Computational Tools for Optimizing Targeted Cancer Treatments and Addressing Bias

Tepeli, Y.I.

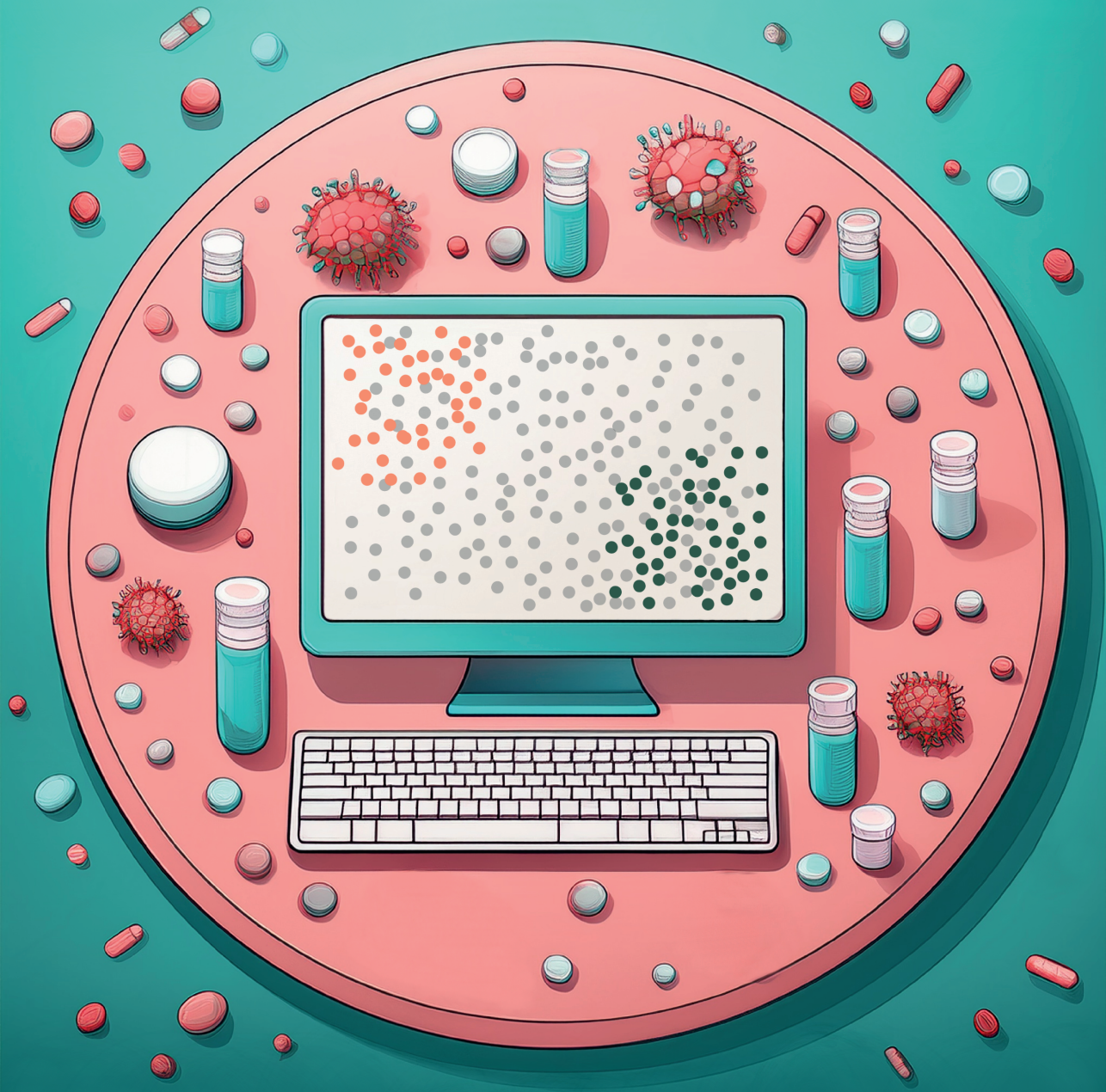
DOI
10.4233/uuid:a51440a5-7bd5-4e3b-b42f-04fa44325a89

Publication date
2025

Document Version
Final published version

Citation (APA)
Tepeli, Y. I. (2025). *Computational Tools for Optimizing Targeted Cancer Treatments and Addressing Bias*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:a51440a5-7bd5-4e3b-b42f-04fa44325a89

Computational Tools for Optimizing Targeted Cancer Treatments and Addressing Bias



Yasin İlkağan Tepeli

Computational Tools for Optimizing Targeted Cancer Treatments and Addressing Bias

COMPUTATIONAL TOOLS FOR OPTIMIZING TARGETED CANCER TREATMENTS AND ADDRESSING BIAS

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, Prof. dr. ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates
to be defended publicly on
Monday 16 June 2025 at 15:00 o'clock

by

Yasin İlkağan TEPELI

Master of Science in Computer Science and Engineering,
Sabancı University, Türkiye
born in Kale, Türkiye

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. ir. M.J.T. Reinders,	Delft University of Technology, *promotor*
Dr. J.S. de Pinho Gonçalves,	Delft University of Technology, *copromotor*
Independent members:	
Prof. dr. ir. B.P.F. Lelieveldt,	Delft University of Technology
Prof. dr. ir. J. de Ridder,	UMC Utrecht
Prof. dr. A.D.J. van Dijk,	University of Amsterdam
Dr. P. Kemmeren,	UMC Utrecht
Dr. W.M. Kouw,	Eindhoven University of Technology
Prof. dr. C.M. Jonker,	Delft University of Technology, reserve member

Keywords: Therapeutic target identification in cancer, genetic interactions, oncogene addiction, fairness, selection bias, semi supervised learning

Printed by: Ridderprint (www.ridderprint.nl)

Cover by: Generated with Adobe Firefly

ISBN 978-94-6384-789-6

An electronic copy of this dissertation is available at
`https://repository.tudelft.nl/`.

Dedicated to all children who have not been given an opportunity to receive a proper education.

CONTENTS

SUMMARY

The shift to precision medicine in cancer focuses on providing therapies targeting vulnerabilities of each individual patient tumor. This approach involves identifying cancer subtypes and discovering targets, such as genetic interactions, to treat patients who lack effective therapy. While computational tools, especially machine learning methods, are essential to analyze complex high-dimensional molecular data and suggest new candidate treatment strategies, their effectiveness is often questioned due to data-related challenges. Specifically, limitations in data collection result in sparse or biased biological data, hindering accurate decision-making and the identification of correct patterns. This thesis proposes state of the art solutions to learn improved prediction models for precision medicine and beyond by leveraging relevant data that was previously ignored, and addressing issues of data sparsity and bias.

Prediction of gene synthetic lethalities to identify novel therapeutic targets has overlooked sequence similarity, which is both a notable indicator of functional relation and available for every gene pair, unlike sparser data sources often used for this prediction task. Existing models also struggle to generalize beyond known synthetic lethalities due to an over reliance on data affected by prominent biases. Similarly, the stratification of cancer cohorts without effective treatments is challenging due to the small sample sizes of cancer (sub)cohorts such as oncogene-driven cohorts. In addition, stratification might not directly uncover an actionable treatment opportunity. The integration of dense protein sequence similarity and comprehensive drug response data each, together with methodological advances, led to significant improvements and revealed promising therapeutic opportunities.

Although these integrations improved the performance of computational methods, selection bias, a nonrandom sampling of training data, remained a significant issue affecting fair evaluation and generalizability. Thus, this thesis also introduces strategies to evaluate and mitigate the impact on model generalizability and fairness when the selected training data is not representative of the underlying population. We first artificially induce multivariate selection bias by favoring the selection of specific clusters of samples to study the fair evaluation of model generalizability. Then, to mitigate selection bias, we advance semi-supervised learning methods that use unlabeled data to gain insight into the distribution of the population beyond the labeled training data and promote sample diversity to counter confirmation bias typical of existing approaches. Our approaches include bias mitigation designed for specific machine learning models, such as forest ensembles and neural networks, and model-agnostic methods that operate under fewer assumptions. We show that diversity-guided semi-supervised learning strategies outperform existing domain adaptation techniques in the presence of various selection biases.

The computational methods proposed in this thesis enhance therapeutic target discovery in cancer and address selection bias in machine learning to advance precision medicine in cancer and improve the generalizability and fairness of bioinformatics models.

SAMENVATTING

De omschakeling naar precisiegeneeskunde in de oncologie richt zich op het bieden van therapieën die inspelen op de kwetsbaarheden van de tumor van iedere individuele patiënt. Deze benadering omvat het identificeren van kankersubtypes en het ontdekken van doelwitten, zoals genetische interacties, voor de behandeling van patiënten die geen effectieve therapie hebben. Hoewel computationele methodes, waaronder vooral machine learning, essentieel zijn voor het analyseren van complexe, hoog-dimensionale moleculaire data en het suggereren van nieuwe kandidaat-behandelstrategieën, wordt hun effectiviteit vaak in twijfel getrokken vanwege data-gerelateerde uitdagingen. Specifiek leiden beperkingen in de dataverzameling tot schaarse of bevooroordeelde biologische data, wat een nauwkeurige besluitvorming en het herkennen van juiste patronen belemmert. Deze thesis stelt state-of-the-art oplossingen voor om verbeterde voorspellingsmodellen te ontwikkelen voor precisiegeneeskunde en daarbuiten, door gebruik te maken van relevante data die voorheen genegeerd werd en door problemen van dataschaarsheid en bias aan te pakken.

De predictie van genetische synthetische lethaliteiten om nieuwe therapeutische doelwitten te identificeren, heeft de sequentiegelijkenis over het hoofd gezien, een opmerkelijke indicator van functionele relatie die voor elk genpaar beschikbaar is, in tegenstelling tot de vaak schaarser aanwezige databronnen die voor deze voorspellingstaak worden gebruikt. Bestaande modellen hebben ook moeite om te generaliseren buiten de bekende synthetische lethaliteiten, vanwege een overmatige afhankelijkheid van data die wordt beïnvloed door prominente biases. Evenzo is de stratificatie van kankercohorten zonder effectieve behandelingen een uitdaging, vanwege de kleine steekproefomvang van kanker-(sub)cohorten, zoals oncogeen-gedreven cohorten. Bovendien onthult stratificatie mogelijk niet direct een bruikbare behandeloptie. De integratie van dichte eiwitsequentiegelijkenis en uitgebreide geneesmiddelresponsdata, samen met methodologische vooruitgangen, leidde tot significante verbeteringen en onthulde veelbelovende therapeutische mogelijkheden.

Hoewel deze integraties de prestaties van computationele methoden verbeterden, bleef selectiebias, een niet-willekeurige sampling van trainingsdata, een significant probleem dat een eerlijke evaluatie en generaliseerbaarheid beïnvloedt. Daarom introduceert deze thesis ook strategieën om de impact op de generaliseerbaarheid en eerlijkheid van modellen te evalueren en te mitigeren wanneer de geselecteerde trainingsdata niet representatief is voor de onderliggende populatie. Allereerst introduceren we kunstmatig multivariate selectiebias door de selectie van specifieke clusters te bevoordelen, om zo de eerlijke evaluatie van de generaliseerbaarheid van modellen te bestuderen. Vervolgens ontwikkelen we, ter mitigatie van selectiebias, semi-gesuperviseerde leermethoden die gebruikmaken van ongelabelde data om inzicht te krijgen in de populatiedistributie

buiten de gelabelde trainingsdata en om de diversiteit van de steekproef te bevorderen ter bestrijding van de typische bias bij bestaande benaderingen. Onze benaderingen omvatten biasmitigatie die is ontworpen voor specifieke machine learning modellen, zoals ensembles van beslissingsbomen en neurale netwerken, evenals model-agnostische methoden die werken onder minder aannames. Wij tonen aan dat op diversiteit gerichte semi-gesuperviseerde leermethoden beter presteren dan bestaande domeinadaptatietechnieken in aanwezigheid van diverse selectiebiases.

De voorgestelde computationele methoden in dit proefschrift verbeteren de identificatie van therapeutische doelwitten in kanker en pakken de selectiebias in machine learning aan, om de precisiegeneeskunde in kanker te bevorderen en de generaliseerbaarheid en billijkheid van bioinformatica-modellen te verbeteren.

1

INTRODUCTION

1.1. ANTI-CANCER THERAPEUTIC TARGET IDENTIFICATION

1.1.1. CANCER AND CONVENTIONAL TREATMENTS

First described by ancient Egyptians, with references to the disease dating back to around 3000 BCE, and then coined by Greek physician Hippocrates due to its phenotypical resemblance to a crab, carcinoma, or cancer [1], is still one of the leading cause of death worldwide, with an estimated 20 million incidents and 10 million deaths in 2022 [2]. It is caused by the accumulation of mutations in the DNA due environmental and damaged molecular processes. Some of these cells with enough accumulated mutations escape cell death, turn into cancer cells, and start spreading. Thus, it is also characterized by the uncontrolled proliferation, division, and growth of cells, leading to the formation of malignant tumors that can invade surrounding tissues and metastasize to distant organs.

Conventional cancer treatments, including surgery, radiation therapy, and chemotherapy, have been the cornerstone of cancer care for many years. These treatments aim to remove or destroy cancer cells but often come with significant limitations and side effects.

Surgery involves the physical removal of cancerous tissue and is most effective for localized early-stage tumors but may be less effective for advanced or metastatic disease [3]. For instance, lumpectomy or mastectomy are common surgical procedures for breast cancer [4, 5], while prostatectomy is used for prostate cancer [6, 7]. Surgery, although effective, also carries risks such as infection, bleeding [8]. Thus, it is not a completely effective treatment option.

Radiation therapy uses high-energy rays to kill cancer cells or shrink tumors, targeting specific areas with precision [9]. Although common for some cancer types such as skin, breast, prostate, and head and neck cancers, radiation based therapies mostly require a combination with another treatment to be effective [10]. However, even then, radiation can damage surrounding healthy tissue, leading to side effects such as fatigue, skin changes, and other organ-specific effects depending on the

treatment site. For example, radiation for head and neck cancer can result in dry mouth, difficulty swallowing, and changes in taste [11].

Chemotherapy involves the use of cytotoxic drugs to kill rapidly dividing cancer cells [12]. It is often used for metastatic cancers and as an adjuvant treatment to reduce the risk of recurrence. Common chemotherapy drugs include doxorubicin [13], cisplatin [14, 15], and paclitaxel[16]. However, chemotherapy lacks specificity, affecting both cancerous and rapidly dividing healthy cells [17]. This nonspecific nature leads to side effects like nausea, hair loss, and increased susceptibility to infections. Patients undergoing chemotherapy for colorectal cancer, for instance, may experience severe gastrointestinal issues that can lead to death [18].

While these treatments have been foundational in cancer care, their associated toxicities and nonspecific nature coupled with the concept of one treatment for all make them ineffective and also highlight the need for more targeted and personalized therapeutic approaches (Fig. 1.1).

1.1.2. CANCER IS NOT A SINGLE DISEASE

Although numerous conventional and targeted therapies have been developed to date, cancer remains a leading cause of death, responsible for approximately 10 million deaths each year [2]. One major challenge in developing a universal treatment is that cancer is not a single disease although often described by one term. Instead, it consists of various diseases with different genetic characteristics, but each of these diseases is characterized by uncontrolled growth and spread of cells. This diversity mainly arises from a unique accumulation of mutations, modulated by the germline genetic landscape of the patient, the tissues and organs it affects, and the environment [19]. As a consequence, these factors cause difference between responses to treatments. In the end, this diversity can result in both inter-tumor heterogeneity which is the variation between different cancer patients, and intra-tumor heterogeneity which is the variation within one tumor or one site in a patient. This heterogeneity makes it difficult to find a single solution that works for all types of cancer. For the effective treatment of cancer patients with different characteristics, it is important to understand the cancer heterogeneity and the factors contributing to it. Although these factors can be presented under different terms, they are usually dependent on each other.

HETEROGENEITY BY TUMOR ENVIRONMENT

Cancer can originate in various locations across the body, thus both the origin of tissue and tumor microenvironment may impact how cancer cells emerge, differentiate, and respond to treatments [20, 21]. Although mainly organized or categorized by the organ or tissue of origin, the behavior of cancer may further differ according to components in the tumor microenvironment such as immune cells, blood vessels, fibroblasts, and extracellular matrix (ECM).

Firstly, each tissue or organ in our body may contain cells that are specialized for

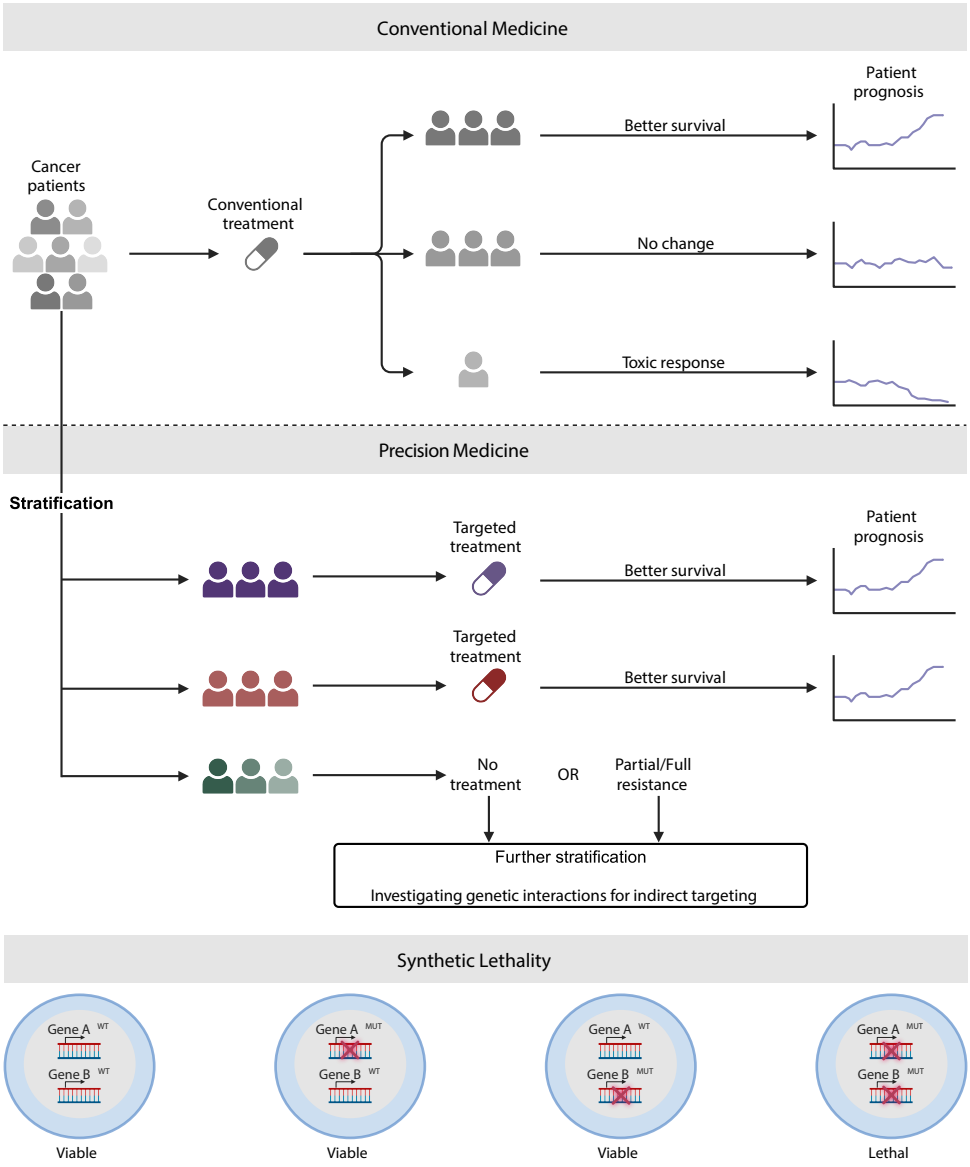


Figure 1.1: Overview of treatment concepts in cancer. **Conventional medicine** uses the same treatment on all patients whereas **precision medicine** stratifies patients to find the correct targeted treatment. Properties of **synthetic lethality**, a genetic interaction, can be used when cancer drivers cannot be targeted directly.

specific tasks and exclusive to that location, besides cells that are common across the body [22]. However, even the types of cells that are common throughout the body will exhibit differing characteristics in different locations, such as macrophages exhibiting different behaviors and functions in the liver (Kupffer cells) [23] with respect to brain (microglia) [24].

The heterogeneity among tissues and the specialized functions of cells significantly influence the process of tumorigenesis. For example, cancer originating from epithelial cells may differ across tissue types. Breast cancer can arise from malignantly transformed mammary epithelial cells[25] and often relies on hormone receptor signaling pathways, such as estrogen receptors [26], that can be targeted with hormone-related therapies like tamoxifen [27]. On the other hand, bronchial epithelial cells, with mutations in genes such as EGFR and KRAS, may lead to different types of lung cancer [28] and often require targeted therapies like tyrosine kinase inhibitors (TKIs) [29, 30]. Colorectal cancer can also originate from the epithelial cells lining the colon or rectum [31], frequently involving mutations in genes like APC [32], KRAS, and TP53, with implications for the Wnt signaling pathway [33].

Beyond epithelial cells, skin cancers like melanoma originate from melanocytes, influenced by UV exposure and commonly involving BRAF mutations, are typically treated with BRAF inhibitors and immune checkpoint inhibitors [34]. Brain cancers such as glioblastoma arise from glial cells and involve alterations in the PI3K/AKT and RAS/MAPK pathways, necessitating treatments that can penetrate the blood-brain barrier. All these examples point at that cancer in each organ or tissue may bear significant differences in terms of the reason and treatment opportunities.

Beyond the macro-level environment, such as the tissue of origin, the interaction between tumor cells and microenvironment components plays a critical role in cancer cell growth, invasion, and response to treatment. Variations in the microenvironment can lead to considerable intra-tumor heterogeneity within one tumor, which may cause resistance to many cancer treatments [35–38]. One well-known component of the tumor microenvironment that causes heterogeneity is tumor-associated macrophages (TAMs), which are immune cells that can polarize into either M1-like TAMs or M2-like TAMs. While M1-like TAMs prevent tumor growth by enhancing cytotoxic activity, M2-like TAMs suppress immune responses and promote drug resistance [39].

Furthermore, colorectal cancers originating in different regions of the colon, such as the right versus the left side, can exhibit significant differences in genetic profiles and treatment responses, influenced by the unique microenvironments of each colon segment [40]. Moreover, while resistance to treatment can result from cancer cells developing new dependencies on alternative genes for survival after the initial target is targeted, it can also arise from changes in the environment that do not directly affect the tumor cells [41].

The varied origins and behaviors of different cancers underscore the need for different diagnostic and therapeutic approaches tailored to the unique cellular environments.

HETEROGENEITY BY MECHANISM

The mechanisms underlying cancer development and progression are diverse, consisting of genetic mutations, epigenetic alterations, and aberrations in cellular signaling pathways. Genetic mutations in cancer genes drive uncontrolled cell growth and spread [42]. Epigenetic changes, including DNA methylation and histone modifications, can silence genes that regulate the cell cycle and apoptosis [43]. Moreover, as a result of epigenetic changes and genetic mutations, aberrant signaling pathways, like the PI3K/AKT/mTOR pathway, cause tumor survival and resistance to therapy [44]. Understanding these varied mechanisms can assist us in discovering novel therapeutic targets and developing treatments tailored for specific cancer types.

For example, in lung cancer, mutations in the EGFR gene cause cancer cell survival, but also make these tumors particularly responsive to EGFR inhibitors like erlotinib and gefitinib [45]. On the other hand, mutations in the KRAS gene commonly found in lung cancer cells, can result in resistance to those same inhibitors, requiring alternative therapeutic strategies [46]. Epigenetic changes may also play a role, such as the hypermethylation of the BRCA1 promoter leading to silencing of the BRCA1 gene and contributing to breast cancer development [47]. Similarly, PTEN mutations, which also fall under genetic alterations, can result in the activation of the PI3K/AKT pathway, promoting cell survival and growth, which can be targeted with PI3K inhibitors [48].

In short, across different tissues and even within the same tissue or tumour, cancer cells can vary and have different characteristics due to the mechanisms it affects.

HETEROGENEITY BY OTHER FACTORS

Cancer heterogeneity is influenced not only by the environment or mechanisms but also by other factors such as clonal evolution, selective pressures, and stochasticity. Clonal evolution refers to the process by which tumor cells acquire new mutations over time, leading to the emergence of diverse subclones within the same tumor. This dynamic process results in a heterogeneous population of cancer cells that can differ in their growth rates, metastatic potential, and response to treatment [49]. For instance, in colorectal cancer, clonal evolution can result in different regions of the same tumor acquiring distinct mutations, such as those in the KRAS, NRAS, or BRAF genes, leading to subclones with varying sensitivity to targeted therapies like EGFR inhibitors [50].

Selective pressures, such as the immune response, hypoxia, and therapeutic interventions, further shape tumor heterogeneity [51]. Mainly, hypoxic conditions within a tumor can select for cells that survive in low-oxygen environments, often resulting in more invasive and therapy-resistant cancer phenotypes [52]. These adaptations create a heterogeneous population of cancer cells with varying abilities to evade immune detection and destruction, leading to more aggressive and treatment-resistant clones.

Moreover, stochasticity in the acquisition of mutations during cell division and tumor

growth can lead to differences in the genetic and phenotypic properties of cancer cells, even among those that originated from the same clone. This randomness can also affect how cancer cells respond to environmental stresses and therapeutic interventions, further complicating treatment strategies.

Understanding these diverse factors is essential for developing more effective and personalized cancer treatments that address the complexity of tumor heterogeneity.

1.1.3. LEVERAGING CANCER FACTORS FOR TARGETED TREATMENT

The heterogeneity of cancer, coupled with the availability of vast genetic data and advanced technologies, have respectively forced and enabled a deeper analysis of tumor development, offering insights into the mechanisms of cancer and highlighting vulnerabilities that can be exploited for therapeutic purposes. The latest developments in cancer research have underscored the dual nature of the factors driving cancer: while they contribute to tumor development, they may also constitute unique targets for treatment.

MECHANISMS UNDERLYING CANCER AND THERAPEUTIC OPPORTUNITIES

The hallmarks of cancer [53] outline the essential traits that enable tumor growth and metastasis, including sustained proliferative signaling, evasion of growth suppressors, resistance to cell death, enabling of replicative immortality, induction of angiogenesis, and activation of invasion and metastasis. These hallmarks, while instrumental in cancer progression, simultaneously offer critical targets for therapeutic intervention. For instance, immune checkpoint inhibitors like pembrolizumab exploit the mechanisms that tumors use to evade immune detection, thereby enhancing the ability of the immune system to recognize and destroy cancer cells [54]. Similarly, venetoclax, a BCL-2 inhibitor, induces apoptosis in cancer cells that evade cell death through BCL-2-mediated survival pathways [55, 56].

While understanding these hallmarks is vital for developing targeted therapies, they do not provide sufficient specificity for individualized treatment. To enhance therapeutic precision, it is necessary to delve deeper into the specific factors that contribute to these hallmarks, such as oncogenes and tumor suppressor genes.

Oncogenes are mutated or overexpressed forms of proto-oncogenes, which are genes that normally regulate cell growth, proliferation, and differentiation. Upon mutation or overexpression, these genes become oncogenes that drive cancer initiation and progression with uncontrolled cell proliferation. Prominent examples include HER2 (ERBB2) in breast cancer [57], BCR-ABL in chronic myeloid leukemia [58], and EGFR and KRAS in various cancers, such as colorectal and lung carcinoma [36, 59, 60]. These oncogenes, while instrumental in cancer development, can also present as targets for therapeutic intervention.

For cancers driven by increased activity of an oncogene, using specific oncogene-targeting inhibitors such as trastuzumab for HER2-positive breast cancer [61, 62] or imatinib for BCR-ABL-positive leukemia [63], can improve treatment outcomes. Tyrosine kinase inhibitors (TKIs) like gefitinib and erlotinib, targeting EGFR-mutated non-small cell lung cancer (NSCLC), also provide a more precise treatment approach with higher response rates and longer progression-free survival over conventional chemotherapy [64]. Similarly, the ALK inhibitor crizotinib has transformed treatment for patients with ALK-positive NSCLC with great improvements compared to chemotherapy [65].

The strategy of precisely targeting oncogenes that drive cancer has resulted in more effective and less toxic treatment options, showing superiority over conventional chemotherapy.

Tumor Suppressing Genes (TSGs) such as TP53, PTEN, and BRCA1, play pivotal roles in regulating cell growth and preventing tumor formation. Contrary to oncogenes, the loss or inactivation of these genes leads to uncontrolled cell proliferation since these genes normally prevent tumor formation when active. Strategies to restore the function of these genes or counteract their loss can provide therapeutic benefits. However, targeting tumor suppressor genes is more challenging than targeting oncogenes, because direct targeting may not be possible if tumor suppressor genes are inactivated or non-functional [66]. To reactivate them or their functions, it is necessary to identify and reverse the cause of their inactivation.

For instance, TP53 [67], a gene previously considered an oncogene, but now recognized as an essential tumor suppressor gene, is often inactivated in multiple cancers through mutations that lead to loss of function. Similarly, TSGs like ARID1A and SETD2 are often mutated in cancers, leading to altered gene expression and tumor progression. Regaining and restoring normal gene function for these genes are not simple but also not impossible. For example, normal functions of p53 can be rescued by various techniques or can also be used as an advantage since these abberations also make tumour cells dependent on other processes that can be targeted [68]. Moreover, some TSGs are implicated in various pathways, further complicating their targeted reactivation. For example, the Wnt/APC pathway involves TSGs like APC and AXIN1, which, when mutated, lead to aberrant cell signaling and cancer progression.

Although more complicated than oncogenes, understanding the complexities of these TSGs and developing innovative approaches to address their inactivation remains as an essential part of cancer research and therapy development.

1.1.4. CHALLENGES AND OPPORTUNITIES IN TARGETED THERAPY

Although significant advances have been made in targeted cancer therapies, several challenges remain that must be addressed to improve their effectiveness. Targeting oncogenes or tumor suppressor genes presents considerable difficulties due to

various factors. Some genes are inherently difficult to target directly (i.e. they are untargetable) because of their specific characteristics, such as the lack of accessible binding sites or their complex cellular context. Additionally, most tumor suppressor genes (TSGs) cannot be directly reactivated once inactivated. Moreover, cancer cells can develop resistance to targeted therapies in many ways, reducing the long-term efficacy of these treatments. Furthermore, targeted therapies can pose big risks, including off-target effects and toxicity, which can have a high impact on patient well-being.

CHALLENGES IN TARGETED TREATMENT

Untargetable cancer drivers. Some factors driving cancer are untargetable or undruggable with current technologies, causing significant challenges in the field of oncology. In the case of oncogenes, often times this issue relates to their structural and chemical characteristics. For example, the protein associated with the KRAS oncogene is difficult to target directly, since it lacks suitable binding pockets for small molecules [69]. This limitation can be overcome by indirect methods, such as targeting downstream effectors in the KRAS pathway or exploiting genetic interactions to inhibit KRAS activity. Similarly, MYC, an oncogene involved in many cancers, is difficult to target directly due to its high-affinity DNA-binding properties and lack of suitable drug-binding sites [70].

For tumor suppressor genes such as BRCA1 and BRCA2, whose mutations are critical in certain breast and ovarian cancers, direct targeting or inhibition is not a solution given that the genes are already inactivated and that is what drives tumor formation in the first place. Alternative strategies must thus be found to indirectly overcome the disruption of TSGs. For instance, therapeutic strategies addressing the loss of the well-known TSG TP53 often focus on compensation mechanisms by reactivating associated pathways. For instance, MDM2 inhibitors aim to restore p53 function by preventing its degradation [71]. However, some mutations affecting TP53, such as the TP53-Y220C mutation, make this important TSG untargetable despite efforts to find indirect methods.

In a majority of cases, while cancer drivers might be known, they cannot be effectively targeted.

Resistance. Cancer cells can develop resistance to targeted therapies through various mechanisms, including secondary mutations in the target, activation of alternative signaling pathways, and phenotypic changes. For example, in non-small cell lung cancer (NSCLC), secondary mutations in the EGFR gene, such as the T790M mutation, can confer resistance to first-generation EGFR inhibitors like gefitinib and erlotinib [72]. This has led to the development of second-generation inhibitors like osimertinib, which can effectively target the T790M mutation and overcome EGFR inhibitor resistance [73].

Another mechanism of resistance involves the activation of alternative signaling

pathways. In melanoma, resistance to BRAF inhibitors such as vemurafenib can occur through the activation of the MAPK pathway allowing cancer cells to bypass BRAF inhibition and continue proliferating [74]. Combining BRAF inhibitors with MEK inhibitors is an effective strategy to overcome this type of resistance [75].

Phenotypic changes can also contribute to resistance. For instance, in prostate cancer, androgen receptor (AR) signaling can adapt and evolve in response to AR-targeted therapies like enzalutamide [76]. This can lead to the emergence of AR splice variants that are constitutively active and do not require androgen for activation, thus rendering AR-targeted therapies ineffective.

Overall, intrinsic resistance combined with acquired resistance over time makes the targeted therapy even more challenging. Overcoming the resistance requires a deep understanding of the underlying mechanisms and the development of novel treatment approaches.

Risks. While targeted therapy offers reduced risks compared to conventional chemotherapy, it can still impact normal cells, leading to toxicity or adverse off-target effects. For example, because EGFR is not only overexpressed in cancer cells but also present in normal epithelial tissues of the skin and intestines, treatments like gefitinib that target EGFR can cause side effects such as skin rashes and diarrhea [77]. Similarly, HER2-targeted therapies like trastuzumab can result in cardiac dysfunction, since HER2 is also expressed in cardiac myocytes, underscoring the difficulty of achieving precise specificity in targeted treatments [78]. Furthermore, some drugs may lack high specificity, potentially binding to unintended targets and causing off-target effects. For instance, tyrosine kinase inhibitors like nilotinib or imatinib, designed to target BCR-ABL, have been found to produce off-target effects with both short- and long-term consequences [79].

Managing these risks involves several strategies, such as optimizing dosing to balance efficacy and toxicity, or developing more selective inhibitors that specifically target the mutant or overexpressed forms of proteins/genes in cancer cells while sparing normal cells. For example, osimertinib, a third-generation EGFR inhibitor, selectively targets the T790M mutant form of EGFR, which is often associated with resistance to first- and second-generation inhibitors, and offers a better safety profile compared to its predecessors [80].

In summary, it is essential to identify the potential risks associated with targeted treatments and understand which patient groups are most affected to enhance the overall effectiveness of cancer therapies.

OPPORTUNITIES TO OVERCOME CHALLENGES IN TARGETED THERAPY

Precision medicine and patient stratification. Precision medicine tailors treatment based on individual patient characteristics and molecular profiles, including genetic and epigenetic features [81]. This approach enhances the likelihood of therapeutic success by selecting treatments that are most likely to be effective for a specific

patient. However, finding a unique treatment for each patient can be extremely expensive and technologically challenging. An alternative strategy typically involves grouping patients with similar characteristics who might also respond similarly to the same treatments (Fig. 1.1).

Historically, precision cancer treatment began by grouping patients based on tissue or organ type, an important decision facilitated first by Rudolf Virchow and the World Health Organization (WHO). The classification system of WHO allows for the systematic categorization of cancers by their tissue of origin to help guide treatment decisions and research [82]. This classification enables oncologists to identify common characteristics within cancer types, such as histological features and growth patterns, pivotal for developing effective treatment protocols. Building on this foundation, precision medicine evolved to consider genetic characteristics, such as specific driver gene mutations. For instance, previous studies discovered that the success of EGFR inhibitors in colorectal cancer patients depends on KRAS mutation status, where those inhibitor drugs work exclusively on KRAS wild-type (WT) patients [83–85]. Thus, colorectal cancer patients can be stratified by KRAS mutation status and the EGFR targeting drugs can be given to patients that benefit from such treatment, mainly those patients whose tumors do not harbor KRAS mutations. Similarly, PARP inhibition therapy can benefit breast and ovarian cancer patients with BRCA1/2 mutations [86], which requires stratifying patients by BRCA1/2 mutation status. Sensitivity of other cancer types such as glioblastoma to PARP inhibitors like olaparib can also be dependent on homologous recombination deficiency and microsatellite instability status of the tumor [86]. Personalized treatment strategies have further advanced to consider specific types of mutations within a gene. In non-small cell lung cancer (NSCLC), patients with EGFR mutations are further stratified by secondary mutations, such as T790M [73]. This stratification allows for drugs like osimertinib, designed to be effective for patients with this specific mutation, to improve treatment efficacy and patient outcomes. Although the advancement is vast, for most cancer types with known mechanisms, there is typically no group of patients that responds with 100% success. Thus, stratifying strategies for cancer patients remain relevant to consider finer details such as the type or location of the mutation of the cancer-driven gene as well as other biomarkers that can be predictive of response to a treatment [87–89].

Leveraging genetic interactions: synthetic lethality. When oncogenes and tumor suppressor genes are undruggable or untargetable, leveraging genetic interactions offers a viable alternative for cancer treatment. It is expected that when cancer originates due to a dysfunction in an oncogene or TSG, the cancer cells may become dependent on other genes, pathways, or factors for survival which can be targeted to kill cancer cells [90]. One such interaction is synthetic lethality, which occurs when the simultaneous impairment of two genes results in cell death, while the impairment of either gene alone is non-lethal (Fig. 1.1). This strategy is particularly useful in cancer therapy because cancer cells often harbor specific mutations that can be exploited by inhibiting interacting partners, leading to selective cancer cell

death without harming normal cells. A prime example of synthetic lethality is the use of PARP inhibitors in BRCA1/2 mutant cancers [91]. The BRCA mutations compromise the homologous directed DNA repair pathway, which increases the activity of alternative repair involving the PARP gene. Inhibiting the activity of PARP in this case leads to an accumulation of DNA damage, promoting cancer cell death.

Beyond synthetic lethality, other genetic interactions can be exploited, such as synthetic rescue and collateral lethality. Synthetic rescue involves restoring cell viability in the case of a harmful mutation by altering another gene or pathway [92, 93]. Although less commonly applied in cancer therapy, it holds potential in genetic disorders where modifying a secondary gene can mitigate the effects of a primary mutation. On the other hand, collateral lethality, a concept related to synthetic lethality, takes advantage of passenger genes deleted alongside tumor suppressor genes [94]. When cancer is driven by genomic deletion of TSGs, there are also additional passenger deletions that are not directly related to cancer. However, if one of these genes is essential for cell survival but redundant due to another backup gene (e.g. synthetic lethal relation), their backup gene can be targeted with the aim of killing the cancer cells. For instance, deletion of the 1p36 locus not only affects multiple cancer driver TSGs but also the neighboring gene ENO1. Although ENO1 is essential for cell survival, cells are still viable due to a backup pathway through ENO2. Targeting this pathway can selectively kill those cancer cells [95].

Synthetic lethality, however, remains the most explored approach, offering a promising avenue for developing targeted cancer therapies. By identifying and targeting specific genetic dependencies of cancer cells, treatments can be designed that are both effective and selective, minimizing harm to normal cells. This approach is further enhanced by advances in genomic profiling and precision medicine, enabling the identification of patient-specific vulnerabilities and tailoring treatments.

1.1.5. COMPUTATIONAL ADVANCES FOR TARGETED THERAPY

While the potential for targeted therapy hinges on a deep biological understanding of cancer mechanisms, our knowledge remains limited as many aspects of cancer development and progression are still not fully understood. With the growing availability of medical records, genotypic data (omics), and phenotypic data, computational tools can play an increasingly important role in uncovering new insights into cancer mechanisms and in advancing the development and implementation of targeted cancer therapies. These tools enable the computational analysis of large datasets for automated and systematic identification of candidate therapeutic targets or patient subgroups, exploration of genetic interactions, and prediction of patient responses to treatments.

COMPUTATIONAL STRATIFICATION FOR PRECISION MEDICINE

Advanced computational methods can greatly facilitate patient stratification for precision cancer treatment, through comprehensive identification of groups that

are most likely to benefit from specific therapies. Machine learning algorithms are essential for analyzing vast amounts of omics data, including genomic, transcriptomic, and proteomic profiles to classify tumors into subtypes and predict their response to targeted therapies. Stratification typically occurs through two primary approaches: classification and clustering.

In classification, patients are grouped into predefined categories based on known characteristics, such as the primary origin of the disease belonging to a well-known tumor subtype, or responsiveness to a specific drug [96]. For example, in the context of EGFR-driven lung cancer, classification algorithms may be used to determine which patients will respond to tyrosine kinase inhibitors by training a machine learning model using previously collected data [97, 98]. These classification tasks are widely applied in precision oncology, enabling more personalized treatment approaches tailored on the basis of genetic profile.

Clustering [99], on the other hand, involves grouping patients without predefined categories, allowing the discovery of stratifications and groups that have not been investigated yet, e.g. a new subtype within a cancer type [100]. Clustering is mainly facilitated through similarity-based methods [101–103], or other methods such as Gaussian mixture models where each sample originates from a Gaussian distribution. In breast cancer, for instance, clustering based on gene expression profiles has revealed subgroups of patients in previously defined subtypes such as HER2-positive, estrogen receptor-positive (ER+), and triple-negative breast cancer [104]. Each of these subtypes requires distinct therapeutic approaches, showing the benefits of clustering in uncovering novel subtypes and tailoring treatments more precisely. Beyond known subtypes, clustering has been particularly effective in identifying previously unrecognized molecular subtypes in cancers such as glioblastoma [105, 106] or large B-cell lymphoma [107], and identifying novel groups within cancer subtypes such as a split in luminal A breast cancer [108].

These discoveries can inform the development of more targeted therapies by addressing the specific biological characteristics of each cancer subtype, ultimately improving treatment outcomes. By leveraging large datasets, computational approaches enable researchers to uncover complex patterns in cancer biology that may not be apparent through traditional methods, thereby enhancing the precision and effectiveness of cancer treatment strategies.

However, computational stratification still often focuses on large patient cohorts, leaving a gap in research on treatment opportunities for smaller cohorts, such as KRAS-addicted colorectal cancer patients, where more targeted analysis is still needed.

COMPUTATIONAL DISCOVERY OF GENETIC INTERACTIONS: SYNTHETIC LETHALITIES

Identifying genetic interactions is crucial to discover new therapeutic possibilities. However, experimentally testing all possible interactions between all genes or other entities in all possible genetic contexts is not feasible. Computational approaches are

powerful tools for uncovering complex interactions between genes, drugs, diseases, and other biological entities by leveraging large datasets and machine learning algorithms. One fundamental use of these methods is constructing and exploiting patterns in protein-protein interaction (PPI) graphs, which provide a framework to understand how proteins interact to carry out cellular functions and influence phenotypes of interest such as response to treatment as well as predicting various interactions within these PPI graphs [71, 109, 110].

While these graphs offer a broad overview of biological networks, refined approaches focus on specific interactions like gene-disease associations and synthetic lethal (SL) interactions. Techniques such as CRISPR-Cas9 knockout and RNA interference (RNAi) are used to experimentally identify and validate synthetic lethal gene pairs in cancer cells. However, the number of possible interactions and the costs involved in testing them are simply prohibitive. As a result, computational methods including conventional statistical approaches and machine learning (ML) models have been proposed to identify promising gene pairs with SL potential. The first methods proposed for computational SL prediction were statistical approaches like DAISY [111], BiSep [112], and ISLE [113], which depend on known SL properties such as mutual exclusivity of mutations. The emergence of powerful ML algorithms and the availability of experimentally identified SL relations further yielded approaches for training ML models that can predict SL relations. We categorize them into two groups of methods, topology- and feature-based. Topology-based methodsconstruct a network of pairwise SL interactions between genes and use techniques like matrix factorization [114–116] or graph-based methods [117, 118] to identify patterns and infer new interactions based on the existing SL topology. Feature-based methods rely on supervised ML algorithms to learn models with complex rules underlying synthetic lethality from a variety of omics data modalities. These feature-based models include DiscoverSL [119], EXP2SL [120], and SBSL [121].

However, effective SL prediction is still limited by various factors. Statistical methods may not be powerful enough to uncover complex SL relationships and they tend to perform worse than machine learning-based methods [121]. On the other hand, the literature presents that topology-based ML methods are susceptible to biases intrinsic to SL networks [121]. Although feature-based ML models look more robust, they do not utilize all available data sources that are relevant for SL prediction such as aminoacid sequences, thus also suffering from data sparsity and low sample size.

1.1.6. INTRINSIC BIAS IN GENETIC INTERACTION DATASETS

Many biological datasets used in machine learning, including SL data, are affected by selection bias and other forms of bias due to limitations in data collection and experimental validation [121–125]. These datasets are often generated by researchers, whose experiments inherently influenced by existing knowledge, research priorities, and funding possibilities. As a result, data collection tends to focus on well-established factors, which may become overrepresented compared to rarer factors that are often neglected. For example, in gene studies, this imbalance makes

datasets biased toward more frequently studied genes, pathways, and interactions, leaving potentially important but lesser-known genes or pathways ignored.

In the context of synthetic lethality (SL), selection bias leads to overrepresentation of interactions of well-researched genes, especially those already associated with cancer pathways [121]. The consequences of these biases can be seen in SL prediction models, which rely on these data to identify SL interactions. When trained on biased datasets, these models tend to perform well in scenarios where the tested genes are similar to the ones in training data, but struggle to make accurate predictions when tested genes or interactions are underrepresented [121]. This can result in an incorrect assessment of model performance and a failure to identify new SL interactions for targeted cancer therapies.

In short, biases undermine the effectiveness of machine learning algorithms, which may overfit to the well-represented data points, reducing their ability to generalize to new or unseen interactions. The issue is particularly problematic in bioinformatics, given the prevalence of selection bias in biology datasets and the impact that prediction models can ultimately have on research and clinical practice.

1.2. MITIGATING SELECTION BIAS IN MACHINE LEARNING AND APPLICATIONS IN BIOINFORMATICS

Machine learning (ML) algorithms have become integral to various bioinformatics tasks, offering powerful tools for analyzing complex biological data and uncovering relationships and properties within biology that would be challenging for humans to figure out. These algorithms are mainly designed to identify patterns and make predictions, enabling applications such as protein structure prediction [126], drug response prediction [127–129], patient survival analysis [130, 131], cell type classification [132, 133], and discovery of genetic interactions [134]. A significant portion of these ML applications involves supervised learning, where models are trained using annotated datasets that combine sample characteristics with corresponding labels. Nevertheless, unsupervised learning methods, such as clustering, remain widely utilized, especially when annotations are unavailable.

In supervised learning, models are typically trained on extensive datasets where each sample is annotated with the correct label associated with the prediction task of interest. For example, in studies that make use of gene expression profiles, samples may be labeled with their respective tissue types or disease states, while in protein-protein interaction prediction data might consist of protein pairs labeled as interacting or non-interacting. Labeled datasets are necessary for building ML models that can predict labels for new, unseen samples. However, the quality and representativeness of the training data are essential to build successful ML models. If the training data is biased or non-representative of the broader population, the resulting models are likely to inherit such biases, leading to suboptimal performance when applied in the real world. Biases usually go unnoticed unless the models are carefully evaluated, underscoring the importance of fair model evaluation.

