, Dongyan Zhao, and Rui Yan. 2021. Predictive adversarial learning from positive and unlabeled data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9):7806–7814.

Shih-Chiang Huang, Chi-Chung Chen, Jui Lan, Tsan-Yu Hsieh, Huei-Chieh Chuang, Meng-Yao Chien, Tao-Sheng Ou, Kuang-Hua Chen, Ren-Chin Wu, Yu-Jen Liu, et al. 2022. Deep neural network trained on gigapixel images improves lymph node metastasis detection in clinical settings. *Nature Communications*, 13(1):1–14.

Jp Hughes, Steven Rees, SB Kalindjian, and KL Philpott. 2011. Principles of early drug discovery. *British Journal of Pharmacology*, 162.

Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110:457–506.

Ian R Humphreys and Sarah Sebastian. 2018. Novel viral vectors in infectious diseases. *Immunology*, 153(1):1–9.

Tomoya Isobe, Masatoshi Takagi, Aiko Sato-Otsubo, Akira Nishimura, Genta Nagae, Chika Yamagishi, Moe Tamura, Yosuke Tanaka, Shuhei Asada, Reina Takeda, et al. 2022. Multi-omics analysis defines highly refractory ras burdened immature subgroup of infant acute lymphoblastic leukemia. *Nature communications*, 13(1):1–16.

Pavel Izmailov, Sharad Vikram, Matthew D. Hoffman, and Andrew Gordon Wilson. 2021a. What are bayesian neural network posteriors really like? In *International Conference on Machine Learning.*

Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Gordon Wilson. 2021b. What are bayesian neural network posteriors really like? In *International conference on machine learning*, pages 4629–4640. PMLR.

Shantanu Jain, Martha White, and Predrag Radivojac. 2017. Recovering true classifier performance in positive-unlabeled learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2020. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2.

Kristen Jaskie and Andreas Spanias. 2019. Positive and unlabeled learning algorithms and applications: A survey. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*. IEEE.

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

Tamalika Kar, Utkarsh Narsaria, Srijita Basak, Debashrito Deb, Filippo Castiglione, David M Mueller, and Anurag P Srivastava. 2020. A candidate multi-epitope vaccine against sars-cov-2. *Scientific reports*, 10(1):10895.

Julian Kates-Harbeck, Alexey Svyatkovskiy, and William Tang. 2019. Predicting disruptive instabilities in controlled fusion plasmas through deep learning. *Nature*, 568(7753):526–531.

Masahiro Kato, Takeshi Teshima, and Junya Honda. 2019. Learning from positive and unlabeled data with a selection bias. In *International Conference on Learning Representations*.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. 2016. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*.

Can Keşmir, Alexander K Nussbaum, Hansjörg Schild, Vincent Detours, and Søren Brunak. 2002. Prediction of proteasome cleavage motifs by neural networks. *Protein engineering*, 15(4):287–296.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Thomas Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations*.

Ryuichi Kiryo, Gang Niu, Marthinus C du Plessis, and Masashi Sugiyama. 2017. Positive-unlabeled learning with non-negative risk estimator. *Advances in Neural Information Processing Systems*.

Vladimir Koltchinskii and Dmitriy Panchenko. 2000. Rademacher processes and bounding the risk of function learning. In *High dimensional probability II*, pages 443–457. Springer.

Lucas Kook, Lisa Herzog, Torsten Hothorn, Oliver Dürr, and Beate Sick. 2022. Deep and interpretable regression models for ordinal outcomes. *Pattern Recognition*, 122:108263.

Philipp Kopper, Sebastian Pölsterl, Christian Wachinger, Bernd Bischl, Andreas Bender, and David Rügamer. 2021. Semi-structured deep piecewise exponential models. In *Proceedings of AAAI Spring Symposium on Survival Prediction – Algorithms, Challenges, and Applications, PMLR*, pages 40–53.

Jeffrey Koury, Mariana Lucero, Caleb Cato, Lawrence Chang, Joseph Geiger, Denise Henry, Jennifer Hernandez, Fion Hung, Preet Kaur, Garrett Teskey, et al. 2018. Immunotherapies: exploiting the immune system for cancer treatment. *Journal of immunology research*, 2018.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NIPS*.

Franziska Lang, Barbara Schrörs, Martin Löwer, Özlem Türeci, and Ugur Sahin. 2022. Identification of neoantigens for individualized therapeutic cancer vaccines. *Nature reviews Drug discovery*, 21(4):261–282.

Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. 2020. Contrastive representation learning: A framework and review. *Ieee Access*, 8:193907–193934.

Further References

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896.