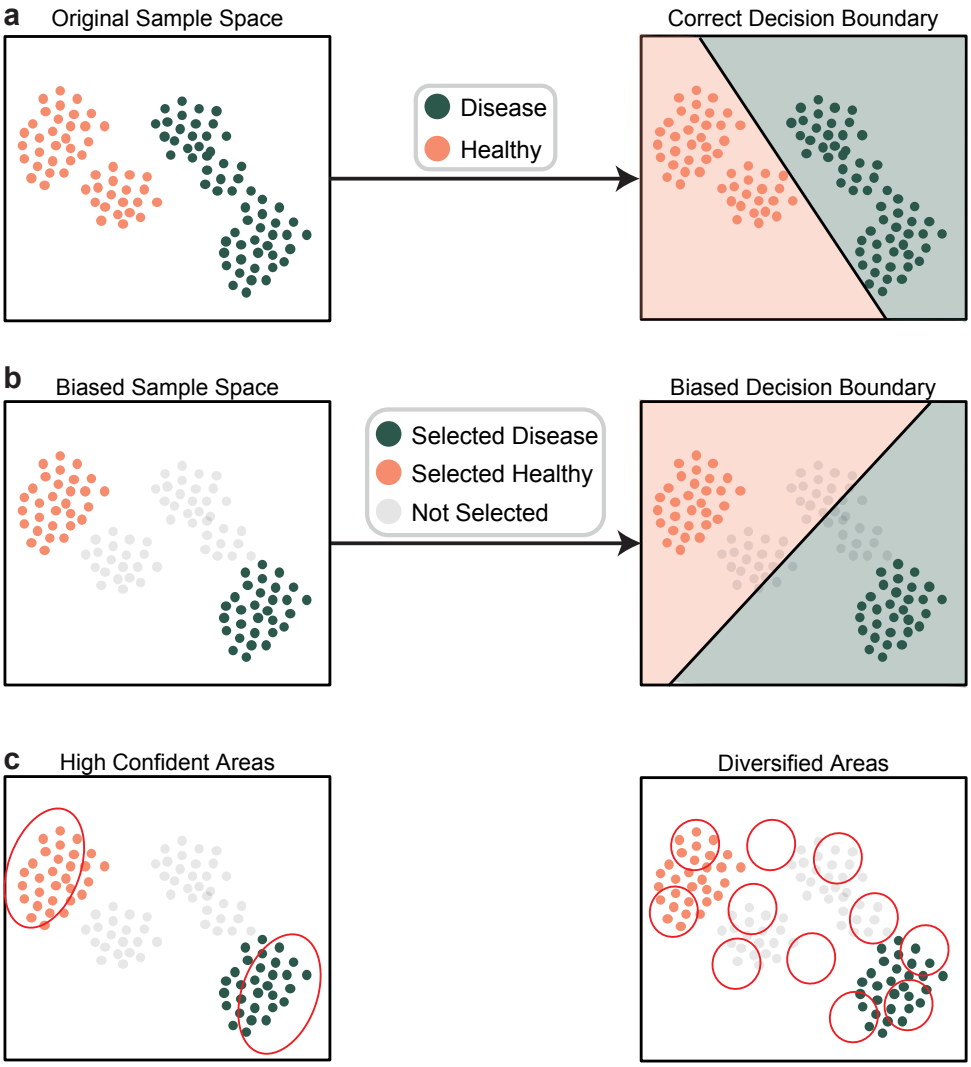


Figure 1.2: **Impact of selection bias to decision making and how to address it. a** A 2D example of sample space of all possible people in a population, colored by being healthy or having disease. On the right, most likely ML model-induced decision boundary is shown for this problem. **b** The same sample space but now only some of the patients are selected. On the right, one of the possible ML decision boundaries is shown for this problem which is different than the original correct decision boundary in **a**. **c** In case self-training is used to include more unlabeled samples to the training process, an ML model would identify high confident areas as on the **left** but ideally representative samples from different locations should be incorporated to avoid strengthening the bias as on the **right**.

1.2.1. CONVENTIONAL EVALUATION OF ML ALGORITHMS

Evaluating the performance of ML models typically involves several key steps, including measuring success metrics (e.g. accuracy, precision, recall), analyzing model behavior, and comparing results against baselines and other state of the art models using established benchmarks from the literature. In bioinformatics ML

applications, the general approach is to split existing datasets uniformly at random into train and test sets, and to use the train set for model development and reserve the test set for evaluation. Ideally, the test set remains entirely independent of the train set and is only used during the evaluation phase to ensure that the generalization ability of the model beyond the specific examples it was trained on is accurately assessed.

Nevertheless, this conventional evaluation approach has a critical limitation: both train and test sets are derived from the same underlying dataset, meaning that they share similar characteristics. If the original dataset is biased in any way, this bias will be reflected in both subsets, compromising the validity of the evaluation. This issue highlights the need for fair and comprehensive evaluation practices in ML applications. To fully understand this challenge, it is essential first to examine the data collection process and the inherent limitations of these datasets.

1.2.2. DATA COLLECTION, BIAS, AND LIMITATIONS

During data collection, several factors can introduce unfairness and bias into the dataset. Selection bias, where the collected samples do not accurately represent the distribution of the actual population, is a significant concern across various fields, including bioinformatics [135]. In genomic studies, high-quality samples are often more readily available from populations with better healthcare access, leading to biased data representation [136]. Similarly, datasets derived from specific labs or geographic regions may fail to capture the true diversity of biological samples [137]. Additionally, technological biases can emerge from the use of certain platforms or protocols that favor particular types of samples or measurements.

When biased datasets are used for training, it can cause ML models to fail when applied to new, diverse datasets, limiting the generalizability of the model (Fig. 1.2a-b). For example, a model trained predominantly on male patients may perform poorly in predicting disease risk for female patients. This issue highlights the critical importance of collecting diverse and representative datasets to ensure the development of fair and effective ML models. However, this problem is often overlooked during model evaluation, also due to the lack of solutions to identify bias and assess models under such conditions. We therefore ask the question: how can ML models be assessed in a truly fair way by taking generalizability in the presence of bias into account?

1.2.3. EVALUATION OF THE REPRESENTATIVENESS OF ML MODELS

Evaluating the representativeness of an ML model involves determining how accurately the patterns learned by ML model generalize to new, unseen data that is independent of the training data. An effective approach to do this consists in testing the prediction performance of the model on a dataset that is collected from a different source that is independent from the train set. For example, for any kind of prediction problem, a model trained on data collected at one hospital could be

tested on data collected by another hospital in a different country to assess the generalizability of the model across clinical cohorts and scenarios.

In the absence of truly independent data, which is often challenging to acquire or unavailable, common evaluation strategies resort to splitting a dataset into two or more non-overlapping subsets, depending on the setup: train and test; train, validation, and test; multiple folds for training and validation, as well as parameter tuning, in a cross-validation setting. While splitting a dataset can be useful when no other option is available, the different partitions will likely inherit any biases of the original dataset relative to the underlying population, which will be reflected in similar biases across train and test evaluation.

EVALUATING THE REPRESENTATIVENESS OF AN ML MODEL WITH A LIMITED DATASET

In the absence of independent test data, the robustness and generalizability of a model can be assessed by introducing artificial bias into the train set. By deliberately sampling training samples based on certain features or classes, we can generate a biased train set with different characteristics and distribution from the originally collected data. Then the ML model can be assessed to see if it can still perform well on the original test set [138–141]. Artificial bias is induced by modifying the training data to overrepresent or underrepresent specific groups or features. For example, we might induce bias by excluding data points from a particular demographic group. The model trained on this biased data can then be tested on a test set to assess its generalizability. This approach helps identify potential weaknesses in the model and provides insights into its ability to handle real-world variability. However, current methods for inducing bias often require prior knowledge of which variables in a dataset could cause bias, and they may not always be effective or reliable in disrupting the expected decision boundary.

1.2.4. MITIGATION OF SELECTION BIAS IN ML MODEL LEARNING

Mitigating selection bias in ML applications involves several strategies, including more comprehensive data collection, incorporating human expertise, and employing computational methods to address biases during training.

EXISTING COMPUTATIONAL APPROACHES TO MITIGATE BIAS DURING ML MODEL LEARNING

Domain adaptation methods aim to adjust models trained on one dataset to perform well on another dataset with different characteristics by correcting the distribution shift between them [142]. These methods encompass various techniques, including importance weighting, subspace alignment, inference-based methods, and deep domain adaptation. Importance weighting (IW) weighs the influence of training samples based on their relevance to the test set, assuming that the support of the test set is included in the train set and that the entire feature space is relevant [140, 143–152]. However, IW techniques can struggle with low sample sizes and

high-dimensional feature spaces and may also suffer when most of the features are not informative for the problem itself. Subspace alignment transforms feature representations to align the conditional probabilities of the train and test sets within a shared subspace, although optimizing these transformations can be challenging when there is a possibility of more than one suitable transformation [153–155]. Inference-based methods, such as minimax estimation, focus on minimizing the loss in worst-case scenarios to ensure the model remains robust under any conditions. Yet, these approaches are often model-specific and may underperform if the chosen model is not well-suited to the test data [141, 156]. Deep domain adaptation (DDA) methods, which use deep neural networks, aim to find domain-invariant representations by penalizing discrepancies between domains. However, these approaches are limited to deep learning models and operate as black-box models [157].

Overall, domain adaptation methods address distribution shifts between a specific training and a test set, which limits their broader applicability and generalization beyond the specific test domains. Essentially, they are not designed to mitigate the bias in the train set, but rather to try to fix the decision boundary just for that specific test set. Furthermore, many of these methods are tailored to specific machine learning models, restricting their usability across different types of models.

OPPORTUNITIES FOR SELECTION BIAS MITIGATION WITH UNLABELED DATA

While supervised learning requires labeled samples and domain adaptation requires additional access to test samples without their labels, for some problems there are vast numbers of unlabeled samples that could better represent the underlying population than available train and test sets but are not annotated due to some limitations such as time, money, and technology. Unlabeled data may present significant opportunities to improve generalizability, since it is not specific to any existing test set but can be assumed to include more general knowledge of the population distribution. We can exploit unlabeled data to learn more generalizable models in a semi-supervised learning manner. Semi-supervised learning (SSL) leverages unlabeled data to enhance representativeness and improve the model's understanding of the underlying population distribution. Techniques such as self-training (ST)[158] and co-training (CT)[159] incorporate additional unlabeled samples into the learning process through pseudo-labeling. In pseudo-labeling, a model is initially trained on labeled samples and then iteratively predicts the labels of unlabeled samples, incorporating those with high-confidence predictions into the train set. Here, "high-confidence" typically refers to predictions with probabilities above a user-defined threshold, such as 0.9, or the top k predictions. However, if the model is trained on a biased dataset, the top predictions will likely also be biased, which may lead to the inclusion of more biased samples in the training process (Fig. 1.2c). Additionally, if one class consistently has higher prediction probabilities, the model may predominantly select samples from that class, leading to class imbalance [160, 161]. While there are methods to reduce the inclusion of redundant samples in pseudo-labeling, such as P3SVM [160], these approaches are typically model-specific,

limiting their applicability to a broader range of machine learning models.

In summary, although there is an opportunity to mitigate selection bias with unlabeled data, it is still not utilized effectively. If applied carefully, semi-supervised learning may be helpful for bioinformatics problems that suffer from non-representative or biased data.

1.3. KEY CHALLENGES TO ADDRESS

This thesis identified several research gaps in therapeutic target identification for cancer treatment and selection bias in machine learning (Fig. 1.3-1.4).

Ineffective synthetic lethality prediction. Research on cancer treatment mainly involves identifying cohorts with common vulnerabilities and determining effective treatment options for them. When suitable cohorts with a potential target are identified but existing treatments are limited or ineffective, genetic interactions can be exploited for an alternative targeting possibilities that seek to sensitize the remaining cohort. Synthetic lethality (SL) between two genes, a genetic interaction used to target cancer cells, is increasingly identified through computational methods. However, current tools often perform poorly or inconsistently across cancer types. In some cases, predictions are worse than random due to the sparsity of data and a lack of diverse biological data sources. Moreover, existing models often follow selection biases in the data, resulting in low generalizability due to their dependence on previously identified SL relationships. Additionally, sequence similarity—an indicator of related gene function that can potentially lead to synthetic lethality (SL) relationships—has not been sufficiently investigated (Fig. 1.3, left panel).

Lack of computational tools to stratify oncogene-addicted cohorts. Another approach to discover new anti-cancer treatments is to further stratify cohorts to find sub-cohorts with similar vulnerabilities. For example, oncogene-driven cohorts often result in oncogene addiction, where tumor cells depend on the oncogene for their survival. Although the target is known, some oncogene-addicted cohorts are not targetable due to factors like the undruggability of the oncogene. However, they may contain sub-cohorts with additional targetable dependencies associated with oncogene addiction. Because these cohorts typically have few samples, they have not traditionally been the focus of computational methods. However, with the availability of cancer cell line omics and extensive drug response data, there is now an opportunity to use these data to computationally identify sub-cohorts of oncogene-addicted patients and potential treatments for them (Fig. 1.3, right panel).

Neglect on selection bias in machine learning. Despite efforts to address discrepancies between sample distributions of source (training) and target (testing) sets through domain adaptation methods, mitigating selection bias in source domains

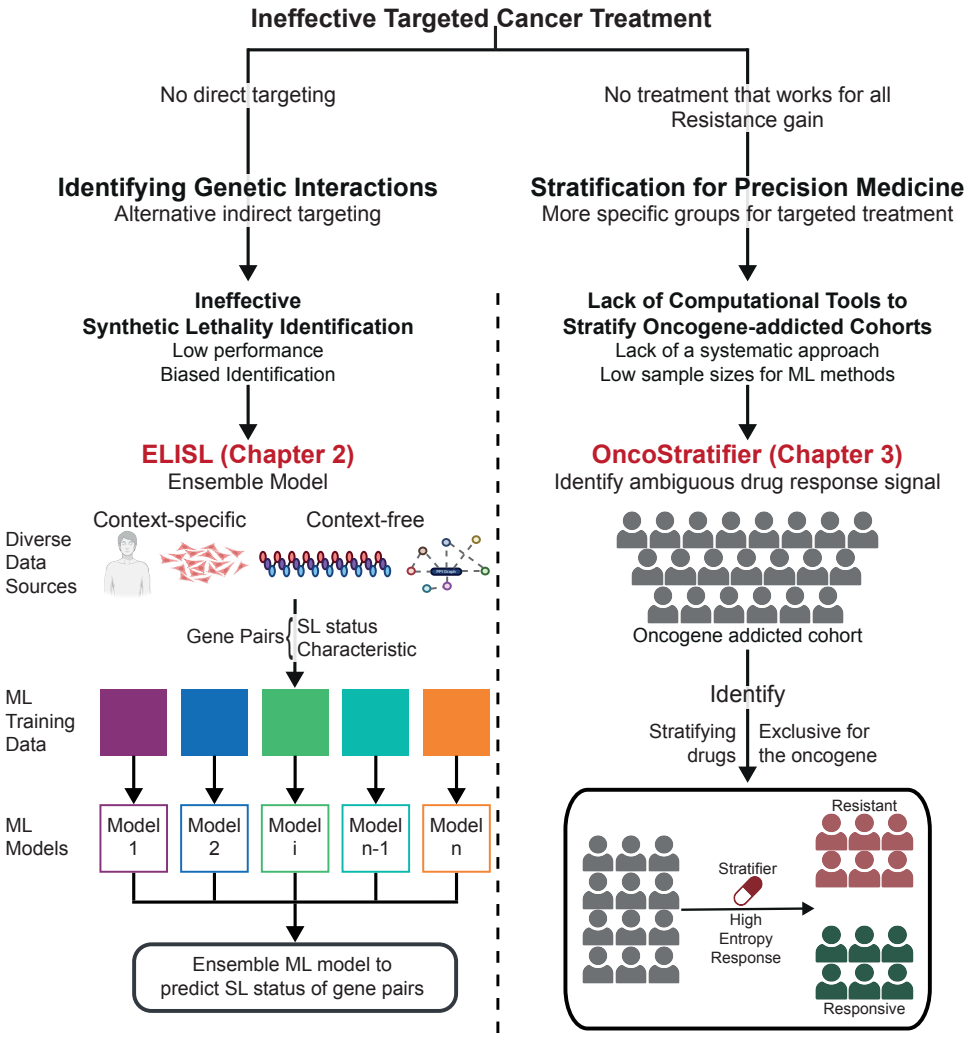


Figure 1.3: Overview of key challenges in targeted therapy that are addressed in this thesis along with proposed solutions: **ELISL** (Chapter 2) for identifying synthetic lethal relationships and **OncoStratifier** (Chapter 3) for stratifying oncogene-addicted cohorts.

to improve generalizability beyond a specific target domain remains underexplored. In bioinformatics, fair evaluation and generalizability are often overlooked unless a separate validation study is available, which is rarely the case.

To address selection bias and enhance generalizability, unlabeled samples, which cannot be used in supervised learning, could be leveraged in semi-supervised settings to learn a better approximation for the distribution and decision boundary associated with a prediction task (Fig. 1.3). Self-training, a semi-supervised method that incorporates unlabeled samples with high prediction confidence into training, shows promise, but it can also reinforce existing biases by favoring already biased samples. Therefore, there is a need for approaches that integrate diverse,

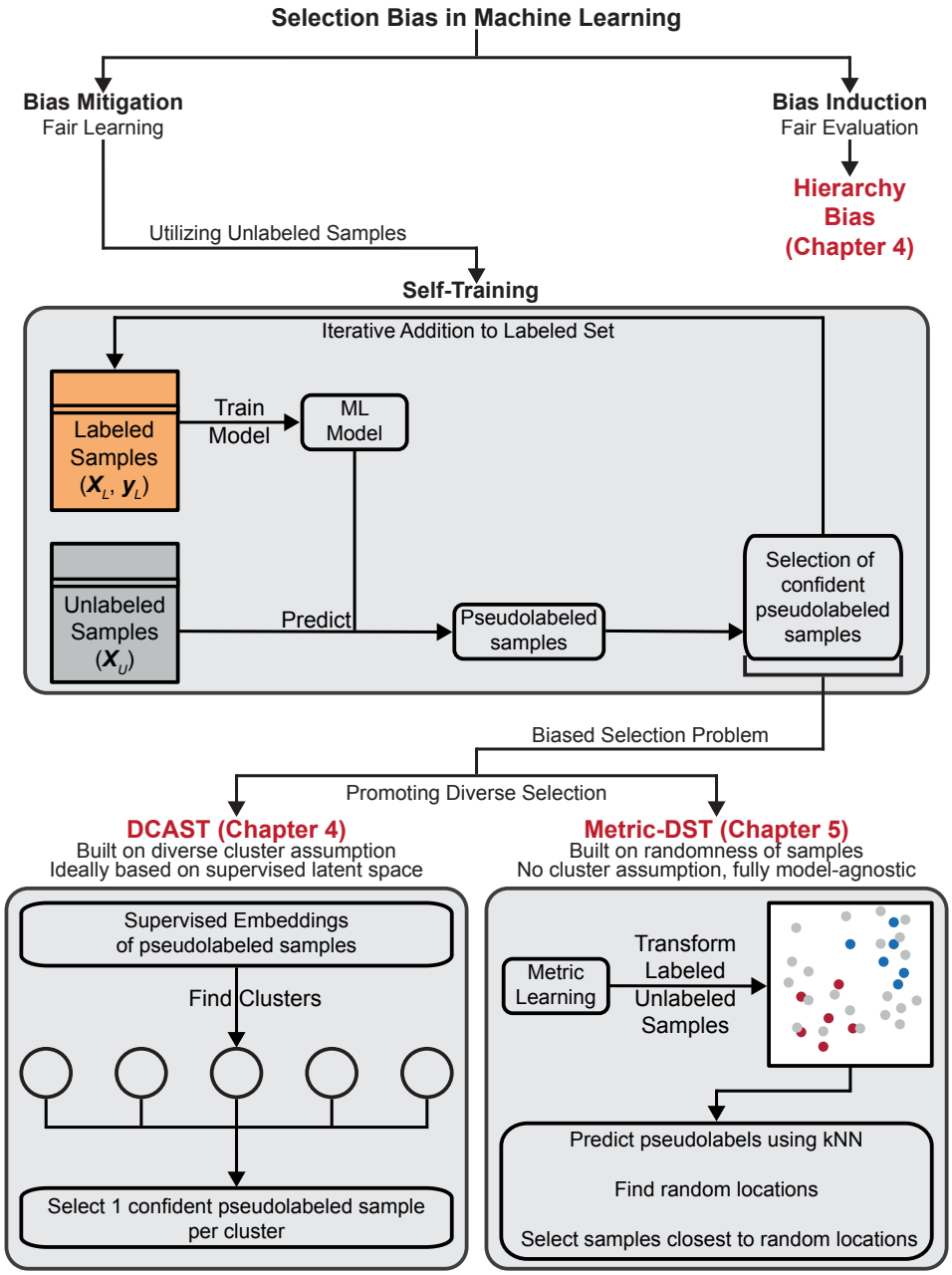


Figure 1.4: Overview of challenges related to selection bias in machine learning, along with vanilla self-training, and proposed solutions: **Hierarchy Bias** (Chapter 4) induction for investigating fair learning, and self-training-based bias mitigation methods, **DCAST** (Chapter 4) and **Metric-DST** (Chapter 5), for achieving fair learning.

representative unlabeled samples into the training process to effectively mitigate selection bias. Diverse samples can be selected from distinct clusters emerging

from the training data, ideally represented according to a supervised latent space: a learned discriminative latent space that is informative of the decision boundary for the prediction task of interest. This strategy works well with ML architectures such as neural networks and random forests, which intrinsically learn supervised latent spaces (Fig. 1.3, bottom left panel). For other methods, such as logistic regression, a supervised latent space is not readily available to promote diversity in this way. Moreover, the assumption that training samples will yield diverse clusters may not hold for every dataset, and thus promoting diversity based on clustering of the data is not always applicable. Therefore, while the diversity gap can be addressed immediately by exploiting intrinsic characteristics of certain models, there is also a clear need for a model-agnostic approach to promote the diverse selection of samples during ML model learning (Fig. 1.3, bottom right panel).

1.4. THESIS CONTRIBUTIONS

This thesis contributes computational methods for therapeutic target identification in cancer and addresses challenges related to selection bias in machine learning models.

As a contribution to finding genetic interactions for targeted therapy, in **Chapter 2**, we introduce ELISL as a machine learning framework designed to predict synthetic lethal (SL) gene pairs using a combination of carefully selected molecular biology data deemed informative for SL prediction (Fig. 1.3). The ELISL models effectively integrate context-free data that reflects functional relatedness of genes, such as amino acid sequences, and context-specific data that varies between tumors, such as tissue omics and survival as well as cancer cell line viability in the presence of gene aberrations, to enhance the SL prediction and generalizability. Furthermore, ELISL does not depend solely on existing SL relations in contrast to some of the previous methods and also utilizes aminoacid sequences of proteins coded by genes as a proxy for functional relatedness of genes. Our results demonstrate that ELISL outperforms existing methods in predicting known SL gene pairs while remaining robust against selection bias inherent in previous SL datasets.

To contribute to precision medicine, in **Chapter 3**, we present OncoStratifier, a computational framework aimed at stratifying oncogene-driven cancer cohorts based on drug response (Fig. 1.3). This approach is the first systematic computational effort to stratify oncogene-driven cohorts and identify possible targets based on available drugs. We show that OncoStratifier can identify subcohorts suitable for specific treatments, particularly in cases where the oncogene driving the cancer is not directly targetable.

To address selection bias in machine learning models, **Chapter 4** introduces two contributions: hierarchy bias, a method that induces bias by selecting samples while favoring specific clusters in sample space; and DCAST, a semi-supervised learning method based on self-training (Fig. 1.4) that incorporates diverse unlabeled samples into the learning process to mitigate selection bias in classification tasks (Fig. 1.4).

DCAST assumes that datasets consist of different clusters formed by diverse samples. Thus, it identifies and includes unlabeled samples from different clusters to promote diversity. Although not a strict requirement, DCAST suggests identification of clusters in supervised latent space informed by the classification task to ensure the sample space is constructed only by informative features. Our experiments show that hierarchy bias is more reliable in inducing bias compared to previous methods, and DCAST consistently improves model performance on biased training sets without the performance degradation seen for other domain adaptation methods. However, DCAST's implementation may vary depending on the machine learning model, as it suggests identifying a supervised latent space per method. Additionally, DCAST assumes that samples can be clustered, with diverse clusters yielding diverse samples to mitigate bias, which may not be true for all datasets.

To overcome the limitations of the cluster assumption and model-specific implementation, **Chapter 5** introduces Metric-DST, a self-training method that uses metric learning to incorporate diverse unlabeled samples into the training process, with the aim of mitigating selection bias (Fig. 1.4). The metric learning model directly creates a bounded supervised latent space that allows for the selection of more diverse samples, randomly distributed throughout the learned space. Metric-DST is model-agnostic, making it applicable to any machine learning model without requiring changes to the implementation. Our experiments show that Metric-DST performs well across biased toy datasets, real-life datasets, and a complex bioinformatics problem, synthetic lethality.

Finally, the thesis concludes with a summary of our contributions, a discussion of their current and potential future impact, and possible extensions in **Chapter 6**.

REFERENCES

[1] A. Sudhakar. “History of Cancer, Ancient and Modern Treatment Methods”. In: *Journal of Cancer Science & Therapy* 01.02 (2009), pp. i–iv. ISSN: 1948-5956.

[2] F. Bray *et al.* “Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries”. In: *CA: A Cancer Journal for Clinicians* 74.3 (Apr. 2024), pp. 229–263. ISSN: 1542-4863.

[3] D. J. Benjamin. “The efficacy of surgical treatment of cancer – 20years later”. In: *Medical Hypotheses* 82.4 (Apr. 2014), pp. 412–420. ISSN: 0306-9877.

[4] B. Fisher *et al.* “Eight-Year Results of a Randomized Clinical Trial Comparing Total Mastectomy and Lumpectomy with or without Irradiation in the Treatment of Breast Cancer”. In: *New England Journal of Medicine* 320.13 (Mar. 1989), pp. 822–828. ISSN: 1533-4406.

[5] X. Sun *et al.* “CD30 ligand is a target for a novel biological therapy against colitis associated with Th17 responses”. en. In: *J. Immunol.* 185.12 (Dec. 2010), pp. 7671–7680.

[6] H. H. Young. “The early diagnosis and radical cure of carcinoma of the prostate. Being a study of 40 cases and presentation of a radical operation which was carried out in four cases”. In: *Johns Hopkins Hosp. Bull.* 16 (1905), pp. 315–321.

[7] T. Millin. *Retropubic urinary surgery*. E. & S. Livingstone, 1947.

[8] P. C. Walsh *et al.* “Campbell’s urology”. In: *Campbell’s urology*. 1998, pp. 3432–3432.

[9] R. Baskar *et al.* “Cancer and Radiation Therapy: Current Advances and Future Directions”. In: *International Journal of Medical Sciences* 9.3 (2012), pp. 193–199. ISSN: 1449-1907.

[10] D. Ishihara *et al.* “Rationale and evidence to combine radiation therapy and immunotherapy for cancer treatment”. In: *Cancer Immunology, Immunotherapy* 66.3 (Oct. 2016), pp. 281–298. ISSN: 1432-0851.

[11] O. B. Wijers *et al.* “Patients with head and neck cancer cured by radiation therapy: A survey of the dry mouth syndrome in long-term survivors”. In: *Head Neck* 24.8 (July 2002), pp. 737–747. ISSN: 1097-0347.

[12] V. T. DeVita and E. Chu. “A History of Cancer Chemotherapy”. In: *Cancer Research* 68.21 (Oct. 2008), pp. 8643–8653. ISSN: 1538-7445.

[13] S. Rivankar. “An overview of doxorubicin formulations in cancer therapy”. In: *Journal of Cancer Research and Therapeutics* 10.4 (2014), p. 853. ISSN: 0973-1482.

[14] S. Ghosh. “Cisplatin: The first metal based anticancer drug”. In: *Bioorganic Chemistry* 88 (July 2019), p. 102925. ISSN: 0045-2068.

[15] A.-M. Florea and D. Büsselberg. “Cisplatin as an Anti-Tumor Drug: Cellular Mechanisms of Activity, Drug Resistance and Induced Side Effects”. In: *Cancers* 3.1 (Mar. 2011), pp. 1351–1371. ISSN: 2072-6694.

[16] B. A. Weaver. “How Taxol/paclitaxel kills cancer cells”. In: *Molecular Biology of the Cell* 25.18 (Sept. 2014). Ed. by W. Bement, pp. 2677–2681. ISSN: 1939-4586.

[17] W. M. C. van den Boogaard, D. S. J. Komninos, and W. P. Vermeij. “Chemotherapy Side-Effects: Not All DNA Damage Is Equal”. In: *Cancers* 14.3 (Jan. 2022), p. 627. ISSN: 2072-6694.

[18] R. M. McQuade, J. C. Bornstein, and K. Nurgali. “Anti-Colorectal Cancer Chemotherapy-Induced Diarrhoea: Current Treatments and Side-Effects”. In: *International Journal of Clinical Medicine* 05.07 (2014), pp. 393–406. ISSN: 2158-2882.

[19] I. Dagogo-Jack and A. T. Shaw. “Tumour heterogeneity and resistance to cancer therapies”. In: *Nature Reviews Clinical Oncology* 15.2 (Nov. 2017), pp. 81–94. ISSN: 1759-4782.

[20] K. M. Haigis, K. Cichowski, and S. J. Elledge. “Tissue-specificity in cancer: The rule, not the exception”. In: *Science* 363.6432 (Mar. 2019), pp. 1150–1151. ISSN: 1095-9203.

[21] F. D. S. E. Melo *et al.* “Cancer heterogeneity—a multifaceted view”. In: *EMBO reports* 14.8 (July 2013), pp. 686–695. ISSN: 1469-3178.

[22] R. Elmentaite *et al.* “Single-cell atlases: shared and tissue-specific cell types across human organs”. In: *Nature Reviews Genetics* 23.7 (Feb. 2022), pp. 395–410. ISSN: 1471-0064.

[23] M. Bilzer, F. Roggel, and A. L. Gerbes. “Role of Kupffer cells in host defense and liver disease”. In: *Liver International* 26.10 (Nov. 2006), pp. 1175–1186. ISSN: 1478-3231.

[24] A. London, M. Cohen, and M. Schwartz. “Microglia and monocyte-derived macrophages: functionally distinct populations that act in concert in CNS plasticity and repair”. In: *Frontiers in Cellular Neuroscience* 7 (2013). ISSN: 1662-5102.

[25] Y. Zhou and X. Liu. “The role of estrogen receptor beta in breast cancer”. In: *Biomarker Research* 8.1 (Sept. 2020). ISSN: 2050-7771.

[26] L. Murphy and E. Leygue. “The Role of Estrogen Receptor-B in Breast Cancer”. In: *Seminars in Reproductive Medicine* 30.01 (Jan. 2012), pp. 05–13. ISSN: 1526-4564.

[27] O. Treeck *et al.* “Estrogen receptor beta exerts growth-inhibitory effects on human mammary epithelial cells”. In: *Breast Cancer Research and Treatment* 120.3 (May 2009), pp. 557–565. ISSN: 1573-7217.

[28] I. I. Wistuba *et al.* “Molecular changes in the bronchial epithelium of patients with small cell lung cancer”. en. In: *Clin. Cancer Res.* 6.7 (July 2000), pp. 2604–2610.

[29] A. Thomas, A. Rajan, and G. Giaccone. “Tyrosine Kinase Inhibitors in Lung Cancer”. In: *Hematology/Oncology Clinics of North America* 26.3 (June 2012), pp. 589–605. ISSN: 0889-8588.

[30] M. Johnson *et al.* “Treatment strategies and outcomes for patients with EGFR-mutant non-small cell lung cancer resistant to EGFR tyrosine kinase inhibitors: Focus on novel therapies”. In: *Lung Cancer* 170 (Aug. 2022), pp. 41–51. ISSN: 0169-5002.

[31] E. R. Fearon and B. Vogelstein. “A genetic model for colorectal tumorigenesis”. In: *Cell* 61.5 (June 1990), pp. 759–767. ISSN: 0092-8674.

[32] L. Zhang and J. W. Shay. “Multiple Roles of APC and its Therapeutic Implications in Colorectal Cancer”. In: *JNCI: Journal of the National Cancer Institute* 109.8 (Apr. 2017). ISSN: 1460-2105.

[33] S.-J. Lee and C. C. Yun. “Colorectal cancer cells – Proliferation, survival and invasion by lysophosphatidic acid”. In: *The International Journal of Biochemistry & Cell Biology* 42.12 (Dec. 2010), pp. 1907–1910. ISSN: 1357-2725.

[34] A. Sample and Y.-Y. He. “Mechanisms and prevention of UV-induced melanoma”. In: *Photodermatology, Photoimmunology & Photomedicine* 34.1 (Aug. 2017), pp. 13–24. ISSN: 1600-0781.

[35] A. Zhang *et al.* “Tumor heterogeneity reshapes the tumor microenvironment to influence drug resistance”. In: *International Journal of Biological Sciences* 18.7 (2022), pp. 3019–3033. ISSN: 1449-2288.

[36] L. Zhu *et al.* “A narrative review of tumor heterogeneity and challenges to tumor drug therapy”. In: *Annals of Translational Medicine* 9.16 (Aug. 2021), pp. 1351–1351. ISSN: 2305-5847.

[37] Q. Jia *et al.* “Heterogeneity of the tumor immune microenvironment and its clinical relevance”. In: *Experimental Hematology & Oncology* 11.1 (Apr. 2022). ISSN: 2162-3619.

[38] M. R. Junttila and F. J. de Sauvage. “Influence of tumour micro-environment heterogeneity on therapeutic response”. In: *Nature* 501.7467 (Sept. 2013), pp. 346–354. ISSN: 1476-4687.

[39] S. Wang *et al.* “Targeting M2-like tumor-associated macrophages is a potential therapeutic approach to overcome antitumor drug resistance”. en. In: *NPJ Precis. Oncol.* 8.1 (Feb. 2024), p. 31.

[40] B. Baran *et al.* “Difference Between Left-Sided and Right-Sided Colorectal Cancer: A Focused Review of Literature”. In: *Gastroenterology Research* 11.4 (2018), pp. 264–273. ISSN: 1918-2813.

[41] M. B. Meads, R. A. Gatenby, and W. S. Dalton. “Environment-mediated drug resistance: a major contributor to minimal residual disease”. In: *Nature Reviews Cancer* 9.9 (Aug. 2009), pp. 665–674. ISSN: 1474-1768.

[42] C. Tomasetti, L. Li, and B. Vogelstein. “Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention”. In: *Science* 355.6331 (Mar. 2017), pp. 1330–1334. ISSN: 1095-9203.

[43] T. K. Kelly, D. D. De Carvalho, and P. A. Jones. “Epigenetic modifications as therapeutic targets”. In: *Nature Biotechnology* 28.10 (Oct. 2010), pp. 1069–1078. ISSN: 1546-1696.

[44] O. Dreesen and A. H. Brivanlou. “Signaling Pathways in Cancer and Embryonic Stem Cells”. In: *Stem Cell Reviews* 3.1 (Jan. 2007), pp. 7–17. ISSN: 1558-6804.

[45] R. Kitadai and Y. Okuma. “Treatment Strategies for Non-Small Cell Lung Cancer Harboring Common and Uncommon EGFR Mutations: Drug Sensitivity Based on Exon Classification, and Structure-Function Analysis”. In: *Cancers* 14.10 (May 2022), p. 2519. ISSN: 2072-6694.

[46] W. Pao *et al.* “KRAS Mutations and Primary Resistance of Lung Adenocarcinomas to Gefitinib or Erlotinib”. In: *PLoS Medicine* 2.1 (Jan. 2005). Ed. by R. Herbst, e17. ISSN: 1549-1676.

[47] M. Esteller. “Promoter Hypermethylation and BRCA1 Inactivation in Sporadic Breast and Ovarian Tumors”. In: *Journal of the National Cancer Institute* 92.7 (Apr. 2000), pp. 564–569. ISSN: 1460-2105.

[48] Y. He *et al.* “Targeting PI3K/Akt signal transduction for cancer therapy”. In: *Signal Transduction and Targeted Therapy* 6.1 (Dec. 2021). ISSN: 2059-3635.

[49] M. Greaves and C. C. Maley. “Clonal evolution in cancer”. In: *Nature* 481.7381 (Jan. 2012), pp. 306–313. ISSN: 1476-4687.

[50] A. S, A. Chakraborty, and S. Patnaik. “Clonal evolution and expansion associated with therapy resistance and relapse of colorectal cancer”. In: *Mutation Research/Reviews in Mutation Research* 790 (July 2022), p. 108445. ISSN: 1383-5742.

[51] D. Sun *et al.* “Differential selective pressure alters rate of drug resistance acquisition in heterogeneous tumor populations”. In: *Scientific Reports* 6.1 (Nov. 2016). ISSN: 2045-2322.

[52] Z. Chen *et al.* “Hypoxic microenvironment in cancer: molecular mechanisms and therapeutic interventions”. In: *Signal Transduction and Targeted Therapy* 8.1 (Feb. 2023). ISSN: 2059-3635.

[53] D. Hanahan and R. A. Weinberg. “The Hallmarks of Cancer”. In: *Cell* 100.1 (Jan. 2000), pp. 57–70. ISSN: 0092-8674.

[54] E. B. Garon *et al.* “Pembrolizumab for the Treatment of Non–Small-Cell Lung Cancer”. In: *New England Journal of Medicine* 372.21 (May 2015), pp. 2018–2028. ISSN: 1533-4406.

[55] J. H. Choi, J. M. Bogenberger, and R. Tibes. “Targeting Apoptosis in Acute Myeloid Leukemia: Current Status and Future Directions of BCL-2 Inhibition with Venetoclax and Beyond”. In: *Targeted Oncology* 15.2 (Apr. 2020), pp. 147–162. ISSN: 1776-260X.

[56] Q. Cao *et al.* “Mechanisms of action of the BCL-2 inhibitor venetoclax in multiple myeloma: a literature review”. In: *Frontiers in Pharmacology* 14 (Nov. 2023). ISSN: 1663-9812.

[57] J. S. Ross and J. A. Fletcher. “The HER-2/ neu Oncogene in Breast Cancer: Prognostic Factor, Predictive Factor, and Target for Therapy”. In: *STEM CELLS* 16.6 (Nov. 1998), pp. 413–428. ISSN: 1549-4918.

[58] M. W. N. Deininger, J. M. Goldman, and J. V. Melo. “The molecular biology of chronic myeloid leukemia”. In: *Blood* 96.10 (Nov. 2000), pp. 3343–3356. ISSN: 0006-4971.

[59] A. Antonicelli *et al.* “EGFR-Targeted Therapy for Non-Small Cell Lung Cancer: Focus on EGFR Oncogenic Mutation”. In: *International Journal of Medical Sciences* 10.3 (2013), pp. 320–330. ISSN: 1449-1907.

[60] T. Lee *et al.* “Non-small Cell Lung Cancer with Concomitant EGFR, KRAS, and ALK Mutation: Clinicopathologic Features of 12 Cases”. In: *Journal of Pathology and Translational Medicine* 50.3 (May 2016), pp. 197–203. ISSN: 2383-7845.

[61] M. D. Pegram *et al.* "Results of Two Open-Label, Multicenter Phase II Studies of Docetaxel, Platinum Salts, and Trastuzumab in HER2-Positive Advanced Breast Cancer". In: *JNCI Journal of the National Cancer Institute* 96.10 (May 2004), pp. 759–769. ISSN: 1460-2105.

[62] N. Robert *et al.* "Randomized Phase III Study of Trastuzumab, Paclitaxel, and Carboplatin Compared With Trastuzumab and Paclitaxel in Women With HER-2–Overexpressing Metastatic Breast Cancer". In: *Journal of Clinical Oncology* 24.18 (June 2006), pp. 2786–2792. ISSN: 1527-7755.

[63] H. Kantarjian *et al.* "Hematologic and Cytogenetic Responses to Imatinib Mesylate in Chronic Myelogenous Leukemia". In: *New England Journal of Medicine* 346.9 (Feb. 2002), pp. 645–652. ISSN: 1533-4406.

[64] L. V. Sequist *et al.* "First-Line Gefitinib in Patients With Advanced Non–Small-Cell Lung Cancer Harboring SomaticEGFRMutations". In: *Journal of Clinical Oncology* 26.15 (May 2008), pp. 2442–2449. ISSN: 1527-7755.

[65] B. J. Solomon *et al.* "First-Line Crizotinib versus Chemotherapy in ALK-Positive Lung Cancer". In: *New England Journal of Medicine* 371.23 (Dec. 2014), pp. 2167–2177. ISSN: 1533-4406.

[66] L. G. T. Morris and T. A. Chan. "Therapeutic targeting of tumor suppressor genes". In: *Cancer* 121.9 (Dec. 2014), pp. 1357–1368. ISSN: 1097-0142.

[67] A. J. Levine, J. Momand, and C. A. Finlay. "The p53 tumour suppressor gene". In: *Nature* 351.6326 (June 1991), pp. 453–456. ISSN: 1476-4687.

[68] S. Nishikawa and T. Iwakuma. "Drugs Targeting p53 Mutations with FDA Approval and in Clinical Trials". In: *Cancers* 15.2 (Jan. 2023), p. 429. ISSN: 2072-6694.

[69] A. R. Moore *et al.* "RAS-targeted therapies: is the undruggable drugged?" In: *Nature Reviews Drug Discovery* 19.8 (June 2020), pp. 533–552. ISSN: 1474-1784.

[70] M. I. Truica *et al.* "Turning Up the Heat on MYC: Progress in Small-Molecule Inhibitors". In: *Cancer Research* 81.2 (Jan. 2021), pp. 248–253. ISSN: 1538-7445.

[71] H. Wang *et al.* "Targeting p53 pathways: mechanisms, structures, and advances in therapy". In: *Signal Transduction and Targeted Therapy* 8.1 (Mar. 2023). ISSN: 2059-3635.

[72] D. Westover *et al.* "Mechanisms of acquired resistance to first- and second-generation EGFR tyrosine kinase inhibitors". In: *Annals of Oncology* 29 (Jan. 2018), pp. i10–i19. ISSN: 0923-7534.

[73] D. A. Cross *et al.* "AZD9291, an Irreversible EGFR TKI, Overcomes T790M-Mediated Resistance to EGFR Inhibitors in Lung Cancer". In: *Cancer Discovery* 4.9 (Sept. 2014), pp. 1046–1061. ISSN: 2159-8290.

[74] K. Trunzer *et al.* "Pharmacodynamic Effects and Mechanisms of Resistance to Vemurafenib in Patients With Metastatic Melanoma". In: *Journal of Clinical Oncology* 31.14 (May 2013), pp. 1767–1774. ISSN: 1527-7755.

[75] S. Y. Lim, A. M. Menzies, and H. Rizos. "Mechanisms and strategies to overcome resistance to molecularly targeted therapy for melanoma". In: *Cancer* 123.S11 (May 2017), pp. 2118–2129. ISSN: 1097-0142.

[76] F. Claessens *et al.* "Emerging mechanisms of enzalutamide resistance in prostate cancer". In: *Nature Reviews Urology* 11.12 (Sept. 2014), pp. 712–716. ISSN: 1759-4820.

[77] E. H.-C. Hsiue *et al.* "Safety of gefitinib in non-small cell lung cancer treatment". In: *Expert Opinion on Drug Safety* 15.7 (June 2016), pp. 993–1000. ISSN: 1744-764X.

[78] A. Seidman *et al.* "Cardiac Dysfunction in the Trastuzumab Clinical Trials Experience". In: *Journal of Clinical Oncology* 20.5 (Mar. 2002), pp. 1215–1221. ISSN: 1527-7755.

[79] J. L. Steegmann *et al.* "Off-target effects of BCR–ABL1 inhibitors and their potential long-term implications in patients with chronic myeloid leukemia". In: *Leukemia; Lymphoma* 53.12 (June 2012), pp. 2351–2361. ISSN: 1029-2403.

[80] S. Narita *et al.* "P2.03-036 Comparing the Efficacy/Toxicity of Osimertinib and First Line EGFR-TKI by Individual Patient Analysis". In: *Journal of Thoracic Oncology* 12.11 (Nov. 2017), S2141. ISSN: 1556-0864.

[81] L. Chin, J. N. Andersen, and P. A. Futreal. “Cancer genomics: from discovery science to personalized medicine”. In: *Nature Medicine* 17.3 (Mar. 2011), pp. 297–303. ISSN: 1546-170X.

[82] A. Carbone. “Cancer Classification at the Crossroads”. In: *Cancers* 12.4 (Apr. 2020), p. 980. ISSN: 2072-6694.

[83] C. J. Allegra *et al.* “American Society of Clinical Oncology provisional clinical opinion: testing for KRAS gene mutations in patients with metastatic colorectal carcinoma to predict response to anti–epidermal growth factor receptor monoclonal antibody therapy”. In: *Journal of clinical oncology* 27.12 (2009), pp. 2091–2096.

[84] A. Lievre *et al.* “KRAS mutations as an independent prognostic factor in patients with advanced colorectal cancer treated with cetuximab”. In: *Journal of clinical oncology* 26.3 (2008), pp. 374–379.

[85] R. G. Amado *et al.* “Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer”. In: *Journal of clinical oncology* 26.10 (2008), pp. 1626–1634.

[86] B. Ganguly *et al.* “Role of Biomarkers in the Development of PARP Inhibitors”. In: *Biomarkers in Cancer* 8s1 (Jan. 2016), BIC.S36679. ISSN: 1179-299X.

[87] A. D. Smith, D. Roda, and T. A. Yap. “Strategies for modern biomarker and drug development in oncology”. In: *Journal of Hematology; Oncology* 7.1 (Oct. 2014). ISSN: 1756-8722.

[88] S. Amin and O. F. Bathe. “Response biomarkers: re-envisioning the approach to tailoring drug therapy for cancer”. In: *BMC Cancer* 16.1 (Nov. 2016). ISSN: 1471-2407.

[89] D.-R. Wang, X.-L. Wu, and Y.-L. Sun. “Therapeutic targets and biomarkers of tumor immunotherapy: response versus non-response”. In: *Signal Transduction and Targeted Therapy* 7.1 (Sept. 2022). ISSN: 2059-3635.

[90] L. H. Hartwell *et al.* “Integrating Genetic Approaches into the Discovery of Anticancer Drugs”. In: *Science* 278.5340 (Nov. 1997), pp. 1064–1068. ISSN: 1095-9203.

[91] P. C. Fong *et al.* “Inhibition of Poly(ADP-Ribose) Polymerase in Tumors fromBRCAMutation Carriers”. In: *New England Journal of Medicine* 361.2 (July 2009), pp. 123–134. ISSN: 1533-4406.

[92] A. D. Sahu *et al.* “Genome-wide prediction of synthetic rescue mediators of resistance to targeted and immunotherapy”. In: *Molecular Systems Biology* 15.3 (Mar. 2019). ISSN: 1744-4292.

[93] A. E. Motter *et al.* “Predicting synthetic rescues in metabolic networks”. In: *Molecular Systems Biology* 4.1 (Jan. 2008). ISSN: 1744-4292.

[94] F. L. Muller, E. A. Aquilanti, and R. A. DePinho. “Collateral Lethality: A New Therapeutic Strategy in Oncology”. In: *Trends in Cancer* 1.3 (Nov. 2015), pp. 161–173. ISSN: 2405-8033.

[95] K.-O. Henrich *et al.* “CAMTA1, a 1p36 Tumor Suppressor Candidate, Inhibits Growth and Activates Differentiation Programs in Neuroblastoma Cells”. In: *Cancer Research* 71.8 (Apr. 2011), pp. 3142–3151. ISSN: 1538-7445.

[96] I. Moon *et al.* “Machine learning for genetics-based classification and treatment response prediction in cancer of unknown primary”. In: *Nature Medicine* 29.8 (Aug. 2023), pp. 2057–2067. ISSN: 1546-170X.

[97] J. Song *et al.* “Development and Validation of a Machine Learning Model to Explore Tyrosine Kinase Inhibitor Response in Patients With Stage IV EGFR Variant–Positive Non–Small Cell Lung Cancer”. In: *JAMA Network Open* 3.12 (Dec. 2020), e2030442. ISSN: 2574-3805.

[98] R. Qureshi *et al.* “Machine learning based personalized drug response prediction for lung cancer patients”. In: *Scientific Reports* 12.1 (Nov. 2022). ISSN: 2045-2322.

[99] A. K. Jain, M. N. Murty, and P. J. Flynn. “Data clustering: a review”. In: *ACM computing surveys (CSUR)* 31.3 (1999), pp. 264–323.

[100] N. Rappoport and R. Shamir. “Multi-omic and multi-view clustering algorithms: review and cancer benchmark”. In: *Nucleic Acids Research* 46.20 (Oct. 2018), pp. 10546–10562. ISSN: 1362-4962.

[101] J. MacQueen *et al.* “Some methods for classification and analysis of multivariate observations”. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 14. Oakland, CA, USA. 1967, pp. 281–297.

[102] F. Murtagh and P. Contreras. “Algorithms for hierarchical clustering: an overview”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2.1 (2012), pp. 86–97.

[103] A. Y. Ng, M. I. Jordan, and Y. Weiss. “On spectral clustering: analysis and an algorithm”. In: *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*. NIPS’01. Vancouver, British Columbia, Canada: MIT Press, 2001, pp. 849–856.

[104] R. Shen, A. B. Olshen, and M. Ladanyi. “Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis”. In: *Bioinformatics* 25.22 (Sept. 2009), pp. 2906–2912. ISSN: 1367-4803.

[105] P. S. Mischel *et al.* “Identification of molecular subtypes of glioblastoma by gene expression profiling”. In: *Oncogene* 22.15 (Apr. 2003), pp. 2361–2373. ISSN: 1476-5594.

[106] R. Shai *et al.* “Gene expression profiling identifies molecular subtypes of gliomas”. In: *Oncogene* 22.31 (July 2003), pp. 4918–4923. ISSN: 1476-5594.

[107] A. A. Alizadeh *et al.* “Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling”. In: *Nature* 403.6769 (Feb. 2000), pp. 503–511. ISSN: 1476-4687.

[108] M. R. Aure *et al.* “Integrative clustering reveals a novel split in the luminal A subtype of breast cancer with impact on outcome”. In: *Breast Cancer Research* 19.1 (Mar. 2017). ISSN: 1465-542X.

[109] Z. Gao *et al.* “Hierarchical graph learning for protein–protein interaction”. In: *Nature Communications* 14.1 (Feb. 2023). ISSN: 2041-1723.

[110] I. A. Kovács *et al.* “Network-based prediction of protein interactions”. In: *Nature Communications* 10.1 (Mar. 2019). ISSN: 2041-1723.

[111] L. Jerby-Arnon *et al.* “Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality”. In: *Cell* 158.5 (2014), pp. 1199–1209.

[112] M. Wappett *et al.* “Multi-omic measurement of mutually exclusive loss-of-function enriches for candidate synthetic lethal gene pairs”. In: *BMC genomics* 17 (2016), pp. 1–15.

[113] J. S. Lee *et al.* “Harnessing synthetic lethality to predict the response to cancer treatment”. In: *Nature communications* 9.1 (2018), p. 2546.

[114] H. Liany, A. Jeyasekharan, and V. Rajan. “Predicting synthetic lethal interactions using heterogeneous data sources”. In: *Bioinformatics* 36.7 (Nov. 2019), pp. 2209–16.

[115] J. Huang *et al.* “Predicting synthetic lethal interactions in human cancers using graph regularized self-representative matrix factorization”. In: *BMC Bioinformatics* 20.S19 (Dec. 2019), p. 657.

[116] Y. Liu *et al.* “SL2MF: Predicting Synthetic Lethality in Human Cancers via Logistic Matrix Factorization”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 17.3 (May 2020), pp. 748–57.

[117] R. Cai *et al.* “Dual-dropout graph convolutional network for predicting synthetic lethality in human cancers”. In: *Bioinformatics* 36.16 (Mar. 2020), pp. 4458–65.

[118] Y. Long *et al.* “Graph contextualized attention network for predicting synthetic lethality in human cancers”. In: *Bioinformatics* 16 (Feb. 2021), pp. 2432–40.

[119] S. Das *et al.* “DiscoverSL: an R package for multi-omic data driven prediction of synthetic lethality in cancers”. In: *Bioinformatics* 35.4 (2019), pp. 701–702.

[120] F. Wan *et al.* “Exp2sl: a machine learning framework for cell-line-specific synthetic lethality prediction”. In: *Frontiers in pharmacology* 11 (2020), p. 112.

[121] C. Seale, Y. Tepeli, and J. P. Gonçalves. “Overcoming selection bias in synthetic lethality prediction”. In: *Bioinformatics* 38.18 (July 2022). Ed. by K. Borgwardt, pp. 4360–4368. ISSN: 1367-4811.

[122] Z. W. Dwyer and J. A. Pleiss. “The problem of selection bias in studies of pre-mRNA splicing”. In: *Nature Communications* 14.1 (Apr. 2023). ISSN: 2041-1723.

[123] G. Orlando, D. Raimondi, and W. F. Vranken. “Observation selection bias in contact prediction and its implications for structural bioinformatics”. In: *Scientific Reports* 6.1 (Nov. 2016). ISSN: 2045-2322.

[124] J. A. Miccio *et al.* “Quantifying treatment selection bias effect on survival in comparative effectiveness research: findings from low-risk prostate cancer patients”. In: *Prostate Cancer and Prostatic Diseases* 24.2 (Sept. 2020), pp. 414–422. ISSN: 1476-5608.

[125] C. J. D. Wallis *et al.* “The effect of selection and referral biases for the treatment of localised prostate cancer with surgery or radiation”. In: *British Journal of Cancer* 118.10 (Mar. 2018), pp. 1399–1405. ISSN: 1532-1827.

[126] J. Jumper *et al.* “Highly accurate protein structure prediction with AlphaFold”. In: *nature* 596.7873 (2021), pp. 583–589.

[127] Y. Chang *et al.* “Cancer drug response profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature”. In: *Scientific reports* 8.1 (2018), p. 8857.

[128] G. Adam *et al.* “Machine learning approaches to drug response prediction: challenges and recent progress”. In: *NPJ precision oncology* 4.1 (2020), p. 19.

[129] S. Mourragui *et al.* “PRECISE: a domain adaptation approach to transfer predictors of drug response from pre-clinical models to tumors”. In: *Bioinformatics* 35.14 (July 2019), pp. i510–i519. ISSN: 1367-4811.

[130] B. Baek and H. Lee. “Prediction of survival and recurrence in patients with pancreatic cancer by integrating multi-omics data”. In: *Scientific reports* 10.1 (2020), p. 18951.

[131] L. A. Vale-Silva and K. Rohr. “Long-term cancer survival prediction using multimodal deep learning”. In: *Scientific Reports* 11.1 (2021), p. 13505.

[132] T. Abdelaal *et al.* “A comparison of automatic cell identification methods for single-cell RNA sequencing data”. In: *Genome biology* 20 (2019), pp. 1–19.

[133] A. Singh *et al.* “scTopoGAN: unsupervised manifold alignment of single-cell data”. In: *Bioinformatics Advances* 3.1 (Jan. 2023). Ed. by M. Rattray. ISSN: 2635-0041.

[134] Y. Roohani, K. Huang, and J. Leskovec. “Predicting transcriptional outcomes of novel multigene perturbations with GEARS”. In: *Nature Biotechnology* 42.6 (Aug. 2023), pp. 927–935. ISSN: 1546-1696.

[135] D. Zhao *et al. Position: Measure Dataset Diversity, Don’t Just Claim It.* 2024. arXiv: 2407.08188 [cs.LG].

[136] T. Schoeler *et al.* “Participation bias in the UK Biobank distorts genetic associations and downstream analyses”. In: *Nature Human Behaviour* 7.7 (2023), pp. 1216–1227.

[137] A. Biddanda, D. P. Rice, and J. Novembre. “A variant-centric perspective on geographic patterns of human allele frequency variation”. In: *eLife* 9 (Dec. 2020). ISSN: 2050-084X.

[138] N. V. Chawla and G. Karakoulas. “Learning from Labeled and Unlabeled Data: An Empirical Study across Techniques and Domains”. In: *J. Artif. Int. Res.* 23.1 (Mar. 2005), pp. 331–366. ISSN: 1076-9757.

[139] A. T. Smith and C. Elkan. “Making Generative Classifiers Robust to Selection Bias”. In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’07. San Jose, California, USA: Association for Computing Machinery, 2007, pp. 657–666. ISBN: 9781595936097.

[140] J. Huang *et al.* “Correcting Sample Selection Bias by Unlabeled Data”. In: *Proceedings of the 19th International Conference on Neural Information Processing Systems*. NIPS’06. Canada: MIT Press, 2006, pp. 601–608.

[141] A. Liu and B. Ziebart. “Robust classification under sample selection bias”. In: *Advances in Neural Information Processing Systems* 1 (Jan. 2014), pp. 37–45.

[142] W. M. Kouw and M. Loog. “A Review of Domain Adaptation without Target Labels”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.3 (Mar. 2021), pp. 766–785.

[143] H. Shimodaira. “Improving predictive inference under covariate shift by weighting the log-likelihood function”. In: *Journal of Statistical Planning and Inference* 90.2 (Oct. 2000), pp. 227–244.

[144] B. Zadrozny. “Learning and Evaluating Classifiers under Sample Selection Bias”. In: *Proceedings of the Twenty-First International Conference on Machine Learning*. ICML ’04. Banff, Alberta, Canada: Association for Computing Machinery, 2004, p. 114. ISBN: 1581138385.

[145] C.-H. Chang and J.-H. Lin. “Decision Support and Profit Prediction for Online Auction Sellers”. In: *Proceedings of the 1st ACM SIGKDD Workshop on Knowledge Discovery from Uncertain Data*. U ’09. Paris, France: Association for Computing Machinery, 2009, pp. 1–8. ISBN: 9781605586755.

[146] C.-W. Seah, I. W.-H. Tsang, and Y.-S. Ong. “Healing Sample Selection Bias by Source Classifier Selection”. In: *2011 IEEE 11th International Conference on Data Mining*. 2011, pp. 577–586.

[147] M. Sugiyama, M. Yamada, and M. C. du Plessis. “Learning under nonstationarity: covariate shift and class-balance change”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 5.6 (Aug. 2013), pp. 465–477.

[148] T. D. Nguyen, M. Christoffel, and M. Sugiyama. “Continuous Target Shift Adaptation in Supervised Learning”. In: *Asian Conference on Machine Learning*. Ed. by G. Holmes and T.-Y. Liu. Vol. 45. Proceedings of Machine Learning Research. Hong Kong: PMLR, Nov. 2016, pp. 285–300.

[149] J. Kremer *et al.* “Nearest neighbor density ratio estimation for large-scale applications in astronomy”. In: *Astronomy and Computing* 12 (Sept. 2015), pp. 67–72.

[150] Z. Shen *et al.* “Causally Regularized Learning with Agnostic Data Selection Bias”. In: *Proceedings of the 26th ACM International Conference on Multimedia*. MM ’18. Seoul, Republic of Korea: Association for Computing Machinery, 2018, pp. 411–419. ISBN: 9781450356657.

[151] M. Diesendruck *et al.* “Importance Weighted Generative Networks”. In: *Machine Learning and Knowledge Discovery in Databases*. Springer International Publishing, 2020, pp. 249–265.

[152] W. Du and X. Wu. “Fair and Robust Classification Under Sample Selection Bias”. In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. CIKM ’21. Virtual Event, Queensland, Australia: Association for Computing Machinery, 2021, pp. 2999–3003. ISBN: 9781450384469.

[153] J. Blitzer, R. McDonald, and F. Pereira. “Domain Adaptation with Structural Correspondence Learning”. In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. EMNLP ’06. Sydney, Australia: Association for Computational Linguistics, 2006, pp. 120–128. ISBN: 1932432736.

[154] B. Fernando *et al.* “Unsupervised Visual Domain Adaptation Using Subspace Alignment”. In: *2013 IEEE International Conference on Computer Vision*. 2013, pp. 2960–2967.

[155] W. M. Kouw *et al.* “Feature-Level Domain Adaptation”. In: *J. Mach. Learn. Res.* 17.1 (Jan. 2016), pp. 5943–5974. ISSN: 1532-4435.

[156] W. M. Kouw and M. Loog. “Robust domain-adaptive discriminant analysis”. In: *Pattern Recognition Letters* 148 (Aug. 2021), pp. 107–113.

[157] G. Wilson and D. J. Cook. “A Survey of Unsupervised Deep Domain Adaptation”. In: *ACM Transactions on Intelligent Systems and Technology* 11.5 (July 2020), pp. 1–46. ISSN: 2157-6912.

[158] G. J. McLachlan. “Iterative Reclassification Procedure for Constructing an Asymptotically Optimal Rule of Allocation in Discriminant Analysis”. In: *Journal of the American Statistical Association* 70.350 (1975), pp. 365–369. eprint: `https://www.tandfonline.com/doi/pdf/10.1080/01621459.1975.10479874`.

[159] A. Blum and T. Mitchell. “Combining Labeled and Unlabeled Data with Co-Training”. In: *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*. COLT’ 98. Madison, Wisconsin, USA: Association for Computing Machinery, 1998, pp. 92–100. ISBN: 1581130570.

[160] C. Persello and L. Bruzzone. “Active and Semisupervised Learning for the Classification of Remote Sensing Images”. In: *IEEE Transactions on Geoscience and Remote Sensing* 52.11 (2014), pp. 6937–6956.

[161] J. W. Richards *et al.* “ACTIVE LEARNING TO OVERCOME SAMPLE SELECTION BIAS: APPLICATION TO PHOTOMETRIC VARIABLE STAR CLASSIFICATION”. In: *The Astrophysical Journal* 744.2 (Dec. 2011), p. 192.

2

ELISL: EARLY-LATE INTEGRATED SYNTHETIC LETHALITY PREDICTION IN CANCER

Yasin I. TEPELI

Colm SEALE

Joana P. GONÇALVES

This chapter is published in: Bioinformatics (2024) 40: Issue 1, doi: 10.1093/bioinformatics/btad764
Supplementary material is available online at: https://academic.oup.com/bioinformatics/article/40/1/btad764/7479688#supplementary-data

Motivation: *Anti-cancer therapies based on synthetic lethality (SL) exploit tumour vulnerabilities for treatment with reduced side effects, by targeting a gene that is jointly essential with another whose function is lost. Computational prediction is key to expedite SL screening, yet existing methods are vulnerable to prevalent selection bias in SL data and reliant on cancer or tissue type-specific omics which can be scarce. Notably, sequence similarity remains underexplored as a proxy for related gene function and joint essentiality.*
Results: *We propose ELISL, Early-Late Integrated forest ensembles for SL prediction using context-free protein sequence embeddings and context-specific omics from cell lines and tissue. Across 8 cancer types, ELISL showed superior robustness to selection bias and recovery of known SL genes, as well as promising cross-cancer predictions. Co-occurring mutations in a BRCA gene and ELISL-predicted pairs from the HH, FGF, WNT, or NEIL gene families were associated with longer patient survival times, revealing therapeutic potential.*
Data*: 10.6084/m9.figshare.23607558*
Code*: github.com/joanagoncalveslab/ELISL*

2.1. INTRODUCTION

Targeted anti-cancer therapy capitalises on tumour-specific molecular changes to selectively kill tumour cells, often resulting in reduced side effects compared to conventional chemotherapy and radiotherapy. Unfortunately, direct drug binding may be prevented by alterations of the drug target, for instance caused by loss of function mutations, amplification, or overexpression [1, 2]. A promising alternative explores synthetic lethality (SL) between a group of genes, whereby co-occurring dysfunction of all genes in the group causes cell death while disruption of only a subset of those genes is non-lethal [3]. Tumours with a known dysfunctional gene can then be treated by targeting its SL partner genes.

The viability of SL-based therapies has been confirmed by the approval of PARP-inhibitor drugs for treatment of BRCA-deficient tumours [4, 5]. Yet, the search for other SL interactions is proving challenging. New SL interactions are identified through expensive and laborious molecular perturbation experiments [6–10], which deem exhaustive screening impractical. Notably, computational SL prediction can greatly help prioritise candidates for follow-up.

Existing SL prediction methods can be categorised into statistical approaches and machine learning (ML) models. Statistical methods such as DAISY [11], BiSep [12], and ISLE [13] select SL pairs by imposing thresholds on statistical properties associated with SL, such as mutual exclusivity of mutations, coexpression, or changes in dependency on a gene for cell survival. Although statistical methods are intuitive, they struggle to capture complex relationships underlying SL interactions and tend to underperform compared to ML-based models [14]. The ML models can be further split into SL-topology and feature-based.

SL-topology methods represent existing SL data as a network of pairwise SL

interactions (edges) between genes (nodes). This network is used to identify shared SL patterns across genes and infer new SL interactions with matrix factorisation (pca-gCMF [15], GRSMF [16], and SL2MF [17]) or graph-based methods (DDGCN [18] and GCATSL [19]). The dependence of SL-topology methods on existing SL interactions typically means that (i) prediction scope is limited to genes with known SL partners, (ii) performance is heavily influenced by connectivity while SL data is reportedly sparse, and (iii) the approach is better suited for transferring SL interactions between genes with similar SL profiles than *de novo* SL discovery. Additionally, SL data shows prevalent selection bias towards functionally related genes with similar SL profiles, which SL-topology methods are designed to exploit. However, such limited set of SL interactions will not generalise to most other genes, making SL-topology methods sensitive to selection bias [14].

Feature-based ML models are built with supervised ML algorithms using omics features (DiscoverSL [20], EXP2SL [21], Lu [22], and SBSL [14]), enabling them to learn complex rules underlying SL interactions and remain more robust to selection bias. Most feature-based methods rely on (regularised) logistic regression or random forests to predict SL based on multiomics features [14, 20, 22]. Alternatively, EXP2SL uses a neural network to learn from a fixed set of genes and their expression in cancer cell lines [21].

Common to feature models is a focus on context-specific data for a tissue type of interest: for lung cancer, this could be omics of lung cancer cell lines and tumour tissue. While valuable for SL prediction, context-specific data may be difficult to obtain for some (rarer) cancer types, limiting the ability to learn useful models.

We argue that context-free metrics of functional similarity between genes could also be informative for SL prediction. The idea is that genes with similar functions have more related or redundant activity, making it more likely that a (cancer) cell would depend on the joint loss of function of those genes for its survival [23]. We consider the homology of protein sequences and similarity of protein-protein interactions (PPIs) as candidate metrics, which have been used successfully as proxies for functional similarity in tasks such as protein function prediction [24, 25]. Of note, the ISLE method has incorporated similarity of gene phylogenetic profiles for SL prediction. While relying on sequence homology to estimate evolutionary conservation across species, the similarity of phylogenetic profiles is ultimately influenced by a number of factors including focus on DNA sequence, choice and homology of other species data, and quality of inferred phylogenies. We thus favour a context-free representation of each gene pair based on direct comparison of the corresponding protein sequences for the organism of interest. Aminoacid sequences are closer to the functional roles of the genes than DNA, and their features can be compared directly for any pair of genes to provide an unbiased view of potential functional relationships for cells of that organism. Our use of vectorised sequence embeddings further enables a fine-grained exploration of sequence features that would otherwise be masked when relying on a single homology value for a pair of genes.

We propose *Early-Late Integrated Synthetic Lethality* (ELISL) prediction models, the first to integrate context-free direct protein sequence relationships and context-specific omics to predict SL for pairs of genes. Context-free features in ELISL encode each gene pair using embeddings of their protein sequences or PPIs. Context-specific features are stratified per tissue and sample type. We consider cancer cell lines because they are well characterised model systems with unique gene dependency data, quantifying cell viability upon gene perturbation, which is notably relevant for SL prediction and unavailable for patient tumours. ELISL looks at the relation between dependency scores and genetic or transcriptional alterations, as increased dependency on a gene in cell lines with altered activity of another gene could signal SL between the two. Separately, we include tissue omics to be able to explore the complexity inherent to human tissues. Here, impact of mutations within a gene on the expression of another gene suggests related function and thus increased SL potential [14]. In addition, correlation in gene expression and copy number aberration in both healthy and tumour tissue could help identify tumour-specific changes in the relationship between a pair of genes [14]. Finally, effect of tumour-specific co-alterations of two genes on patient survival could be indicative of SL, as simultaneous loss of function of SL genes might prolong survival by inducing cancer cell death, even if co-alterations are rare due to natural selection [13, 26, 27]. To effectively learn from low and high-dimensional data across sparser and denser representations, ELISL combines early (concatenation) and late (output ensembling) integration [28] using a collection of forest ensembles.

2.2. METHODS

The aim of the proposed ELISL framework is to predict if a given gene pair is synthetic lethal by leveraging context-free and context-specific omics that represent different relationships between the pair of genes at the molecular level (Fig. 2.1a). To do this, ELISL makes use of an early-late integration strategy comprising six regularised forest ensembles. Five models learn each from one individual context-free/specific source for later integration, and one early integration model learns from all concatenated features, enabling interactions across data sources (Fig. 2.1a). The final ELISL prediction probability is calculated as a weighted average of the probabilities of its six submodels.

2.2.1. DATA COLLECTION AND FEATURE GENERATION

ELISL models learn from two categories of features: context-free relations between genes based on protein sequence or PPIs, and context-specific features based on cell line and tissue omics. A featurised representation of each gene pair is derived per category and data source as an f_i-dimensional vector, where f_i is the number of features for data source i. For a set of N samples or gene pairs, this yields a matrix of dimensions $N \times f_i$, where each row refers to a gene pair and columns denote the different features.

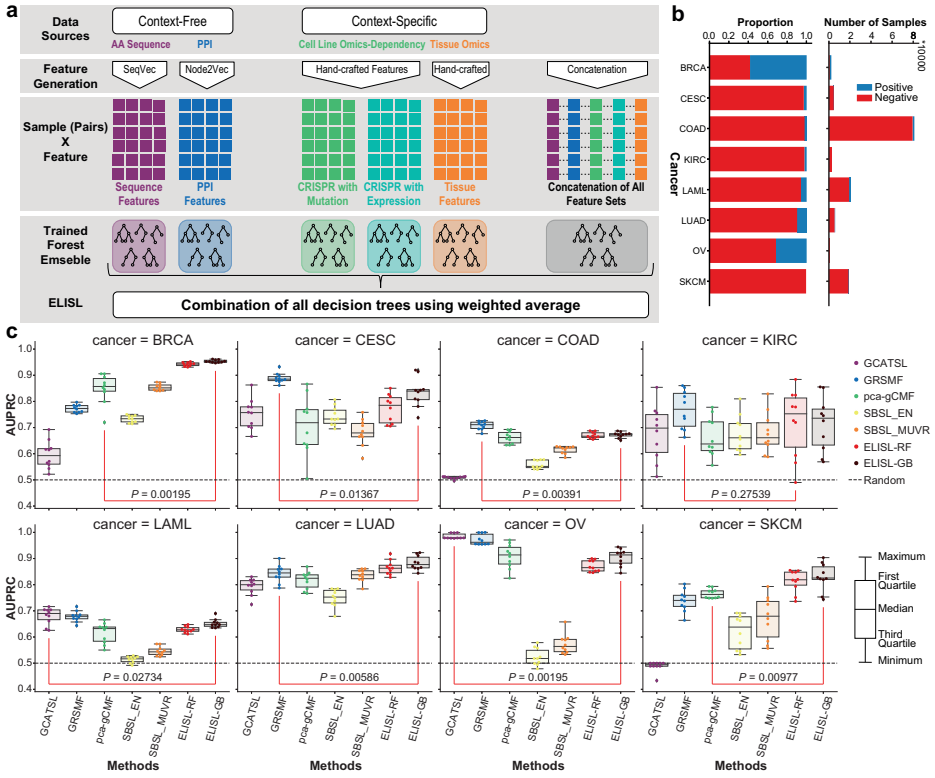


Figure 2.1: **ELISL framework, SL label imbalance, and within cancer prediction performance.** **a,** The ELISL framework **b,** Number and ratio of positive and negative samples in the train set for each cancer type. **c,** Prediction performance (AUPRC) of SL prediction methods within a cancer type over 10 runs. *P*: significance of the difference in performance between the best of other models and the best ELISL model over 10 runs (red lines).

Protein sequence and protein-protein interactions. We retrieved reviewed protein sequences from UniProt [29] and used the SeqVec pretrained model [30] to extract a 1024-dimensional embedding vector for every protein sequence. The sequence-based feature vector of each gene pair was then calculated as the absolute difference between the vectors of the proteins encoded by the two genes in the pair. We collected protein-protein interactions (PPIs) from the STRING database [31], considering only manually curated or experimentally validated interactions. Using these data, we built a network graph of genes (nodes) and undirected interactions between them (edges), and extracted a 64-dimensional embedding vector for each gene in the network using the Node2Vec method with default parameters [32]. To obtain the PPI feature vector for each pair of genes, we took the absolute difference between the embedding vectors of the two genes.

Cancer cell line omics. We retrieved dependency scores of cancer cell lines measured upon gene perturbation from the Cancer Dependency Map portal (public release 2018Q3 [33, 34]). Gene expression and mutation data from the Cancer Cell

Line Encyclopedia (CCLE) [35, 36] were obtained from the cBioPortal repository (Broad 2019) [37]. Based on these omics data, we defined alterations as encompassing non-silent mutations, gene expression z-scores larger than 1.96 or smaller than −1.96 (95% confidence), and discrete copy number aberration score equal to 2 (amplification) or −2 (deep loss). For gene expression, we used log-transformed mRNA z-scores compared to the expression distribution of all samples (RNA-seq RPKM). For copy number scores, we used discrete values generated by the GISTIC algorithm [37, 38]. Two feature sets were created based on cell line omics: *CRISPR with mutation* and *CRISPR with expression* based on CRISPR gene dependency scores and mutation data or gene expression, respectively. Each of these comprised four features: average dependency of the first (or second) gene across cell lines where the second (or first) gene was unaltered, and average dependency of the first (or second) gene across cell lines where the second (or first) gene was altered.

Tissue omics. We collected gene expression, mutation, copy number aberration, and clinical data for patient tissue samples in The Cancer Genome Atlas ([39] from the cBio portal [37]. We used two different gene expression scores: log-transformed mRNA z-scores relative to the distribution of all samples (RNA Seq RPKM) to identify expression based alterations, and mRNA gene expression (RNA Seq V2 RSEM) to quantify expression level. Additionally, we collected healthy donor tissue gene expression data as transcript per million (TPM) from the GTEx portal [40] (dbGaP Accession phs000424.v8.p2). Alterations were defined as encompassing non-silent somatic mutations, gene expression z-scores larger than 1.96 or smaller than −1.96 (95% confidence), and discrete copy number score of 2 (amplification) or −2 (deep loss). Using these alterations, we categorised patient tumour samples into two groups: with alterations in both genes, where an alteration in one of the omics was sufficient; and without simultaneous alterations in both genes. From tissue omics, we generated the following sets of features: patient survival, average gene expression in altered or unaltered tumour patient samples, gene coexpression in patient tumour/normal tissue or in healthy donor tissue, and correlation of copy number aberrations in patient tumour samples. The survival feature was the p-value of a Wald significance test for the patient group variable based on co-mutation status using a Cox proportional hazards (CoxPH) model of survival time, including covariates for age, sex, and cancer type. Four average gene expression features were defined as the average gene expression of the first (or second) gene in tumour samples where the second (or first) gene was: unaltered (2 features) or altered (2 features). Additionally, six coexpression features were calculated as the Pearson's correlation and respective p-value between the expression levels of the two genes in a gene pair in the following sets of samples: TCGA tumour samples from cancer patients (2 features), TCGA normal samples from cancer patients (2 features), and GTEx healthy donor tissue samples (2 features). Finally, two features expressing the correlation and p-value of copy number aberrations between the two genes in a gene pair were calculated using Spearman's correlation.

Synthetic lethality labels. We obtained experimentally derived SL labels from four studies: DiscoverSL [20], ISLE [13], EXP2SL [21], and Lu et al. [22]. These aggregate the results of 25 original experimental studies (Supplementary Table 2.S1), providing positive (SL) and negative (non-SL) labelled pairs. We note that there is no consensus on the criteria used to identify SL and non-SL pairs, with each study employing its own methodology. Positive SL relationships are typically identified based on statistical tests to detect an effect of simultaneous alterations to two genes, endogenous or induced, as a reduction in cell survival ability. As for non-SL pairs, some studies use statistical tests to determine if the interaction between the two genes improves cell survival or growth (opposite of an SL effect), while others label any gene pairs tested but not significant for an SL relationship as non-SL pairs. This makes non-SL pairs less reliable, which we consider during model evaluation. From these 4 studies, we found SL labels for 8 different cancer types (Fig. 2.1b), and removed all gene pairs with any disagreements in SL label across studies (Supplementary Table 2.S2). Unless otherwise specified, we used one SL dataset containing all unique gene pairs found across the four SL label sets.

2.2.2. ELISL MODELS

ELISL models (Fig. 2.1a) take as input a featurised representation of a given gene pair, and generate an SL prediction score denoting the probability that such gene pair is synthetic lethal. Models are learned using SL labelled gene pairs, and the representation comprises features from context-free and -specific omics data.

EARLY LATE INTEGRATION FRAMEWORK

The early-late integrated framework is designed to learn models from a given number k of data sources, with $k \in \mathbb{N}$ and $k \geq 2$, as follows. We build k models, each learning from the feature set created for one of the k individual data sources of interest. We also train an additional model using the feature set obtained by concatenating the features generated from all the individual k data sources. The predictions of the $k+1$ models are aggregated using weighted average, with weights based on the validation performances of the individual models. More formally, each individual dataset X_i, with $i \in \mathbb{N}$ and $\{1,...,k\}$, is a feature matrix $X_i \in \mathbb{R}^{N \times f_i}$, with N denoting the number of examples or gene pairs (rows in X_i) and f_i the number of features (columns in X_i). The concatenated dataset is defined as $X_{k+1} \in \mathbb{R}^{N \times \sum_{i=1}^{k} f_i}$ and results from concatenating the sets of feature matrices of all k individual data sources, $\{X_1,...,X_k\}$. Each model is an ensemble of trees learned using a given dataset X_i with the corresponding labels for its N examples (gene pairs). Models are trained together with shared hyperparameters. Finally, the prediction score of a pair is calculated as $\hat{y} = \sum_{i=1}^{k+1} w_i \hat{y}_i$, where w_i is the weight of model i and $\hat{y}_i$ is the prediction probability score of the gene pair according to model i. The weight w_i of each model in the final score is determined as the prediction performance on the validation set, normalised over all models: $w_i = \frac{p_i}{\sum_{i=1}^{k+1} p_i}$, where p_i denotes the

performance of model i (see Supplementary Materials).

2.2.3. MODEL TRAINING AND EVALUATION

We built ELISL models using two types of ensembles of decision trees: random forests (ELISL-RF [41]) and gradient-boosted decision trees (ELISL-GB [42]).

Single-cancer models. For each cancer type, we first split the labelled pairs into disjoint train (80%) and test (20%) sets. Then we generated ten runs: per run, pairs of train and test were drawn by random undersampling of the majority class to ensure balance of positive and negative SL labels. All SL prediction models were evaluated in ten runs, each using one of the generated train/test splits (runtimes in Supplementary Table 2.S3). Per run, models were learned on the train set and evaluated on the test set using area under the precision-recall curve (AUPRC) and receiver-operating characteristic curve (AUROC) as performance metrics. For ELISL, the hyperparameters and the weight of each submodel were determined with Bayesian grid search and 5-fold cross-validation, using validation AUPRC as performance metric (Supplementary Materials)We assessed significance of the difference in performance between the best ELISL and the best of the other models using two-sided Wilcoxon signed-rank tests.

Comparison with other SL prediction methods. We trained the pca-gCMF, GCATSL, and GRSMF methods using the parameters suggested by the authors. For SBSL-EN, and SBSL-MUVR, we found hyperparameters using grid search as described in the original paper (Supplementary Materials). All models were trained and evaluated on the same train and test sets.

Pan-cancer models. Pan-cancer models were obtained by ensembling the already trained models from each cancer type, where the weight of each model in the final prediction was attributed based on validation performance. Combining the predictions of the different models in this way allowed us to bypass challenges of training with large imbalances in number of samples across cancer types. This would have required us to balance the data across cancer types, which could also severely limit the number of pairs available for training.

Importance of feature categories. We calculated the importance of each feature category for the ELISL-RF models of the six cancer types with the smallest variance in AUPRC scores across runs (BRCA, CESC, COAD, LAML, LUAD, and OV). To calculate the importance score for a given feature set, we permuted the values of all of its features across the gene pairs in the test set, so as to break the relation between features and labels. When permuting a given feature set, the concatenated features also changed accordingly. We calculated the prediction errors for the

original test set and each of 20 different permuted test sets as (1-AUPRC) scores. The importance score was then defined as the ratio between the prediction errors obtained for the permuted test set and the original test set.

Detailed analysis of predicted SL pairs

To evaluate predictions for gene pairs with known labels, we ranked all gene pairs found in at least one of the ten tests sets based on their average prediction probability scores of the single-cancer models obtained over the ten runs.

Predictions for gene pairs with unknown SL labels. We created a set of gene pairs with unknown SL labels for breast cancer by generating all possible pairs of genes found in cancer and DNA repair pathways, using KEGG, PID, and Reactome pathway gene sets from the molecular signatures database v7.1 [43]. From the total of 572 genes found across all pathways (Supplementary Materials), we generated 163,306 gene pairs. After excluding the pairs already present in the train or test sets, we ended up with 163,118 gene pairs. The SL scores of the pairs with unknown labels were determined as the average prediction probability over the 10 runs of the single cancer experiment.

Survival analysis of newly predicted SL gene pairs. To validate predicted SL gene pairs without known labels, we investigated differences in survival time between patients with or without simultaneous alterations (co-mutation) in both genes. Given that only a small number of patient tumours typically carried simultaneous mutations, we looked at the relation between gene families rather than individual genes. We stratified the patient tumour samples into two groups based on co-mutation status, denoting presence or absence of alterations in genes of both families. Specifically, for a given pair of genes (Gene 1, Gene 2), we denote the group of samples with co-mutations in both a member from the family of Gene 1 (Fam 1) and a member from the family of Gene 2 (Fam 2) as (Fam 1 and Fam 2), while the group without co-mutations is expressed by ~(Fam 1 and Fam 2). Survival times of both groups were estimated using a Cox proportional hazards (CoxPH) model, including covariates for age, sex, and cancer type in addition to co-mutation status. The significance of each variable in the CoxPH model (p-value) was calculated using Wald significance tests. We also generated plots of Kaplan-Meier survival curves for the patient groups. Additionally, we represented two subgroups of the group without co-mutations, namely: the subgroup with mutation in only one of the families but not both (Fam 1 xor Fam 2), and the subgroup with no mutation in any of the genes from both families (Unaltered). Note that, although the ELISL-RF model included a survival-based feature as part of the tissue-specific model, the contribution of tissue features overall was reportedly small (1.09). One reason for this could be the fact that survival data was very sparse due to the rare occurrence of co-mutations in both genes.

2.3. RESULTS AND DISCUSSION

2.3.1. CANCER-SPECIFIC SYNTHETIC LETHALITY PREDICTION

We first evaluated the ability of ELISL models to generalise within a cancer type, for eight distinct cancer types. We compared ELISL-RF and ELISL-GB to five other recently published ML models with high performances in their categories, namely: pca-gCMF, GRSMF, and GCATSL as SL-topology methods, and SBSL-MUVR and SBSL-EN as supervised ML models.

Supervised ELISL models significantly outperformed the other methods in breast (BRCA), lung (LUAD), and skin (SKCM) cancers (Wilcoxon $p \leq 0.01$). Graph-based matrix factorisation GRSMF took the lead in cervix (CESC) and colon (COAD), and was close second to GCATSL in leukemia (LAML) and ovarian (OV) cancers (AUPRC Fig. 2.1c, AUROC Supplementary Fig. 2.S1a, 2.S2a), with ELISL models remaining competitive as well. The performance of GCATSL varied widely across cancer types, and was notably poor in BRCA, COAD, and SKCM. For kidney (KIRC) cancer, all methods showed high variance, and there was no clear best performing model. Overall, across all cancer types and runs, ELISL-GB was the most successful method (average AUPRC 0.805), while GRSMF and ELISL-RF were second and third (average AUPRCs 0.796 and 0.785), respectively (Supplementary Fig. 2.S2a). SL-topology models showed strikingly high performances in OV. This is consistent with the previous report that SL-topology methods might excel on OV due to the strong selection bias in SL labelled pairs, which span a limited set of functionally related genes [14].

2.3.2. ROBUSTNESS OF SL PREDICTION TO GENE SELECTION BIAS

To assess the impact of gene selection bias on the SL prediction methods, we performed experiments with induced or inherent differences in selection bias between the train and test sets.

Double gene holdout. To induce differences in gene selection bias, we enforced zero overlap in genes between each train and corresponding test set (Fig. 2.2a). This differs from the original experiment (Fig. 2.1c), where matched train/test sets were disjoint in terms of gene pairs but not individual genes. All methods were evaluated in four cancer types: BRCA, CESC, LUAD, and OV. We excluded KIRC and SKCM due to the limited number of gene pairs, and COAD and LAML due to poor performances in the original experiment (Fig. 2.1c).

Using double gene holdout, the performances of all models decreased significantly for all cancer types (AUPRC Fig. 2.2a, AUROC Supplementary Fig. 2.S1b, 2.S2b), possibly owing to the reduction in the number of training gene pairs imposed by the train/test set construction (Supplementary Table 2.S4). For BRCA, the two ELISL models performed the best (median AUPRC: ELISL-RF 0.67, ELISL-GB 0.69), while the performance of SL-topology methods dropped to nearly random (Fig. 2.2a, top left). For CESC, GRSMF had outperformed ELISL in the original single cancer

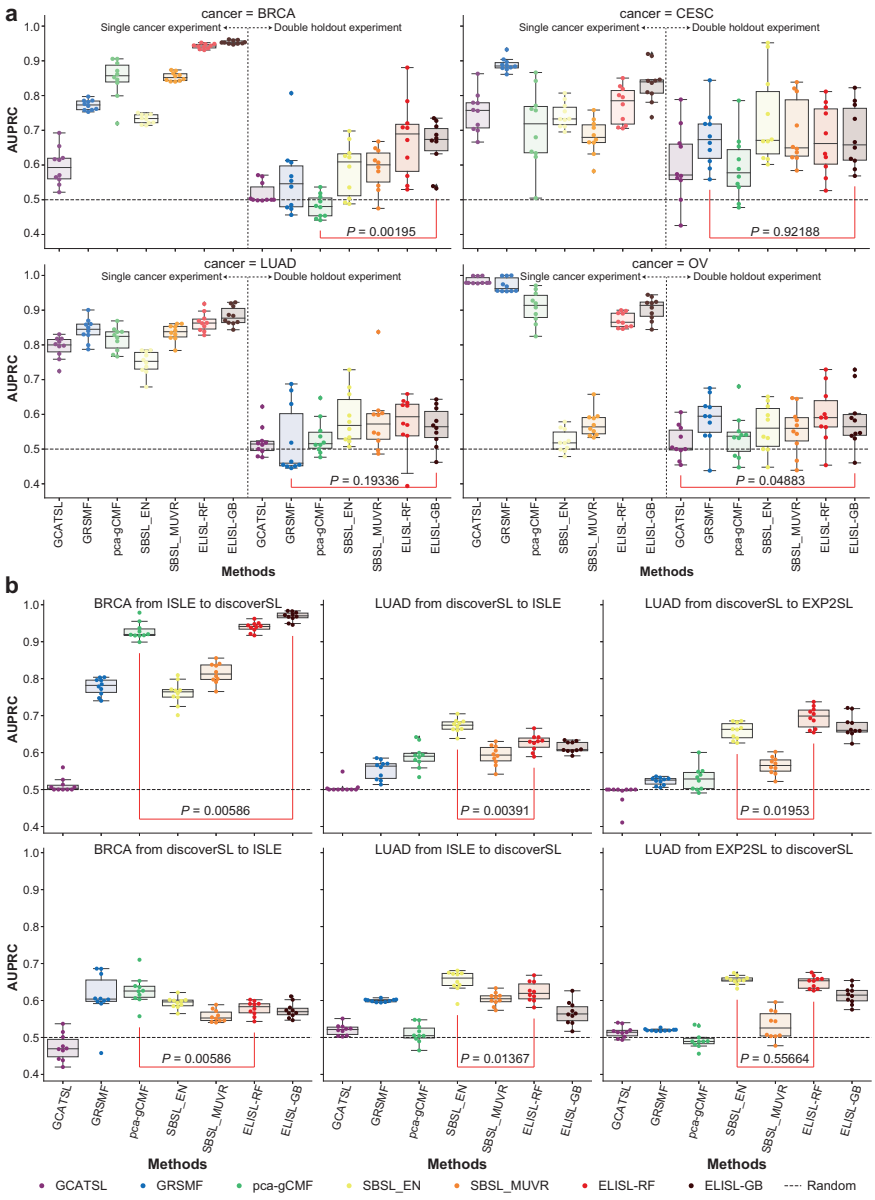


Figure 2.2: **Impact of gene selection bias on SL prediction performance.** **a,** *Left panels*: performance under similar train/test bias (same as in Fig. 2.1c); *Right panels*: double gene holdout inducing differences in gene selection bias between train and test set. Performance (AUPRC) per cancer type and for 10 runs where each pair of train and test sets does not share any genes. *P*: significance of the difference between the double holdout performances of the two models that performed best under similar bias. **b,** Cross-SL label source. Performance (AUPRC) reported for models trained using labels from one SL source and evaluated on another SL source (10 runs). *P*: significance of the difference between the best ELISL model and the best of the other models.

experiment, but this difference was no longer apparent or significant using double gene holdout (Wilcoxon $p \approx 0.92$, Fig. 2.2a, top right). For LUAD, most methods

struggled with double gene holdout Fig. 2.2a, bottom left). However, supervised ML models SBSL and ELISL retained above random performances, with ELISL-RF achieving the best median AUPRC (0.59). For OV, we saw the largest decrease in performance using double holdout compared to the original experiment, which was expected given the prominent SL label bias. ELISL-RF and GRSMF performed the best in OV (median AUPRC 0.59 for both) using double gene holdout, while SBSL models retained their originally modest performances (Fig. 2.2a, bottom right). The GCATSL method performed poorly with double gene holdout in all cancers (near 0.5 median AUPRC), including in OV for which it was the best model in the original experiment (0.98 median AUPRC).

Overall, supervised ML models SBSL and ELISL performed better than the remaining models using double gene holdout. SL-topology methods delivered inconsistent performances across cancer types, and were thus more sensitive to selection bias. ELISL models outperformed the other methods in BRCA and LUAD, and were comparable to the best performing models in CESC and OV.

Cross-SL label prediction. Since the double holdout is an extreme scenario, we also evaluated SL prediction models with inherently occurring differences in gene selection bias between train and test sets. To do this, we trained the models using SL labelled pairs from one data source and tested them on labelled pairs from another source for the same cancer type. We used the following (and reverse) SL labelled sources, yielding between 78 and 1146 train samples (Supplementary Table 2.S5): for BRCA, train on ISLE and test on DiscoverSL; for LUAD, train on DiscoverSL and test on EXP2SL or Lu et al.

ELISL models outperformed the other methods when training on ISLE and predicting on DiscoverSL for BRCA, as well as when training on DiscoverSL and predicting on EXP2SL for LUAD (AUPRC Fig. 2.2b, AUROC Supplementary Fig. 2.S1c, 2.S2c). For the remaining LUAD experiments, one of the ELISL models ranked second, whereas the linear SBSL-EN model took the lead. ELISL was not competitive when training on DiscoverSL and predicting on ISLE for BRCA: this was the combination where models had the least number of gene pairs to train on, 78, which could be challenging for models using larger numbers of features such as ELISL. Overall, across all cancer types and runs, ELISL-RF was the most successful method in both the double holdout and cross-dataset experiments (average AUPRCs 0.631 and 0.685), while SBSL-EN was second best with average AUPRCs 0.617 and 0.665, respectively (Supplementary Fig. 2.S2b-c). Thus, supervised ML models emerged as the most robust to selection bias, with SBSL-EN and ELISL-RF standing out.

2.3.3. CROSS-CANCER SL PREDICTION USING ELISL-RF MODELS

There is evidence that some SL interactions may occur in multiple cancer types. For instance, PARP inhibitor drugs are approved for the treatment of BRCA-deficient breast, ovarian, prostate [44], and pancreatic [45] tumours [46]. This suggests that there could be some benefit in leveraging successful models trained on cancer types

with sufficient data (BRCA, LUAD, OV) to predict SL in other cancers, for which samples are either not available or difficult to obtain (CESC, KIRC, and SKCM). To investigate, we evaluated the performance of cancer-specific ELISL-RF models against each of the remaining cancer types using the corresponding train and test sets over ten runs from the original single cancer experiment.

The success of cross-cancer SL predictions was modest for most pairwise cancer combinations, to which the quality and biases of the labels could have contributed as well (AUPRC Fig. 2.3a, (AUROC Supplementary Fig. 2.S3a)). Nevertheless, we saw some promising results. For the prediction of CESC pairs, the LUAD-trained model performed better than the CESC-trained model itself (0.85 vs. 0.77 mean AUPRC). Models trained on COAD or KIRC also achieved reasonable performances in CESC (0.69 and 0.71 mean AUPRC, respectively). For SL prediction in KIRC, the best model was trained using KIRC labeled pairs (0.72 mean AUPRC), followed by the model trained on CESC (0.68 mean AUPRC), and by the models trained on BRCA and LUAD (0.63 mean AUPRC). Overall, the results indicate that there could be potential in identifying SL relationships across cancer types.

We further investigated if models learned using SL labels from multiple cancer types (pan-cancer) would provide any benefit compared to cross-cancer predictions. For every cancer type T, we trained models using labelled pairs from all other cancer types except T, and then evaluated the predictions for labelled pairs in T. (see Methods). Pan-cancer models showed promising performance for CESC (0.74 mean AUPRC) and reasonable results for KIRC (0.65; Fig. 2.4a, bottom row). Performances of pan-cancer models were not better than those of cancer-specific and cross-cancer models, indicating that prior selection of relevant cancer types could be needed to effectively enable pan-cancer models to predict SL for cancer types with limited sample sizes.

2.3.4. Feature contributions to ELISL-RF models

To quantify the contribution of the different feature categories to the predictions of the ELISL-RF model, we used permutation feature importance [47] (see Methods). Sequence embeddings emerged as the most important feature in five cancer types (BRCA, COAD, LAML, LUAD, and OV), and second most important in CESC (mean importance: sequence 1.18) behind dependency with mutation (mean importance: 1.23). We note that importance values were more prominent for BRCA, CESC, LUAD, and OV because the performance of ELISL-RF was also higher for these cancer types (between 0.77 and 0.94 mean AUPRC) compared to COAD and LAML (0.67 and 0.63). High performance means low errors, which can result in larger ratios (importances) for small changes in performance. Beyond sequence, PPI and the interaction of CRISPR dependency and mutation were the second most important feature categories overall. Ultimately, all data sources contributed to the ELISL-RF model (mean importance > 1) in at least two cancer types, with the variation in importance across cancers suggesting that the integration of multiomics could be beneficial for cross-cancer SL prediction. We checked if the high-dimensionality of sequence embeddings influenced ELISL-RF, but using embedding sizes between 32

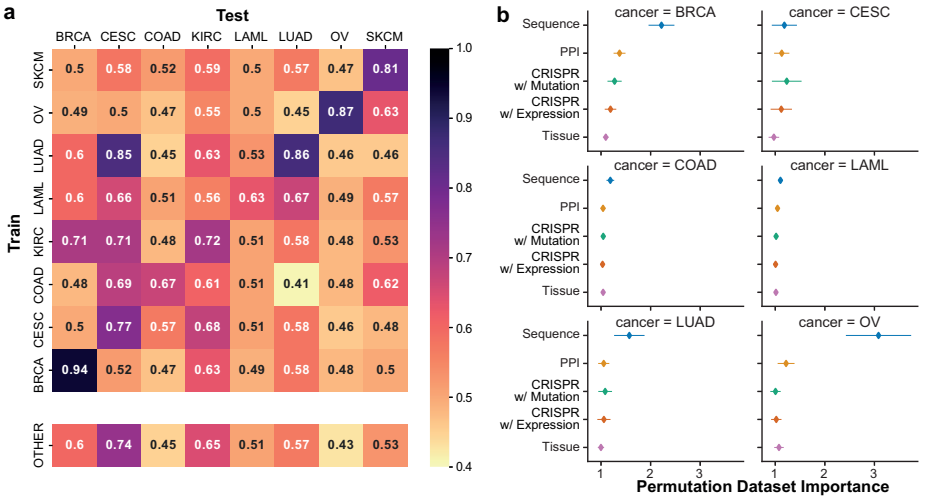


Figure 2.3: **ELISL-RF SL prediction within/across cancer types and feature contribution.** **a,** Performance of cancer-specific models and pan-cancer models, measured as average AUPRC over 10 runs. Pan-cancer model performances are reported in a separate row at the bottom, where models are trained on all other cancer types except the one the model is supposed to predict on. **b,** Contribution of each data source to the predictions of the ELISL-RF model within the same cancer type.

and 1024 led to comparable performances (Supplementary Materials and Supplementary Fig. 2.S3b).

2.3.5. POTENTIAL OF SL PAIRS PREDICTED BY ELISL-RF MODELS

To further assess the potential of ELISL-RF models, we first analysed the top known gene pairs ranked by prediction probability in BRCA, LUAD, and OV. The top three pairs for BRCA and OV were labelled as synthetic lethal (SL, Fig. 2.4a). In fact, all top 82 pairs for BRCA and top 16 pairs for OV had positive labels, confirming that ELISL-RF can recover known SL interactions. For LUAD, we counted six SL and four non-SL pairs amongst the top 10 predictions (Supplementary Table 2.S6). Notably, the highest ranked gene pair in LUAD, KRAS-MRPL28, had a non-SL label. However, an independent study found that disruption of *MRPL28* was lethal in *KRAS*-mutant cancer cell lines [48]. The finding was for colorectal cell lines, but lung cancer could share underlying mechanisms given that KRAS mutations are frequent in lung and colorectal cancers, and colorectal cancers often metastasise to lung [49, 50]. Therefore, we cannot discard the possibility that KRAS-MRPL28 could be mislabelled for LUAD.