Wee Sun Lee and B. Liu. 2003. Learning with positive and unlabeled examples using weighted logistic regression. In *ICML*.

Fuyi Li, Shuangyu Dong, André Leier, Meiya Han, Xudong Guo, Jing Xu, Xiaoyu Wang, Shirui Pan, Cangzhi Jia, Yang Zhang, Geoffrey I Webb, Lachlan J M Coin, Chen Li, and Jiangning Song. 2021. Positive-unlabeled learning in bioinformatics and computational biology: a brief review. *Briefings in Bioinformatics*, 23(1).

Tingting Li, Dongxia Liu, Yadi Yang, Jiali Guo, Yujie Feng, Xinmo Zhang, Shilong Cheng, and Jie Feng. 2020a. Phylogenetic supertree reveals detailed evolution of sars-cov-2. *Scientific reports*, 10(1):1–9.

Wenkai Li, Qinghua Guo, and Charles Elkan. 2010. A positive and unlabeled learning algorithm for one-class classification of remote-sensing data. *IEEE Transactions on geoscience and remote sensing*, 49(2):717–725.

Ying Li, Lingfei Ma, Zilong Zhong, Fei Liu, Michael A Chapman, Dongpu Cao, and Jonathan Li. 2020b. Deep learning for lidar point clouds in autonomous driving: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 32(8):3412–3432.

Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2020. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18.

Weiping Liu, Jia Sun, Wanyi Li, Ting Hu, and Peng Wang. 2019. Deep learning on point clouds and its application: A survey. *Sensors*, 19(19):4188.

Brian D Livingston, Mark Newman, Claire Crimi, Denise McKinney, Robert Chesnut, and Alessandro Sette. 2001. Optimization of epitope processing enhances immunogenicity of multiepitope dna vaccines. *Vaccine*, 19(32):4652–4660.

Malte D Luecken and Fabian J Theis. 2019. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746.

Claus Lundegaard, M Buggert, Ac Karlsson, Ole Lund, Carina Perez, and Morten Nielsen. 2010. Popcover: a method for selecting of peptides with optimal population and pathogen coverage. In *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology*, pages 658–659.

Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Atlanta, Georgia, USA.

David JC MacKay. 2003. *Information theory, inference and learning algorithms*. Cambridge university press.

Wesley J. Maddox, T. Garipov, Pavel Izmailov, Dmitry P. Vetrov, and Andrew Gordon Wilson. 2019. A simple baseline for bayesian uncertainty in deep learning. In *Neural Information Processing Systems*.

Naresh Manwani and PS Sastry. 2013. Noise tolerance under risk minimization. *IEEE transactions on cybernetics*, 43(3):1146–1151.

James Martens and Roger Grosse. 2015. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pages 2408–2417. PMLR.

Julie Maupin-Furlow. 2012. Proteasomes and protein conjugation across domains of life. *Nature Reviews Microbiology*, 10(2):100–111.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

Cornelis JM Melief, Thorbald van Hall, Ramon Arens, Ferry Ossendorp, Sjoerd H van der Burg, et al. 2015. Therapeutic cancer vaccines. *The Journal of clinical investigation*, 125(9):3401–3412.

Ira Mellman, George Coukos, and Glenn Dranoff. 2011. Cancer immunotherapy comes of age. *Nature*, 480(7378):480–489.

Aditya Menon, Brendan Van Rooyen, Cheng Soon Ong, and Bob Williamson. 2015. Learning from corrupted binary labels via class-probability estimation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 125–134, Lille, France. PMLR.

Rhiannon Michelmore, Matthew Wicker, Luca Laurenti, Luca Cardelli, Yarin Gal, and Marta Kwiatkowska. 2020. Uncertainty quantification with statistical guarantees in end-to-end autonomous driving control. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE.

Clair E Miller, Albert W Tucker, and Richard A Zemlin. 1960. Integer programming formulation of traveling salesman problems. *Journal of the ACM (JACM)*, 7(4):326–329.

Marvin Minsky and Seymour A. Papert. 1969. *Perceptrons: an introduction to computational geometry.* The MIT Press.

Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. 2018. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246.

Ishan Misra and Laurens van der Maaten. 2020. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6707–6717.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. 2018. *Foundations of machine learning.* MIT press.

Christoph Molnar, Gunnar König, Julia Herbinger, Timo Freiesleben, Susanne Dandl, Christian A Scholbeck, Giuseppe Casalicchio, Moritz Grosse-Wentrup, and Bernd Bischl. 2022. General pitfalls of model-agnostic interpretation methods for machine learning models. In *xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, pages 39–68. Springer.