Predictions for gene pairs with unknown SL status. Finally, we used ELISL-RF to make predictions for unknown gene pairs. We focused on BRCA, for which ELISL-RF models achieved the highest performance across experiments with varying gene selection bias. Since we aimed to assess the impact of top SL and non-SL predictions on patient survival, we also trained a separate ELISL-RF model on BRCA

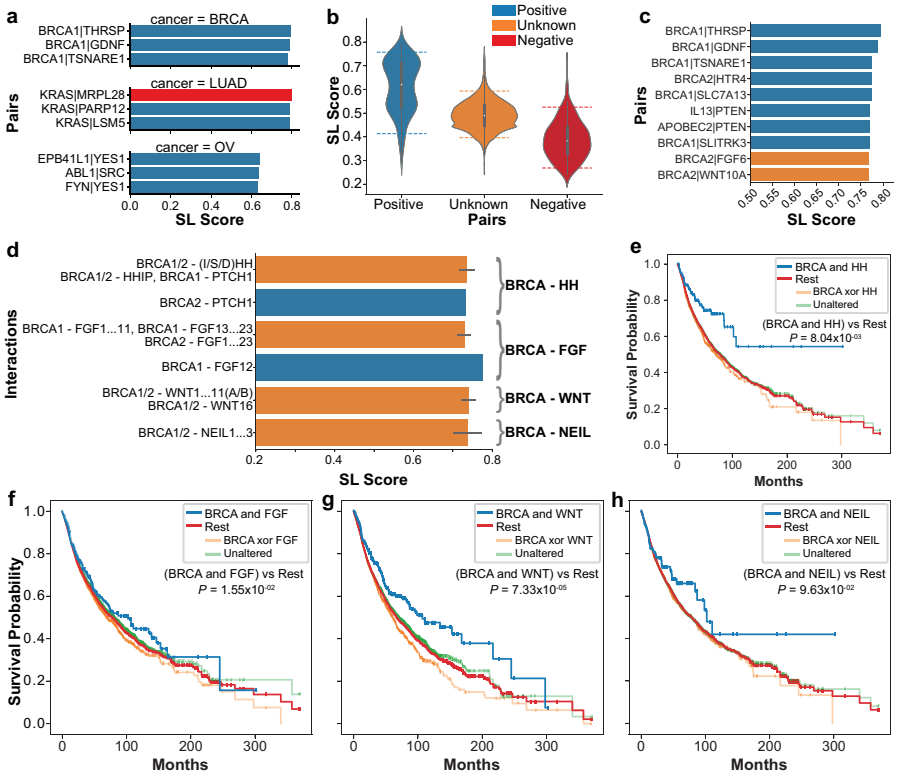


Figure 2.4: **Analysis of top SL gene pairs predicted by ELISL-RF. a,** Top 3 pairs ranked by SL prediction score for BRCA, LUAD, and OV (average across 10 test sets). Figures **b-c** show results for prediction of unknown gene pairs (not in test sets) using ELISL-RF trained on BRCA data without the survival feature. **b,** Distribution of SL scores for unknown pairs compared to known SL and non-SL pairs. Dashed lines denote 5% and 95% percentiles. **c,** Prediction scores of ELISL-RF without survival for the top 10 pairs in the BRCA test set and the unknown set. **d,** Prediction scores of ELISL-RF without survival for pairs involving BRCA1/2 and HH, FGF, or WNT family members. Bar length denotes average SL score and black line length represents standard deviation for the set of pairs of interest. Figures **e-h** show differences in survival between patient tumours with and without simultaneous alterations in both families of a gene pair, using Kaplan-Meier curves and Wald test p-values of survival differences based on CoxPH models of co-mutation status adjusted for age, sex, cancer type. For pairs involving BRCA genes and members of the **(e)** HH, **(f)** FGF, **(g)** WNT, and **(h)** NEIL families.

data without the survival feature for fairer analysis. We predicted labels for all pairs of genes involved in cancer and DNA repair pathways from KEGG, Reactome, and PID (Supplementary Materials) using both models. Overall, ELISL-RF without survival feature assigned higher SL prediction scores to pairs with known SL labels (median 0.62), compared to pairs with known non-SL labels (median 0.38), as expected (Fig. 2.4b). The distribution of SL prediction scores for unknown pairs showed no particular tendency (median 0.49).

Without the survival feature, we found two unknown gene pairs among the ten pairs with the highest ELISL-RF prediction scores, BRCA2-FGF6 and BRCA2-WNT10A (Fig. 2.4c), immediately followed by BRCA1-NEIL2 and BRCA1-NEIL1 among unknown pairs (Supplementary Fig. 2.S4). Using the survival feature, ELISL-RF ranked three unknown gene pairs in the top ten: BRCA1-HHIP, BRCA2-FGF6, and

BRCA1-FGF8 (Fig. 2.4c, Supplementary Fig. 2.S4). Of note, BRCA1-HHIP also ranked highly without the survival feature (15th among unknowns). We investigated the functional roles of these genes and their families, as well as association with patient survival. We extended our analysis to gene families to obtain more robust estimates of survival time, given that genes were infrequently co-altered.

Concerning the BRCA1-HHIP interaction, the hedgehog interacting protein (HHIP) binds to all three hedgehog family members (IHH, SHH, DHH) with affinity to the PTCH1 receptor, and regulates the hedgehog (HH) signaling pathway [51–53]. The HH pathway is SL with the PI3K/AKT/mTOR pathway in rhabdomyosarcoma [54], and the inhibition of PI3K is known to strengthen BRCA-PARP synthetic lethality in BRCA1-deficient breast cancer [55]. We thus reason that the HHIP gene or HH family could be an SL partner for BRCA1/2. Notably, the BRCA2-PTCH1 pair had a positive SL label [56], and all pairs between BRCA genes and HH family members yielded high prediction scores (>0.7 without survival feature, Fig. 2.4d). Analysis of TCGA tumour samples showed that patients whose tumours carried alterations in a BRCA gene (BRCA1 or BRCA2) and a HH family member (IHH, SHH, DHH, PTCH1) had longer survival times than the rest (difference in median >220 months and $p \approx 8.04 \times 10^{-3}$, Fig. 2.4e, Supplementary Table 2.S7).

We assessed the BRCA2-FGF6 and BRCA1-FGF8 pairs together, as involving a BRCA gene and FGF family member (FGF1 to FGF23). The fibroblast growth factor (FGF) family regulates cell differentiation and proliferation, taking part in cancer pathogenesis [57]. The BRCA1-FGF12 pair had a positive SL label, and all pairs between a BRCA gene and FGF family members had prediction scores higher than 0.7 (Fig. 2.4d). The median survival time for patients whose tumours had alterations in both families, BRCA1/2 and FGF1 to FGF23, was 23 months longer than for other patients with $p \approx 1.55 \times 10^{-2}$ (Fig. 2.4f, Supplementary Table 2.S7).

The top 5% of gene pairs (SL score > 7.57), also included several interactions between BRCA genes and WNT family members, eight and six when using and not using the survival feature, respectively(Fig. 2.4d, Supplementary Fig. 2.S4). The WNT pathway regulates various processes including cell fate determination [58, 59], and its inhibition could induce a BRCA-like state that makes cells vulnerable to PARP inhibition [60]. This might suggest interactions between WNT, BRCA, and PARP. Patients with tumours carrying mutations in BRCA and WNT genes lived (median) 89 months longer than the rest ($p \approx 7.35 \times 10^{-5}$, Fig. 2.4g, Supplementary Table 2.S7).

The NEIL gene family (comprising NEIL1-3) encodes DNA glycosylases involved in DNA repair via the base excision repair (BER) mechanism [61, 62]. Prior literature has suggested that specific SNPs in the NEIL2 gene could establish a synthetic lethal relationship with BRCA1/2 genes [63, 64]. Our analysis of TCGA tumour samples unveiled that patients with alterations in a BRCA gene (BRCA1/2) and a member of the NEIL family (NEIL1-3) experienced 24-month longer median survival times than others, although this difference did not reach statistical significance, likely due to the infrequency of co-occurring alterations ($p \approx 9.63 \times 10^{-2}$; Fig. 2.4f and Supplementary Table 2.S7).

For comparison with the known BRCA-PARP interaction, alterations in both BRCA and PARP (PARP1-16) genes led to 20 months longer median survival ($p \approx 3.14 \times 10^{-3}$, Supplementary Fig. 2.S5). For contrast, we looked at the four gene pairs with the lowest ELISL-RF scores for both models, with and without the survival feature. the union yielded 5 unique gene pairs: three pairs with non-SL label, PARP1|RIPK1 (both models), MAP3K7|PARP1 (both models), and GRK4|PARP1 (without survival); and two pairs with unknown SL status, namely MAP2K2|PARP1 (with survival) and DAPK2|PARP1 (both models) (Supplementary Fig. 2.S6). For PARP|RIPK, MAP3K|PARP, DAPK|PARP, and GRK|PARP, survival of patients with alterations in both gene families was respectively 8, 3, 9, and 8 months shorter (p-values 3.83×10^{-1}, 4.07×10^{-6}, 2.15×10^{-1}, 2.09×10^{-2} (Supplementary Fig. 2.S6a-d). For MAP2K|PARP, alteration in both gene families was associated with 17 months longer survival and $p \approx 2.41 \times 10^{-3}$ (Supplementary Fig. 2.S6e).

Overall, the significant association between patient survival times and co-alteration in families of highly ranked gene pairs suggests that ELISL-RF prioritises promising SL interactions.

2.4. Conclusion

We proposed ELISL, forest ensemble models that leverage gene functional relationships to predict SL in cancer. To our knowledge, ELISL models are the first to use context-free direct protein sequence relationships as a proxy for functional association for SL prediction, in addition to context-specific omics. The ELISL early-late integration strategy effectively enabled learning from high-dimensional sequence embeddings and tailored omics features.

ELISL models outperformed existing SL prediction methods, emerging as the most robust models overall under varying gene selection bias. Nevertheless, learning from biased data remains a fundamental ML challenge that merits further research. Some SL-topology models (GRSMF, pca-gCMF) performed well when train and test set followed similar distributions, but struggled to make useful predictions under different bias, confirming previous work [14]. Other feature-based models, SBSL, showed inconsistent performances across cancer types. This result exposed the issue of relying on context-specific features alone, which can be sparse or unavailable for some cancer types.

Sequence embeddings contributed the most to the predictions of ELISL models, and thus were responsible for the advantage of ELISL over context-specific SBSL models. Sequence embeddings also make ELISL models less dependent on context-specific features like gene dependencies, which are exclusively available for cellular models and may not directly translate to patient tumours.

Predicting across cancer types revealed challenging, but it was encouraging to see that ELISL models trained on colon, kidney, or lung cancer performed reasonably well on cervix cancer. Cross-cancer prediction should improve as higher quality, less biased, SL data becomes available. Nevertheless, a few successful cases point to the

existence of SL interactions across cancer types, which could bring benefit to a larger number of patients in the future.

Using ELISL to make predictions for unknown gene pairs, we investigated promising SL interactions. Survival analysis showed that simultaneous mutations in a BRCA gene and at least one member of the HH, FGF, WNT, or NEIL families associated with longer median patient survival times, reinforcing the ability of ELISL to predict SL interactions with therapeutic potential.

FUNDING

This work was supported by Holland Proton Therapy Center [2019020 to C.S.], US National Institutes of Health [U54EY032442, U54DK134302, U01DK133766, R01AG078803 to J.P.G.]. Authors are solely responsible, funders were not involved in this work.

REFERENCES

[1] J. Setton *et al.* “Synthetic Lethality in Cancer Therapeutics: The Next Generation”. In: *Cancer Discovery* 11.7 (Apr. 2021), pp. 1626–35.

[2] Y. Zhang *et al. Bias-Tolerant Fair Classification.* 2021.

[3] D. A. Chan and A. J. Giaccia. “Harnessing synthetic lethal interactions in anticancer drug discovery”. In: *Nature Reviews Drug Discovery* 10.5 (Apr. 2011), pp. 351–64.

[4] P. C. Fong *et al.* “Inhibition of Poly(ADP-Ribose) Polymerase in Tumors fromBRCAMutation Carriers”. In: *New England Journal of Medicine* 361.2 (July 2009), pp. 123–134. ISSN: 1533-4406.

[5] L. Hutchinson. “PARP inhibitor olaparib is safe and effective in patients with BRCA1 and BRCA2 mutations”. In: *Nature Reviews Clinical Oncology* 7.10 (Sept. 2010), p. 549.

[6] C. Jacquemont *et al.* “Non-specific chemical inhibition of the Fanconi anemia pathway sensitizes cancer cells to cisplatin”. In: *Molecular Cancer* 11.1 (2012), p. 26.

[7] C. M. Toledo *et al.* “Genome-wide CRISPR-Cas9 Screens Reveal Loss of Redundancy between PKMYT1 and WEE1 in Glioblastoma Stem-like Cells”. In: *Cell Reports* 13.11 (Dec. 2015), pp. 2425–39.

[8] D. Etemadmoghadam *et al.* “Synthetic lethality between CCNE1 amplification and loss of BRCA1”. In: *Proceedings of the National Academy of Sciences* 110.48 (Nov. 2013), pp. 19489–94.

[9] C. G. Hubert *et al.* “Genome-wide RNAi screens in human brain tumor isolates reveal a novel viability requirement for PHF5A”. In: *Genes & Development* 27.9 (May 2013), pp. 1032–45.

[10] D. Kranz and M. Boutros. “A synthetic lethal screen identifies FAT1 as an antagonist of caspase-8 in extrinsic apoptosis”. In: *The EMBO Journal* 33 (Jan. 2014), pp. 181–97.

[11] L. Jerby-Arnon *et al.* “Predicting Cancer-Specific Vulnerability via Data-Driven Detection of Synthetic Lethality”. In: *Cell* 158.5 (Aug. 2014), pp. 1199–209.

[12] M. Wappett *et al.* “Multi-omic measurement of mutually exclusive loss-of-function enriches for candidate synthetic lethal gene pairs”. In: *BMC Genomics* 17.1 (Jan. 2016), p. 65.

[13] J. S. Lee *et al.* “Harnessing synthetic lethality to predict the response to cancer treatment”. In: *Nature Communications* 9.1 (June 2018), p. 2546.

[14] C. Seale, Y. Tepeli, and J. P. Gonçalves. “Overcoming selection bias in synthetic lethality prediction”. In: *Bioinformatics* 38.18 (July 2022), pp. 4360–8.

[15] H. Liany, A. Jeyasekharan, and V. Rajan. “Predicting synthetic lethal interactions using heterogeneous data sources”. In: *Bioinformatics* 36.7 (Nov. 2019), pp. 2209–16.

[16] J. Huang *et al.* “Predicting synthetic lethal interactions in human cancers using graph regularized self-representative matrix factorization”. In: *BMC Bioinformatics* 20.S19 (Dec. 2019), p. 657.

[17] Y. Liu *et al.* “SL2MF: Predicting Synthetic Lethality in Human Cancers via Logistic Matrix Factorization”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 17.3 (May 2020), pp. 748–57.

[18] R. Cai *et al.* “Dual-dropout graph convolutional network for predicting synthetic lethality in human cancers”. In: *Bioinformatics* 36.16 (Mar. 2020), pp. 4458–65.

[19] Y. Long *et al.* “Graph contextualized attention network for predicting synthetic lethality in human cancers”. In: *Bioinformatics* 16 (Feb. 2021), pp. 2432–40.

[20] S. Das *et al.* “DiscoverSL: an R package for multi-omic data driven prediction of synthetic lethality in cancers”. In: *Bioinformatics* 35.4 (July 2018), pp. 701–2.

[21] F. Wan *et al.* “EXP2SL: A Machine Learning Framework for Cell-Line-Specific Synthetic Lethality Prediction”. In: *Frontiers in Pharmacology* 11 (Feb. 2020), p. 112.

[22] X. Lu *et al.* “Predicting Human Genetic Interactions from Cancer Genome Evolution”. In: *PLOS ONE* 10.5 (May 2015), e0125795.

[23] J. K. Dhanjal, N. Radhakrishnan, and D. Sundar. “Identifying synthetic lethal targets using CRISPR/Cas9 system”. In: *Methods* 131 (Dec. 2017), pp. 66–73.

[24] R.-S. Wang *et al.* “Analysis on multi-domain cooperation for predicting protein-protein interactions”. In: *BMC Bioinformatics* 8.1 (2007), p. 391.

[25] M. Kulmanov, M. A. Khan, and R. Hoehndorf. “DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier”. In: *Bioinformatics* 34.4 (Oct. 2017), pp. 660–8.

[26] S. Srihari *et al.* “Inferring synthetic lethal interactions from mutual exclusivity of genetic events in cancer”. eng. In: *Biology Direct* 10 (Oct. 2015), p. 57. ISSN: 1745-6150.

[27] X. Feng *et al.* “A platform of synthetic lethal gene interaction networks reveals that the GNAQ uveal melanoma oncogene controls the hippo pathway through FAK”. In: *Cancer cell* 35.3 (2019), pp. 457–472.

[28] M. Zitnik *et al.* “Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities”. In: *Information Fusion* 50 (Oct. 2019), pp. 71–91.

[29] A. Bateman *et al.* “UniProt: the universal protein knowledge base in 2021”. In: *Nucleic Acids Research* 49.D1 (Nov. 2021), pp. D480–9.

[30] M. Heinzinger *et al.* “Modeling aspects of the language of life through transfer-learning protein sequences”. In: *BMC Bioinformatics* 20.1 (Dec. 2019), p. 723.

[31] L. J. Jensen *et al.* “STRING 8–a global view on proteins and their functional interactions in 630 organisms”. In: *Nucleic Acids Research* 37.Database (Jan. 2009), pp. D412–6.

[32] A. Grover and J. Leskovec. “Node2vec: Scalable Feature Learning for Networks”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 855–864. ISBN: 9781450342322.

[33] R. M. Meyers *et al.* “Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells”. In: *Nature Genetics* 49.12 (Oct. 2017), pp. 1779–84.

[34] J. M. Dempster *et al.* “Extracting Biological Insights from the Project Achilles Genome-Scale CRISPR Screens in Cancer Cell Lines”. In: *bioRxiv* 720243 (July 2019).

[35] J. Barretina *et al.* “The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity”. In: *Nature* 483.7391 (Mar. 2012), pp. 603–7.

[36] M. Ghandi *et al.* “Next-generation characterization of the Cancer Cell Line Encyclopedia”. In: *Nature* 569.7757 (May 2019), pp. 503–8.

[37] E. Cerami *et al.* “The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data: Figure 1.” In: *Cancer Discovery* 2.5 (May 2012), pp. 401–4.

[38] R. Beroukhim *et al.* “Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma”. In: *Proceedings of the National Academy of Sciences* 104.50 (Dec. 2007), pp. 20007–12.

[39] TCGA GDAC. *Analysis-ready standardized TCGA data from Broad GDAC Firehose 2016_01_28 run*. 2016.

[40] J. Lonsdale *et al.* “The genotype-tissue expression (GTEx) project”. In: *Nature Genetics* 45.6 (2013), p. 580.

[41] T. K. Ho. “Random decision forests”. In: *Proceedings of 3rd international conference on document analysis and recognition*. Vol. 1. IEEE. 1995, pp. 278–282.

[42] J. H. Friedman. “Greedy function approximation: A gradient boosting machine.” In: *The Annals of Statistics* 29.5 (Oct. 2001), pp. 1189–232.

[43] A. Liberzon *et al.* “Molecular signatures database (MSigDB) 3.0”. In: *Bioinformatics* 27.12 (May 2011), pp. 1739–40. ISSN: 1367-4803.

[44] D. Teyssonneau *et al.* “Prostate cancer and PARP inhibitors: progress and challenges”. In: *Journal of Hematology & Oncology* 14.1 (Mar. 2021), p. 51.

[45] T. J. Brown and K. A. Reiss. “PARP Inhibitors in Pancreatic Cancer”. In: *The Cancer Journal* 27.6 (Nov. 2021), pp. 465–75.

[46] A. Ashworth and C. J. Lord. “Synthetic lethal therapies for cancer: what’s next after PARP inhibitors?” In: *Nature Reviews Clinical Oncology* 15.9 (June 2018), pp. 564–76.

[47] A. Fisher, C. Rudin, and F. Dominici. *All Models are Wrong, but Many are Useful: Learning a Variable’s Importance by Studying an Entire Class of Prediction Models Simultaneously*. 2019. arXiv: 1801.01489 [stat.ME].

[48] T. D. Martin *et al.* “A Role for Mitochondrial Translation in Promotion of Viability in K-Ras Mutant Cells”. In: *Cell Reports* 20.2 (July 2017), pp. 427–38.

[49] E. Mitry *et al.* “Epidemiology, management and prognosis of colorectal cancer with lung metastases: a 30-year population-based study”. In: *Gut* 59.10 (Aug. 2010), pp. 1383–8.

[50] C. Penna and B. Nordlinger. “Colorectal metastasis (liver and lung)”. In: *Surgical Clinics of North America* 82.5 (Oct. 2002), pp. 1075–90.

[51] P. W. Ingham. “Hedgehog signaling in animal development: paradigms and principles”. In: *Genes & Development* 15.23 (Dec. 2001), pp. 3059–87.

[52] V. Marigo *et al.* “Biochemical evidence that Patched is the Hedgehog receptor”. In: *Nature* 384.6605 (Nov. 1996), pp. 176–9.

[53] Y. Chen and G. Struhl. “Dual Roles for Patched in Sequestering and Transducing Hedgehog”. In: *Cell* 87.3 (Nov. 1996), pp. 553–63.

[54] U. Graab, H. Hahn, and S. Fulda. “Identification of a novel synthetic lethality of combined inhibition of hedgehog and PI3K signaling in rhabdomyosarcoma”. In: *Oncotarget* 6.11 (Mar. 2015), pp. 8722–35.

[55] A. Juvekar *et al.* “Combining a PI3K Inhibitor with a PARP Inhibitor Provides an Effective Therapy for BRCA1-Related Breast Cancer”. In: *Cancer Discovery* 2.11 (Aug. 2012), pp. 1048–63.

[56] X. Wang *et al.* “Widespread genetic epistasis among cancer genes”. In: *Nature Communications* 5.1 (Nov. 2014), p. 4828.

[57] A. Beenken and M. Mohammadi. “The FGF family: biology, pathophysiology and therapy”. In: *Nature Reviews Drug Discovery* 8.3 (Mar. 2009), pp. 235–53.

[58] R. Nusse. “Wnt signaling in disease and in development”. In: *Cell Research* 15.1 (Jan. 2005), pp. 28–32.

[59] S. Patel *et al.* “Wnt Signaling and Its Significance Within the Tumor Microenvironment: Novel Therapeutic Insights”. In: *Frontiers in Immunology* 10 (Dec. 2019), p. 2872.

[60] A. Kaur *et al.* “WNT inhibition creates a BRCA-like state in Wnt-addicted cancer”. In: *EMBO Molecular Medicine* 13.4 (Mar. 2021), e13349.

[61] A. Prakash, S. Doublié, and S. S. Wallace. “Chapter 4 - The Fpg/Nei Family of DNA Glycosylases: Substrates, Structures, and Search for Damage”. In: *Mechanisms of DNA Repair*. Ed. by P. W. Doetsch. Vol. 110. Progress in Molecular Biology and Translational Science. Academic Press, 2012, pp. 71–91.

[62] J. Parsons and M. Edmonds. “The Base Excision Repair Pathway”. In: *Encyclopedia of Cell Biology*. Ed. by R. A. Bradshaw and P. D. Stahl. Waltham: Academic Press, 2016, pp. 442–450. ISBN: 978-0-12-394796-3.

[63] C. Benítez-Buelga *et al.* “Genetic variation in the NEIL2 DNA glycosylase gene is associated with oxidative DNA damage in BRCA2 mutation carriers”. In: *Oncotarget* 8.70 (Nov. 2017), pp. 114626–114636.

[64] A. Osorio *et al.* “DNA Glycosylases Involved in Base Excision Repair May Be Associated with Cancer Risk in BRCA1 and BRCA2 Mutation Carriers”. In: *PLoS Genetics* 10.4 (Apr. 2014). Ed. by M. S. Horwitz, e1004256.

2.5. Supplementary Materials

2.5.1. Other Synthetic Lethality Prediction Methods

PCA-gCMF. The PCA-gCMF method is a version of collective matrix factorization (CMF) proposed for synthetic lethality (SL) prediction [1]. It first uses principal component analysis (PCA) to reduce the dimensionality of rows and columns across multiple matrices, including a (genes × genes) matrix of SL interactions and other matrices containing different types of molecular data. After applying PCA, the method uses group-sparse CMF to decompose these multiple matrices and make SL predictions. Although different versions of the method were originally proposed (CMF, PCA-CMF, gCMF, and PCA-gCMF), PCA-gCMF was the one that showed the best performance and was therefore selected for comparison with ELISL models. We applied the PCA-gCMF method using the following matrices: a matrix of SL interaction labels (genes × genes); a pairwise gene co-dependency matrix (genes × genes), containing significance p-values for the change in dependency score of one gene in cell lines with vs. without a mutation in the other gene in each pair (Wilcoxon rank-sum test), using CCLE data; a gene expression matrix (genes × samples), containing expression values measured across the available TCGA patient tumour samples; a co-expression matrix (genes × genes), containing the Spearman's correlation coefficient between the expression of each pair of genes across the TCGA patient tumour samples; and a CNV profile matrix (genes × samples), containing continuous copy number values for the TCGA patient tumour samples. We used the hyperparameter values suggested in the PCA-gCMF paper.

GRSMF. Graph regularized self-representative matrix factorization (GRSMF) is another SL prediction method based on matrix factorization [2]. The GRSMF approach learns a self-representation from a matrix of pairwise SL interaction labels, regularized by a matrix of pairwise functional similarities between genes based on Gene Ontology (GO) annotations. We constructed the similarity matrix as described in the original work, using annotations from the biological process ontology. We used the hyperparameter values suggested in the GRSMF paper.

GCATSL. The GCATSL method creates a graph of known SL interactions, as well as graphs for other types of functional similarity relationships between genes (additional data modalities), and uses these to predict new SL interactions [3]. The approach learns node representations for local and global neighbours in each data modality using graph attention networks, then aggregates local and global representations to obtain modality-level features using multilayer perceptrons, and finally optimizes the weights of the different modalities in a regularized linear model to reconstruct the matrix of SL interactions. The prediction probabilities obtained for unknown gene pairs are used for prediction. Originally, the authors used three different functional similarity matrices: two based on GO annotations (biological process and cellular component), and one based on protein-protein interactions (PPIs) from the BioGRID database [4]. To ensure a fairer comparison, we applied GCATSL using the same

manually curated or experimentally validated PPIs from the STRING [5] database that we also used with ELISL models. We set the hyperparameter values as suggested in the GCATSL paper.

SBSL. The SBSL framework uses conventional supervised machine learning algorithms to learn SL prediction models based on a collection of 27 features [6]. Four SBSL models were originally proposed: two linear models using regularized logistic regression (L0L2 [7] and Elastic Net [8]), and two non-linear models using regularized random forests (Regularized Random Forest (RRF) [9] and Multivariate methods with Unbiased Variable (MUVR) [10]). The 27 features used by SBSL models are mainly context-specific and derived from different types of molecular profiles for cancer cell lines, healthy donor tissues, and patient tissue samples. Data modalities include, for instance: mutation and copy-number data, gene expression, gene dependency scores, and patient survival data. For the SBSL methods, feature calculation and hyperparameter optimization were performed as described in the original paper.

2.5.2. DATA AND CODE

DATA SOURCES

Tissue data

Tumour patient tissue omics and clinical data (TCGA):

TCGA combined study containing samples from 8 studies: cBioPortal - TCGA Firehose.

Healthy donor tissue data (GTEx):

GTEx gene expression: GTEx Portal - Gene TPMs (v8).

GTEx sample annotation: GTEx Portal - dbGaP de-identified open access version (v8).

Cell line data

Cell line omics: CCLE Broad Institute & Novartis 2019.

CRISPR dependency scores: CCLE Broad Institute & Novartis 2019.

PPI data and embeddings

Protein-protein interaction data: STRING (v11).

PPI node embedding tool: Node2Vec.

Protein sequence data and embeddings

Human proteins with reviewed amino acid sequence data: UniProt.

Protein sequence embedding tool: SeqVec.

Pathway gene sets used to create unknown pairs for SL prediction
Names and numbers of genes in cancer and DNA repair pathway gene sets used to generate gene pairs with unknown SL status for prediction of promising SL pairs:

- KEGG PATHWAYS IN CANCER (325)
- KEGG BASE EXCISION REPAIR (35)
- REACTOME BASE EXCISION REPAIR (91)
- WP NUCLEOTIDE EXCISION REPAIR (44)
- KEGG NUCLEOTIDE EXCISION REPAIR (44)
- REACTOME NUCLEOTIDE EXCISION REPAIR (110)
- KEGG MISMATCH REPAIR (23)
- REACTOME MISMATCH REPAIR (15)
- WP DNA MISMATCH REPAIR (23)
- WP HOMOLOGOUS RECOMBINATION (13)
- KEGG HOMOLOGOUS RECOMBINATION (28)
- KEGG NON HOMOLOGOUS END JOINING (13)
- PID FANCONI PATHWAY (47)

CODE AND LIBRARIES

The code for the different experiments was written and integrated with Python 3.6. Only PCA-gCMF [1] and the SBSL methods [6] were run in R. We used *LightGBM* [11] and *scikit-learn* [12] together for the regularized random forest and regularized gradient boosting decision tree models. We optimized the models using bayesian optimization with gaussian process from the *scikit-optimize* [13] package. For plotting, we made use of the *seaborn* [14] and *matplotlib* [15] libraries. Additionally, we used the *lifelines* [16] package for the Kaplan-Meier plots and survival tests.

2.5.3. HYPERPARAMETER SETTINGS FOR ELISL MODELS

Here we report the settings and default values used for hyperparameter optimization of ELISL models with cross-validation on the train set.

- Number of leaves: 165
- Max depth: [10, 15, 20, . . . , 100, 105, 110, Inf]
- Learning rate: 0.1
- No of estimators: [100, 110, 120, . . . , 1180, 1190, 1200]
- Subsample for bin: 200000
- Minimum split gain: 0
- Minimum child weight: 5
- Minimum child samples: {1, 2, 4, 10}
- Subsample: {0.632, 0.8, 0.99}
- Subsample frequency: 1
- Colsample by tree: {0.5, 0.8, 1}
- Alpha regularization: 0
- Lambda regularization: {5, 10}

2.5.4. IMPACT OF SEQUENCE EMBEDDING DIMENSION

The high-dimensionality of the sequence feature embeddings used by ELISL models (1024) could lead to overfitting [17]. To investigate the impact of sequence embedding dimension, we retrained the single-cancer ELISL-RF models using different sequence embedding sizes and re-evaluated the performance. To reduce the dimension, we first applied PCA to the matrix of protein sequence embeddings, and then regenerated the sequence feature vector for each gene pair as the absolute difference between the embedding vectors of the proteins encoded by the two genes in the pair in the new PCA-transformed feature space. We evaluated the following embedding sizes: 32, 64, 128, 256, 512, and 1024. Except for OV (ovarian cancer), the changes in performance were small and within the standard deviation of the original experiment (Supplementary Figure 2.S3b). For OV, the performance remained similar using embedding sizes of 1024 (average 0.87 AUPRC), 512 and 256, and modestly dropped using smaller embedding sizes of 128, 64, and 32 (average 0.8 AUPRC). These results show that the high-dimensionality of sequence embeddings did not play a major role in the contribution of sequence data towards the performance of ELISL models.

2.5.5. Supplementary Figures

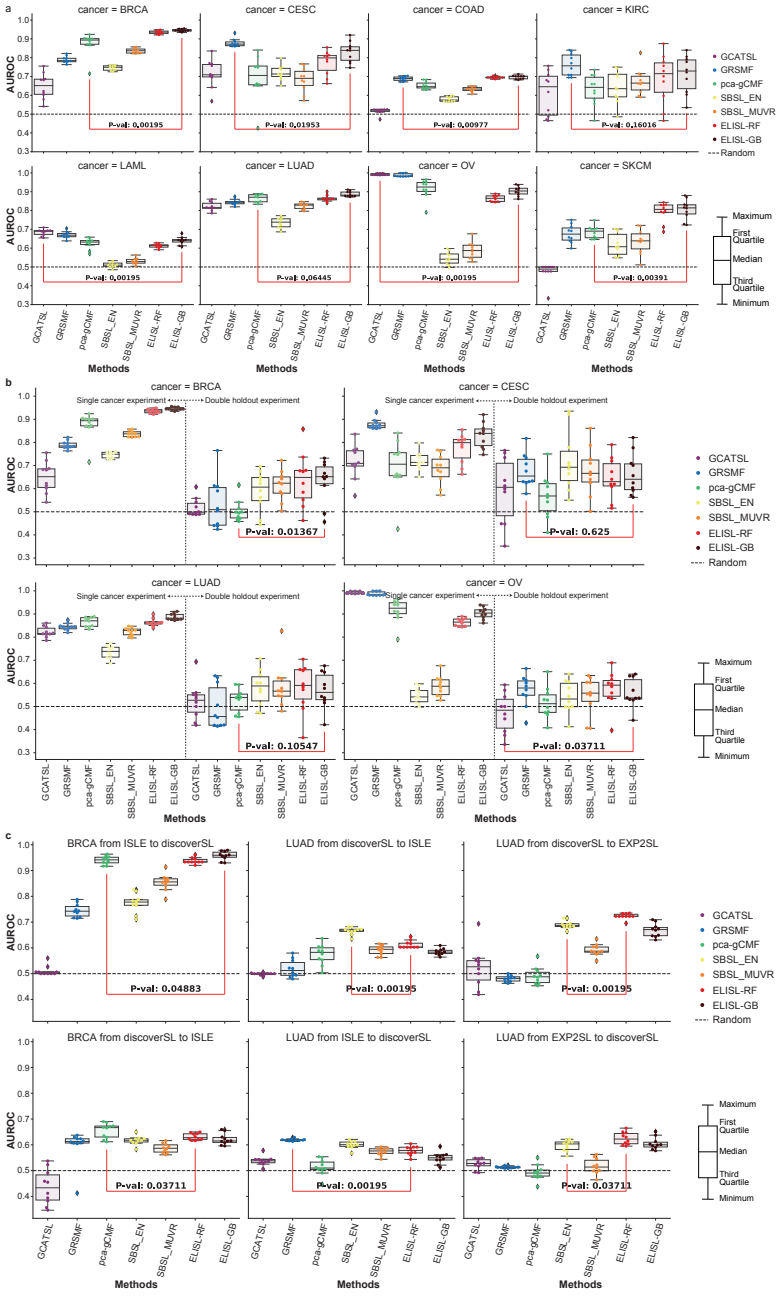


Figure 2.S1: **Within cancer prediction with similar and distinct bias between train and test sets.** Per cancer AUROC performance of cancer-specific SL prediction models on unseen gene pairs from the same cancer (10 different train/test splits), under three scenarios: **(a)** conventional non-overlapping-pair train/test sets, thus allowed to follow similar selection bias; **(b)** double gene holdout to induce distinct selection biases between train and test sets, with left side reporting the original performance (similar bias) and the right side using double gene holdout (distinct bias); and **(c)** cross-SL dataset prediction under inherently occurring differences in selection bias between three sources of SL data (ISLE, DiscoverSL, EXP2SL), with models trained using labels from one SL dataset and evaluated on another SL dataset considering the combinations of cancer type (BRCA, LUAD) and SL dataset with sufficient numbers of samples. Methods: matrix factorization and graph-based (GCATSL, GRSMF, pca-gCMF); supervised learning, including existing models (SBSL-EN/MUVR), and proposed ELISL models (ELISL-RF/GB). Boxplots: boxes indicate the range between lower (first) and upper (third) quartiles, or interquartile range (IQR), with a horizontal line across the box denoting the median; whiskers extend from the box to the largest (or smallest) value within 1.5 times the IQR of the upper (lower) quartile, and points beyond the whiskers are outliers. Red lines compare the best ELISL model with the best among the other models in single cancer experiment using a Wilcoxon signed rank test.

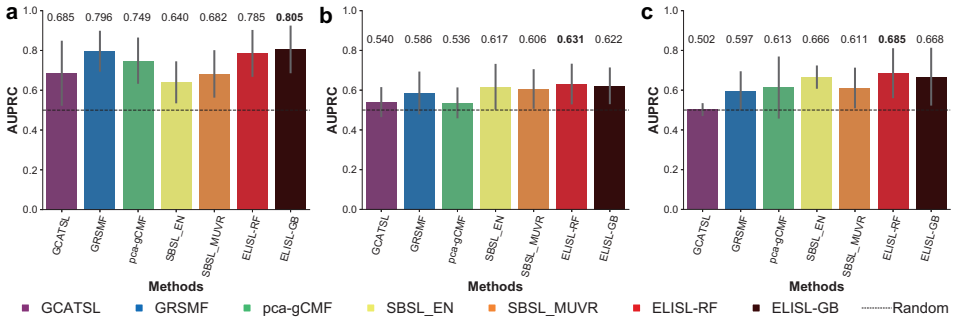


Figure 2.S2: **Aggregated results of within cancer prediction with similar and distinct bias between train and test sets.** Average AUPRC performance of cancer-specific SL prediction models in an aggregated style over cancer types on unseen gene pairs from the same cancer (10 different train/test splits), under three scenarios: **(a)** conventional non-overlapping-pair train/test sets, thus allowed to follow similar selection bias; **(b)** double gene holdout to induce distinct selection biases between train and test sets; and **(c)** cross-SL dataset prediction under inherently occurring differences in selection bias between three sources of SL data (ISLE, DiscoverSL, EXP2SL), with models trained using labels from one SL dataset and evaluated on another SL dataset considering the combinations of cancer type (BRCA, LUAD) and SL dataset with sufficient numbers of samples. Methods: matrix factorization and graph-based (GCATSL, GRSMF, pca-gCMF); supervised learning, including existing models (SBSL-EN/MUVR), and proposed ELISL models (ELISL-RF/GB). The vertical black lines refer to the standard deviation of AUPRC performances.

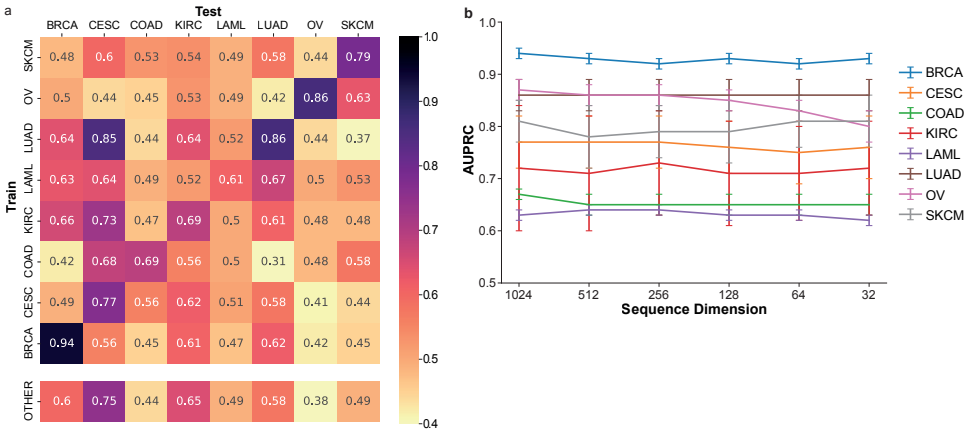


Figure 2.S3: **ELISL-RF SL prediction within/across cancer types and impact of sequence embedding dimension. (a)** Performance of cancer-specific models and pan-cancer models, measured as average AUPRC over 10 runs using undersampled 80/20 train/test splits. For cancer-specific models, presented in a matrix, the diagonal reports prediction performance within the same cancer type, and the remaining cells show performance for prediction on other cancer types. Pan-cancer model performances are reported in a separate row at the bottom, where models are trained on all other cancer types except the one the model is supposed to predict on. Rows denote the cancer type used for training, columns indicate the cancer type used for prediction and evaluation. **(b)** Within-cancer prediction performance (AUPRC) of cancer-specific ELISL-RF models as the dimension of the sequence embedding is gradually reduced from 1024 to 32. The horizontal lines connect the average AUPRC performance for different embedding dimensions over 10 runs using independently drawn train/test set splits. The vertical lines for each embedding dimension indicate the standard deviation over the 10 runs.

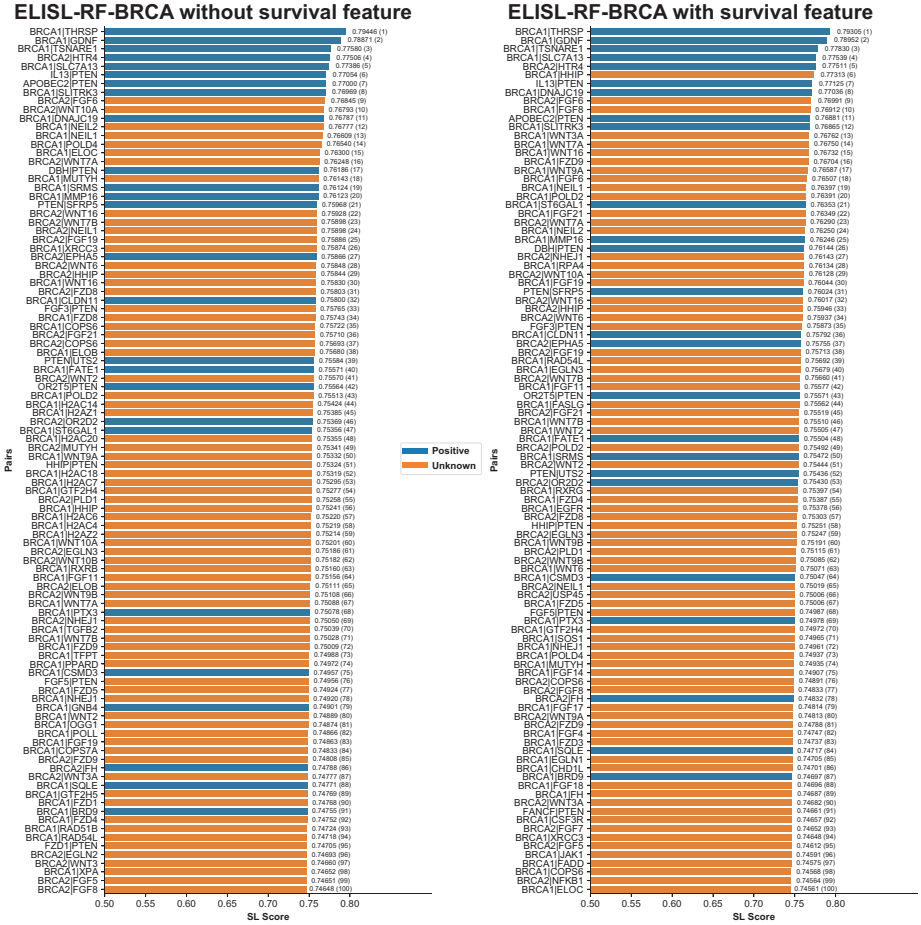
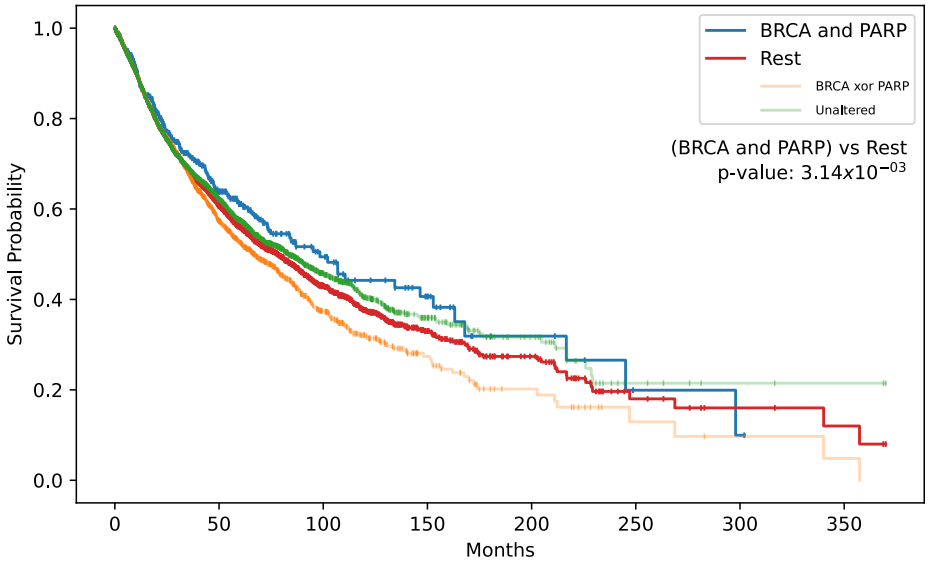


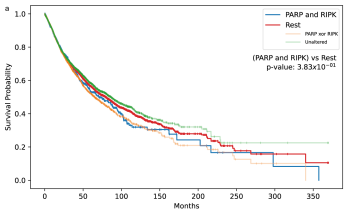
Figure 2.S4: **Top predictions of ELISL-RF BRCA model without and with the survival feature.** Prediction scores of ELISL-RF without the survival feature (left) and with the survival feature (right) for the top 100 pairs in the BRCA test set and the unknown set.



	# of Cases	# of Events	Median Survival Time
BRCA and PARP	501	160	98.4
Rest	10318	3365	78.21
BRCA xor PARP	3092	1102	66.67
Unaltered	7226	2263	83.31

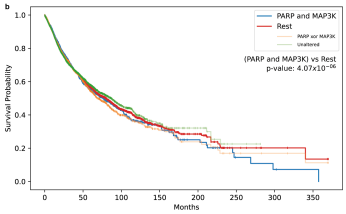
	coef	exp(coef)	z	p
Diagnosis Age	0.01	1.01	15.02	$5.05x10^{-51}$
Cancer Type	-0.01	0.99	-5.03	$5.01x10^{-07}$
Mutation	0.15	1.16	2.95	$3.14x10^{-03}$
Gender	0.09	1.09	4.31	$1.66x10^{-05}$

Figure 2.S5: **Survival of patients with *vs.* without simultaneous alterations in BRCA and PARP genes.** Survival analysis between groups of patients with and without simultaneous alterations in both genes of the BRCA (1-2) and PARP (1 to 16) gene families. The plot shows Kaplan-Meier survival curves for the group with simultaneous alterations ("BRCA and PARP", blue) and for the group without simultaneous alterations (all other patients, "Rest", red), where the latter is further divided into two subgroups: patients with alterations on only one of the two genes (orange), and patients with both genes unaltered (green). The survival p-value included in the plot is based on a Wald significance test of the co-mutation status variable in a Cox proportional hazards (PH) model of survival time considering co-mutation status ("BRCA and PARP" or blue vs. "Rest" or red groups) and adjusted for age, sex, and cancer type. Detailed analysis of the Cox PH model is shown in the tables below the KM plot. The top table shows the number of patients as "# of Cases", number of deaths as "# of Events", and median survival time for the groups of patients described above. The bottom table shows, for each of the four variables of the Cox PH model, the coefficient ($coef$) and hazard ratio ($exp(coef)$) of the variable in the model, as well as its effect size and significance based on a Wald test.



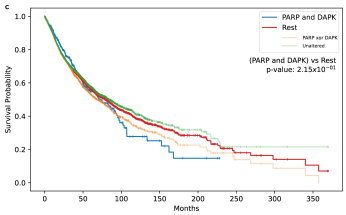
	# of Cases	# of Events	Median Survival Time
PARP and RIPK	836	296	70.82
Rest	9983	3229	79.46
PARP xor RIPK	2847	1025	65.0
Unaltered	7136	2204	86.27

	coef	exp(coef)	z	p
Diagnosis Age	0.01	1.01	14.91	$2.76x10^{-50}$
Cancer Type	-0.01	0.99	-5.16	$2.41x10^{-07}$
Mutation	0.03	1.03	0.87	$3.83x10^{-01}$
Gender	0.09	1.09	4.34	$1.43x10^{-05}$



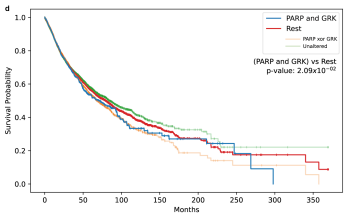
	# of Cases	# of Events	Median Survival Time
PARP and MAP3K	1823	645	76.24
Rest	8996	2880	79.53
PARP xor MAP3K	3422	1130	69.07
Unaltered	5574	1750	86.76

	coef	exp(coef)	z	p
Diagnosis Age	0.01	1.01	15.16	$6.63x10^{-52}$
Cancer Type	-0.01	0.99	-4.95	$7.41x10^{-07}$
Mutation	0.12	1.13	4.61	$4.07x10^{-06}$
Gender	0.09	1.09	4.30	$1.67x10^{-05}$



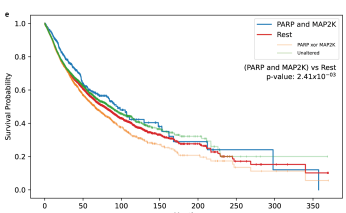
	# of Cases	# of Events	Median Survival Time
PARP and DAPK	416	147	71.01
Rest	10403	3378	79.07
PARP xor DAPK	3143	1116	68.48
Unaltered	7260	2262	83.31

	coef	exp(coef)	z	p
Diagnosis Age	0.01	1.01	14.94	$1.92x10^{-50}$
Cancer Type	-0.01	0.99	-5.05	$4.38x10^{-07}$
Mutation	0.06	1.06	1.24	$2.15x10^{-01}$
Gender	0.09	1.09	4.34	$1.40x10^{-05}$



	# of Cases	# of Events	Median Survival Time
PARP and GRK	787	292	70.16
Rest	10032	3233	79.53
PARP xor GRK	3017	1054	68.15
Unaltered	7015	2179	84.49

	coef	exp(coef)	z	p
Diagnosis Age	0.01	1.01	14.98	$9.45x10^{-51}$
Cancer Type	-0.01	0.99	-5.04	$4.75x10^{-07}$
Mutation	0.09	1.09	2.31	$2.09x10^{-02}$
Gender	0.09	1.09	4.33	$1.51x10^{-05}$



	# of Cases	# of Events	Median Survival Time
PARP and MAP2K	576	174	95.14
Rest	10243	3351	77.65
PARP xor MAP2K	3218	1167	65.46
Unaltered	7025	2184	85.51

	coef	exp(coef)	z	p
Diagnosis Age	0.01	1.01	15.02	$5.45x10^{-51}$
Cancer Type	-0.01	0.99	-4.98	$6.37x10^{-07}$
Mutation	0.13	1.14	3.03	$2.41x10^{-03}$
Gender	0.09	1.09	4.29	$1.79x10^{-05}$

Figure 2.S6: **Survival analysis of patients with vs. without simultaneous alterations in the families of non-SL predicted gene pairs.** Survival of groups of patients with and without simultaneous alterations in genes of both gene families, for the four gene pairs that ranked the lowest in ELISL-RF predictions: **(a)** RIPK (RIPK1-4, DSTYK) and PARP1-16 gene families; **(b)** MAP3K (MAP3K1-15, TAOK1-2, RAF1, BRAF, ARAF, MAP3K20) and PARP1-16 gene families; **(c)** DAPK (DAPK1-3, STK17A, STK17B) and PARP1-16 gene families; **(d)** MAP2K1-7 and PARP1-16 gene families. For each survival analysis **(a)-(d)**, the figure shows a Kaplan-Meier (KM) plot (left) and survival analysis tables (right). The plot shows KM survival curves for the groups with simultaneous alterations (blue) and for all other patients (red), with the latter further split into two subgroups: patients with alterations on only one of the families (orange), and patients with both families unaltered (green). The survival p-value is based on a Wald test of co-mutation status in a Cox proportional hazards (CoxPH) model considering co-mutation status (blue vs. red groups) and adjusted for age, sex, and cancer type. A CoxPH model summary is shown in the tables besides the KM plot. The top table shows number of patients as "# of Cases", number of deaths as "# of Events", and median survival time for the patient groups. The bottom table shows, for each CoxPH variable: coefficient ($coef$) and hazard ratio ($exp(coef)$), as well as effect size and p-value based on a Wald test.

2.5.6. SUPPLEMENTARY TABLES

TABLES S1-S2 - DETAILED SL LABELS

Table 2.S1: Details of the experimental SL screens included in each SL label dataset (ISLE, EXP2SL, LU, and dSL). For each experimental SL screen (row of the table), the table shows: a short name used internally by us to identify the screen ("Screen" column), cross symbols "x" identifying the specific SL label datasets including the gene pairs of the given screen ("ISLE", "EXP2SL", "LU", and "dSL" columns), the cancer type of the cell lines used in the screen or the targeted gene ("Cancer or target" column), the reference to the study ("PMID or DOI" column), and the type of the experiment ("Type" column). Double-gene-knockout (gRNA): DGKO, double-gene-knockdown (siRNA or shRNA): DGKD, single-gene-knockout (gRNA): SGKO or single-gene-knockdown (siRNA or shRNA): SGKD, chemical inhibitor: CI, PARP inhibitor: PARPi. SGKO and SGKD are either applied to a cell line with an existing mutation in a specific gene or used with an inhibitor such as CI, PARPi, or a drug to cause aberration in another gene so that simultaneous mutation can be simulated.

Screen	ISLE	EXP2SL	LU	dSL	Cancer or target	PMID or DOI	Type
Zhao		x			CESC, LUAD	29452643	11,475 DGKO 459 SGKO
Big Papi		x			RCC, SKCM, LUAD, COAD, OV	29251726	DGKO
Han	x				LAML	28319085	DGKO
Shen	x	x			CESC, LUAD, KIRC	28319113	23,652 DGKO 657 SGKO
ISLE1	x				LAML	28162770	SGKO
ISLE2	x				CESC	27453043	SGKO Drug
ISLE4	x				OV	26637171	SGKD Drug
ISLE5	x				CESC	26437225	SGKD CI
ISLE6	x				BRCA	25407795	DGKD
ISLE7	x		x		COAD	24104479	SGKD
ISLE8	x				SKCM	22623531	SGKD
ISLE9	x			x	KRAS gene	22613949	SGKD
ISLE10	x			x	KRAS gene	19490893	SGKD
ISLE11	x				CESC	20049736	CI
ISLE12	x			x	BRCA	18388863	SGKD PARPi
ISLE13	x			x	BRCA	18832051	SGKD PARPi
ISLE14	x			x	KIRC	18948595	SGKD
ISLE15	x				LUAD	17429401	SGKD
LU1			x		COAD	23563794	DGKD
dSL1				x	BRCA, OV, PDAC	22585861	SGKD
dSL2				x	HLRCC	24568598	2 SGKDs
dSL3				x	BRCA, PDAC, OV and UTE	26427375	SGKD
dSL4				x	All	10.1146/annurev-acancerbio-042016-073434	Curated

Table 2.S2: Number of labelled gene pairs available for each cancer type. Number of positive (synthetic lethal, "+") and negative (non-synthetic lethal, "−") gene pairs per cancer type (rows) and SL label dataset before removing duplicates between these SL label datasets (columns "Exp2SL", "Lu15", "ISLE", "dSL"), the total number of labelled gene pairs for each cancer type after removing disagreeing duplicates and combining the SL datasets. Duplicates columns: "Agree", contains the number of duplicates with the same label across datasets, "Disagree", contains the number of duplicates with different labels across the datasets.

Cancer	**Exp2SL**		**Lu15**		**ISLE**		**dSL**		**Total**		**Duplicates**	
	+	**−**	**+**	**−**	**+**	**−**	**+**	**−**	**+**	**−**	**Agree**	**Disagree**
BRCA	0	0	0	0	590	1012	885	75	1444	1037	53	14
CESC	0	0	0	0	145	4762	0	0	145	4762	0	0
COAD	18	155	231	5621	2100	74244	0	0	1728	79323	350	484
KIRC	0	0	0	0	60	2514	0	0	60	2514	0	0
LAML	0	0	0	0	1191	19308	0	0	1191	19308	0	0
LUAD	307	2369	0	0	169	4735	372	339	597	5515	1695	242
OV	0	0	0	0	255	554	0	0	255	554	0	0
SKCM	18	72	0	0	89	18630	0	0	107	18702	0	0

TABLES S3 - RUN TIME OF ELISL-RF MODEL

Table 2.S3: Average run time (in seconds) of ELISL-RF for the single cancer experiment, per cancer type over 10 runs. It does not include the feature generation, only hyperparameter tuning and training of the final model.

Runtime (sec)	**BRCA**	**CESC**	**COAD**	**KIRC**	**LAML**	**LUAD**	**OV**	**SKCM**
Grid-search	2981.8	756.8	6257.0	419.8	4962.4	2159.8	689.1	384.4
Final Training	139.0	20.9	296.0	19.2	242.9	92.3	37.9	20.0
Total	3120.8	777.7	6553.0	439.0	5205.3	2252.2	727.1	404.4

TABLES S4-S5 - NUMBER OF SAMPLES USED IN SL PREDICTION EXPERIMENTS

Table 2.S4: Number of samples used in single cancer and double holdout experiments, average and standard deviation over 10 runs using independently drawn train/test set splits.

Cancer	**Single Cancer Experiment**		**Double-Holdout Experiment**	
	Training Samples	**Testing Samples**	**Training Samples**	**Testing Samples**
BRCA	1658.0 ± 0.0	416.0 ± 0.0	808.8 ± 113.3	239.8 ± 75.4
CESC	232.0 ± 0.0	58.0 ± 0.0	116.0 ± 18.4	36.0 ± 12.8
COAD	2764.0 ± 0.0	692.0 ± 0.0	1571.4 ± 185.1	299.8 ± 147.3
KIRC	96.0 ± 0.0	24.0 ± 0.0	56.0 ± 12.1	12.0 ± 4.8
LAML	1906.0 ± 0.0	476.0 ± 0.0	1048.6 ± 75.5	255.6 ± 41.4
LUAD	956.0 ± 0.0	238.0 ± 0.0	449.8 ± 241.7	147.2 ± 95.5
OV	408.0 ± 0.0	102.0 ± 0.0	215.0 ± 15.4	60.2 ± 9.5
SKCM	172.0 ± 0.0	42.0 ± 0.0	61.8 ± 35.8	21.0 ± 13.9

TABLES S6 - TOP TEN ELISL-RF SL GENE PAIR PREDICTIONS IN THE TEST SET

Table 2.S5: Number of samples used in cross-dataset experiment, average and standard deviation over 10 runs using independently drawn train/test set splits.

Cancer	From	To	Training Samples	Testing Samples
BRCA	ISLE	DiscoverSL	1146.0 ± 0.0	150.0 ± 0.0
	DiscoverSL	ISLE	78.0 ± 0.0	1180.0 ± 0.0
LUAD	ISLE	DiscoverSL	338.0 ± 0.0	678.0 ± 0.0
	DiscoverSL	ISLE	678.0 ± 0.0	338.0 ± 0.0
	DiscoverSL	Exp2SL	678.0 ± 0.0	614.0 ± 0.0
	Exp2SL	DiscoverSL	614.0 ± 0.0	678.0 ± 0.0

Table 2.S6: Top 10 scored gene pairs using cancer-specific ELISL-RF models for BRCA, LUAD, and OV. Gene pairs were ranked based on the average ELISL-RF SL prediction score over 10 runs. Column "Score" contains the average SL prediction score. Column "Label" contains the known synthetic lethality status, where "+" denotes synthetic lethal and "−" denotes non-synthetic lethal.

Gene Pair	Score	Label	Gene Pair	Score	Label
BRCA			**LUAD**		
BRCA1|THRSP	0.793051	+	KRAS|MRPL28	0.797100	−
BRCA1|GDNF	0.789517	+	KRAS|PARP12	0.790102	+
BRCA1|TSNARE1	0.778300	+	KRAS|LSM5	0.789647	+
BRCA1|SLC7A13	0.775389	+	KRAS|POLR2G	0.786567	+
BRCA2|HTR4	0.775112	+	KRAS|TEAD2	0.783438	+
IL13|PTEN	0.771248	+	KRAS|POLL	0.782135	−
BRCA1|DNAJC19	0.770358	+	KRAS|MTA2	0.779881	+
APOBEC2|PTEN	0.768812	+	KRAS|SERPINI1	0.778718	+
BRCA1|SLITRK3	0.768645	+	KRAS|NR1D2	0.777780	−
BRCA1|ST6GAL1	0.763534	+	KRAS|OSM	0.771441	−
OV					
EPB41L1|YES1	0.640591	+			
ABL1|SRC	0.635870	+			
FYN|YES1	0.629252	+			
ABL1|YES1	0.629216	+			
GAB1|YES1	0.623906	+			
FYN|NEDD9	0.617254	+			
ABL1|BCAR3	0.616596	+			
ABL1|LCK	0.614998	+			
ABL2|EPB41L1	0.604245	+			
LCK|PLCG2	0.596859	+			

TABLE S7 - SURVIVAL ANALYSIS FOR BRCA-HH, BRCA-FGF, BRCA-WNT, AND BRCA-NEIL

Table 2.S7: **Survival tables and CoxPH models for the gene families of promising unknown SL pairs predicted by ELISL-RF for breast cancer (BRCA-HH, BRCA-FGF, BRCA-WNT, BRCA-NEIL).** For each of the three gene pairs, one survival table (left) and one CoxPH model table (right) are provided. ***Survival tables (left):*** show the number of patients as "Cases", the number of deaths as "Events", and the median survival time in months for different groups of patients. The group "GeneFam1 and GeneFam2" includes patients with simultaneous alterations on genes of both gene families GeneFam1 and GeneFam2 (see paper for the definition of alteration). The group "Rest" includes all the other patients, that is, those that do not have simultaneous alterations in genes from both families. This latter group is further divided into two groups: the "GeneFam1 xor GeneFam2" group, containing patients with alterations on either gene family but not both; and the group "Unaltered", containing patients without alterations on any genes of the two families. Note that "Median survival time" denotes the time point at which the probability of survival for the group of patients is 0.5, meaning that half of the patients in that group are expected to be alive. ***CoxPH models and significance tests (right):*** show details of the CoxPH functions to model the association between survival time with and without simultaneous alterations in the two gene families ("Co-Mutation"), adjusted for age, cancer type, and sex. The "Co-Mutation" status variable is defined based on the two groups of interest: "GeneFam1 and GeneFam2" as Co-Mutation= 0 and "Rest" or "~(GeneFam1 and GeneFam2)" as Co-Mutation= 1, also highlighted in bold in the corresponding Survival Tables. For each of the four variables in the CoxPH model (age, cancer type, co-mutation, and sex), the table includes the corresponding coefficient ($coef$) and hazard ration ($exp(coef)$) of the variable according to the model. Additionally, the table shows the effect of each variable in the model as z, together with its statistical significance as p, determined using a Wald test. Values $coef > 0.0$ (or $exp(coef) > 1.0$) indicate longer survival time for the patient group with simultaneous alterations in the genes of both families, "GeneFam1 and GeneFam2".

Survival Tables				**CoxPH Functions**				
Group	**Cases**	**Events**	**Median survival time**	**Variable**	$coef$	$exp(coef)$	z	p
BRCA and HH	152	34	Inf	Age	0.01	1.01	14.91	2.76e-50
BRCA xor HH	1256	437	68.66	Cancer type	-0.01	0.99	-4.96	7.19e-07
Unaltered	9411	3054	78.44	**Co-Mutation**	0.21	1.24	2.65	8.04e-03
Rest	10667	3491	77.65	Sex	0.09	1.09	4.27	1.97e-05
BRCA and FGF	451	142	102.1	Age	0.01	1.01	14.98	9.89e-51
BRCA xor FGF	3038	1050	70.13	Cancer type	-0.01	0.99	-5.12	3.09e-07
Unaltered	7330	2333	81.2	**Co-Mutation**	0.12	1.13	2.42	1.55e-02
Rest	10368	3383	78.44	Sex	0.09	1.09	4.35	1.39e-05
BRCA and WNT	161	41	167.9	Age	0.01	1.01	14.93	2.13e-50
BRCA xor WNT	1291	419	81.73	Cancer type	-0.01	0.99	-5.01	5.55e-07
Unaltered	9366	3065	77.19	**Co-Mutation**	0.23	1.26	2.88	3.96e-03
Rest	10658	3483	78.21	Sex	0.09	1.09	4.32	1.56e-05
BRCA and NEIL	122	35	102.1	Age	0.01	1.01	14.94	1.75e-50
BRCA xor WNT	1358	458	78.97	Cancer type	-0.01	0.99	-5.07	4.03e-07
Unaltered	9339	3032	78.18	**Co-Mutation**	0.17	1.18	1.66	9.63e-02
Rest	10697	3490	78.21	Sex	0.09	1.09	4.36	1.30e-05

REFERENCES

[1] H. Liany, A. Jeyasekharan, and V. Rajan. “Predicting synthetic lethal interactions using heterogeneous data sources”. In: *Bioinformatics* 36.7 (Nov. 2019), pp. 2209–16.

[2] J. Huang *et al.* “Predicting synthetic lethal interactions in human cancers using graph regularized self-representative matrix factorization”. In: *BMC Bioinformatics* 20.S19 (Dec. 2019), p. 657.

[3] Y. Long *et al.* “Graph contextualized attention network for predicting synthetic lethality in human cancers”. In: *Bioinformatics* 16 (Feb. 2021), pp. 2432–40.

[4] R. Oughtred *et al.* “The BioGRID interaction database: 2019 update”. In: *Nucleic Acids Research* 47.D1 (Nov. 2018), pp. D529–41.

[5] L. J. Jensen *et al.* “STRING 8–a global view on proteins and their functional interactions in 630 organisms”. In: *Nucleic Acids Research* 37.Database (Jan. 2009), pp. D412–6.

[6] C. Seale, Y. Tepeli, and J. P. Gonçalves. “Overcoming selection bias in synthetic lethality prediction”. In: *Bioinformatics* 38.18 (July 2022), pp. 4360–8.

[7] H. Hazimeh and R. Mazumder. *Fast Best Subset Selection: Coordinate Descent and Local Combinatorial Optimization Algorithms*. 2018. arXiv: `1803.01454 [stat.CO]`.

[8] J. Friedman, T. Hastie, and R. Tibshirani. “Regularization Paths for Generalized Linear Models via Coordinate Descent”. In: *Journal of Statistical Software* 33.1 (2010), pp. 1–22.

[9] H. Deng and G. Runger. “Feature selection via regularized trees”. In: *The 2012 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2012, pp. 1–8.

[10] L. Shi *et al.* “Variable selection and validation in multivariate modelling”. In: *Bioinformatics* 35.6 (2019), pp. 972–80.

[11] G. Ke *et al.* “Lightgbm: A highly efficient gradient boosting decision tree”. In: *Advances in neural information processing systems* 30 (2017), pp. 3146–54.

[12] F. Pedregosa *et al.* “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–30.

[13] T. Head *et al. scikit-optimize/scikit-optimize*. Version v0.8.1. Sept. 2020.

[14] M. L. Waskom. “seaborn: statistical data visualization”. In: *Journal of Open Source Software* 6.60 (2021), p. 3021.

[15] J. D. Hunter. “Matplotlib: A 2D graphics environment”. In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–5.

[16] C. Davidson-Pilon. “lifelines: survival analysis in Python”. In: *Journal of Open Source Software* 4.40 (2019), p. 1317.

[17] R. Clarke *et al.* “The properties of high-dimensional data spaces: implications for exploring gene and protein expression data”. In: *Nature Reviews Cancer* 8.1 (Jan. 2008), pp. 37–49.

3

ONCOSTRATIFIER: STRATIFYING ONCOGENE-ADDICTED COHORTS BY DRUG RESPONSE

Yasin I. TEPELI

Lucia TRASTULLA

Joana P. GONÇALVES

Francesco IORIO

Further supplementary material is available online at:
https://doi.org/10.4121/c269140b-9d5f-411b-bafb-72ab8dba45da.v1

Oncogenes, when mutated or overexpressed, drive tumorigenesis and can lead to oncogene addiction, where cancer cells rely on such genes for survival and proliferation. Stratifying oncogene-addicted cohorts is essential to uncover alternative therapeutic avenues, for instance in cases of untargetable oncogenes or resistance to treatment. We propose Oncostratifier, a framework to identify drugs that specifically target oncogene-addicted cancer cohorts, differentiating between drugs that induce sensitivity or resistance. Our results reveal 21,020 stratifying drugs spanning 267 oncogenes in cell lines of 31 different cancer types. We identify 59 mutational markers associated with 36 of the stratifying drugs, which can possibly be used to stratify patient tumors in the absence of drug response data. These findings reveal the potential of Oncostratifier as a tool to generate candidate hypotheses for precision cancer treatment strategies. Explore our detailed results at `https://edu.nl/6hnct`.

3.1. INTRODUCTION

Cancer is characterized by a complex interplay of genetic and environmental factors that drive uncontrolled cell proliferation and enable cells to escape mechanisms designed to control their survival [1–3]. Central to this process are oncogenes, that is, genes whose alteration via mutations or overexpression contributes to tumorigenesis [4]. Well-known oncogenes such as KRAS [5], EGFR [6], and MYC [7] play pivotal roles in regulating cell growth, division, and differentiation. As a result, tumor-driving alterations can also lead to oncogene addiction, where cancer cells become reliant on the activity of the oncogene for their survival and proliferation [4]. Oncogene addiction provides an opportunity for therapeutic targeting using drugs that inhibit the function of the oncogene to selectively promote the death of oncogene-addicted cancer cells. For example, inhibitors targeting BCR-ABL in chronic myeloid leukemia (CML) [8] or EGFR inhibitors in non-small cell lung cancer (NSCLC) [9] have shown significant clinical success. However, not all oncogene-addicted cancers can be effectively targeted. Some oncogenes are challenging to inhibit directly due to their structural properties, for instance KRAS lacks deep binding pockets and has a high affinity for GTP/GDP, making it difficult to develop effective inhibitors [10]. Moreover, cancers can develop resistance to targeted therapies through secondary mutations or activation of alternative signaling pathways [11, 12]. Resistance mechanisms often result from dynamic changes in the mutational landscape and phenotypic characteristics of cancer cells that lead to variable drug sensitivity and the need for novel therapeutic strategies. Patient stratification offers a solution in this regard to optimize treatment outcomes by tailoring therapies to individual tumor profiles. Traditionally, stratification has been performed either on the entire cancer patient population or within specific cancer types, delineating subtypes based on genetic, molecular, and clinical features [13]. While these approaches have improved treatment efficacy and patient survival, they often overlook the unique and specific characteristics of oncogene-addicted cohorts. Given the distinctive nature of oncogene addiction and the associated treatment challenges, there is a pressing need to stratify these cohorts with high resolution. Stratifying oncogene-addicted tumors who currently lack effective treatments can

uncover specific vulnerabilities and guide the development of more targeted and effective therapies. Despite its potential, the stratification of oncogene-addicted cohorts has not been systematically implemented. To address this gap, we propose the Oncostratifier framework, a systematic approach to identify drugs that specifically stratify oncogene-addicted cohorts. This framework aims to uncover drugs that can effectively target these unique patient populations by examining changes in drug response specific to oncogene addiction. It further categorizes drugs based on whether they induce sensitivity or resistance within the oncogene-addicted cohort, ensuring that the observed effects are specific to the oncogene in question. Using Oncostratifier, we can better understand the landscape of drug response in oncogene-addicted cancers and identify promising therapeutic candidates. This approach not only enhances our ability to provide effective treatments for oncogene-addicted tumors but also contributes to the broader field of precision oncology, where the goal is to deliver the right treatment to the right patient at the right time. The Oncostratifier can be directly applied to stratify patient tumors provided the availability of drug response data. Here, we showcase its use to stratify oncogene-addicted cohorts of cancer cell lines and generate new leads for possible treatments which might not have yet been explored in a patient tumor setting, thus providing valuable insights for precision treatment strategies.

3.2. Results and Discussion

We used the Oncostratifier framework to identify drugs that stratified an oncogene addicted (mutated) cohort into sensitive (responder) vs. resistant (non-responder) cell lines more strongly than in the wild-type cohort for a given cancer type. The analysis was performed for 665 drugs and 6681 unique pairs of oncogene and cancer type. Specifically, for each oncogene-cancer type pair, cell lines were split into two cohorts based on the mutation status of the oncogene: $Oncogene_{Mut}$ (mutated) and $Oncogene_{WT}$ (wild-type). We relied on the Oncostratifier score to quantify the stratification within each cohort. This score was defined as the entropy of the proportions of sensitive and resistant cell lines, grouped based on binarized drug response values, where the natural logarithm of the drug concentration enabling 50% of the maximal inhibitory effect ($\ln(IC50)$) was respectively smaller or greater than the peak plasma concentration $\ln(CMax)$ (Methods). Entropy was used to denote the ambiguity of the drug response and thus also the stratification potential of the drug within a cohort. To quantify the effect of oncogene addiction on stratification potential, we expressed the change between oncogene mutated and wild-type cohorts using the difference of entropies, $\Delta H = H(Oncogene_{Mut}) - H(Oncogene_{WT})$, with $\Delta H \approx 1$ denoting drug response stratification in $Oncogene_{Mut}$ but not $Oncogene_{WT}$, and $\Delta H \approx -1$ referring to stratification in $Oncogene_{WT}$ but not $Oncogene_{Mut}$. Finally, we selected the candidate drugs yielding more pronounced changes in stratification based on the Oncostratifier permutation-based p-value (Methods). We categorized drugs showing significant impact (Oncostratifier p-value below 0.05) into four groups, based on their stratifying characteristics and changes in response rates between the mutated

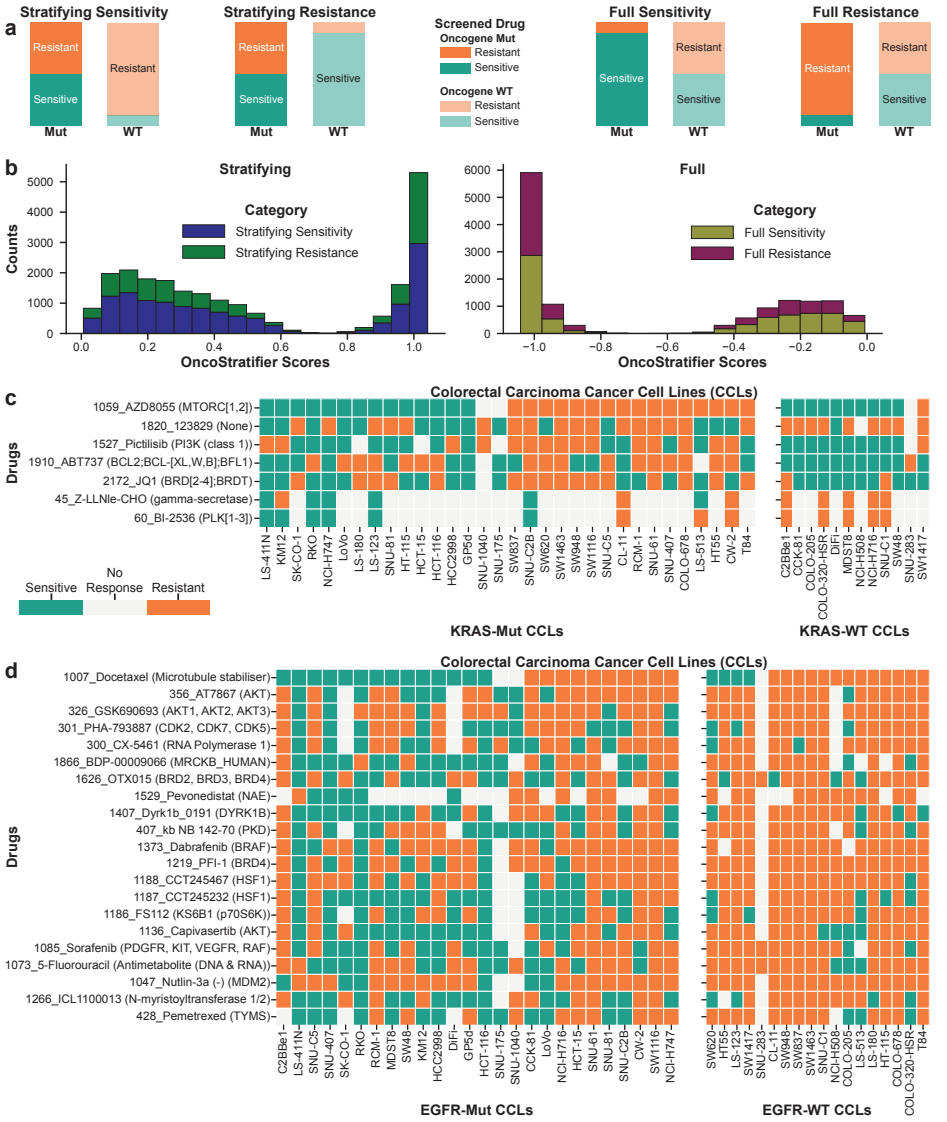


Figure 3.1: **Oncostratifier drug categories, distribution of significant drugs, and KRAS & EGFR addictions in colorectal carcinoma.** **(a)** Categories of drugs found by Oncostratifier: stratifying categories include drugs that stratify the oncogenic addicted (Mut) cohort whereas full categories include drugs that stratify the wild-type (WT) oncogene cohort. Sensitivity categories refer to cases where the oncogene addicted cohort gains sensitivity to the selected drug whereas Resistance categories refer to drugs gaining resistance with oncogene addiction compared to the wild-type. **(b)** Histogram of drugs found significant per Oncostratifier category. **(c-d)** Drug response of cancer cell lines to stratifying drugs found in **(c)** KRAS or **(d)** EGFR addiction in colorectal carcinoma split by oncogene addicted and oncogene wild-type cohorts.

and wild-type cohorts (Fig. 3.1a). Two categories exhibited stronger stratification in the mutated cohort than in the wild-type, namely: stratifying sensitivity (*SS*) and stratifying resistance (*SR*), denoting drugs that stratified the mutated cohort respectively with a gain or loss in sensitivity, quantified respectively by a larger or

smaller proportion of sensitive cell lines compared to the wild-type cohort. The remaining two categories showed weaker stratification in the mutated cohort than in the wild-type, including full sensitivity (*FS*) and full resistance (*FR*), where the mutated cohort respectively gained or lost sensitivity compared to the wild-type.

3.2.1. Oncogenic Addiction Influences the Response to Targeted Therapy

We identified 21,020 instances where drugs significantly stratified the oncogene addicted cohort, comprising 13,094 *SS* and 7,926 *SR* cases across 267 oncogenes and 31 different cancer types (Fig. 3.1b, Supplementary Tables 3.S1-3.S2). While most significant drugs showed near maximal changes in stratification ability (Oncostratifier score ≈ 1 or ≈ -1), there were notable exceptions with minimal but statistically significant changes (Oncostratifier Score ≈ 0).

Further examination focused on prominent oncogene addictions, such as KRAS and EGFR in colorectal carcinoma (Fig. 3.1c-d). Mutations in the KRAS gene, prevalent in various cancers, pose challenges to direct drug targeting due to the complexity of the binding site and high affinity for downstream effectors that often lead to off-target effects [14]. We identified three stratifying drugs where the KRAS-WT cohort was predominantly resistant and the KRAS-Mut cohort showed increased sensitivity associating with KRAS addiction (Fig. 3.1c): *123829*, *Z-LLNle-Cho*, and *BI-2536*. One of these drugs, the gamma secretase inhibitor (GSI) *Z-LLNle-Cho*, interferes with Notch signaling [15], which is known to be required for the survival of KRAS induced lung cancer cells [16]. This finding highlights how the sensitivity gain associated with KRAS mutations could be exploited therapeutically, even if observed in a different cancer type. Conversely, drugs like *AZD8055*, *Pictilisib*, *ABT737*, and *JQ1* demonstrated increased resistance in KRAS-Mut cohorts. Drugs found in both categories underscore the intricate relationship between genetic mutations and drug response, and could inform targeted treatment strategies for KRAS-addicted patients.

The EGFR (Epidermal Growth Factor Receptor) gene is a well-documented oncogene in various cancers, including colorectal [17–19], where the mutational activation of EGFR leads to uncontrolled cell proliferation and survival. The role of EGFR in colorectal cancer makes it a prime target for anticancer therapies, with several anti-EGFR therapies such as *Cetuximab* [20, 21] and *Panitumumab* [22] already demonstrating clinical efficacy in managing disease progression in patients with wild-type KRAS [23, 24]. However, the development of resistance to EGFR-targeted therapies, often through secondary mutations or alternative signaling pathways [20, 24], remains a challenge, underscoring the need to identify additional drugs that can bypass or overcome such resistance mechanisms. We identified 21 drugs that stratified the EGFR-Mut cohort with increased sensitivity (*SS*), indicating a potential for these drugs to enhance treatment outcomes in patients exhibiting EGFR-driven oncogenesis (Fig. 3.1d). Five of these drugs achieved a score close to the maximum of 1, indicative of strong stratification among EGFR-addicted patients, while all EGFR-WT cancer cell lines were fully resistant (*Pevonedistat*, *PFI-1*, *GSK690693*,

Nutlin-3a (-), and *Dabrafenib*). The observed sensitivity to the NEDD8 inhibitor *Pevonedistat* in EGFR-mutated colorectal cancer cell lines was corroborated by literature reporting that the combined blockade of NEDD8 and EGFR pathways significantly enhances growth arrest and apoptosis in colorectal cancer models [25].

In total, we identified stratifying drugs for 164 different oncogene addictions where the relation between the oncogene and cancer type was supported by the literature, including KRAS and EGFR in colorectal carcinoma, ERBB2 in breast carcinoma, and MYC in ovarian carcinoma (Supplementary Table 3.S1). The drugs identified in both *SS* and *SR* categories illustrate the complexity and variety of responses based on KRAS or EGFR mutational status, and reinforce the importance of precision medicine approaches in the treatment of cancer. Understanding these dynamics can help refine therapeutic strategies with the aim of developing more effective second-line treatments or combination therapies to address resistance or druggability issues associated with oncogene addiction.

3.2.2. ONCOGENE-ADDICTED COHORTS GAIN SENSITIVITY TO ONCOGENE-TARGETED THERAPIES

The robustness of the Oncostratifier approach was assessed by examining if oncogene-addicted cohorts gained sensitivity against drugs that specifically targeted the oncogene. We hypothesized that such drugs would predominantly fall into the stratifying sensitivity (*SS*) and full sensitivity (*FS*) categories, denoting a gain in sensitivity when the oncogene was mutated. Confirming this hypothesis, our analysis showed that 44.1% of the drugs targeting the tested oncogene were categorized as *SS*, and 22.5% as *FS*. In contrast, fewer drugs were categorized under stratifying resistance (*SR*, 14.4%) and full resistance (*FR*, 18.9%), with the latter category showing minimal numbers of responsive cell lines in the oncogene addicted cohort (Fig. 3.2a).

We attributed the emergence of resistance in the oncogene-addicted cohort for drugs categorized under *SR* and *FR* to two main factors. First, the existence of secondary targets of those drugs might confound their efficacy. Second, the addiction effect of the oncogene could vary across cancer types. A case in point involves drugs targeting EGFR, a crucial oncogene in glioblastoma [26], non-small cell lung [27], head and neck [28], colorectal [17], and pancreatic cancers [29]. In these cancer types, we found 7 drugs significantly associated with a gain in drug sensitivity in the oncogene-addicted cohort, of which 4 in the *SS* and 3 in the *FS* categories.However, only 2 drugs (*Pelitinib* and *CUDC-101*) were identified in the *SR* category with score less than 0.4, both in non-small cell lung cancer, where nearly half of the oncogene-addicted cohort still responded to both drugs. Moreover, *CUDC-101* also targeted the HAC1-10 and ERBB2 genes as well as EGFR, and this low selectivity could make *CUDC-101* less reliable.

We further looked into oncogene-targeting drugs focusing on cancer types with at least 10 significant drugs identified in any category. Results indicated notable sensitivity gains in breast cancer (51.72% *SS*, 44.83% *FS*), colorectal carcinoma

(57.69% *SS*, 19.23% *FS*), ovarian cancer (80% *SS*, 20% *FS*), melanoma (30.77% *SS*, 38.46% *FS*), B-cell non-Hodgkin's lymphoma (45.45% *SS*, 27.27% *FS*), and gastric carcinoma (53.85% *SS*, 23.08% *FS*). For non-small (16.67% *SS*, 25% *FS*) and small cell lung carcinomas (36.84% *SS*, 10.53% *FS*), the drugs identified did not predominantly fall into categories associated with a gain in sensitivity, highlighting the variable efficacy of oncogene-targeting drugs across tumor types.

3.2.3. Drugs Repeatedly Impacted by Oncogene Addiction

Oncogene addiction influences tumor development through complex interactions involving multiple genes and pathways. This dependency often modulates the effectiveness of drugs targeting various molecular pathways. Our Oncostratifier framework identified several drugs with stratifying characteristics frequently observed for multiple cancer types and oncogenes.

Specifically, seven drugs were repeatedly categorized under stratifying sensitivity (*SS*) for over 100 distinct oncogene-cancer type pairs (Supplementary Fig. 3.S1): *CX-5461*, *Oxaliplatin*, *Cisplatin*, *5-Fluorouracil*, *Mirin*, *Afatinib*, and *Methotrexate*. Additionally, *Gemcitabine* also consistently appeared in the *SS* category. Five of these drugs primarily interfere with RNA and/or DNA synthesis and cause DNA damage (*CX-5461* [30, 31], Oxaliplatin [32–34], *Cisplatin* [35–37], *Methotrexate* [38], and *Gemcitabine* [39]). As for the three others, *Mirin* prevents homology-dependent repair by affecting G2/M checkpoint [40], *Afatinib* targets tumor growth factors important in multiple cancer types [41–44], and *5-Fluorouracil* [45, 46] impairs the synthesis of pyrimidine which in turn induces apoptosis. These mechanisms suggest a broader impact on cellular processes crucial for oncogene-addicted cells, reflecting why disruptions in nucleic acid metabolism are particularly effective. The efficacy of the highlighted inhibitors could rely on additional requirements beyond oncogene addiction, such as *Cisplatin* needing ERK activation to induce apoptosis [37] or the knockdown of NFBD1 and MDC1 enhancing the impact of *Cisplatin* and *5-Fluorouracil* [46].

The pervasive stratification ability of these drugs in oncogene mutated cohorts is consistent with their frequent use alone or in combination therapies in multiple cancer types. The link between oncogenic addiction and enhanced sensitivity to drugs affecting DNA/RNA synthesis suggests that oncogene-addicted cells could rely more heavily on these fundamental processes, making them more vulnerable to such interventions. Further investigation into the specific pathways and oncogene interactions with these drugs could provide deeper insight into the mechanisms by which oncogene addiction alters drug sensitivity. Additionally, exploring patterns of resistance development and the efficacy of combination therapies involving these drugs could inform more effective treatment strategies for oncogene-addicted cancers.

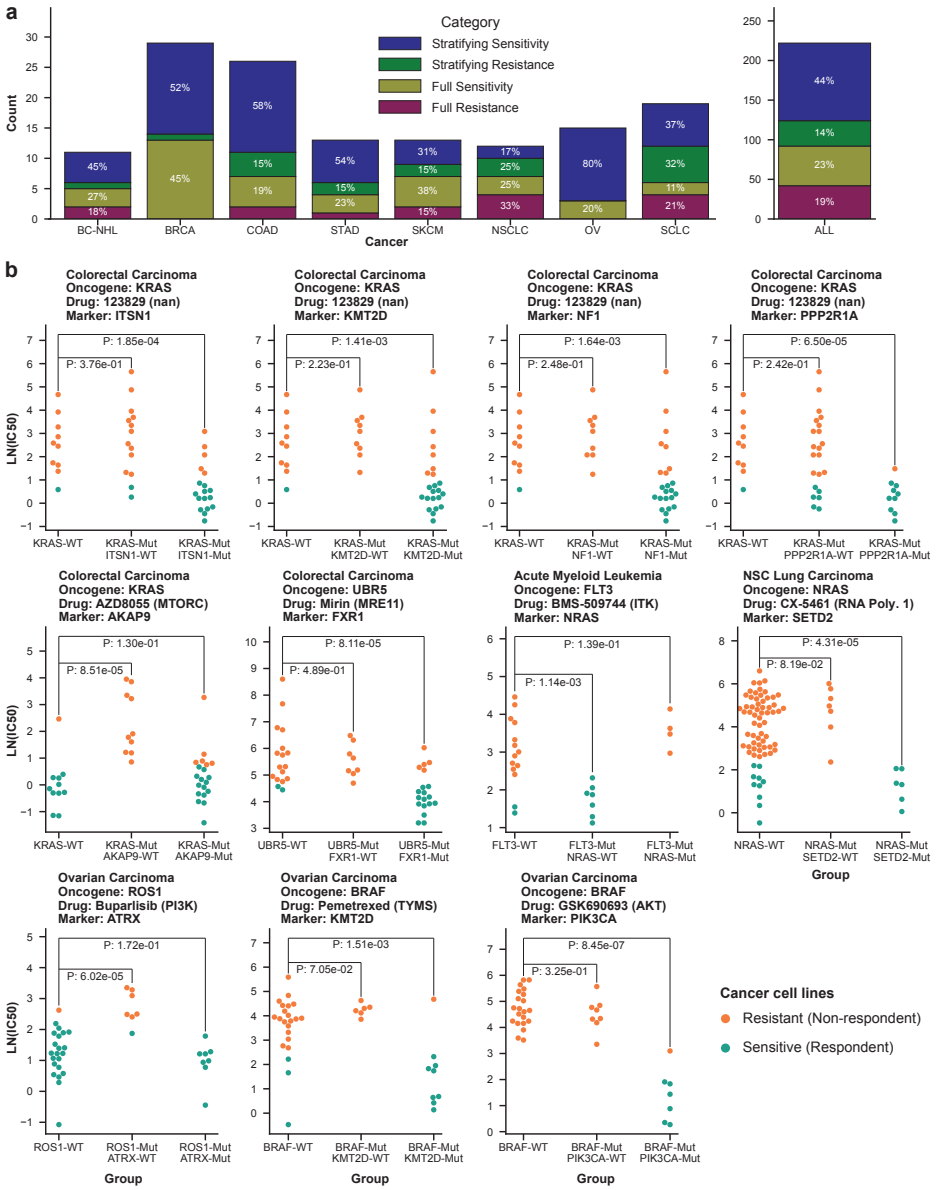


Figure 3.2: **Distribution of significant drugs and associated mutational markers stratifying well-known oncogene addictions.** **(a)** Distribution of significant drugs found in an oncogene addiction where the oncogene is also targeted by the drug. **(b)** Response of oncogene-WT cohort as well as oncogene addicted cohorts with and without mutation in the mutational markers of oncostratifiers found for well-known oncogene addictions. P values refer to the significance of the change in proportions between the oncogene-WT cohort and other cohorts using Fisher's exact test with Benjamini-Hochberg correction.

3.2.4. IDENTIFYING MUTATIONAL MARKERS FOR STRATIFYING DRUGS

In the pursuit of translating in vitro cell line results to in vivo patient tumors or PDX models, it is crucial to identify biomarkers correlating with drug stratification in

oncogene-addicted cohorts that can be used to generate hypotheses for additional drug targets. We focused on gene mutational markers (Methods).

We identified 59 mutational markers associated with 36 stratifying drugs across 36 distinct genes and 11 cancer types, encompassing 46 cancer type-oncogene pairings (Fig. 3.2b, Supplementary Fig. 3.S2-3.S5). For 7 of these pairings, existing literature supporting the link between the oncogene and cancer provided reassurance of the relevance of our findings.

Acute Myeloid Leukemia (AML) and FLT3 Addiction [47]: We identified *BMS-509744* as a stratifying sensitivity drug (*SS*) with NRAS as a mutational marker in the context of *FLT3* oncogene addiction in AML cell lines (Fig. 3.2b, Supplementary Fig. 3.S6). Within the FLT3 oncogene-addicted cohort, the subcohort with NRAS WT was sensitive to the ITK inhibiting drug *BMS-509744* ($P \approx 1.14 \times 10^{-3}$) while the NRAS mutated subcohort was resistant ($P \approx 1.39 \times 10^{-1}$). This delineation underscores the important role of NRAS status in modulating response to *BMS-509744* and highlights the potential this drug to target FLT3-addicted leukemia with specific genetic backgrounds.

Colorectal Cancer: For colorectal cancer cell lines characterized by KRAS addiction, we identified two stratifying drugs: *123829* and *AZD8055*. The drug *123829* , categorized as *SS* (stratifying sensitivity), was associated with four mutational markers: ITSN1, KMT2D, NF1, and PPP2R1A. With KRAS addiction, the subcohorts harboring mutations in the ITSN1, KMT2D, and/or NF1 markers showed a significantly larger proportion of cell lines sensitive to the *123829* drug ($P \approx 1.85 \times 10^{-4}$ $P \approx 1.41 \times 10^{-3}$ and $\approx 1.64 \times 10^{-3}$, respectively) than the subcohorts with the corresponding wild-type markers, which were predominantly resistant. Interestingly, almost all cell lines with PPP2R1A mutations showed sensitivity to *123829* ($P \approx 6.50 \times 10^{-5}$), suggesting PPP2R1A mutation status as a robust predictor of sensitivity to this drug in KRAS addicted cohorts.

On the other hand, the drug *AZD8055*, categorized as *SR* (stratifying resistance), showed association with a single mutational marker, AKAP9. In KRAS-addicted colorectal cancer cell lines, the whole subcohort with the wild-type AKAP9 gene exhibited resistance to *AZD8055*. This differed significantly from the KRAS wild-type cohort ($P \approx 8.51 \times 10^{-5}$) and the subcohort of KRAS mutated cell lines where AKAP9 was also mutated, both containing more than half of lines sensitive to *AZD8055*, indicating that AKAP9 mutations might mitigate the loss of sensitivity usually associated with KRAS addiction.

Furthermore, we identified *Mirin* as a stratifying drug with FXR1 mutation status as a marker in UBR5-addicted colorectal cancer cell lines (Fig. 3.2b, Supplementary Fig. 3.S7). The UBR5 gene is involved in damage response and apoptosis [48], and has been advanced as a potential oncogene in colorectal carcinoma [49]. With the UBR5 oncogene addicted cohort, only the subcohort with mutated FXR1 showed a significantly larger proportion of cell lines sensitive to *Mirin* ($P \approx 8.11 \times 10^{-5}$) compared to the UBR WT cohort. Thus, FXR1 could potentially be used as a marker to decide which UBR5 addicted colorectal carcinomas could be more effectively

targeted using *Mirin*.

Non-Small Cell Lung Cancer (NSCLC): For NSCLC, the mutational status of SETD2 fully differentiated the response to the drug *CX-5461*, targeting RNA polymerase I, in the NRAS-addicted cohort (Fig. 3.2b, Supplementary Fig. 3.S8). Specifically, all cell lines were sensitive in the NRAS-addicted subcohort with mutated SETD2 ($P \approx 4.31 \times 10^{-5}$), suggesting a potential synthetic lethal interaction between NRAS and SETD2 associating with treatment using the *CX-5461* drug.

Ovarian Cancer: In ovarian cancer, *Buparlisib* (targeting ROS1 addiction), *Pemetrexed* (targeting BRAF addiction), and *GSK690693* (targeting BRAF addiction) showed differential effectiveness in oncogene-addicted cohorts based on ATRX, KMT2D, and PIK3CA mutations, respectively. The effectiveness of *Buparlisib* was compromised in ROS1-mutated cohorts with wild-type ATRX ($P \approx 6.02 \times 10^{-5}$), whereas *Pemetrexed* and *GSK690693* showed increased sensitivity in BRAF-addicted cohorts with mutated KMT2D ($P \approx 1.51 \times 10^{-3}$) and mutated PIK3CA ($P \approx 8.45 \times 10^{-7}$), respectively, illustrating the potential for KMT2D and PIK3CA as stratifying markers in BRAF-driven ovarian cancers.

These findings underscore the complexity of drug response in oncogene-addicted cohorts and highlight the importance of genetic markers for predicting therapeutic outcomes. Marker identification offers valuable insight into the mechanisms of drug response and could assist in tailoring targeted treatment strategies for cancer patients in the future.

3.2.5. STRATIFICATION OF ONCOGENE-ADDICTED PATIENT TUMORS BASED ON MUTATIONAL MARKERS

To assess translation potential, we used the TCGA patient cohort to stratify oncogene-addicted patient tumors according to the mutational status of the previously identified cell line markers. For each cancer type-oncogene-marker pairing, we assessed the number of patients exhibiting oncogene addiction, and its stratification by mutational marker into WT and mutated subcohorts with at least three patients each (Table 3.1).

In most cases, the marker-WT subcohort was larger than the marker-mutated subcohort. This finding could be interesting for markers showing increased sensitivity in the oncogene-addicted compared to the non-addicted cohort, suggesting that a larger number of patients could benefit from the stratifying drug if it revealed effective in patient tumors. Notable instances included: FLT3 addiction in AML, where 96% (50/52) of the patient tumors had wild-type NRAS (Table 3.1), and the cell line wild-type NRAS subcohort showed increased sensitivity to the drug *BMS-509744* (Fig. 3.2b); JAK3 and MYBL addictions in colorectal cancer with 80% (14/17) and 82% (16/20) of the oncogene-addicted tumors showing wild-type MAP2K1 which is associated with increased sensitivity to *AZD7969*, respectively; and NUP98 addiction also in colorectal cancer where over 50% of the patient tumors had wild-type markers CNOT1, PIK3CA, or PTEN associated with increased

Table 3.1: **Number of oncogene addicted TCGA patient tumors (per cancer type-oncogene pair) stratified by mutational marker, and genes differentially expressed (DEGs) between the subcohorts defined by marker status**. Boldface font is used to indicate cases for which the drug response in cell lines was significantly different in the specific oncogene-addicted mutational marker subcohort compared to the cohort without oncogene addiction regardless of marker status ($P < 0.05$). An asterisk $*$ before or after the value indicates a larger number of either resistant or sensitive cell lines, respectively.