Kenneth Murphy and Casey Weaver. 2016. *Janeway's immunobiology.* Garland science.

Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective.* MIT press.

Raymond H Myers, Douglas C Montgomery, G Geoffrey Vining, and Timothy J Robinson. 2012. *Generalized linear models: with applications in engineering and the sciences.* John Wiley & Sons.

Radford M. Neal. 2011. Mcmc using hamiltonian dynamics. *arXiv: Computation*, pages 139–188.

Tiza Ng'uni, Caroline Chasara, and Zaza M Ndhlovu. 2020. Major scientific hurdles in hiv vaccine development: historical perspective and future directions. *Frontiers in immunology*, 11:590780.

Morten Nielsen, Claus Lundegaard, Thomas Blicher, Kasper Lamberth, Mikkel Harndahl, Sune Justesen, Gustav Røder, Bjoern Peters, Alessandro Sette, Ole Lund, et al. 2007. Netmhcpan, a method for quantitative predictions of peptide binding to any hla-a and-b locus protein of known sequence. *PloS one*, 2(8):e796.

Beau Norgeot, Benjamin S Glicksberg, and Atul J Butte. 2019. A call for deep-learning healthcare. *Nature medicine*, 25(1):14–15.

Mikel Olazaran. 1996. A sociological study of the official history of the perceptrons controversy. *Soc Stud Sci*, 26(3):611–659.

Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. 2018. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in neural information processing systems*, 31.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Further References

Timothy J O'Donnell, Alex Rubinsteyn, and Uri Laserson. 2020. Mhcflurry 2.0: improved pan-allele prediction of mhc class i-presented peptides by incorporating antigen processing. *Cell systems*, 11(1):42–48.

Norbert Pardi, Michael J Hogan, Frederick W Porter, and Drew Weissman. 2018. mrna vaccines—a new era in vaccinology. *Nature reviews Drug discovery*, 17(4):261–279.

Sepideh Parvizpour, Mohammad M Pourseif, Jafar Razmara, Mohammad A Rafi, and Yadollah Omidi. 2020. Epitope-based vaccine design: a comprehensive overview of bioinformatics approaches. *Drug Discovery Today*, 25(6):1034–1042.

Debleena Paul, Gaurav Sanap, Snehal Shenoy, Dnyaneshwar Kalyane, Kiran Kalia, and Rakesh Kumar Tekade. 2020. Artificial intelligence in drug discovery and development. *Drug Discovery Today*, 26:80 – 93.

Alexander H Pearlman, Michael S Hwang, Maximilian F Konig, Emily Han-Chung Hsiue, Jacqueline Douglass, Sarah R DiNapoli, Brian J Mog, Chetan Bettegowda, Drew M Pardoll, Sandra B Gabelli, et al. 2021. Targeting public neoantigens for cancer immunotherapy. *Nature cancer*, 2(5):487–497.

Bjoern Peters, Morten Nielsen, and Alessandro Sette. 2020. T cell epitope predictions. *Annu. Rev. Immunol.*, 38(1):123–145.

Nico Pfeifer and Oliver Kohlbacher. 2008. Multiple instance learning allows mhc class ii epitope predictions across alleles. In *Algorithms in Bioinformatics: 8th International Workshop, WABI 2008, Karlsruhe, Germany, September 15-19, 2008. Proceedings 8*, pages 210–221. Springer.

Marthinus Christoffel du Plessis, Gang Niu, and Masashi Sugiyama. 2014. Analysis of learning from positive and unlabeled data. In *NIPS*.

Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. 2021. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974.

Julien Racle, Justine Michaux, Georg Alexander Rockinger, Marion Arnaud, Sara Bobisse, Chloe Chong, Philippe Guillaume, George Coukos, Alexandre Harari, Camilla Jandus, et al. 2019. Robust prediction of hla class ii epitopes by deep motif deconvolution of immunopeptidomes. *Nature biotechnology*, 37(11):1283–1286.

Harish Ramaswamy, Clayton Scott, and Ambuj Tewari. 2016. Mixture proportion estimation via kernel embeddings of distributions. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2052–2060, New York, New York, USA. PMLR.

Rashika Ramola, Shantanu Jain, and Predrag Radivojac. 2018. Estimating classification accuracy in positive-unlabeled learning: characterization and correction strategies. In *BIOCOMPUTING 2019: Proceedings of the Pacific Symposium*, pages 124–135. World Scientific.

Daniele Ravì, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang-Zhong Yang. 2016. Deep learning for health informatics. *IEEE journal of biomedical and health informatics*, 21(1):4–21.