Cancer	Oncogene	Marker	Oncogene Addicted Tumors	Marker Mut	Marker WT	DEGs	Tested Genes
AML	FLT3	NRAS	52	2	**50***	12	16758
BRCA	BRD4	PIK3R1	8	**3***	5	40	19170
BRCA	MSI2	NR4A2	5	**1***	4	120	19004
BRCA	ROS1	CHD9	22	**6***	16	79	19556
COAD	JAK3	MAP2K1	17	3	**14***	170	19121
COAD	KRAS	AKAP9	220	16	***204**	1140	19943
COAD	KRAS	ITSN1	220	**14***	206	70	19943
COAD	KRAS	KMT2D	220	**29***	191	2028	19943
COAD	KRAS	NF1	220	**8***	212	14	19943
COAD	KRAS	PPP2R1A	220	**6***	214	93	19943
COAD	KSR2	FXR1	25	**3***	22	17	19263
COAD	MECOM	FXR1	36	**7***	29	7	19362
COAD	MTOR	FXR1	43	**7***	36	5	19425
COAD	MYBL1	MAP2K1	20	4	**16***	17	19201
COAD	NTRK3	FXR1	24	**5***	19	14	19149
COAD	NUP98	CNOT1	26	13	**13***	79	19241
COAD	NUP98	PIK3CA	26	11	**15***	20	19241
COAD	NUP98	PTEN	26	7	**19***	30	19241
COAD	UBR5	FXR1	33	**4***	29	25	19408
COAD	WWP1	FXR1	27	**6***	***21**	17	19300
COAD	WWP1	PIK3CA	27	**13***	***14**	79	19300
GBM	IGF1R	EGFR	3	***1**	2	477	18610
NSCLC	KDM5A	ARID1A	13	**2***	11	340	19498
NSCLC	LGR5	CLSPN	24	***1**	23	793	19682
NSCLC	MAP3K13	CLSPN	14	***1**	13	18	19493
NSCLC	MGAM	TP53BP1	58	3	55	4	19863
NSCLC	NTRK3	KEAP1	48	15	**33***	1392	19832
NSCLC	PDGFRB	MAP4K3	17	2	**15***	125	19673
SCLC	CARD11	TJP1	32	**3***	29	18	19719
SCLC	INSR	MED12	14	**1***	13	13	19435
SKCM	JAK3	ANK3	20	14	***6**	38	19722

sensitivity to *HG558801* in cell lines.

Cases where the mutated marker associated with increased sensitivity in the oncogene-addicted cell line cohort displayed more modest numbers of tumors assigned to the mutated marker subcohort. For example, in COAD with KRAS addiction, out of 220 tumors, only 14 and 29 harbored mutated ITSN1 and KMT2D markers, respectively. In another case, 13 out of 27 WWP1-addicted colorectal cancer patient tumors carried a mutated PIK3CA.

Additionally, PIK3CA was identified as a mutational marker for response to the drug *Mirin* in WWP1-addicted colorectal cancer cell line cohort. In this group, the cohort with mutated PIK3CA demonstrated significantly increased sensitivity to *Mirin*, whereas those with WT PIK3CA lost sensitivity, highlighting the mutation's dual role in modifying drug response. This clear dichotomy makes PIK3CA a valuable predictive marker for therapeutic strategies in WWP1-addicted COAD, with nearly equal division of tumors into mutated and WT subcohorts.

DIFFERENTIALLY EXPRESSED GENES WITHIN ONCOGENIC ADDICTED COHORT AND THEIR ENRICHMENT

We further investigated the oncogene-addicted patient cohorts stratified by the mutational marker to identify changes in gene expression between the two subcohorts defined by mutational marker status (Table 3.1). We identified differentially expressed genes (DEGs) and performed functional and pathway enrichment, focusing on cases where both cohorts had at least 10 patients. Here we highlight KRAS addiction in colorectal cancer, for which we identified 220 KRAS-addicted tumors (Supplementary Fig. 3.S9).

Previously, we reported a significant loss in sensitivity against the mTORC-targeting drug *AZD8055* in the subcohort of KRAS addicted tumors with wild-type AKAP9, including 204 of 220 patients. This suggested that AKAP9 mutation status could be used as a marker to exclude patients with KRAS addicted tumors who might not benefit from an mTORC-targeting drug that promotes anti-tumor immunity. We identified 1140 DEGs between the mutated and wild-type AKAP9 subcohorts of the KRAS addicted cohort, which were mostly associated with the major histocompatibility complex (MHC), as well as interferon gamma response. Consistent with this, the deterioration in MHC class 1 molecules is known to cause resistance to immunotherapy [50]. Similarly, interferon gamma (IFN-γ) plays a vital role in boosting the ability of the immune system to identify and destroy cancer cells [51], making it essential for the efficacy of immunotherapies [52].

For another marker, KMT2D, the KMT2D mutated subcohort of the KRAS addicted cohort was associated with gain in sensitivity against the drug *123829* in colorectal carcinoma. The composition and name of compound *123829* have not been disclosed in the GDSC studies, so it is not possible to provide further interpretation on the underlying mechanisms of this drug. In any case, 29 of 220 patients in the KRAS addicted cohort had KMT2D mutations, which could potentially benefit from

treatment with *123829*. We found 2028 genes differently expressed between the mutated and wild-type KMT2D subcohorts, which were also enriched with the MHC and interferon gamma response.

3.2.6. Drug Sets to Cover Oncogene Addicted Cohort

Stratifying ,oncogene addicted patient tumors is crucial to define subcohorts that might be sensitive to treatment due to the impact of mutations in the oncogene. However, stratification based on mutational markers usually implies that one subcohort is sensitive while the other one is resistant to the drug. Thus, not all oncogene-addicted patients are treatable by the stratifying drug. To bring the potential treatment option to all of the oncogene-addicted subcohorts, we investigated an alternative strategy where we looked for drugs that could cover or sensitize the whole set of oncogene-addicted cancer cell lines (CCLs). We identified 62 cases in various cancer types with at least 10 oncogene addicted cell lines, where the entire oncogene addicted cohort could be sensitized by only 2 drugs (Supplementary Table 3.S3). Out of the 62 cases, we found literature support for 3 pairs of oncogene and cancer type.

The ERBB2 gene, widely known as “HER2”, is strongly associated with poor prognosis in breast cancer [53] and impacts cell growth, differentiation, and migration together with other EGF receptors [54]. The ERBB2 addicted cohort of 12 cell lines in breast cancer was covered by *Afatinib* (9 CCLs) and *Telomerase Inhibitor IX* (8 CCLs). The drug *Afatinib* is already used to target ERBB2 [55], whereas the *Telomerase Inhibitor IX* drug presents an opportunity to sensitize the ERBB2 addicted subcohort resistant to *Afatinib* [56].

The MDM2 gene is an oncogene due to highly expressed MDM2 supressing TP53, which increases the risk of cancer [57]. Inhibition of MDM2 selectively targets PTEN-deficient colorectal cancer cells, activating p53 and inhibiting tumor growth [26]. The MDM2 addicted colorectal cancer cohort (11 CCLs) was covered by HDAC1 targeting *AR-42* (8 CCLs) and AURKA targerting *Alisertib* (6 CCLs) drugs. The HDAC1 enzyme can deacetylate p53 by binding to MDM2 [58] and AURKA enhances the p53 degradation effect of MDM2 [59]. As a result, inhibiting both genes could potentially reduce p53 degradation and lead to better prognosis in patients with MDM2 mutated tumors. Moreover, UBR5 is an oncogene associated with poor prognosis in gastric carcinoma [60]. Our analysis showed that the UBR5 addicted gastric cancer cohort (12 CCLs) could be covered by the PIK3CG targeting *PIK-93* (8 CCLs) and FEN1 targeting *FEN1* (6 CCLs) drugs.

Furthermore, we found cases where 1 or 2 cell lines in the oncogene addicted cohort were never sensitive to any of the stratifying drugs we identified. To reveal partial stratification opportunities for these cases, we excluded such cell lines and repeated the analysis to identify the minimal set of drugs that sensitized the remaining lines (Supplementary Table 3.S4). We were then able to identify drug set covers with only 2 stratifying drugs also for other well-known oncogene addictions such as FLT3 addiction in Acute Myeloid Leukemia (AML); EGFR addiction in Glioblastoma; AKT1,

DDR2, MET, NRAS, PIK3CB, RET, and ROS1 addiction in non-small cell lung carcinoma (NSCLC); and EGFR and MET in small cell lung carcinoma.

3.3. CONCLUSION

In this study, we proposed Oncostratifier, a statistical framework that leverages drug response data on cancer cell lines to find therapy-associated stratification opportunities for oncogene addicted cohorts. Although stratification methods are used extensively in cancer, to our knowledge Oncostratifier is the first method to stratify oncogene addicted cohorts computationally and systematically. Importantly, Oncostratifier suggests possible therapeutic avenues to target oncogene addicted subcohorts based on existing drugs, which can thus be directly tested using drug cell line screens or eventual clinical trials.

Oncostratifier identified stratifying drugs for multiple oncogene addictions in cell lines of different cancer types, including well-known EGFR, KRAS, and UBR5 addictions in colorectal carcinoma, MYC addiction in ovarian carcinoma, and ERBB2 addiction in breast cancer. Finding drugs that sensitize part of the oncogene addicted cohort is particularly valuable for some cases, such as when the oncogene itself is untargetable (KRAS in colorectal carcinoma) or when there is acquired resistance to treatment (EGFR in colorectal carcinoma).

To attempt a translation from cancer cell lines to patient tumors, we identified mutational markers for the stratifying drugs found by Oncostratifier. The markers enable similar stratification to that achieved by the corresponding drug, and are therefore used as a proxy to stratify patient tumors in the absence of drug response. Using the identified mutational markers, we demonstrated the potential to stratify tumors that are presumed to be oncogene-addicted into subgroups that may benefit from more targeted therapeutic strategies. This assumes that the identified markers retain their predictive value for drug response, as observed in cancer cell lines. However, actual tumors are more complex than cancer cell lines, and it is not always certain which oncogene drives a given patient's tumor. It would be in great interest to focus on tumor patient cohorts with a well-defined oncogene dependency and available drug response data to validate our findings directly. Since clinical drug response data in patients remain limited, testing these markers in patient-derived xenograft (PDX) models, where human tumor samples are implanted into mice to more closely reflect the tumor's natural microenvironment, offers a promising alternative. Such models could be used to confirm that our mutational markers are predictive of drug response for oncogene addiction and thus strengthen the rationale for using them potentially in clinic.

Oncostratifier also uncovered treatment opportunities for oncogene addicted cell line cohorts. For some oncogene addictions, all of the oncogene addicted subcohorts can be sensitized using 2 or more of the identified stratifying drugs.

3.4. METHODS

3.4.1. DATA COLLECTION AND PROCESSING

DRUG RESPONSE

We used drug response data for cancer cell lines from the Genomics of Drug Sensitivity in Cancer (GDSC) database [61, 62] (version 27Oct23). This included data for a total of 665 compounds across two different drug screens, 367 in GDSC1 and 198 in GDSC2, together with putative drug targets. Drugs overlapping between GDSC1 and GDSC2 were treated as distinct. The drug efficacy for a given cancer cell line was quantified using the natural logarithm of the 50% growth inhibition value ($ln(IC50)$). We further categorized each cell line into sensitive (responder) or resistant (non-responder) by binarizing the ln($IC50$) response score. For this we used the natural logarithm of the peak plasma concentration (ln($CMax$)) as a threshold, where the cell line was deemed sensitive if ln($IC50$) was smaller than ln($CMax$) or resistant if ln($IC50$) was greater than ln($CMax$).

CANCER CELL LINES AND ONCOGENES

We obtained annotations and lists of mutations for the cancer cell lines from the Cell Model Passports repository [63] (version 20220510). The oncogenes to be tested were obtained from the Memorial Sloan Kettering Precision Oncology Knowledge Base [64] (latest version as of 10/02/2023). We considered only those genes recognized as oncogenes in at least one tissue ("Is Oncogene" == "Yes") and not classified as tumor suppressor genes in any tissue ("Is Tumor Suppressor Gene" == "No"). The mutated status of a gene was determined based on the presence of non-silent mutations. To ensure a minimum representation for statistical analysis, we required that the mutated and wild-type groups of cell lines versions of the oncogene were each represented by at least three distinct cell lines. The original data comprised 42 cancer types and 268 oncogenes, yielding 11256 pairs of cancer type and oncogene. Of these, 6681 pairs met the aforementioned requirements and were subsequently tested using Oncostratifier.

PRIMARY TUMOURS

We collected gene expression, mutation, and clinical data for primary tumor samples from the Tumor Cancer Genome Atlas (TCGA) patient tissue samples, courtesy of the Pan-Cancer Atlas [13]. The processed and curated data was accessed through the cBio portal [65, 66] (on 15/12/2023). We considered only primary tumor samples and non-silent mutation data. For differential gene expression analysis, we leveraged RSEM (Batch normalized from Illumina HiSeq RNASeqV2) mRNA expression data. Moreover, to visualize the normalized expression values of differentially expressed genes, we used log-transformed mRNA expression (RNASeqV2 RSEM) z-scores compared to the expression distribution of all samples.

3.4.2. ONCOSTRATIFIER

IDENTIFICATION OF STRATIFYING DRUGS

We used Oncostratifier to identify drugs with promising differential stratification potential in the oncogene addicted (mutated) cohort compared to the wild-type cohort. For a given cancer type, the approach first split the available cancer cell lines into two cohorts, $Oncogene_{Mut}$ (mutated) and $Oncogene_{WT}$ (wild-type), based on the non-silent mutation status of the relevant oncogene within each cell line. Then the stratification within each cohort was characterized by quantifying the variability or uncertainty in the drug response proportions, of sensitive and resistant cancer cell lines, using the Shannon entropy (Eq. 3.1)

$$\begin{aligned} H\big(Oncogene_{Mut}\big) &= -p_S \times log_2(p_S) - p_R \times log_2(p_R) \\ H\big(Oncogene_{WT}\big) &= -q_S \times log_2(q_S) - q_R \times log_2(q_R) \end{aligned} \tag{3.1}$$

where p_S and p_R (q_S and q_R) represent the proportions of sensitive and resistant cancer cell lines within the $Oncogene_{Mut}$ ($Oncogene_{WT}$) cohort, respectively. To discern drugs that selectively stratified one cohort more strongly than the other, we examined the differential entropy (ΔH) between the cohorts for every pair of oncogene and cancer type (Eq. 3.2).

$$\Delta H = H(Oncogene_{Mut}) - H(Oncogene_{WT}) \tag{3.2}$$

A positive or negative ΔH value would indicate that the drug showed respectively stronger and weaker stratification in the $Oncogene_{Mut}$ cohort compared to the $Oncogene_{WT}$ cohort. To select candidate drugs with a more promising differential stratification effect, a permutation-based significance test p-value was obtained for each $\Delta H_{observed}$ value as follows. We performed 10000 permutations, each involving the random shuffling of the drug response labels assigned to the cancer cell lines followed by the computation of the corresponding differential stratification score, $\Delta H_{permutation}$. The p-value was then defined as the probability of obtaining an equally or more extreme $\Delta H_{permutation}$ value than the actual $\Delta H_{observed}$ value (Eq. 3.3), under the assumption that the null hypothesis positing no significance difference in entropy between the two cohorts was true.

$$p-value = \frac{\text{Number of permutations where } \left|\Delta H_{permutation}\right| \geq |\Delta H_{observed}|}{\text{Total number of permutations}} \tag{3.3}$$

CATEGORIZATION OF STRATIFYING DRUGS

Drugs with a permutation p-value lower than the arbitrarily chosen threshold 0.05 were considered significant, and were further categorized according to the

differential entropy values ΔH and proportions of sensitive and resistant cell lines of the cohorts as follows:

- **Stratifying sensitivity (*SS*)**: $\Delta H > 0$; p-value < 0.05; $q_S < q_R$
- **Stratifying resistance (*SR*)**: $\Delta H > 0$; p-value < 0.05; $q_S > q_R$
- **Full sensitivity (*FS*)**: $\Delta H < 0$; p-value < 0.05; $p_S > p_R$
- **Full resistance (*FR*)**: $\Delta H < 0$; p-value < 0.05; $p_S < p_R$

The two stratifying categories denoted cases where oncogene addiction was associated with increased cancer cell line stratification potential, acquired through acquisition (*SS*) or loss (*SR*) of sensitivity in the oncogene addicted cohort $Oncogene_{Mut}$, given a predominantly resistant (*SS*) or sensitive (*SR*) wild-type cohort $Oncogene_{WT}$. Conversely, the two other categories denoted a reduction in stratification ability in the oncogene addicted cohort $Oncogene_{Mut}$ due to cell lines gaining (*FS*) or losing (*FR*) sensitivity compared to the wild-type cohort $Oncogene_{WT}$.

3.4.3. Analysis of Stratifying Drugs Found by OncoStratifier

Mutational markers for stratifying drugs in cell lines

We revealed markers for drugs that stratified oncogene mutated cell lines through association analysis between drug response (sensitive or resistant) and the mutational status of each cancer-related gene. We quantified the association separately for the $Oncogene_{Mut}$ and $Oncogene_{WT}$ cohorts using Fisher's exact tests, focusing on genes previously identified as cancer genes (CGs) amid cancer functional events (CFEs) [62]. The Fisher's test p-values were corrected for multiple testing using the Benjamini-Hochberg procedure [67]. A given gene was deemed a mutational marker of a stratifying drug if its corrected p-values were respectively significant in the $Oncogene_{Mut}$ (below 0.05) and non-significant in the $Oncogene_{WT}$ cohort (not below 0.05).

Differential gene expression analysis in primary tumors

Considering primary tumor samples where the oncogene was mutated, we identified differentially expressed genes between subcohorts where the marker gene was either mutated or wild-type. We first excluded genes with variance in rounded RSEM read count estimates below 0.0001. We used DESeq2 to identify outliers based on the Cook's distance, imput new values for the filtered outliers, and identify differentially expressed genes [68] using the pyDeSeq2 implementation [69]. As suggested by DESeq2, only the genes whose p-values passed the independent filtering stage [68] were identified and had their p-values adjusted using the Benjamini-Hochberg procedure [67]. Genes with adjusted p-values below 0.05 were identified as differentially expressed.

FUNCTIONAL AND PATHWAY ENRICHMENT FOR DIFFERENTIALLY EXPRESSED GENES

We investigated whether the differentially expressed genes associated with each cancer-oncogene-marker pair were significantly enriched within various functional and pathway annotations. We included annotations from the Gene Ontology (version 2023; encompassing biological processes, cellular components, and molecular functions) [70] and MSigDB Hallmark [71], as well as canonical pathways from databases KEGG (2021, Human) [72], WikiPathway (2023, Human) [73], and Reactome (2022) [74]. We performed the enrichment analysis using EnrichR (version June 8, 2023, accessed on 17/01/2023) [75], which tests for enrichment with Fisher's exact tests and adjusts p-values for multiple testing using the Benjamini-Hochberg procedure [67].

IDENTIFICATION OF MINIMAL STRATIFYING DRUG SETS FOR ONCOGENE ADDICTED COHORTS

For each oncogene-cancer type pair, we identified a minimal set of stratifying drugs such that each cell line in the oncogene-addicted cohort was sensitive to at least one drug in the set. We formulated the minimal drug set selection problem as a set cover optimization. Let C denote the set of all oncogene-addicted cell lines and $C_j \subseteq C$, a subset of the cell lines that are sensitive to drug j. The objective is to identify the smallest number of such subsets whose union covers the entire cohort C. Formally, we seek a minimal collection such that:

$$\bigcup_{j}^{K} C_j = C.$$

where K is a set of drugs in the minimal set cover.

To find the minimal set cover, we implemented a recursive algorithm that exhaustively explores combinations of candidate drugs. In the case where multiple minimal solutions exist, the algorithm reports the first solution it discovers.

For some oncogene–tumor type pairs, one or two oncogene-addicted cell lines were never sensitive to any of the stratifying drugs, making it impossible to find a set that covered the entire cohort. Thus, we repeated the analysis including only the cell lines that were sensitive to at least one candidate drug. This allowed us to identify minimal set covers even when a complete cover set was not possible, to provide additional insight into partially actionable therapeutic strategies.

REFERENCES

[1] D. Hanahan and R. A. Weinberg. “The Hallmarks of Cancer”. In: *Cell* 100.1 (Jan. 2000), pp. 57–70.

[2] B. Vogelstein and K. W. Kinzler. “Cancer genes and the pathways they control”. In: *Nat. Med.* 10.8 (July 2004), pp. 789–799.

[3] D. Hanahan and R. A. Weinberg. “Hallmarks of Cancer: The Next Generation”. In: *Cell* 144.5 (Mar. 2011), pp. 646–674.

[4] I. B. Weinstein and A. K. Joe. “Mechanisms of Disease: oncogene addiction—a rationale for molecular targeting in cancer therapy”. In: *Nat. Clin. Pract. Oncol.* 3.8 (Aug. 2006), pp. 448–457.

[5] J. Downward. “Targeting RAS signalling pathways in cancer therapy”. In: *Nat. Rev. Cancer* 3.1 (Jan. 2003), pp. 11–22.

[6] M. Scaltriti and J. Baselga. “The Epidermal Growth Factor Receptor Pathway: A Model for Targeted Therapy”. In: *Clin. Cancer Res.* 12.18 (Sept. 2006), pp. 5268–5272.

[7] R. Dhanasekaran *et al.* “The MYC oncogene - the grand orchestrator of cancer growth and immune evasion”. en. In: *Nat. Rev. Clin. Oncol.* 19.1 (Jan. 2022), pp. 23–36.

[8] B. J. Druker *et al.* “Efficacy and Safety of a Specific Inhibitor of the BCR-ABL Tyrosine Kinase in Chronic Myeloid Leukemia”. In: *N. Engl. J. Med.* 344.14 (Apr. 2001), pp. 1031–1037.

[9] T. J. Lynch *et al.* “Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non–Small-Cell Lung Cancer to Gefitinib”. In: *N. Engl. J. Med.* 350.21 (May 2004), pp. 2129–2139.

[10] K. Parikh *et al.* “Drugging KRAS: current perspectives and state-of-art review”. In: *J. Hematol. Oncol.* 15.1 (Oct. 2022).

[11] J. A. Engelman *et al.* “MET Amplification Leads to Gefitinib Resistance in Lung Cancer by Activating ERBB3 Signaling”. In: *Science* 316.5827 (May 2007), pp. 1039–1043.

[12] S. Kobayashi *et al.* “EGFRMutation and Resistance of Non–Small-Cell Lung Cancer to Gefitinib”. In: *N. Engl. J. Med.* 352.8 (Feb. 2005), pp. 786–792.

[13] Tcga Gdac. *Analysis-ready standardized TCGA data from Broad GDAC Firehose 2016_01_28 run.* 2016.

[14] P. Liu, Y. Wang, and X. Li. “Targeting the untargetable KRAS in cancer therapy”. In: *Yao Xue Xue Bao* 9.5 (Sept. 2019), pp. 871–879.

[15] X. Meng *et al.* “GSI-I (Z-LLNle-CHO) inhibits γ-secretase and the proteosome to trigger cell death in precursor-B acute lymphoblastic leukemia”. en. In: *Leukemia* 25.7 (July 2011), pp. 1135–1146.

[16] S. Licciulli *et al.* “Notch1 is required for Kras-induced lung adenocarcinoma and controls tumor cell survival via p53”. en. In: *Cancer Res.* 73.19 (Oct. 2013), pp. 5974–5984.

[17] C. Messa *et al.* “EGF, TGF-alpha, and EGF-R in human colorectal adenocarcinoma”. en. In: *Acta Oncol.* 37.3 (1998), pp. 285–289.

[18] I. Porebska, A. Harlozińska, and T. Bojarowski. “Expression of the tyrosine kinase activity growth factor receptors (EGFR, ERB B2, ERB B3) in colorectal adenocarcinomas and adenomas”. en. In: *Tumour Biol.* 21.2 (2000), pp. 105–115.

[19] D. S. Salomon *et al.* “Epidermal growth factor-related peptides and their receptors in human malignancies”. en. In: *Crit. Rev. Oncol. Hematol.* 19.3 (July 1995), pp. 183–232.

[20] C. S. Karapetis *et al.* “K-ras mutations and benefit from cetuximab in advanced colorectal cancer”. en. In: *N. Engl. J. Med.* 359.17 (Oct. 2008), pp. 1757–1765.

[21] E. Van Cutsem *et al.* “Cetuximab and chemotherapy as initial treatment for metastatic colorectal cancer”. en. In: *N. Engl. J. Med.* 360.14 (Apr. 2009), pp. 1408–1417.

[22] F. Battaglin *et al.* “Anti-EGFR monoclonal antibody panitumumab for the treatment of patients with metastatic colorectal cancer: an overview of current practice and future perspectives”. en. In: *Expert Opin. Biol. Ther.* 17.10 (Oct. 2017), pp. 1297–1308.

[23] C. Bokemeyer *et al.* “Fluorouracil, leucovorin, and oxaliplatin with and without cetuximab in the first-line treatment of metastatic colorectal cancer”. en. In: *J. Clin. Oncol.* 27.5 (Feb. 2009), pp. 663–671.

[24] R. G. Amado *et al.* “Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer”. en. In: *J. Clin. Oncol.* 26.10 (Apr. 2008), pp. 1626–1634.

[25] F. Invrea *et al.* “Synthetic Lethality Screening Highlights Colorectal Cancer Vulnerability to Concomitant Blockade of NEDD8 and EGFR Pathways”. en. In: *Cancers* 13.15 (July 2021).

[26] C. W. Brennan *et al.* “The somatic genomic landscape of glioblastoma”. en. In: *Cell* 155.2 (Oct. 2013), pp. 462–477.

[27] J.-C. Soria *et al.* “EGFR-mutated oncogene-addicted non-small cell lung cancer: current trends and future prospects”. en. In: *Cancer Treat. Rev.* 38.5 (Aug. 2012), pp. 416–430.

[28] S. Nair, J. A. Bonner, and M. Bredel. “EGFR Mutations in Head and Neck Squamous Cell Carcinoma”. en. In: *Int. J. Mol. Sci.* 23.7 (Mar. 2022).

[29] P. A. Philip and M. P. Lutz. “Targeting Epidermal Growth Factor Receptor-Related Signaling Pathways in Pancreatic Cancer”. en. In: *Pancreas* 44.7 (Oct. 2015), pp. 1046–1052.

[30] D. Drygin *et al.* “Targeting RNA polymerase I with an oral small molecule CX-5461 inhibits ribosomal RNA synthesis and solid tumor growth”. en. In: *Cancer Res.* 71.4 (Feb. 2011), pp. 1418–1430.

[31] M. J. Bywater *et al.* “Inhibition of RNA polymerase I as a therapeutic strategy to promote cancer-specific activation of p53”. en. In: *Cancer Cell* 22.1 (July 2012), pp. 51–65.

[32] E. Raymond *et al.* “Oxaliplatin: a review of preclinical and clinical studies”. en. In: *Ann. Oncol.* 9.10 (Oct. 1998), pp. 1053–1071.

[33] M. Q. Mohammed and S. Retsas. “Oxaliplatin is active in vitro against human melanoma cell lines: comparison with cisplatin and carboplatin”. en. In: *Anticancer Drugs* 11.10 (Nov. 2000), pp. 859–863.

[34] L. Pendyala and P. J. Creaven. “In vitro cytotoxicity, protein binding, red blood cell partitioning, and biotransformation of oxaliplatin”. en. In: *Cancer Res.* 53.24 (Dec. 1993), pp. 5970–5976.

[35] B. S. Cummings and R. G. Schnellmann. “Cisplatin-induced renal cell apoptosis: caspase 3-dependent and -independent pathways”. en. In: *J. Pharmacol. Exp. Ther.* 302.1 (July 2002), pp. 8–17.

[36] H.-R. Park *et al.* “Enhanced antitumor efficacy of cisplatin in combination with HemoHIM in tumor-bearing mice”. en. In: *BMC Cancer* 9 (Mar. 2009), p. 85.

[37] X. Wang, J. L. Martindale, and N. J. Holbrook. “Requirement for ERK activation in cisplatin-induced apoptosis”. en. In: *J. Biol. Chem.* 275.50 (Dec. 2000), pp. 39435–39443.

[38] H. Tian and B. N. Cronstein. “Understanding the mechanisms of action of methotrexate: implications for the treatment of rheumatoid arthritis”. en. In: *Bull. NYU Hosp. Jt. Dis.* 65.3 (2007), pp. 168–173.

[39] H. Wang, B. R. Word, and B. D. Lyn-Cook. “Enhanced efficacy of gemcitabine by indole-3-carbinol in pancreatic cell lines: the role of human equilibrative nucleoside transporter 1”. en. In: *Anticancer Res.* 31.10 (Oct. 2011), pp. 3171–3180.

[40] A. Dupré *et al.* “A forward chemical genetic screen reveals an inhibitor of the Mre11-Rad50-Nbs1 complex”. en. In: *Nat. Chem. Biol.* 4.2 (Feb. 2008), pp. 119–125.

[41] D. Li *et al.* “BIBW2992, an irreversible EGFR/HER2 inhibitor highly effective in preclinical lung cancer models”. en. In: *Oncogene* 27.34 (Aug. 2008), pp. 4702–4711.

[42] C. H. Wong *et al.* “Preclinical evaluation of afatinib (BIBW2992) in esophageal squamous cell carcinoma (ESCC)”. en. In: *Am. J. Cancer Res.* 5.12 (Nov. 2015), pp. 3588–3599.

[43] X.-K. Wang *et al.* “Afatinib circumvents multidrug resistance via dually inhibiting ATP binding cassette subfamily G member 2 in vitro and in vivo”. en. In: *Oncotarget* 5.23 (Dec. 2014), pp. 11971–11985.

[44] T. Yoshioka *et al.* “Antitumor activity of pan-HER inhibitors in HER2-positive gastric cancer”. en. In: *Cancer Sci.* 109.4 (Apr. 2018), pp. 1166–1176.

[45] R. Han *et al.* “Amphiphilic dendritic nanomicelle-mediated co-delivery of 5-fluorouracil and doxorubicin for enhanced therapeutic efficacy”. en. In: *J. Drug Target.* 25.2 (Feb. 2017), pp. 140–148.

[46] Q. Zeng *et al.* “Knockdown of NFBD1/MDC1 enhances chemosensitivity to cisplatin or 5-fluorouracil in nasopharyngeal carcinoma CNE1 cells”. en. In: *Mol. Cell. Biochem.* 418.1-2 (July 2016), pp. 137–146.

[47] E. Weisberg *et al.* “Drug resistance in mutant FLT3-positive AML”. en. In: *Oncogene* 29.37 (Sept. 2010), pp. 5120–5134.

[48] R. F. Shearer *et al.* “Functional Roles of the E3 Ubiquitin Ligase UBR5 in Cancer”. en. In: *Mol. Cancer Res.* 13.12 (Dec. 2015), pp. 1523–1532.

[49] Z. Xie *et al.* “Significance of the E3 ubiquitin protein UBR5 as an oncogene and a prognostic biomarker in colorectal cancer”. en. In: *Oncotarget* 8.64 (Dec. 2017), pp. 108079–108092.

[50] K. Dhatchinamoorthy, J. D. Colbert, and K. L. Rock. “Cancer Immune Evasion Through Loss of MHC Class I Antigen Presentation”. en. In: *Front. Immunol.* 12 (Mar. 2021), p. 636568.

[51] E. Alspach, D. M. Lussier, and R. D. Schreiber. “Interferon γ and Its Important Roles in Promoting and Inhibiting Spontaneous and Therapeutic Cancer Immunity”. en. In: *Cold Spring Harb. Perspect. Biol.* 11.3 (Mar. 2019).

[52] E. Song and R. D. Chow. “Mutations in IFN-γ signaling genes sensitize tumors to immune checkpoint blockade”. en. In: *Cancer Cell* 41.4 (Apr. 2023), pp. 651–652.

[53] D. J. Slamon *et al.* “Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene”. en. In: *Science* 235.4785 (Jan. 1987), pp. 177–182.

[54] B. C. Browne *et al.* “HER-2 signaling and inhibition in breast cancer”. en. In: *Curr. Cancer Drug Targets* 9.3 (May 2009), pp. 419–438.

[55] Y. Harada *et al.* “Anti-cancer effect of afatinib, dual inhibitor of HER2 and EGFR, on novel mutation HER2 E401G in models of patient-derived cancer”. en. In: *BMC Cancer* 23.1 (Jan. 2023), p. 77.

[56] T. Takeda *et al.* “Yes1 signaling mediates the resistance to Trastuzumab/Lap atinib in breast cancer”. en. In: *PLoS One* 12.2 (Feb. 2017), e0171356.

[57] Y. Zhao, H. Yu, and W. Hu. “The regulation of MDM2 oncogene and its impact on human cancers”. en. In: *Acta Biochim. Biophys. Sin.* 46.3 (Mar. 2014), pp. 180–189.

[58] A. Ito *et al.* “MDM2-HDAC1-mediated deacetylation of p53 is required for its degradation”. en. In: *EMBO J.* 21.22 (Nov. 2002), pp. 6236–6245.

[59] H. Katayama *et al.* “Phosphorylation by aurora kinase A induces Mdm2-mediated destabilization and inhibition of p53”. en. In: *Nat. Genet.* 36.1 (Jan. 2004), pp. 55–62.

[60] F. Ding *et al.* “UBR5 oncogene as an indicator of poor prognosis in gastric cancer”. en. In: *Exp. Ther. Med.* 20.5 (Nov. 2020), p. 7.

[61] W. Yang *et al.* “Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells”. In: *Nucleic Acids Res.* 41.D1 (Nov. 2012), pp. D955–D961.

[62] F. Iorio *et al.* “A Landscape of Pharmacogenomic Interactions in Cancer”. In: *Cell* 166.3 (July 2016), pp. 740–754.

[63] D. van der Meer *et al.* “Cell Model Passports—a hub for clinical, genetic and functional datasets of preclinical cancer models”. In: *Nucleic Acids Res.* 47.D1 (Sept. 2018), pp. D923–D929.

[64] S. P. Suehnholz *et al.* “Quantifying the Expanding Landscape of Clinical Actionability for Patients with Cancer”. In: *Cancer Discov.* (Oct. 2023).

[65] E. Cerami *et al.* “The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data: Figure 1”. In: *Cancer Discov.* 2.5 (May 2012), pp. 401–404.

[66] J. Gao *et al.* “Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal”. In: *Sci. Signal.* 6.269 (Mar. 2013), l1–pl1.

[67] Y. Benjamini and Y. Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. In: *J. R. Stat. Soc. Series B Stat. Methodol.* 57.1 (Jan. 1995), pp. 289–300.

[68] M. I. Love, W. Huber, and S. Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome Biol.* 15.12 (Dec. 2014).

[69] B. Muzellec *et al.* “PyDESeq2: a python package for bulk RNA-seq differential expression analysis”. In: (Dec. 2022).

[70] M. Ashburner *et al.* “Gene Ontology: tool for the unification of biology”. In: *Nat. Genet.* 25.1 (May 2000), pp. 25–29.

[71] A. Liberzon *et al.* “The Molecular Signatures Database Hallmark Gene Set Collection”. In: *Cell Systems* 1.6 (Dec. 2015), pp. 417–425.

[72] M. Kanehisa. “KEGG: Kyoto Encyclopedia of Genes and Genomes”. In: *Nucleic Acids Res.* 28.1 (Jan. 2000), pp. 27–30.

[73] A. R. Pico *et al.* “WikiPathways: Pathway Editing for the People”. In: *PLoS Biol.* 6.7 (July 2008), e184.

[74] A. Fabregat *et al.* “The Reactome pathway Knowledgebase”. In: *Nucleic Acids Res.* 44.D1 (Dec. 2015), pp. D481–D487.

[75] E. Y. Chen *et al.* “Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool”. In: *BMC Bioinformatics* 14.1 (Apr. 2013).

3.5. SUPPLEMENTARY MATERIALS

3.5.1. SUPPLEMENTARY TABLES

The Supplementary Tables accompanying this chapter are available via the private 4TU data repository [1].

- **Table 3.S1:** Detailed results of every drug found significant over all cancer type-oncogene pairs. The table includes the found drug categories, entropy scores in both cohorts, Oncostratifier score (ΔH), Oncostratifier p-value, mutational marker if found, and additional statistics on oncogene addicted patients for each case, as well as differentially expressed genes for the mutational markers found.
- **Table 3.S2:** Number of Oncostratifier drug and oncogene pairs found for each cancer type and their assigned categories.
- **Table 3.S3:** For each cancer type-oncogene pair, set of stratifying drugs that can cover or sensitize all oncogene addicted cancer cell lines.
- **Table 3.S4:** For each cancer type-oncogene pair, set of stratifying drugs that can cover or sensitize all oncogene addicted cancer cell lines known to be sensitive to at least 1 stratifying drug.

3.5.2. SUPPLEMENTARY FIGURES

The Supplementary Figures are included in the subsequent pages of this document.

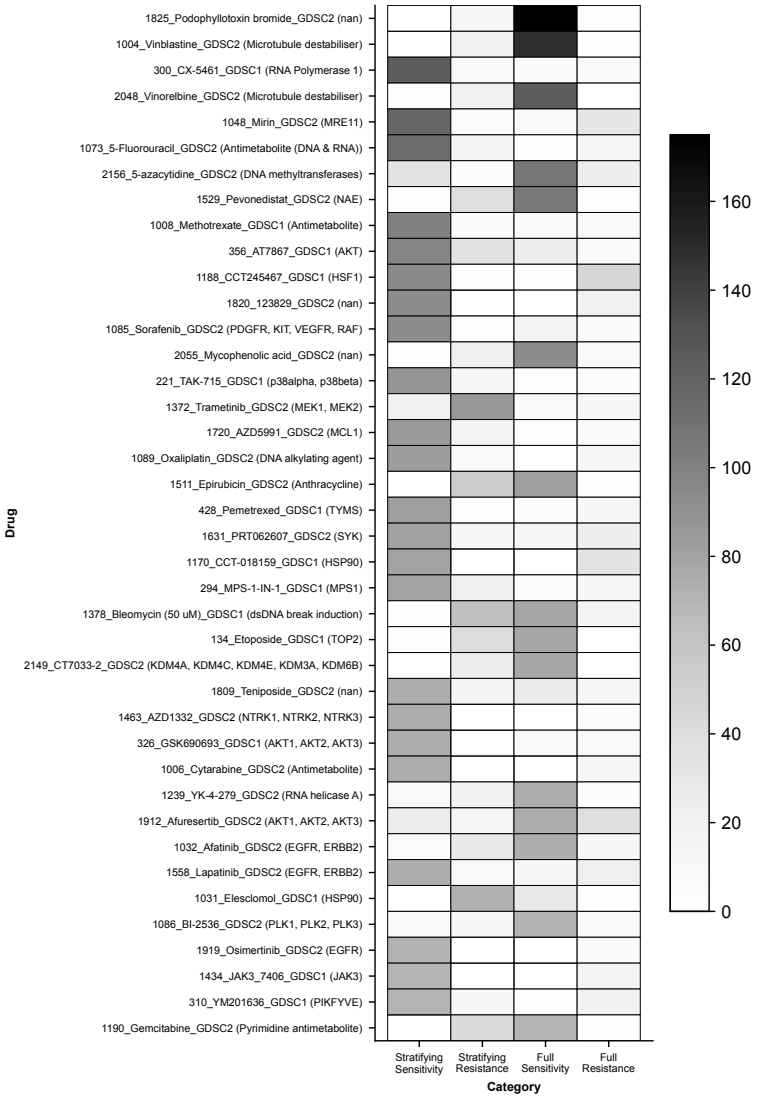


Figure 3.S1: **Top drugs repeatedly found significant for oncogene-cancer type pairs.** Number of times each drug is found significantly in one of the Oncostratifier categories across oncogenes and cancer types. Only the top 40 drugs are shown.

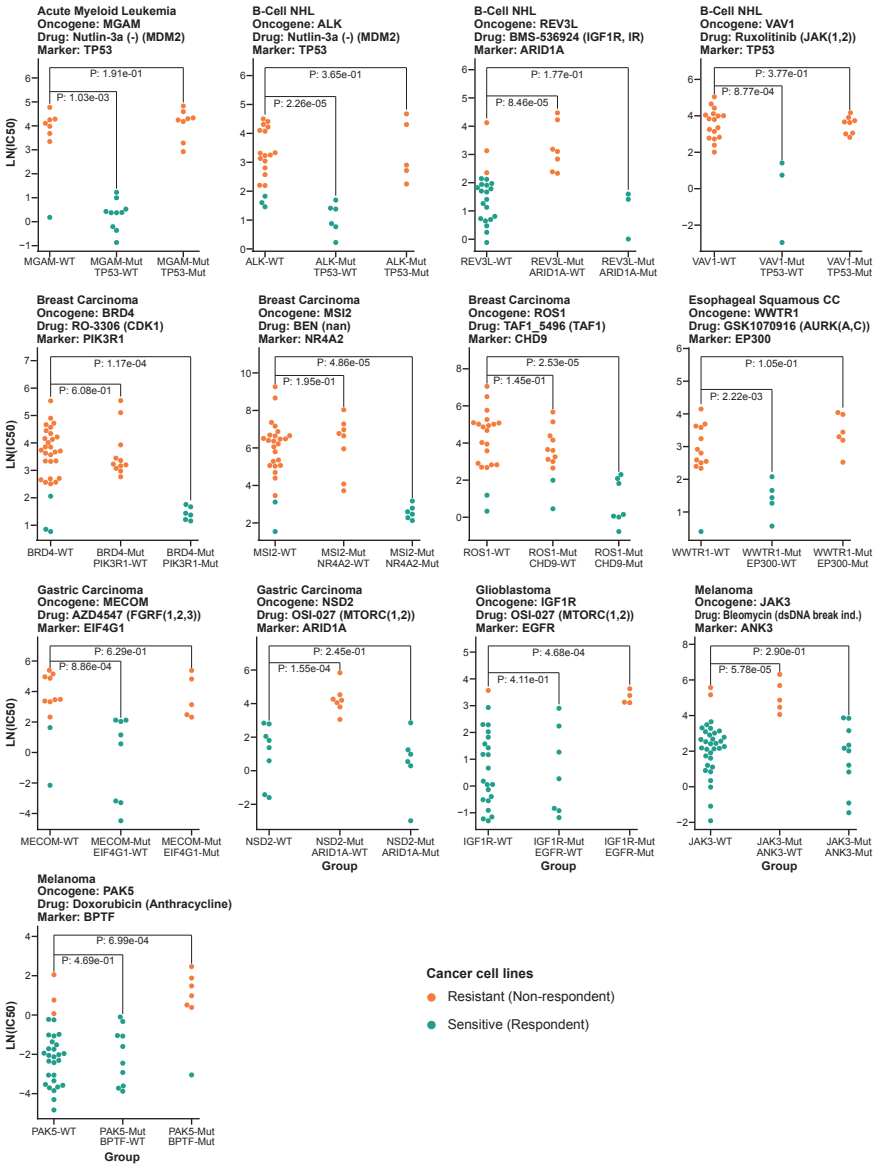


Figure 3.S2: **Additional mutational markers for stratifying drugs.** Distribution of the cancer cell line drug response grouped by oncogene-WT, oncogene-Mut & marker-WT, and oncogene-Mut & marker-Mut for the drugs that are found as stratifying in acute myeloid leukemia, b-cell non-hodgkin leukemia, breast carcinoma, esophageal squamous cell carcinoma, gastric carcinoma, glioblastoma, and melanoma. $LN(IC50)$ scores are used as drug response. The target of each drug is pointed out in the parentheses. We tested the significance of response change in oncogene-Mut & marker-WT and oncogene-Mut & marker-Mut groups against the oncogene-WT group using Fisher's exact test (corrected by the Benjamini-Hochberg [2] procedure).

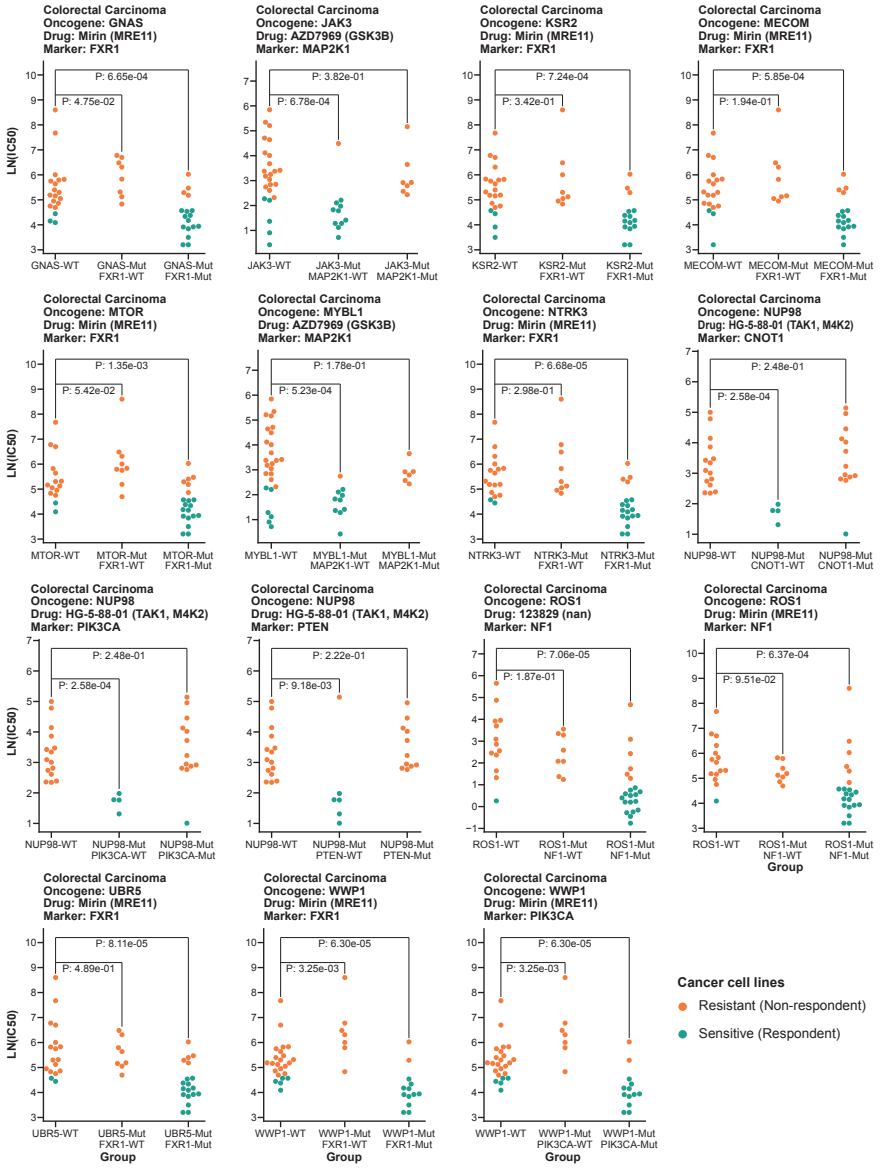


Figure 3.S3: **Additional mutational markers for stratifying drugs in colorectal carcinoma.** Distribution of the cancer cell line drug response grouped by oncogene-WT, oncogene-Mut & marker-WT, and oncogene-Mut & marker-Mut for the drugs that are found as stratifying in colorectal carcinoma. $LN(IC50)$ scores are used as drug response. The target of each drug is pointed out in the parentheses. We tested the significance of response change in oncogene-Mut & marker-WT and oncogene-Mut & marker-Mut groups against the oncogene-WT group using Fisher's exact test (corrected by the Benjamini-Hochberg [2] procedure).

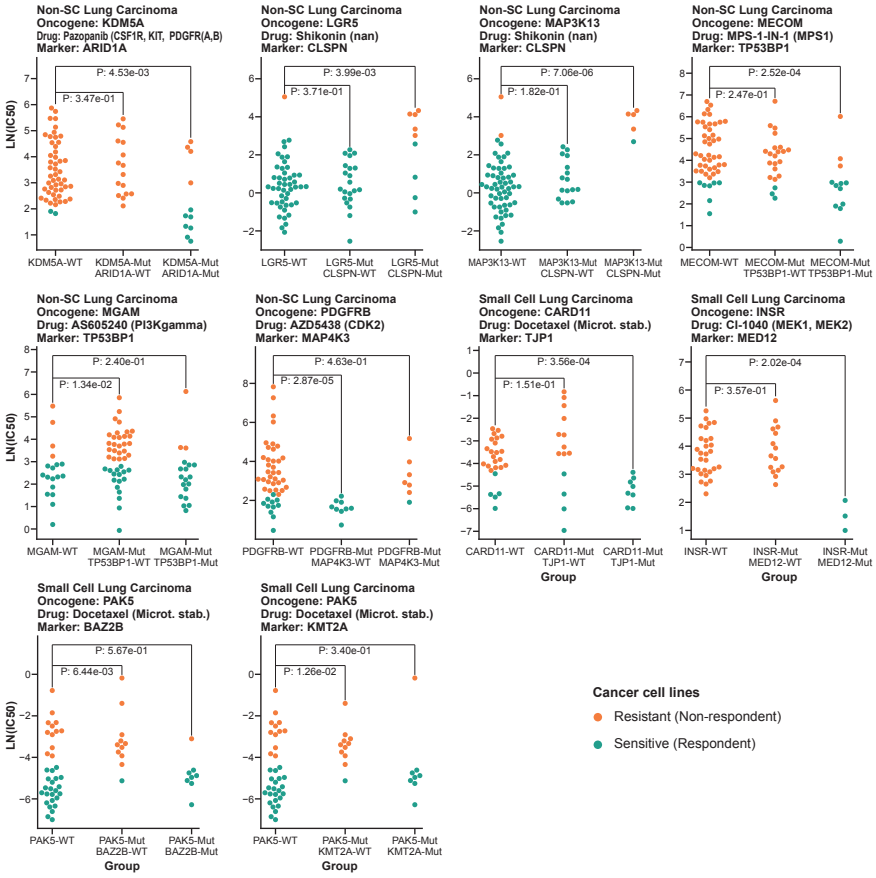


Figure 3.S4: **Additional mutational markers for stratifying drugs in lung carcinoma.** Distribution of the cancer cell line drug response grouped by oncogene-WT, oncogene-Mut & marker-WT, and oncogene-Mut & marker-Mut for the drugs that are found as stratifying in lung carcinoma. $LN(IC50)$ scores are used as drug response. The target of each drug is pointed out in the parentheses. We tested the significance of response change in oncogene-Mut & marker-WT and oncogene-Mut & marker-Mut groups against the oncogene-WT group using Fisher's exact test (corrected by the Benjamini-Hochberg [2] procedure).

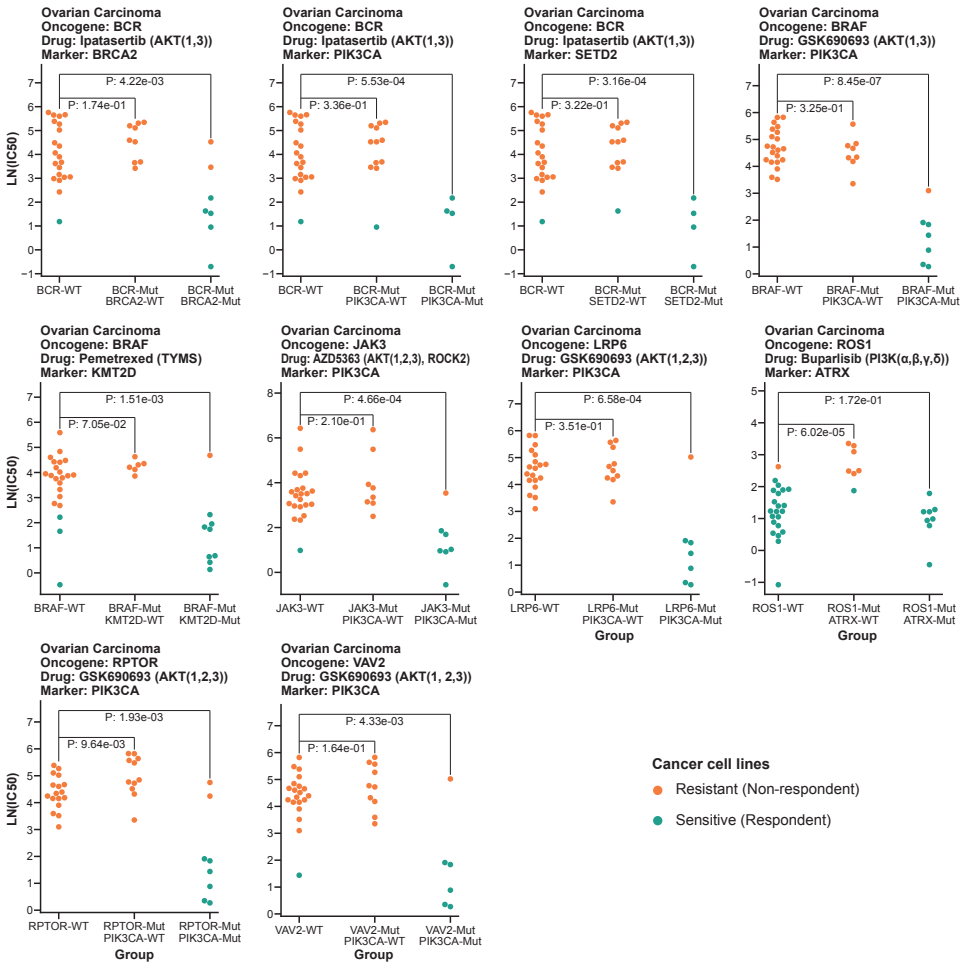


Figure 3.S5: **Additional mutational markers for stratifying drugs in ovarian carcinoma.** Distribution of the cancer cell line drug response grouped by oncogene-WT, oncogene-Mut & marker-WT, and oncogene-Mut & marker-Mut for the drugs that are found as stratifying in ovarian carcinoma. $LN(IC50)$ scores are used as drug response. The target of each drug is pointed out in the parentheses. We tested the significance of response change in oncogene-Mut & marker-WT and oncogene-Mut & marker-Mut groups against the oncogene-WT group using Fisher's exact test (corrected by the Benjamini-Hochberg [2] procedure).

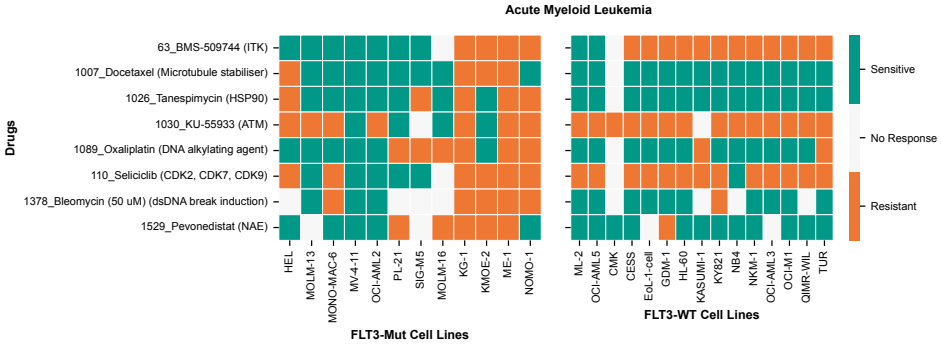


Figure 3.S6: Drug response of cancer cell lines to stratifying drugs found in FLT3 addiction in acute myeloid leukemia divided by FLT3 addicted and FLT3 -WT cohorts

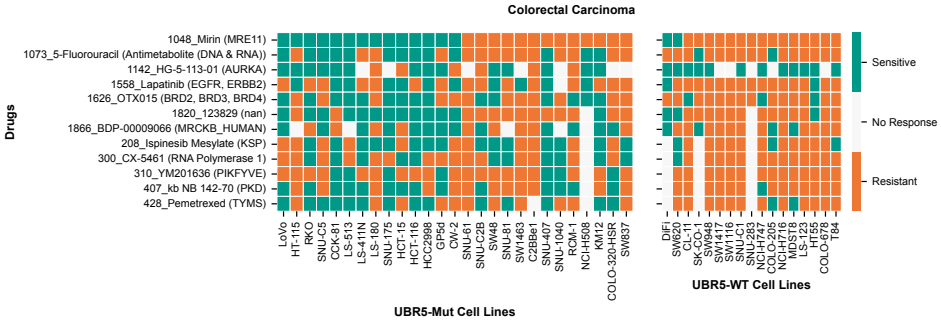


Figure 3.S7: Drug response of cancer cell lines to stratifying drugs found in UBR5 addiction in colorectal carcinoma divided by UBR5 addicted and UBR5-WT cohorts.

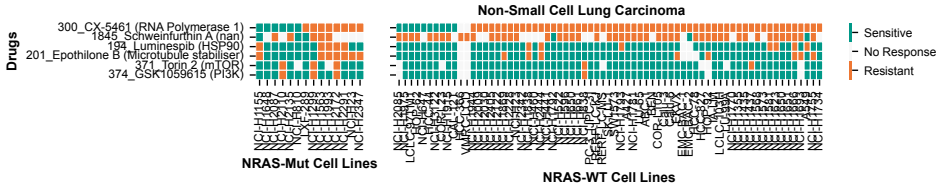


Figure 3.S8: Drug response of cancer cell lines to stratifying drugs found in NRAS addiction in non-small cell lung carcinoma divided by NRAS addicted and NRAS-WT cohorts.

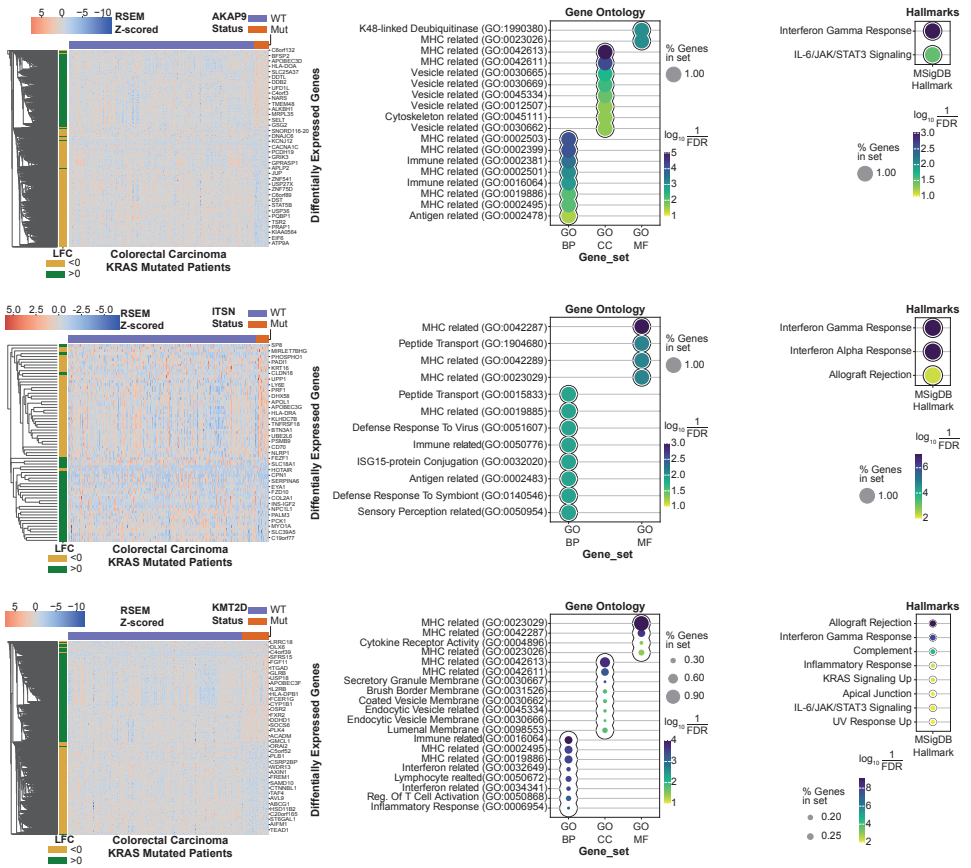


Figure 3.S9: **KRAS-addicted colorectal patients stratified by mutational marker and their respective DGEs.** Differentially Expressed Genes' expression between oncogene addicted patients where marker gene is mutated or WT. Expression values are standardized. Middle column shows the enriched GO Terms and right column shows enriched hallmarks for each set of DGEs.

REFERENCES

[1] Y. Tepeli *et al. Data and code underlying the publication: Stratifying Oncogene-addicted Cohorts by Drug Response*. en. 2024.

[2] Y. Benjamini and Y. Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. In: *J. R. Stat. Soc. Series B Stat. Methodol.* 57.1 (Jan. 1995), pp. 289–300.

4

DCAST: Diverse Class-Aware Self-Training Mitigates Selection Bias for Fairer Learning

Yasin I. Tepeli

Joana P. Gonçalves

This chapter is published in: arXiv (2024), doi: doi.org/10.48550/arXiv.2409.20126 (submitted)
Supplementary material is also available online at: https://doi.org/10.48550/arXiv.2409.20126

Fairness in machine learning seeks to mitigate model bias against individuals based on sensitive features such as sex or age, often caused by an uneven representation of the population in the training data due to selection bias. Notably, bias unascribed to sensitive features is challenging to identify and typically goes undiagnosed, despite its prominence in complex high-dimensional data from fields like computer vision and molecular biomedicine. Strategies to mitigate unidentified bias and evaluate mitigation methods are crucially needed, yet remain underexplored. We introduce: (i) Diverse Class-Aware Self-Training (DCAST), model-agnostic mitigation aware of class-specific bias, which promotes sample diversity to counter confirmation bias of conventional self-training while leveraging unlabeled samples for an improved representation of the underlying population; (ii) hierarchy bias, multivariate and class-aware bias induction without prior knowledge. Models learned with DCAST showed improved robustness to hierarchy and other biases across eleven datasets, against conventional self-training and six prominent domain adaptation techniques. Advantage was largest on multi-class classification, emphasizing DCAST as a promising strategy for fairer learning in different contexts.

4.1. INTRODUCTION

As predictive machine learning (ML) increasingly makes its way to applications with an impact on society, one major concern is to ensure that ML models deliver fair predictions and do not discriminate against individuals in the population. Selection bias is one of the most prominent sources of unfairness in ML, whereby the data used to build ML models is not representative of the real-world and thus violates the fundamental assumption of ML that it is independently drawn and identically distributed to the underlying population.

Research on fairness in ML has focused on mitigating (selection) bias associated with legally protected or sensitive features, such as sex, age, or skin color [1, 2]. However, biases can be indirectly linked to sensitive features via proxies not recognized as sensitive [1, 2], or they can be unrelated to sensitive features and still lead to unfairness. Ultimately, biases are likely to remain undiagnosed and be propagated by ML models without scrutiny when a link to sensitive features is challenging to identify. Unknown biases are often present when data is complex and high-dimensional, data collection is non-random, and knowledge of the domain is incomplete. We argue that unfairness mitigation should thus address bias more generally, beyond what can be ascribed to sensitive features. This issue has deserved attention across fields, including computer vision [3, 4], astronomy [5, 6], biomedicine and healthcare [7–10], finance and economics [11–13], information retrieval [14, 15], and language [7, 16]. Nevertheless, its impact is typically overlooked, resulting in models with optimistic performances due to bias-unaware evaluation. We identify two key areas for improvement, namely evaluation of ML model robustness to bias, and ML bias mitigation.

Evaluation is crucial to ensure that ML models generalize and are robust to bias, but assessing performance on data representative of the real-world distribution is rarely

achievable. Independent test data is not always available or guaranteed to be unbiased, and conventional data splits do not create train-test distribution shifts suitable for model bias evaluation. A viable alternative is to induce bias to the train set and assess the learned model on the original test set. Common bias induction approaches include subsampling using univariate selection probabilities, based on values or the distribution of one feature [17, 18]. This is however not representative of multivariate biases typically present in complex high-dimensional data. Existing methods to induce multivariate bias include: joint bias [19], which favors the selection of samples closer to the mean; and Dirichlet bias, [20], which assigns sample selection likelihoods based on a Dirichlet distribution. Both methods ignore class labels and thus do not generate class-specific biases. They might also cause class imbalances for otherwise balanced data.

We propose hierarchy bias, a multivariate class-aware bias induction technique to produce complex class-specific biases. Hierarchy bias identifies distinctly distributed groups of samples in the original data using clustering, and then generates a biased selection by influencing the representation of one group of samples relative to the others. Selection is performed per class to induce class-specific bias, aiming for an identical number of samples per class to ensure class balance.

Methods to mitigate bias in ML generally fall in the scope of domain adaptation (DA, [21]), seeking to adapt a model to the distribution shift between the source training domain and a target prediction domain. Relevant DA categories span importance weighting, subspace alignment, inference-based, and semi-supervised learning methods. Importance weighting (IW) weighs training samples based on their relevance to the test set, using probability ratios or discrepancy measures [6, 11, 19, 22–28]. Since IW assumes that the train set contains the support of the test set and most features contribute to the prediction, it can be less effective with high-dimensional data or small sample sizes. Subspace alignment (SA) transforms the data representation [29–31], assuming there is a common subspace where transformed train and test sets exhibit matching conditional probabilities, which may be difficult to optimize if many transformations fit. Inference-based (IB) methods include minimax estimation [20, 32], where loss minimization is coupled with an adversarial maximization objective that steers the model to fit more conservatively, aiming for improved generalization. The IB methods may underperform if the model choice is less suitable for the test set. Overall, most IW, SA, and IB methods adapt the model for one target test set, which can hamper generalizability. Semi-supervised learning (SSL) leverages unlabeled samples to provide model learning with insight into the underlying population distribution. The most benefit can in principle be achieved by using as much unlabeled data as available, though some SSL approaches still adapt to individual test sets [33, 34]. Unlabeled samples are typically incorporated by SSL using self-training (ST) [35] or co-training (CT) [36], which assigns predicted pseudo-labels to unlabeled samples and selects a subset of these to include at each training iteration. Sample selection is often based on prediction confidence according to the model trained thus far, which may strengthen existing bias or create other biases such as class imbalance

for originally balanced data [4, 5]. Attempts to mitigate this behavior include, for instance, the P3SVM support vector machine (SVM) [4] that selects pseudo-labeled samples distant from each other and located within the margins furthest away from the decision boundary. This method is however SVM-specific, and its sample selection dependent on the size of the margin may limit the contribution of unlabeled data. In summary, most DA methods mitigate distribution shifts for one test set at a time, leading to ML models with limited generalizability beyond the train and test domains. It remains to be investigated if generalization could be improved by training on additional unlabeled data. Semi-supervised learning offers this possibility, but existing methods fall short in actively mitigating bias present in the data or further induced during model learning. Finally, many DA methods are model-specific and cannot be applied to different types of ML models.

To improve bias mitigation, we propose Diverse Class-Aware Self-Training (DCAST), a model-agnostic semi-supervised learning framework that gradually incorporates unlabeled data in a class-aware manner, guided by two active bias mitigation strategies. The core CAST strategy addresses class-specific bias by selecting a set of pseudo-labeled samples to include separately per class, using a relaxed confidence threshold, with options to preserve the class ratios of the original labeled train set or to add the same number of pseudo-labeled samples per class at each iteration. The extended DCAST strategy seeks to counter confidence-induced bias by further selecting diverse pseudo-labeled samples, as measured by inter-sample distances in the learned discriminative embedding or the original feature space.

We evaluate both hierarchy bias induction and (D)CAST bias mitigation across eleven datasets, against competing approaches including Dirichlet and joint bias as well as conventional self-training and six domain adaptation techniques. Specifically, we investigate which bias induction method induces the most challenging type of selection bias, leading to the strongest impact on ML model prediction performance. We further assess to what extent the class-awareness and diversity in (D)CAST improve robustness to bias, both across datasets and compared to the alternative bias mitigation strategies, while coupling model-agnostic (D)CAST with three types of ML models.

4.2. RESULTS AND DISCUSSION

The proposed hierarchy bias induction and (D)CAST bias mitigation methods aim to provide, respectively: (i) a more realistic type of class-aware multivariate selection bias for the evaluation of ML model robustness to bias, and (ii) class-aware and diversity-guided strategies to learn ML models with improved generalizability in the presence of selection bias. We briefly introduce these techniques and discuss their evaluation across 11 datasets using logistic regression (LR), random forest (RF), and 2-hidden layer neural network (NN) prediction models. Every dataset was randomly partitioned into 80% train and 20% test, with the test data reserved for prediction model evaluation (Methods). Effects of bias induction on the data and model prediction performance were assessed over 30 runs, each relying on a random split

of the train set into labeled (30%) and unlabeled (70%) train sets. The labeled train set was used for bias induction and for training ML models, either intact or upon bias induction. For bias mitigation, unlabeled data was additionally used during training, where conventional self-training (ST) and (D)CAST leveraged the unlabeled train set, and other domain adaptation techniques exploited the unlabeled test set instead (Methods).

4.2.1. Hierarchy bias induces effective multivariate and class-specific selection bias

Hierarchy bias generates a biased selection of samples for a given dataset, aiming to deviate from the original data distribution by skewing the representation of a group of samples that is deemed closer together in feature space than the remaining samples (Fig. 4.1). The approach selects k samples per class and controls group representation using bias ratio b as follows. A class-specific group of at least k closely related samples is first identified using agglomerative hierarchical clustering. To obtain the biased selection, $k \times b$ samples are chosen uniformly at random from the identified group and $k \times (1-b)$ samples are chosen uniformly at random from the remaining samples (Methods).

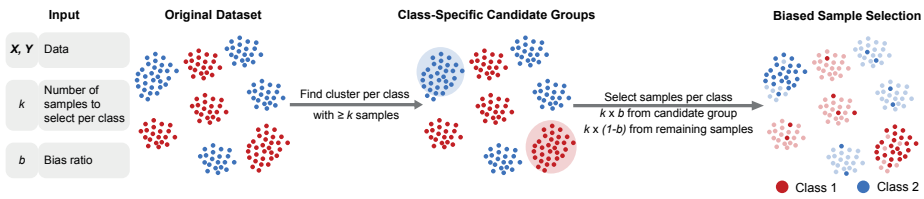


Figure 4.1: **Hierarchy bias approach for induction of selection bias.** Given input data $\boldsymbol{X}$ with labels $\boldsymbol{Y}$, number of samples to select k, and bias ratio $b \in [0,1]$, hierarchy bias selects k samples per class c: $k \times b$ from a specific group and $k \times (1-b)$ from the remaining samples. Each class-specific candidate group (for class c) is identified via agglomerative hierarchical clustering with Euclidean distances and Ward linkage of the c-labeled samples until a cluster of size $\geq k$ is obtained, from which $k \times b$ samples are drawn uniformly at random. The $k \times (1-b)$ samples are drawn uniformly at random from the remaining c-labeled samples.

To evaluate bias induction, we assessed the ability to generate a distribution shift between the biased selection and the original data, as well as the impact of the induced shift on ML model prediction performance. We compared hierarchy bias with $b = 0.9$ to random subsampling and two alternative bias induction techniques: joint bias [19] and Dirichlet bias [20]. Hierarchy bias and random subsampling were set to select 30 samples per class, whereas Dirichlet targeted 60 and 300 samples in total respectively for binary and multiclass labeled datasets. Note that Dirichlet and joint bias do not take class labels into account when performing their selection, and joint bias does not allow control over the selected number of samples.

Effect on data distribution. We first assessed the effect of bias induction on the distribution of distances between samples. The underlying idea is that a biased selection would exclude portions of the original data that deviate from the rest of

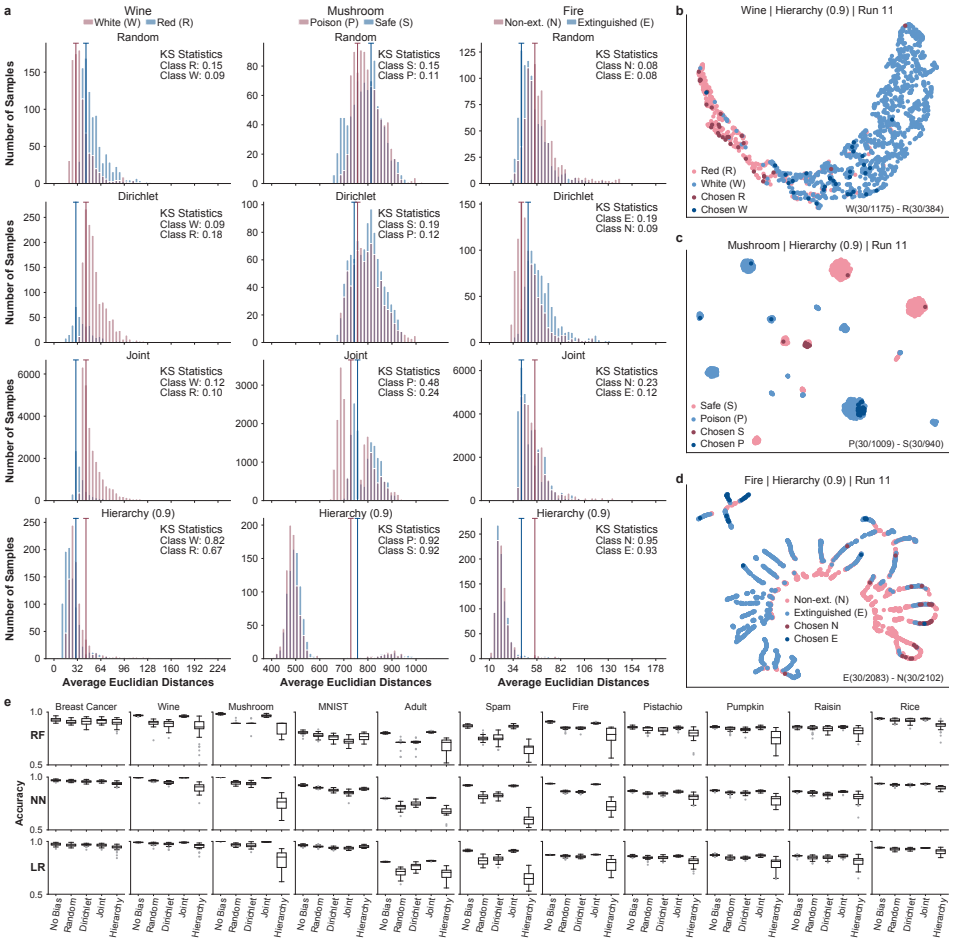


Figure 4.2: **Bias induction impact on sample distances, latent space, and classifier performance. (a)** Class-specific distributions of per sample average Euclidean distances to all other samples, for the biased selection (histograms) and for all samples in the labeled train set (histogram peaks denoted by lines ending in a "T" shape), using three bias induction techniques (hierarchy with $b = 0.9$, joint, and Dirichlet) and random subsampling on three datasets (wine, mushroom, and fire). Kolmogorov-Smirnov (KS) effect sizes quantify the distribution shift between the biased selection vs. all samples. **(b-d)** Samples selected by hierarchy bias ($b = 0.9$), highlighted on the respective latent UMAP space of the labeled train set for the wine, mushroom, and fire datasets (arbitrarily chosen run 11). **(e)** Accuracy of supervised RF, NN, and LR models on the test set after training on the original or biased labeled train set, over 30 distinct train runs. Box height delimits the interquartile range ($IQR = Q3 - Q1$), with a line across the box denoting the median; whiskers indicate the largest and smallest values within $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$, with points beyond the range as outliers.

the samples to some extent, thus making inter-sample distances closer on average. For each dataset, we obtained class-specific distributions of the per sample average Euclidean distance to all other samples. We further quantified the deviation between the class-specific distance distributions obtained for the biased selection and the original labeled set using Kolmogorov-Smirnov (KS) tests. Hierarchy bias ($b = 0.9$) induced the most significant shift in the distance distributions for all 11 datasets (KS

effect sizes > 0.65, p-values < 0.05; Fig. 4.2a and Supplementary Fig. 4.S1-4.S2), and primarily towards smaller average inter-sample distances, in line with the selection of close samples that hierarchy bias is designed to produce. Random selection resulted in the most similar distance distributions to the original data, with the smallest KS effect for 8 datasets. Dirichlet and joint bias led to modest shifts than hierarchy bias, with joint bias generally showing larger KS effects than Dirichlet (9 of 11 datasets). We also examined the samples selected from each labeled train set in the feature space, reduced to 2 dimensions (2D) using Uniform Manifold Approximation and Projection (UMAP) for an example run 11. Hierarchy bias selected samples from specific clusters or regions of the feature space. This was apparent across datasets (Supplementary Fig. 4.S3), for instance hierarchy bias ignored samples in the top right area of the 2D space for the wine dataset (Fig. 4.2b), selected from specific clusters of the mushroom dataset (Fig. 4.2c), and focused on the top left and bottom right areas of the 2D space for the fire dataset (Fig. 4.2d). In contrast, samples selected by random selection, as well as by the Dirichlet and joint biases, were spread throughout the 2D space and thus more representative of the original labeled train set for all datasets (Supplementary Fig. 4.S4k-4.S6). For random sampling, this was expected, given that no particular bias was introduced. For joint bias the result was also unsurprising, seeing that it selected the largest proportions of samples across datasets and thus captured most of the data (overall mean average 63%, minimum 44%, and maximum 80%; for hierarchy bias: 17%, 0.4%, and 67%; Supplementary Table 4.S1).

Impact on prediction performance. We evaluated the impact of bias induction on the classification accuracy of supervised ML models for the 11 datasets across 30 runs. Per run, we trained 2-hidden layer neural network (NN), random forest (RF), and logistic regression (LR) models using the original labeled train set (No Bias) or a selection of its samples. The latter was obtained either by random subsampling or using Dirichlet, joint, or hierarchy bias induction. All models were evaluated on the original test set. The induced bias led to a decrease in accuracy with every technique except joint bias (Fig. 4.2e), which as previously mentioned selected most of the original samples and thus did not induce particularly strong bias. Hierarchy bias caused the largest decrease in accuracy for all datasets except MNIST, where the most impact was seen with joint bias (Fig. 4.2e). Note that the preset targets on the number of samples to select for hierarchy bias, Dirichlet bias, and random selection led these methods to select 64-70% of the MNIST samples per class compared to 46-60% with joint bias. This larger coverage of the original data likely influenced the ability of hierarchy and Dirichlet to produce a more effective biased selection for MNIST. Overall, hierarchy bias consistently selected samples in close proximity, leading to a significant shift in inter-sample distances and a bias towards class-specific parts of the original distribution. This caused a marked decrease in prediction accuracy of supervised ML models relative to other bias induction techniques.

4.2.2. DIVERSE CLASS-AWARE SELF-TRAINING (DCAST) FOR SELECTION BIAS MITIGATION

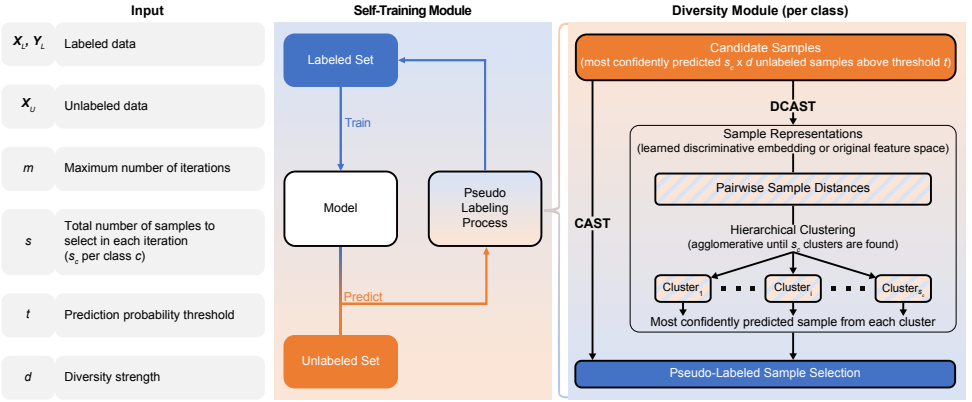


Figure 4.3: **Diverse Class-Aware Self-Training (DCAST) framework.** (Left) Input to DCAST. Labeled data $\boldsymbol{X_L}$ (with labels $\boldsymbol{Y_L}$) and unlabeled data $\boldsymbol{X_U}$, maximum number of iterations m, number of pseudo-labeled samples s to select per iteration, confidence or prediction probability threshold $t \in [0,1]$, and integer diversity strength parameter $d \geq 1$. (Middle) Self-training module. At each iteration, a model trained with labeled samples is used to predict pseudo-labels for unlabeled samples, from which a subset is newly selected and added to the labeled set for the next iteration. (Right) Diversity module. Selects the subset of $s_c = s \times class\_ratio(c)$ confidently predicted and diverse pseudo-labeled samples per class c, as follows: (i) select the top $s_c \times d$ samples from the unlabeled set with confidence or prediction probability larger than t (or $1.2/C$, whichever is largest); and (ii) reduce this $s_c \times d$ selection to a set of s_c diverse samples by identifying s_c clusters using hierarchical clustering (agglomerative single-linkage) and selecting the most confidently predicted sample from each cluster. Note that $class\_ratio$ can otherwise be fixed to be equal across classes. Distance between samples is based on either learned discriminative embeddings, relating samples with respect to prediction output, or alternatively an unsupervised embedding or the original feature space. When $d = 1$, DCAST becomes CAST, without the diversity strategy.

The proposed (D)CAST semi-supervised learning strategies (Fig. 4.3) aim to mitigate selection bias by leveraging insight from unlabeled data about the underlying distribution of the population. Both rely on self-training to gradually incorporate unlabeled data: at each training iteration, the learnt model is used to predict pseudo-labels for all unlabeled samples, from which a subset of s samples (s_c per class) is selected to be included in the labeled set for the next iteration. To address class-related bias, sample selection is done separately per class as follows. First, a set of $s \times d$ candidates is selected as the most confidently predicted samples with prediction probability above a threshold t, where s and d denote the number of samples to select and diversity strength. For CAST ($d = 1$), this directly results in the final set of s pseudo-labeled samples to add for the next iteration. The DCAST selection ($d > 1$) extends upon CAST to mitigate confidence-related bias through sample diversity, reducing the set of $s \times d$ candidates to a final set of s diverse pseudo-labeled samples. Capturing diverse sample groups is achieved via hierarchical clustering of the candidate samples into s clusters (s_c per class), followed by selection of diverse samples comprising the most confidently predicted sample per cluster. To ensure (D)CAST remains model-agnostic, sample distances for clustering can be based on discriminative embeddings learnt by the model or the original feature space.

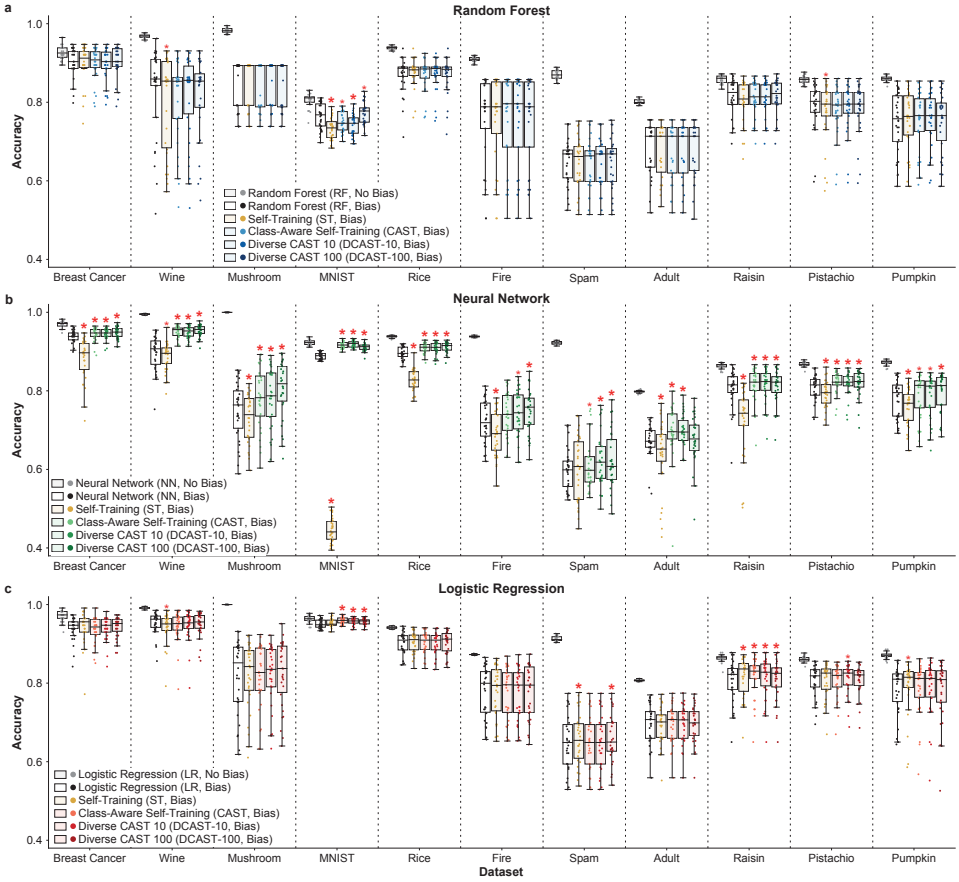


Figure 4.4: **Bias mitigation by semi-supervised (D)CAST in the presence of hierarchy bias (ratio $b = 0.9$).** Accuracy of supervised and semi-supervised learning methods with **(a)** RF, **(b)** NN, and **(c)** LR models across 11 datasets. Results for 30 runs: each training on a different split of the train set into labeled and unlabeled sets, all evaluated on the same original test set. Models included (top to bottom): supervised RF/NN/LR models trained on the original (No Bias) or biased (Bias) labeled set; and semi-supervised RF/NN/LR models, using conventional self-training (ST) on the biased labeled train set plus the unlabeled test set, or (D)CAST on the biased labeled train set plus the unlabeled train set. Red asterisks (*) denote statistically significant changes in accuracy over 30 runs for each semi-supervised approach compared to supervised learning on the biased labeled set, using one-sided Wilcoxon signed-rank tests (larger asterisks indicate $p < 0.01$ and smaller asterisks $0.01 < p < 0.05$).

4.2.3. Diversity and class-awareness in (D)CAST improve bias mitigation via self-training

To evaluate (D)CAST bias mitigation, we first assessed its test prediction accuracy against supervised learning and conventional self-training (ST) [35] on the biased labeled train set, with additional unlabeled samples for self-training strategies. Training and evaluation were performed for 11 datasets over 30 runs as previously described, using RF, NN, and LR models. We induced hierarchy bias with ratio $b = 0.9$, as this type of selection bias showed the most impact on supervised models compared to Dirichlet and joint bias (Fig. 4.2e). The (D)CAST method was assessed

without diversity (CAST, $d = 1$) or with diversities $d = \{10, 100\}$ (CAST-10, DCAST-100), and was set to include $s = 3\times$(number of classes) pseudo-labeled samples per iteration, for at most $m = 100$ iterations, using prediction threshold $t = 0.9$ (or the 85th or 93rd percentile in the case of RF models). Conventional ST selected the $3\times$(number of classes) most confidently predicted samples per iteration (Methods, Bias mitigation strategies). Concerning the mitigation of hierarchy bias with ratio $b = 0.9$, with NN models the semi-supervised (D)CAST strategies significantly improved generalizability over supervised learning across all 11 datasets ($p < 0.05$ with one-sided Wilcoxon signed-rank tests, Fig. 4.4b). Specifically, class-awareness with moderate diversity (DCAST-10) was significantly better than supervised learning on the 11 datasets, whereas class-awareness alone (CAST) or coupled with stronger diversity (DCAST-100) both improved on 10 datasets and remained comparable respectively on the fire and adult datasets. By contrast, conventional ST was significantly worse than supervised learning on 10 datasets with NN models. Using RF and LR models, mitigation of hierarchy bias with ratio $b = 0.9$ was more modest. Semi-supervised (D)CAST and ST performed comparably to supervised learning on most datasets (8 with RF and 7 with LR models; Fig. 4.4a,c), possibly due to the use of regularization, which could hamper model adaptation. We thus saw occasional statistically significant changes and smaller effect sizes with RF and LR models. Notably, the higher diversity strategy DCAST-100 led to the only significant improvement of semi-supervised over supervised learning using RF models, on the MNIST dataset (Fig. 4.4a). Also with RF models, CAST and DCAST-10 decreased accuracy on MNIST, while ST decreased accuracy on 3 datasets (wine, MNIST, and pistachio; Fig. 4.4a). With LR models, (D)CAST strategies improved over supervised learning on 4 datasets (MNIST, spam, raisin, and pistachio), whereas ST improved on 3 datasets (spam, raisin, and pumpkin) but also caused a decrease on the wine dataset (Fig. 4.4c).

Experiments with alternative bias induction techniques revealed similar findings, where (D)CAST bias mitigation consistently outperformed ST across datasets under random subsampling (Supplementary Fig. 4.S7), and under induced Dirichlet or joint bias (Supplementary Fig. 4.S8-4.S9). Again, we saw the largest performance differences with NN models, coinciding with the most improvement of (D)CAST and weakest results of ST over supervised learning.

In summary, (D)CAST effectively mitigated selection bias induced by different techniques when paired with non-regularized NN models, and was not outperformed by supervised learning or conventional ST with regularized RF and LR models. In contrast, conventional ST struggled to recover from the bias with all three types of models, especially NNs. These results suggest that the class-awareness and diversity features introduced to the pseudo-labeling procedure in (D)CAST provide a promising semi-supervised learning strategy to mitigate selection bias.

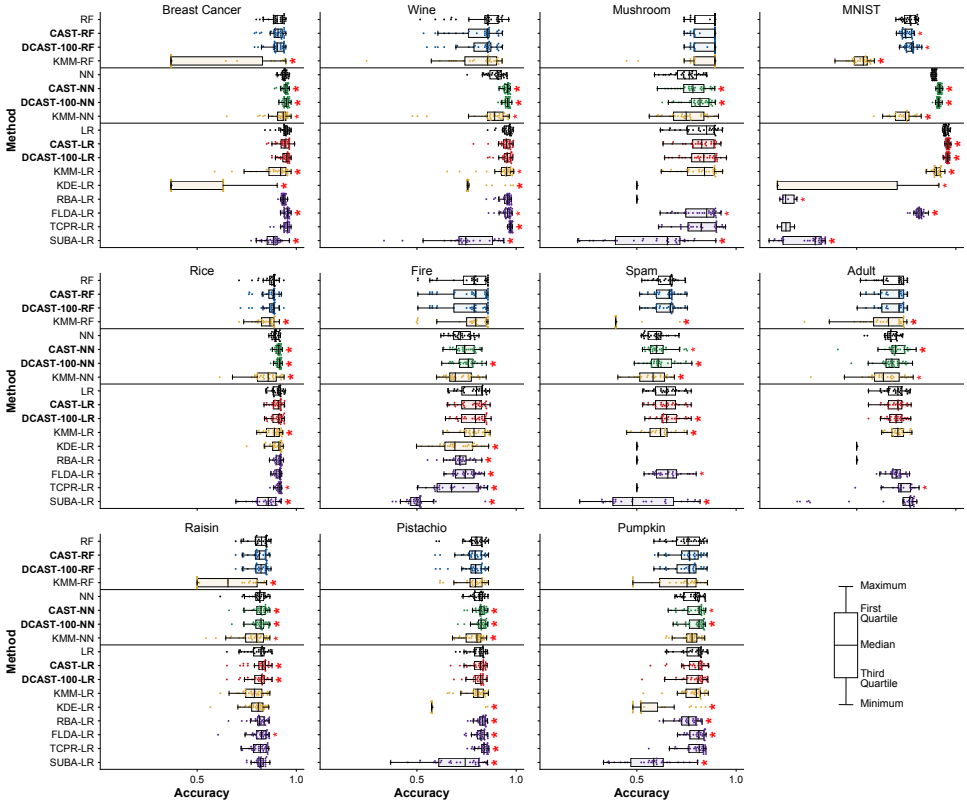


Figure 4.5: **Bias mitigation by (D)CAST or domain adaptation beyond semi-supervised learning under hierarchy bias ($b = 0.9$).** Accuracy of semi-supervised (D)CAST strategies against alternative bias mitigation techniques with 3 different types of ML models for 11 datasets over 30 runs. Per run, each model was trained using a different labeled train set with induced hierarchy bias. We included a supervised learning model as baseline per ML model type (RF, NN, LR), together with bias mitigation models incorporating additional unlabeled samples from either the unlabeled train set ((D)CAST) or the unlabeled test set (remaining methods). All models were evaluated on the same original test set. Bias mitigation methods per category: semi-supervised (CAST and DCAST-100); importance weighting (KMM, KDE); minimax estimation (RBA, TCPR); and subspace alignment (FLDA, SUBA). The (D)CAST and KMM methods were coupled with RF, NN, and LR models, whereas the remaining methods used LR models only. For clarity, horizontal lines group bias mitigation strategies by model type. The "x" symbol indicates model training was unsuccessful across all 30 runs.

4.2.4. SEMI-SUPERVISED (D)CAST BIAS MITIGATION IS SUPERIOR TO COMPETING DOMAIN ADAPTATION

We also evaluated (D)CAST against bias mitigation techniques beyond semi-supervised learning. This included importance weighting methods KMM [19] and KDE [22], minimax approaches RBA [20] and TCPR [32], and subspace alignment methods FLDA [31] and SUBA [30]. All methods were trained on the biased labeled train set and evaluated on the original test set, with (D)CAST further incorporating samples from the unlabeled train set and the remaining methods using unlabeled test samples during training. The (D)CAST and KMM approaches were coupled with RF, NN, and LR models, while the remaining methods used LR

only as per the original work.

Similar to our previous findings, CAST and DCAST-100 were the most robust bias mitigation methods. Overall, these strategies preserved or significantly improved over the supervised learning performance across the 3 model types and 11 datasets, with the exception of CAST showing a decrease in accuracy for MNIST when used with RF models. (Fig. 4.4-4.5). In contrast, KMM led to significant decreases in accuracy for 8 datasets with NN models, as well as for 5 and 6 datasets respectively with LR and RF models. As for the remaining bias mitigation methods using only LR models, KDE resulted in significant decreases in performance for all except the rice dataset. Apart from an improvement with RBA for the pistachio dataset, the RBA and SUBA methods degraded performance significantly for 6 and 9 datasets, respectively. The best competing methods were FLDA and TCPR, which showed significant improvements respectively for 5 and 4 datasets (FLDA: breast cancer, spam, raisin, pistachio, and pumpkin; TCPR: wine, rice, adult, and pistachio). The FLDA approach also led to significant decreases for 4 datasets (wine, mushroom, MNIST, and fire), while TCPR caused a significant decrease for the fire dataset. Concerning the MNIST dataset, TCPR failed to build models for most runs and caused a clear performance drop for the few remaining ones, resulting in insufficient power to determine statistical significance. Overall, CAST and DCAST-100 demonstrated consistent ability to match or outperform supervised learning in the presence of hierarchy bias compared to other bias mitigation methods. The gap was most evident on the multi-class classification problem (MNIST), where the other methods resulted in drastic decreases in performance.

4.3. CONCLUSION

We put forth two contributions to improve the learning of prediction models in the presence of selection bias. First, a bias induction approach termed hierarchy bias to enable the evaluation of complex multivariate bias effects on the generalizability of prediction models. Second, a model-agnostic semi-supervised learning framework named (D)CAST that exploits unlabeled data in a class-aware manner and promotes sample diversity to mitigate selection bias.

Hierarchy bias uses clustering to isolate one distinct group of samples per class and then skews the representation of such group during sample selection to induce class-specific multivariate bias, allowing control over the level of bias through a bias ratio parameter. Induced hierarchy bias showed a stronger impact on the distribution of inter-sample distances and proved more challenging for prediction models to overcome, compared to joint and Dirichlet bias.

The (D)CAST model learning strategy progressively incorporates unlabeled samples using self-training, which is further made class-aware in CAST by pseudo-labeling confidently predicted unlabeled samples over a given threshold per class. Its extended variant, DCAST, seeks to counter confidence-associated bias with sample diversity by clustering and selecting pseudo-labeled samples from distinct groups,

using distances based on either the discriminative embeddings provided by the underlying model or the original feature representation.

Both class-awareness and diversity proved effective, leading to significant improvements in the bias mitigation ability of (D)CAST over conventional self-training across datasets and bias induction techniques. Models trained by (D)CAST also outperformed other models built using six alternative domain adaptation methods, comprising different importance weighting, minimax estimation, and subspace alignment approaches.

Diversity strength was shown to influence the extent of (D)CAST bias mitigation, where a larger value resulted in improved robustness to selection bias. More generally, we recommend setting the diversity strength parameter such that the number of candidate samples considered for selection at each iteration is significantly larger than the number of samples to select. We further suggest choosing a number of samples to select per iteration comfortably below the size of the training set to promote a gradual adaptation of the model, but not too small so that the added samples can have an impact: a possible choice could be the closest even number to $\lfloor\sqrt{N}\rfloor$, with N denoting the size of the training set. The confidence threshold can be adjusted according to the distribution of prediction probabilities of the model to allow (D)CAST to consider at least as many samples as the number to add at each iteration.