Yanrong Ren, Xiaolin Chen, Ming Feng, Qiang Wang, and Peng Zhou. 2011. Gaussian process: a promising approach for the modeling and prediction of peptide binding affinity to mhc proteins. *Protein and peptide letters*, 18(7):670–678.

Birkir Reynisson, Bruno Alvarez, Sinu Paul, Bjoern Peters, and Morten Nielsen. 2020. Netmhcpan-4.1 and netmhciipan-4.0: improved predictions of mhc antigen presentation by concurrent motif deconvolution and integration of ms mhc eluted ligand data. *Nucleic acids research*, 48(W1):W449–W454.

Mohammad Rezaee, Masoud Mahdianpari, Yun Zhang, and Bahram Salehi. 2018. Deep convolutional neural network for complex wetland classification using optical remote sensing imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(9):3030–3039.

Philip Rhodes, William Archibald Robson Thomson, Douglas James Guthrie, E. Ashworth Underwood, and Robert G. Richardson. 2023. History of medicine. In *Encyclopedia Britannica*. Encyclopaedia Britannica, Inc.

Robert A Rigby and D Mikis Stasinopoulos. 2005. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554.

Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. 2021. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*.

Daniel A Roberts, Sho Yaida, and Boris Hanin. 2022. *The principles of deep learning theory*. Cambridge University Press Cambridge, MA, USA.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer.

F. Rosenblatt. 1958. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. 2018. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR.

David Rügamer, Chris Kolb, and Nadja Klein. 2023. Semi-structured distributional regression – extending structured additive models by arbitrary deep neural networks and data modalities. *The American Statistician*, 0(ja):1–25.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1985. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

Stuart J Russell. 2010. *Artificial intelligence a modern approach*. Pearson Education, Inc.

Ugur Sahin and Özlem Türeci. 2018. Personalized vaccines for cancer immunotherapy. *Science*, 359(6382):1355–1360.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. 2019. A theoretical analysis of contrastive unsupervised representation learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5628–5637. PMLR.

Mansi Saxena and Nina Bhardwaj. 2018. Re-emergence of dendritic cell vaccines for cancer treatment. *Trends in cancer*, 4(2):119–137.

Benjamin Schubert and Oliver Kohlbacher. 2016. Designing string-of-beads vaccines with optimal spacers. *Genome medicine*, 8(1):1–10.

Martin G Schultz, Clara Betancourt, Bing Gong, Felix Kleinert, Michael Langguth, Lukas Hubert Leufen, Amirpasha Mozaffari, and Scarlet Stadtler. 2021. Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A*, 379(2194):20200097.

Alessandro Sette, Antonella Vitiello, Barbara Reherman, Patricia Fowler, Ramin Nayersina, W Martin Kast, CJ Melief, Carla Oseroff, Lunli Yuan, Jorg Ruppert, et al. 1994. The relationship between class i binding affinity and immunogenicity of potential cytotoxic t cell epitopes. *Journal of immunology (Baltimore, Md.: 1950)*, 153(12):5586–5592.

Shabnam Shamriz, Hamideh Ofoghi, and Nasrin Moazami. 2016. Effect of linker length and residues on the structure and stability of a fusion protein with malaria vaccine application. *Computers in biology and medicine*, 76:24–29.

Xiaoshan M Shao, Rohit Bhattacharya, Justin Huang, IK Ashok Sivakumar, Collin Tokheim, Lily Zheng, Dylan Hirsch, Benjamin Kaminow, Ashton Omdahl, Maria Bonsack, et al. 2020. High-throughput prediction of mhc class i and ii neoantigens with mhcnuggets. *Cancer immunology research*, 8(3):396–408.

Keerthi Shetty and Patrick A Ott. 2021. Personal neoantigen vaccines for the treatment of cancer. *Annual Review of Cancer Biology*, 5:259–276.

Further References

Ravid Shwartz-Ziv and Amitai Armon. 2022. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90.

Craig L. Slingluff. 2011. The present and future of peptide vaccines for cancer. *The Cancer Journal*, 17(5):343–350.

Mervyn Stone. 1977. An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):44–47.

Guangxin Su, Weitong Chen, and Miao Xu. 2021. Positive-unlabeled learning from imbalanced data. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization.

Ruoyu Sun. 2019. Optimization for deep learning: theory and algorithms. *arXiv preprint arXiv:1912.08957*.

Richard Sutton. 2019. The bitter lesson. *Incomplete Ideas (blog)*, 13(1).

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2018. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5552–5560.

Erico Tjoa and Cuntai Guan. 2020. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, 32(11):4793–4813.

Nora C. Toussaint, Pierre Dönnes, and Oliver Kohlbacher. 2008. A Mathematical Framework for the Selection of an Optimal Set of Peptides for Epitope-Based Vaccines. *PLoS Computational Biology*, 4(12):e1000246.

Nora C Toussaint and Oliver Kohlbacher. 2009. Optitope—a web server for the selection of an optimal set of peptides for epitope-based vaccines. *Nucleic acids research*, 37(suppl_2):W617–W622.

Nora C. Toussaint, Yaakov Maman, Oliver Kohlbacher, and Yoram Louzoun. 2011. Universal peptide vaccines – Optimal peptide vaccine design based on viral sequence conservation. *Vaccine*, 29(47):8745–8753.

Isaac Triguero, Salvador García, and Francisco Herrera. 2015. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information systems*, 42:245–284.

Jesper E Van Engelen and Holger H Hoos. 2020. A survey on semi-supervised learning. *Machine learning*, 109(2):373–440.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*. ACM.

Tal Vider-Shalit, Shai Raffaeli, and Yoram Louzoun. 2007. Virus-epitope vaccine design: informatic matching the hla-i polymorphism to the virus genome. *Molecular immunology*, 44(6):1253–1261.

Antonella Vitiello and Maurizio Zanetti. 2017. Neoantigen prediction and the need for validation. *Nature biotechnology*, 35(9):815–817.

Hao Wang and Dit-Yan Yeung. 2016. A survey on bayesian deep learning. *ACM Computing Surveys (CSUR)*, 53:1 – 37.

Chih-Jen Wei, Michelle C Crank, John Shiver, Barney S Graham, John R Mascola, and Gary J Nabel. 2020. Next-generation influenza vaccines: opportunities and challenges. *Nature reviews Drug discovery*, 19(4):239–252.

Max Welling and Yee Whye Teh. 2011. Bayesian learning via stochastic gradient langevin dynamics. In *International Conference on Machine Learning*.

KA Wetterstrand. 2021. Dna sequencing costs: Data from the nhgri genome sequencing program (gsp). www.genome.gov/sequencingcostsdata. Accessed: 2023-04-20.

Andrew G Wilson and Pavel Izmailov. 2020. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33:4697–4708.

Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P. Xing. 2015. Deep kernel learning. *ArXiv*, abs/1511.02222.

Tom Nuno Wolf, Sebastian Pölsterl, and Christian Wachinger. 2022. Daft: A universal module to interweave tabular data and 3d images in cnns. *NeuroImage*, 260:119505.

Simon N Wood. 2003. Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114.

Simon N Wood. 2006. Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, 62(4):1025–1036.

Simon N Wood. 2017. *Generalized additive models: an introduction with R*. chapman and hall/CRC.

Zhu Xiaojin and Ghahramani Zoubin. 2002. Learning from labeled and unlabeled data with label propagation. *Tech. Rep., Technical Report CMU-CALD-02–107*.

Yongliang Yang, S. James Adelstein, and Amin I. Kassis. 2009. Target discovery from data mining approaches. *Drug discovery today*, 14 3-4:147–54.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196.

Daniel Zeiberg, Shantanu Jain, and Predrag Radivojac. 2020. Fast nonparametric estimation of class proportions in the positive-unlabeled classification setting. In *AAAI Conference on Artificial Intelligence*.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. Mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*.

Lifang Zhang. 2018. Multi-epitope vaccines: a promising strategy against tumors and viral infections. *Cellular & molecular immunology*, 15(2):182–184.

Xiao-Meng Zhang, Li Liang, Lin Liu, and Ming-Jing Tang. 2021. Graph neural networks and their current applications in bioinformatics. *Frontiers in genetics*, 12:690049.

Yunrui Zhao, Qianqian Xu, Yangbangyan Jiang, Peisong Wen, and Qingming Huang. 2022. Dist-pu: Positive-unlabeled learning from a label distribution perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14461–14470.

Xueyi Zheng, Zhao Yao, Yini Huang, Yanyan Yu, Yun Wang, Yubo Liu, Rushuang Mao, Fei Li, Yang Xiao, Yuanyuan Wang, et al. 2020. Deep learning radiomics can predict axillary lymph node status in early-stage breast cancer. *Nature communications*, 11(1):1–9.

Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. Graph neural networks: A review of methods and applications. *AI open*, 1:57–81.

Laurence Zitvogel and Guido Kroemer. 2017. *Oncoimmunology: a practical guide for cancer immunotherapy*. Springer.

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12. Juli 2011, § 8 Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den 17.07.2023

Emilio Dorigatti