We demonstrated that (D)CAST is model-agnostic through application with random forests (RF), neural networks (NN), and logistic regression (LR) models. The success of bias mitigation differed across architectures, with the most benefit achieved using NN models. We hypothesized that the use of regularization could also have played a role, by restricting model adaptation and thus limiting the contribution of unlabeled samples in the RF and LR models. Further investigation would be needed to obtain conclusive evidence.

Overall, our results present (D)CAST and hierarchy bias as promising strategies to improve the learning and evaluation of machine learning models in the presence of selection bias, as an essential step in striving towards fairness in machine learning.

4.4. METHODS

4.4.1. HIERARCHY BIAS INDUCTION AND (D)CAST BIAS MITIGATION

Notation. We denote the input data (sample × feature) matrix as $\boldsymbol{X} \in \mathbb{R}^{N\times F}$, the input label matrix as $\boldsymbol{Y} \in \{0,1\}^{N\times C}$, and output prediction probability matrix as $\tilde{\boldsymbol{Y}} \in \mathbb{R}^{N\times C}$, where N is the number of samples, F is the number of features, and C is the number of classes. Following this notation, $\boldsymbol{x}_n \in \mathbb{R}^{1\times F}$ is the feature vector of sample $n \in \{1,2,...,N-1,N\}$, y_n^c is the binary label of sample n for class $c \in \{1,2,...,C-1,C\}$ (1 if assigned, 0 otherwise), and $\tilde{y}_n^c$ is the prediction probability of sample n being of class c where $\sum_{c=1}^{C} y_n^c = 1$ and $\sum_{c=1}^{C} \tilde{y}_n^c = 1$.

HIERARCHY BIAS

Hierarchy bias induction generates a biased selection of samples from a given dataset in a class-aware and multivariate manner. The idea is that the samples belonging to each class in the dataset can be seen as originating from a mixture of multivariate distributions. Based on this, the goal is to identify one of the mixtures and then make a skewed selection of samples by controlling the representation of the target mixture over the remaining samples. Hierarchy bias induction takes as input a data matrix $\boldsymbol{X}$, a label matrix $\boldsymbol{Y}$, a parameter k denoting the number of samples to select per class, and a bias parameter $b \in [0,1]$ denoting the ratio of samples that should be selected from the identified mixture (Alg. 1). The output is a biased selection of samples, generated as follows. Agglomerative hierarchical clustering is first applied to identify a mixture of interest per class c, corresponding to a cluster of at least k samples. We perform the clustering for class c using all samples from matrix $\boldsymbol{X}$ labeled with class c, with Euclidean inter-sample distances on the original feature vectors and Ward linkage between clusters (Alg. 1, lines 4-5). Once the cluster is identified, the final biased selection is obtained by choosing $k \times b$ samples uniformly at random from the cluster and choosing another $k - k \times b$ samples uniformly at random from the remaining samples not in the cluster (Alg. 1, lines 6-8).

Algorithm 1 Hierarchy Bias

Require: $\boldsymbol{X}$, $\boldsymbol{Y}$, k, b.
Ensure: $Selection \leftarrow \emptyset$
1: $k_{cluster} \leftarrow k \times b$
2: $k_{rest} \leftarrow k - k \times b$
3: **for** `each class` $c \in C$ **do**
4: Apply agglomerative clustering with Euclidean distance and Ward linkage to $\boldsymbol{X}_{S_c}$, $S_c = \{n : n \in y_n^c == 1\}$.
5: $Cluster \leftarrow$ Set of samples from the first cluster that reaches a number of samples $\geq k$.
6: $S_{cluster} \leftarrow$ Select set of $k_{cluster}$ samples uniformly at random from $Cluster$.
7: $S_{rest} \leftarrow$ Select set of k_{rest} samples uniformly at random from the remaining samples (not in $Cluster$).
8: $Selection \cup S_{cluster} \cup S_{rest}$
9: **end for**
10: **return** $Selection$

(D)CAST - DIVERSE CLASS-AWARE SELF-TRAINING

The proposed semi-supervised model learning framework, Diverse Class-Aware Self-Training (DCAST), leverages unlabeled data to gain insight into the underlying distribution of the population that may not be well represented by the labeled data. It does this using self-training (ST), and actively addresses selection bias by preserving class ratios or balance (CAST), and optionally also incorporating sample

diversity into the pseudo-labeling process to counter biases present in the data or introduced during training (DCAST).

More formally, the (D)CAST method takes as input the labeled data $\{\boldsymbol{X_L}, \boldsymbol{Y_L}\}$ and unlabeled data $\boldsymbol{X_U}$ to learn from, validation data $\{\boldsymbol{X_V}, \boldsymbol{Y_V}\}$ for early stopping, and the following four additional parameters: maximum number of iterations m, number of pseudo-labeled samples s to select per iteration, confidence or prediction probability threshold $t \in [0,1]$, and integer diversity parameter $d \geq 1$. Model learning in (D)CAST is then performed by self-training as follows. At iteration i, model $M^{(i)}$ is trained on the labeled data $\{\boldsymbol{X_{L^{(i)}}}, \boldsymbol{Y_{L^{(i)}}}\}$, and used to make predictions $\tilde{\boldsymbol{Y}}_{\boldsymbol{U^{(i)}}}$ for all samples in the unlabeled set $U^{(i)}$ (and matrix $\boldsymbol{X_{U^{(i)}}}$). As with regular self-training, a pseudo-labeling procedure then selects a subset of the unlabeled samples, $S^{(i)} \subseteq U^{(i)}$, to be incorporated into model learning (Fig. 4.3). The selected samples $S^{(i)}$ are pseudo-labeled and included in the set of labeled samples for training in the subsequent iteration, $L^{(i+1)} = L^{(i)} \cup S^{(i)}$, as well as removed from the unlabeled set $U^{(i+1)} = U^{(i)} \setminus S^{(i)}$. Matrices $\boldsymbol{X_{L^{(i+1)}}}$, $\boldsymbol{Y_{L^{(i+1)}}}$, and $\boldsymbol{X_{U^{(i+1)}}}$ are also updated for the next iteration accordingly.

Pseudo-labeling in (D)CAST: class-aware with and without diversity. The (D)CAST-specific pseudo-labeling is accomplished by the Diversity Module (Fig. 4.3). The core CAST strategy addresses class-specific bias by performing the pseudo-labeling separately per class, offering to either preserve the class ratios found in the original labeled set or select an equal number of samples per class at each iteration. Its extension, DCAST, aims for further bias mitigation by promoting sample diversity. In conventional self-training, the pseudo-labeling procedure tends to confirm and follow biases potentially present in the labeled set: either by selecting unlabeled samples similar to the original labeled samples (in feature space) or by selecting unlabeled samples whose prediction the model is most confident about. In contrast, (D)CAST seeks to mitigate this behavior and work against the strengthening of existing bias during training. To achieve this, (D)CAST selects and pseudo-labels samples that are diverse amongst each other and also more dissimilar to the possibly biased labeled samples. The (D)CAST pseudo-labeling (Alg. 2) comprises the following steps per training iteration:

Step 1. (D)CAST - Select candidate samples for pseudo-labeling based on model confidence. The goal of Step 1 is to select a set of candidate unlabeled samples for pseudo-labeling and inclusion in model training. This corresponds to the $s \times class\_ratio(c) \times d$ most confidently predicted unlabeled samples per class c, with corresponding probabilities in $\tilde{\boldsymbol{Y}}_{\boldsymbol{U^{(i)}}}$ larger than a user-defined threshold t (or a baseline threshold $r = 1.2/C$, whichever is largest) (Alg. 2, lines 9-11). For CAST, with $d = 1$ and thus no diversity strategy, this selection automatically leads to the final set of s pseudo-labeled samples ($s_c = s \times class\_ratio(c)$ per class) to incorporate during learning in the subsequent iteration. For DCAST, with $d > 1$ (Alg. 2, lines 13-15), the selected set of $s \times d$ samples ($s_c \times d$ per class) represents a larger pool of candidates to consider and narrow down further to obtain the final selected set of s samples (s_c per class) using the diversity strategy. Our recommendation for DCAST is to set the

confidence threshold t and diversity parameter d not too strictly, so as to allow for a sufficient number (and diversity) of candidate samples.

Step 2. DCAST - Diversity: Create representations of candidate samples for distance calculation. From the set of $s \times d$ candidate samples selected in Step 1, DCAST aims to extract the subset of s diverse samples. Diversity is assessed based on pairwise sample distances, calculated using a specific sample vector representation or embedding (denoted for all candidate samples as matrix $\boldsymbol{E^{(i)}} \in \mathbb{R}^{(s \times d) \times v}$, where v is the embedding vector size). Preferably, DCAST uses discriminative embeddings based on the learnt model $M^{(i)}$, where two types are currently supported. For a random forest, each sample representation corresponds to a one-hot encoded vector of the prediction of that sample across all the leaves of the decision trees in the forest; for a neural network, the sample representation corresponds to the embedding based on the hidden layer closest to the output layer. For models without discriminative embeddings, such as SVM or LR, DCAST uses the original feature vector representation.

Step 3. DCAST - Diversity: Calculate pairwise distances between candidate samples. To assess diversity, we use distances between samples: the larger the distances amongst samples in a given set, the more diverse the set will be considered. Distances are calculated by DCAST based on sample embeddings or original feature vector representations (Alg. 2, line 13). With discriminative embeddings, DCAST calculates normalized distances as $1 - (E \cdot E^T) / \max(E \cdot E^T)$, given an embedding matrix $E \in \mathbb{R}^{(s \times d) \times v}$. Specifically, for a random forest model, these distances represent the normalized frequency of non co-occurrence of a pair of samples in the leaves of the decision trees. With original feature vectors, DCAST uses Euclidean distances between sample vectors instead.

Step 4. DCAST - Diversity: Identify distinct clusters and select diverse samples to pseudo-label. The distances calculated in Step 3 are used in Step 4 to select diverse samples, potentially capturing different aspects of the pool of candidates and its underlying distribution. To do this, DCAST first identifies s (or s_c per class) distinct groups of candidate samples using a clustering algorithm (Alg. 2, line 14). The current implementation relies on agglomerative hierarchical clustering with single linkage, however any other algorithm of choice could be employed. Given that clustering is designed to maximize inter-cluster distances, samples across the different clusters are likely to yield the largest distances and thus the most diversity under the employed clustering strategy. Accordingly, DCAST selects a single sample per identified cluster to pseudo-label, namely the candidate sample with the highest confidence $\bar{y}_n^c$ value (sample n and class c, Alg. 2, line 15).

Step 5. (D)CAST - Pseudo-label selected samples. At the end of each iteration, selected samples in the set S_c are added to the labeled data matrices $\{\boldsymbol{X_L}, \boldsymbol{Y_L}\}$ and removed from the unlabeled data matrix $\boldsymbol{X_U}$.

Time Complexity of (D)CAST. To derive an upper bound for the worst-case time complexity of the (D)CAST algorithm, we assume the following time complexities for

an input of n samples defined over v features: training a base prediction model is $O(T(n,v))$, making predictions using the trained model is $O(P(n,v))$, and calculating pairwise sample distances and applying hierarchical clustering is $O((n \times v)^2)$.

At iteration i, the time complexity of (D)CAST is dominated by the following operations: retraining the model with $l + i \times s$ labeled samples in $O(T(l + i \times s, v))$ time (Alg. 2, line 4), making predictions for $l - i \times s$ unlabeled samples in $O(P(l - i \times s, v))$ time (Alg. 2, line 5), and applying hierarchical clustering with pairwise distances to at most $s \times d$ candidate unlabeled samples in $O((s \times d \times v)^2)$ time (Alg. 2, lines 11-12). Note that l denotes the number of labeled samples in the input matrices $\{\boldsymbol{X_L}, \boldsymbol{Y_L}\}$ at the start of the execution, and $i \times s$ denotes the number of samples that are pseudo-labeled up to iteration i (thus also added and removed respectively from the labeled and unlabeled data). The maximum possible number of samples for prediction at any one iteration is equal to the number of unlabeled samples u in the input matrix $\boldsymbol{X_U}$ before any pseudo-labeling has occurred, leading to the upper bound $O(P(u,v))$ on the prediction time per iteration. Similarly, u is the maximum number of samples that can be added to the input labeled data (initially containing l samples) over all iterations, which determines the upper bound $O(T(l+u,v))$ on the training time per iteration. Combining all together, each iteration takes $O(T(l+u,v) + P(u,v) + (s \times d \times v)^2)$ time, and therefore the upper bound on the worst-case time complexity of m iterations is $O(m \times (T(l+u,v) + P(u,v) + (s \times d \times v)^2))$.

4.4.2. DATA

In addition to 8 datasets from the UCI Data Repository (breast cancer, adult, spam, wine, raisin, rice, mushroom, and MNIST; https://archive.ics.uci.edu), we also used 3 datasets from other sources, including the pistachio [37], fire [38], and pumpkin [39] datasets (Supplementary Table 4.S2). All datasets had binary class labels, except for MNIST with 10 different class labels. The breast cancer, wine, spam, rice, raisin, pistachio, pumpkin and MNIST datasets comprised between 7 to 64 continuous features. The fire and adult datasets included mixed types of features, of which 1 and 7 were respectively categorical features. The mushroom dataset only had categorical features. For the fire, adult, and mushroom datasets, all categorical features were one-hot encoded.

4.4.3. EVALUATION OF BIAS INDUCTION AND BIAS MITIGATION METHODS

We performed experiments across 11 ML benchmark datasets with different characteristics to assess the effectiveness of (i) selection bias induction using the proposed hierarchy bias technique, and (ii) selection bias mitigation using the proposed (D)CAST strategies. Hierarchy bias was compared to other bias induction techniques concerning both the distribution shift produced by the data selection procedure and its effect on the performance of prediction models built using supervised learning. The (D)CAST semi-supervised bias mitigation strategies were evaluated against conventional semi-supervised self-training (ST), as well as a range

Algorithm 2 (D)CAST - Diverse Class-Aware Self-Training

Require: T (model type); $\boldsymbol{X_L}$, $\boldsymbol{Y_L}$ (labeled train data); $\boldsymbol{X_V}$, $\boldsymbol{Y_V}$ (labeled validation data); $\boldsymbol{X_U}$ (unlabeled data); s (number of samples to select per iteration); t (prediction probability threshold); d (diversity strength); m (maximum number of iterations).

$terminate \leftarrow False$
$i \leftarrow 0$
while $terminate$ is $False \vee i = m$ **do**
    $M^{(i)} \leftarrow$ train model instance of type T with $\boldsymbol{X_L}$, $\boldsymbol{Y_L}$
    $\bar{Y} \leftarrow$ predict class probability for samples in $\boldsymbol{X_U}$ using $M^{(i)}$
    for `each class` $c \in C$ **do**
        $s_c \leftarrow s \times class\_ratio(c)$
        $t_c \leftarrow \max(t, r)$
        $S_c \leftarrow$ top $s_c \times d$ confidently predicted samples with $max(\bar{y}_n^c) > t_c$
        if $d > 1$ **then**
            $E \leftarrow$ calculate pairwise distances for samples in S_c
            $Clusters \leftarrow$ apply agglomerative clustering to obtain s_c clusters using distances E and single linkage
            $S_c \leftarrow$ choose the sample with the highest prediction probability from each cluster in $Clusters$
        end if
        for `each` selected sample $n \in S_c$ **do**
            $\boldsymbol{X_L}$.add($\boldsymbol{x_n}$), $\boldsymbol{Y_L}$.add($\boldsymbol{y_n}$), $\boldsymbol{X_U}$.remove($\boldsymbol{x_n}$)
        end for
    end for
    ▷ Stopping conditions: maximum number of iterations m is reached OR all unlabeled samples have been incorporated OR validation accuracy did not improve for the last 5 iterations.
    if ($i == m$) $\vee$
      ($len(\boldsymbol{X_U}) == 0$) $\vee$
      ($\exists z \in \{i-6, \ldots, i-1\}$ such that $Accuracy(M^{(i)}, \boldsymbol{X_V}, \boldsymbol{Y_V}) <$
      $Accuracy(M^{(z)}, \boldsymbol{X_V}, \boldsymbol{Y_V})$) **then**
            $terminate \leftarrow True$
            $M_{best} \leftarrow argmax_{z=0,\ldots,i}(Accuracy(M^{(z)}, \boldsymbol{X_V}, \boldsymbol{Y_V}))$
    end if
    $i \leftarrow i + 1$
end while
return M_{best}

of alternative domain adaptation methods, on their ability to build prediction models from biased data with better generalization than using supervised learning.

Data splits and bias induction. For each dataset, 20% of the samples were uniformly selected at random, stratified by class, and reserved as test data to evaluate prediction models (Fig. 4.6). The adult dataset already had its own separate test set, which we reserved. Additionally, we created 30 distinct train runs per dataset, each by randomly splitting the remaining 80% of the samples into two train sets, stratified by class: a labeled train set, containing 30% of the samples, from which we also generated biased labeled sets by applying different bias induction techniques; and an unlabeled train set, comprising the remaining 70% of the samples. The original and biased labeled train sets were later used to build prediction models with supervised learning or bias mitigation strategies, while the unlabeled train set was used to learn prediction models with the semi-supervised bias mitigation strategies (D)CAST and conventional ST (other bias mitigation methods used test data without labels). When necessary for model training, a validation set was further extracted from each biased train set, given that unbiased labeled data would not be available for this purpose in a realistic setting.

Bias induction impact on data distribution

Bias induction methods were first assessed on their ability to cause a distribution shift in the biased selection relative to the original labeled train set. Quantitatively, we analyzed the change in the distribution of inter-sample distances as follows. We first calculated class-specific distributions of the per sample average Euclidean distance to all other samples in either the biased selection or the original labeled train set. We then determined the class-specific distribution shifts between the biased selection and the original data using two-sample Kolmogorov-Smirnov (KS) statistical tests. We report KS effect sizes, as well as histograms of inter-sample distances for the biased selection distribution and histogram peaks for the original data distribution.

Visually, we analyzed to what extent the biased selection was representative of the original labeled train set by inspecting 2D dimension reductions of the original data using the Uniform Manifold Approximation and Projection (UMAP) algorithm. We applied UMAP to the original labeled set with four different nearest neighbor parameter values (15, 50, 100, and 200) to obtain a reasonable representation of the sample space for each dataset.

Bias induction and mitigation effects on prediction performance

Furthermore, to evaluate bias induction and bias mitigation techniques, we investigated how prediction models trained on data affected or not by selection bias generalized to test data that was more representative of the original distribution. All models built using supervised learning or bias mitigation techniques were trained and evaluated as follows.

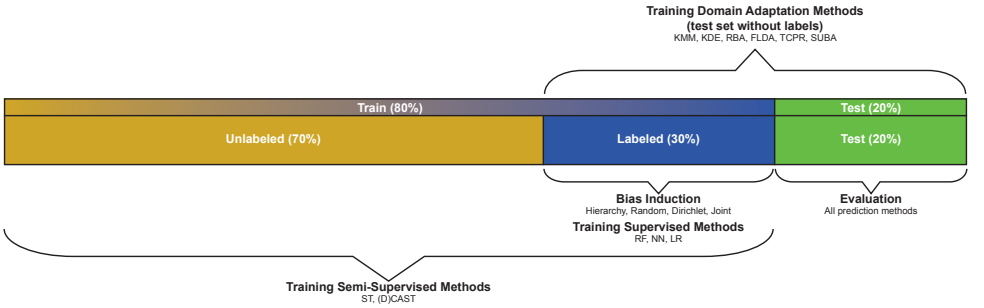


Figure 4.6: **Data split for evaluation of bias induction and bias mitigation effects on prediction performance.** Each dataset is randomly split into train (80%) and test (20%) sets, and 30 different train runs are created by splitting the samples in the train set randomly into labeled (30%) and unlabeled (70%) train sets. Bias induction is further applied to the labeled train sets to generate corresponding biased labeled train sets. Supervised learning is used to build models separately from the original labeled train set and from the biased labeled train set, which serve as baselines to assess the effects of bias induction and bias mitigationl on prediction performance. For bias mitigation, CAST and DCAST learn prediction models using both the unlabeled and labeled train sets, while domain adaptation methods learn from the labeled train set together with the test set (without labels). All models are evaluated on the labeled test set.

Training of models using supervised learning or bias mitigation. To quantify the baseline prediction performance, without bias induction, we built models using supervised learning on the original labeled train set. To assess the effect of bias induction compared to the baseline, we built models using supervised learning on the biased labeled train set. Additionally, to assess the bias mitigation strategies and investigate if they could generalize better than supervised learning on the biased labeled train set, we used them to train models on the biased labeled train set together with unlabeled data (namely the unlabeled train set for semi-supervised (D)CAST and conventional ST, or the unlabeled test set for the remaining methods). The prediction models we trained using supervised learning or bias mitigation strategies were based on three different model types: L2-regularized random forests (RF, [40]), 2 hidden-layered (input, 8-node, 12-node, output) neural networks (NN), and L2-regularized logistic regression (LR) [41]. We used default parameter values (Supplementary Table 4.S3), since fine-tuning with a biased validation set could further reinforce the bias. To account for variation introduced by randomness in the training procedures of the RF and NN models, we used different seeds to train 10 prediction models instead of one per run for any given combination of dataset, model type, bias induction technique, and model learning strategy.

Evaluation of models trained using supervised learning or bias mitigation. The performance of all prediction models was evaluated on the test set. We focused on quantifying prediction accuracy rather than loss, since the loss could often be improved by increasing model confidence without a measurable improvement in accuracy, which is ultimately the goal of the models under study. We report the performance results as the median test accuracy of the 10 models using different seeds per run, with a total of 30 runs, for every combination of dataset, model type, bias induction technique, and model learning strategy. Some model learning strategies did not successfully build prediction models for all runs, which is

necessarily reflected in the results and corresponding figures.

4.4.4. Experimental settings of bias induction and mitigation methods

Bias induction and sample selection methods

We compared the proposed hierarchy bias induction method against the joint and Dirichlet bias induction techniques, as well as random subsampling. Hierarchy bias was used with a fixed target of $k = 30$ samples to select per class, and a bias ratio of $b = 0.9$ across experiments. Random subsampling consisted in selecting k samples uniformly at random per class, where k was similarly set to 30. Joint bias assigns a selection probability to each sample based on its proximity to the sample mean over the labeled train data, and then independently selects samples according to their selection probabilities [19]. Joint bias induction does not include any parameter to control the number of selected samples, and it was therefore used without a fixed target number of selected biased samples. Dirichlet bias selects a subset of samples without replacement, where the biased selection probability of each sample is determined based on a random likelihood function sampled from a Dirichlet distribution [20]. This method does not consider class labels in its biased selection and was therefore set to select a total of $k \times |C|$ samples, with $|C|$ denoting the number of classes and $k = 30$. Of note, hierarchy bias and random subsampling generate a biased selection that is balanced across classes, whereas joint and Dirichlet bias induction do not offer such guarantee.

Bias mitigation strategies

We assessed the proposed semi-supervised (D)CAST methods against competing bias mitigation techniques, including semi-supervised conventional self-training and alternative domain adaptation strategies.

The semi-supervised methods, (D)CAST and conventional ST, learned models using the labeled and unlabeled train sets. Additionally, (D)CAST relied on early stopping based on validation performance to make training more efficient and robust. To be fair to other methods, (D)CAST used a portion of the labeled train set for validation rather than a separate validation set. We set the following parameter values for (D)CAST across experiments: maximum number of iterations $m = 100$, number of pseudo-labeled samples to include per iteration s as $3 \times |C|$ (or 3 times the number of classes), and three different diversity strengths $d = \{1, 10, 100\}$. In addition, the confidence threshold t used by (D)CAST to select candidate samples for pseudo-labeling was set to a prediction probability of 0.9 for NN and LR models. Since RF models generally showed lower prediction probabilities, possibly due to regularization, we defined the threshold for binary RF classification models as the 93rd percentile of all prediction probabilities on unlabeled data. This threshold was not fully optimized, only considered sufficient to allow pseudo-labeling of some samples across all datasets with binary class labels. For MNIST, probabilities were

even lower given the multiclass nature of the problem, thus we set the threshold of RF models as the 85th percentile instead.

Given that most semi-supervised learning approaches designed to mitigate sample selection bias are not model agnostic and do not have readily available implementations, we compared (D)CAST with the closely related conventional self-training (ST) methods. We implemented and tested two variants of conventional ST, which pseudo-labeled either the $3 \times |C|$ samples with the highest prediction probabilities or all samples with prediction probabilities over 0.9. The former variant performed better and was thus selected.

We included domain adaptation methods beyond semi-supervised learning across three categories, using Python implementations available in the libTLDA Python library [42]: importance weighting approaches Kernel Mean Matching (KMM [19]) and Kernel Density Estimation (KDE [22]), minimax estimation strategies Robust Bias-Aware classifier (RBA [20]) and Target Contrastive Pessimistic Risk (TCPR [32]), and subspace alignment methods Feature-Level Domain Adaptation (FLDA [31]) and Subspace Alignment classifier (SUBA [30]). All of these methods were applied as originally proposed by their authors to learn models based on the labeled train set together with the test set without labels. In addition, all methods except KMM were used exclusively with L2-regularized LR models. The KMM importance weighting approach is ML model-agnostic, since it independently calculates a weight for each sample based exclusively on the train and test data, and was therefore applied with RF, NN, and LR models.

Data availability

The data used in this article were obtained from publicly available sources, detailed in the Methods section. The raw data necessary to reproduce the experiments, along with the main experimental results for CAST and DCAST, are accessible via Figshare at doi.org/10.6084/m9.figshare.27003601.

Code availability

An implementation of the hierarchy bias and the (D)CAST methods in Python has been made available under an open source license at github.com/joanagoncalveslab/DCAST.

Acknowledgements

The authors received funding from the US National Institutes of Health [U54EY032442, U54DK134302, U01DK133766, R01AG078803 to J.P.G.]. Authors are solely responsible for the research, the funders were not involved in the work. The authors further acknowledge the High-Performance Compute (HPC) cluster of the Department of Intelligent Systems at the Delft University of Technology.

REFERENCES

[1] N. Mehrabi *et al.* “A Survey on Bias and Fairness in Machine Learning”. In: *ACM Computing Surveys* 54.6 (July 2021). ISSN: 0360-0300.

[2] D. Pessach and E. Shmueli. “A Review on Fairness in Machine Learning”. In: *ACM Computing Surveys* 55.3 (Feb. 2022), pp. 1–44. ISSN: 1557-7341.

[3] D. Wu *et al.* “Correcting sample selection bias for image classification”. In: *2008 3rd International Conference on Intelligent System and Knowledge Engineering*. Vol. 1. 2008, pp. 1214–1220.

[4] C. Persello and L. Bruzzone. “Active and Semisupervised Learning for the Classification of Remote Sensing Images”. In: *IEEE Transactions on Geoscience and Remote Sensing* 52.11 (2014), pp. 6937–6956.

[5] J. W. Richards *et al.* “ACTIVE LEARNING TO OVERCOME SAMPLE SELECTION BIAS: APPLICATION TO PHOTOMETRIC VARIABLE STAR CLASSIFICATION”. In: *The Astrophysical Journal* 744.2 (Dec. 2011), p. 192.

[6] J. Kremer *et al.* “Nearest neighbor density ratio estimation for large-scale applications in astronomy”. In: *Astronomy and Computing* 12 (Sept. 2015), pp. 67–72.

[7] R. Romero, E. L. Iglesias, and L. Borrajo. “Building Biomedical Text Classifiers under Sample Selection Bias”. In: *Advances in Intelligent and Soft Computing*. Springer Berlin Heidelberg, 2011, pp. 11–18.

[8] J. Y. Chan and J. A. Cook. “Inferring Zambia’s HIV prevalence from a selected sample”. In: *Applied Economics* 52.39 (Mar. 2020), pp. 4236–4249.

[9] C. Seale, Y. Tepeli, and J. P. Gonçalves. “Overcoming selection bias in synthetic lethality prediction”. In: *Bioinformatics* 38.18 (July 2022). Ed. by K. Borgwardt, pp. 4360–4368. ISSN: 1367-4811.

[10] Y. I. Tepeli, C. Seale, and J. P. Gonçalves. “ELISL: early-late integrated synthetic lethality prediction in cancer”. en. In: *Bioinformatics* 40.1 (Jan. 2024).

[11] C.-H. Chang and J.-H. Lin. “Decision Support and Profit Prediction for Online Auction Sellers”. In: *Proceedings of the 1st ACM SIGKDD Workshop on Knowledge Discovery from Uncertain Data*. U ’09. Paris, France: Association for Computing Machinery, 2009, pp. 1–8. ISBN: 9781605586755.

[12] C. Castagnetti, L. Rosti, and M. Töpfer. “The Age Pay Gap between Young and Older Employees in Italy: Perceived or Real Discrimination against the Young?” In: *Research in Labor Economics*. Emerald Publishing Limited, Nov. 2020, pp. 195–221.

[13] F. Shen *et al.* “Reject inference in credit scoring using a three-way decision and safe semi-supervised support vector machine”. In: *Information Sciences* 606 (Aug. 2022), pp. 614–627.

[14] M. Melucci. “Investigating sample selection bias in the relevance feedback algorithm of the vector space model for Information Retrieval”. In: *2014 International Conference on Data Science and Advanced Analytics (DSAA)*. 2014, pp. 83–89.

[15] M. Melucci. “Impact of Query Sample Selection Bias on Information Retrieval System Ranking”. In: *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. 2016, pp. 341–350.

[16] G. Zhang *et al.* “Selection Bias Explorations and Debias Methods for Natural Language Sentence Matching Datasets”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019, pp. 4418–4429.

[17] N. V. Chawla and G. Karakoulas. “Learning from Labeled and Unlabeled Data: An Empirical Study across Techniques and Domains”. In: *J. Artif. Int. Res.* 23.1 (Mar. 2005), pp. 331–366. ISSN: 1076-9757.

[18] A. T. Smith and C. Elkan. “Making Generative Classifiers Robust to Selection Bias”. In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’07. San Jose, California, USA: Association for Computing Machinery, 2007, pp. 657–666. ISBN: 9781595936097.

[19] J. Huang *et al.* “Correcting Sample Selection Bias by Unlabeled Data”. In: *Proceedings of the 19th International Conference on Neural Information Processing Systems*. NIPS’06. Canada: MIT Press, 2006, pp. 601–608.

[20] A. Liu and B. Ziebart. “Robust classification under sample selection bias”. In: *Advances in Neural Information Processing Systems* 1 (Jan. 2014), pp. 37–45.

[21] W. M. Kouw and M. Loog. “A Review of Domain Adaptation without Target Labels”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.3 (Mar. 2021), pp. 766–785.

[22] H. Shimodaira. “Improving predictive inference under covariate shift by weighting the log-likelihood function”. In: *Journal of Statistical Planning and Inference* 90.2 (Oct. 2000), pp. 227–244.

[23] B. Zadrozny. “Learning and Evaluating Classifiers under Sample Selection Bias”. In: *Proceedings of the Twenty-First International Conference on Machine Learning*. ICML ’04. Banff, Alberta, Canada: Association for Computing Machinery, 2004, p. 114. ISBN: 1581138385.

[24] C.-W. Seah, I. W.-H. Tsang, and Y.-S. Ong. “Healing Sample Selection Bias by Source Classifier Selection”. In: *2011 IEEE 11th International Conference on Data Mining*. 2011, pp. 577–586.

[25] M. Sugiyama, M. Yamada, and M. C. du Plessis. “Learning under nonstationarity: covariate shift and class-balance change”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 5.6 (Aug. 2013), pp. 465–477.

[26] Z. Shen *et al.* “Causally Regularized Learning with Agnostic Data Selection Bias”. In: *Proceedings of the 26th ACM International Conference on Multimedia*. MM ’18. Seoul, Republic of Korea: Association for Computing Machinery, 2018, pp. 411–419. ISBN: 9781450356657.

[27] M. Diesendruck *et al.* “Importance Weighted Generative Networks”. In: *Machine Learning and Knowledge Discovery in Databases*. Springer International Publishing, 2020, pp. 249–265.

[28] W. Du and X. Wu. “Fair and Robust Classification Under Sample Selection Bias”. In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. CIKM ’21. Virtual Event, Queensland, Australia: Association for Computing Machinery, 2021, pp. 2999–3003. ISBN: 9781450384469.

[29] J. Blitzer, R. McDonald, and F. Pereira. “Domain Adaptation with Structural Correspondence Learning”. In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. EMNLP ’06. Sydney, Australia: Association for Computational Linguistics, 2006, pp. 120–128. ISBN: 1932432736.

[30] B. Fernando *et al.* “Unsupervised Visual Domain Adaptation Using Subspace Alignment”. In: *2013 IEEE International Conference on Computer Vision*. 2013, pp. 2960–2967.

[31] W. M. Kouw *et al.* “Feature-Level Domain Adaptation”. In: *J. Mach. Learn. Res.* 17.1 (Jan. 2016), pp. 5943–5974. ISSN: 1532-4435.

[32] W. M. Kouw and M. Loog. “Robust domain-adaptive discriminant analysis”. In: *Pattern Recognition Letters* 148 (Aug. 2021), pp. 107–113.

[33] W. Fan and I. Davidson. “Reverse Testing: An Efficient Framework to Select amongst Classifiers under Sample Selection Bias”. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’06. Philadelphia, PA, USA: Association for Computing Machinery, 2006, pp. 147–156. ISBN: 1595933395.

[34] J. Ren *et al.* “Type Independent Correction of Sample Selection Bias via Structural Discovery and Re-balancing”. In: *Proceedings of the 2008 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, Apr. 2008.

[35] G. J. McLachlan. “Iterative Reclassification Procedure for Constructing an Asymptotically Optimal Rule of Allocation in Discriminant Analysis”. In: *Journal of the American Statistical Association* 70.350 (1975), pp. 365–369. eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1975.10479874.

[36] A. Blum and T. Mitchell. “Combining Labeled and Unlabeled Data with Co-Training”. In: *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*. COLT’ 98. Madison, Wisconsin, USA: Association for Computing Machinery, 1998, pp. 92–100. ISBN: 1581130570.

[37] I. A. Ozkan, M. Koklu, and R. Saraçoğlu. “Classification of Pistachio Species Using Improved K-NN Classifier”. In: *Progress in Nutrition* 23.2 (July 2021), e2021044. ISSN: 1129-8723.

[38] M. Koklu and Y. S. Taspinar. “Determining the Extinguishing Status of Fuel Flames With Sound Wave by Machine Learning Methods”. In: *IEEE Access* 9 (2021), pp. 86207–86216.

[39] M. Koklu, S. Sarigil, and O. Ozbek. “The use of machine learning methods in classification of pumpkin seeds (Cucurbita pepo L.)” In: *Genetic Resources and Crop Evolution* 68.7 (June 2021), pp. 2713–2726.

[40] T. K. Ho. “Random decision forests”. In: *Proceedings of 3rd international conference on document analysis and recognition*. Vol. 1. IEEE. 1995, pp. 278–282.

[41] F. Pedregosa *et al.* “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–30.

[42] W. Kouw. *wmkouw/libTLDA v0.1*. Version v0.1. Apr. 2018.

4.5. SUPPLEMENTARY MATERIALS

4.5.1. SUPPLEMENTARY FIGURES

BIAS INDUCTION TO OTHER DATASETS

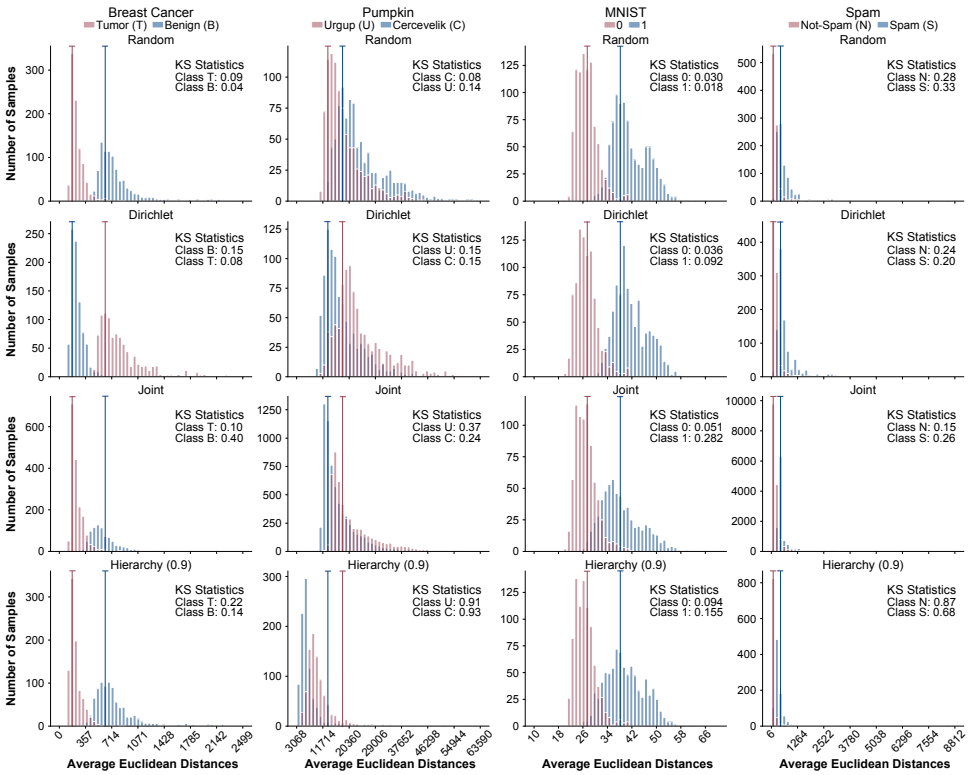


Figure 4.S1: **Bias induction impact on sample distances for the breast cancer, pumpkin, MNIST, and spam datasets.** Class-specific distributions of per sample average Euclidean distances to all other samples, for the biased selection (histograms) and for all samples in the labeled train set (histogram peaks denoted by lines ending in a "T" shape), using three bias induction techniques (hierarchy with $b = 0.9$, joint, and Dirichlet) and random subsampling on four datasets (breast cancer, pumpkin, MNIST, and spam). Kolmogorov-Smirnov (KS) effect sizes quantify the distribution shift between the biased selection vs. all samples distributions.

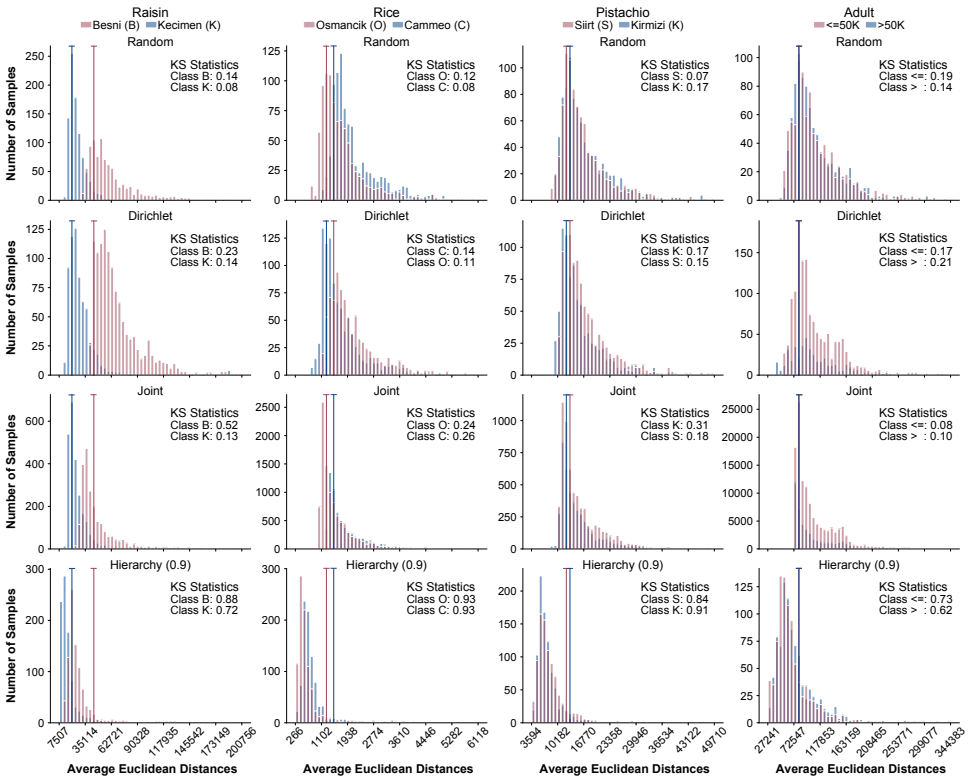


Figure 4.S2: **Bias induction impact on sample distances for the raisin, rice, pistachio, and adult datasets.** Class-specific distributions of per sample average Euclidean distances to all other samples, for the biased selection (histograms) and for all samples in the labeled train set (histogram peaks denoted by lines ending in a "T" shape), using three bias induction techniques (hierarchy with $b = 0.9$, joint, and Dirichlet) and random subsampling on four datasets (raisin, rice, pistachio, and adult). Kolmogorov-Smirnov (KS) effect sizes quantify the distribution shift between the biased selection vs. all samples distributions.

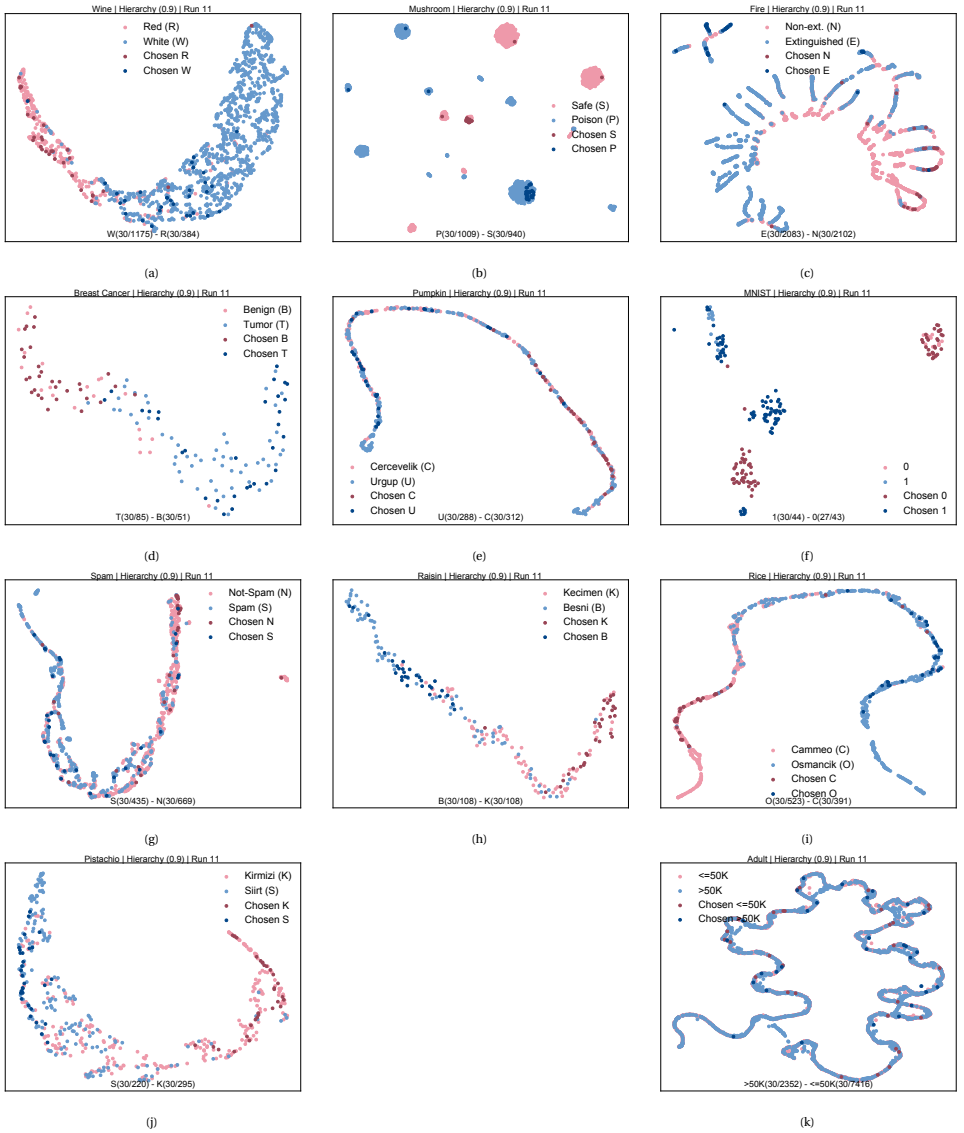


Figure 4.S3: **Impact of hierarchy bias induction on the UMAP latent space.** Samples selected by hierarchy bias ($b = 0.9$) highlighted on the respective latent UMAP space of the labeled train set for each of the 11 datasets: (a) wine, (b) mushroom, (c) fire, (d) breast cancer, (e) pumpkin, (f) MNIST, (g) spam, (h) raisin, (i) rice, (j) pistachio, and (k) adult. Results are shown for run 11 (arbitrarily chosen).

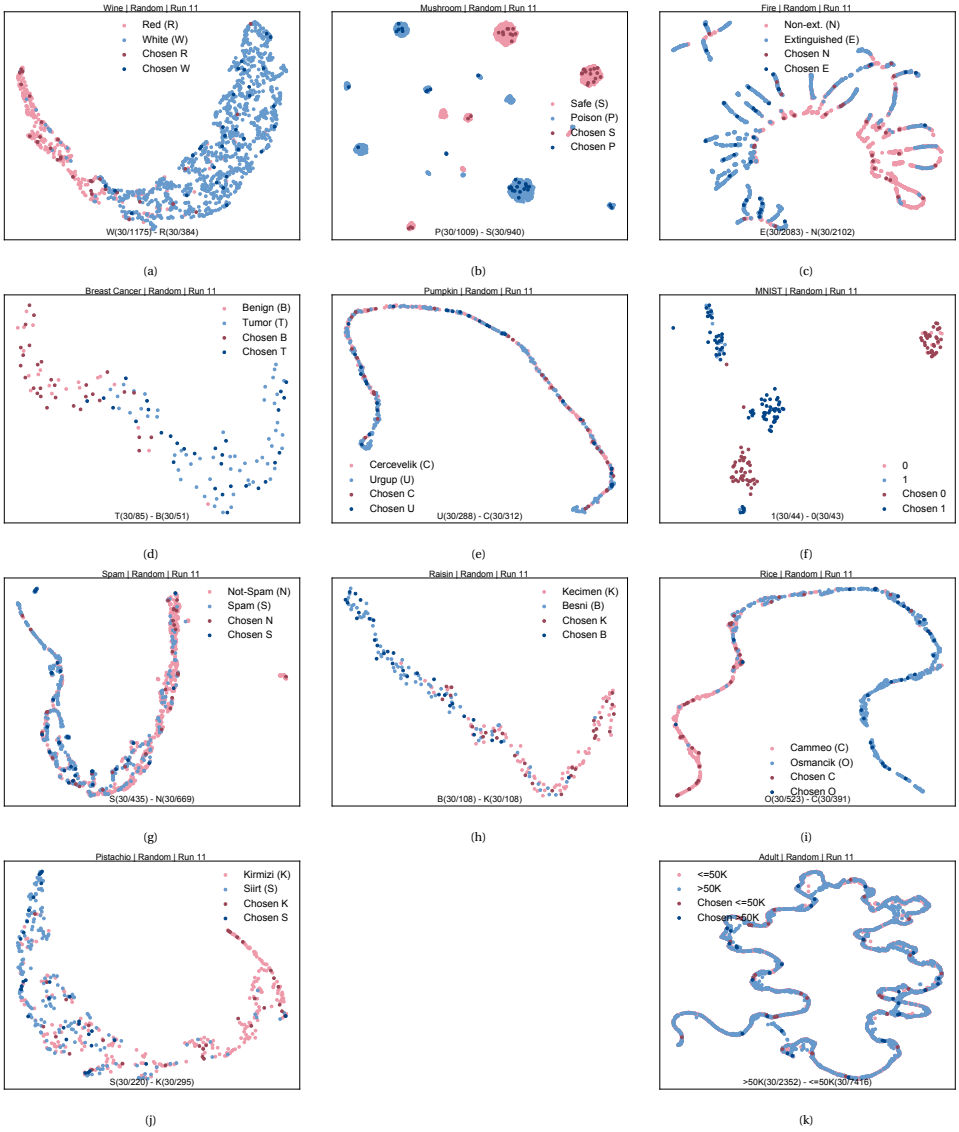


Figure 4.S4: **Impact of random subsampling on the UMAP latent space.** Samples selected by random subsampling highlighted on the respective latent UMAP space of the labeled train set for each of the 11 datasets: (a) wine, (b) mushroom, (c) fire, (d) breast cancer, (e) pumpkin, (f) MNIST, (g) spam, (h) raisin, (i) rice, (j) pistachio, and (k) adult. Results are shown for run 11 (arbitrarily chosen).

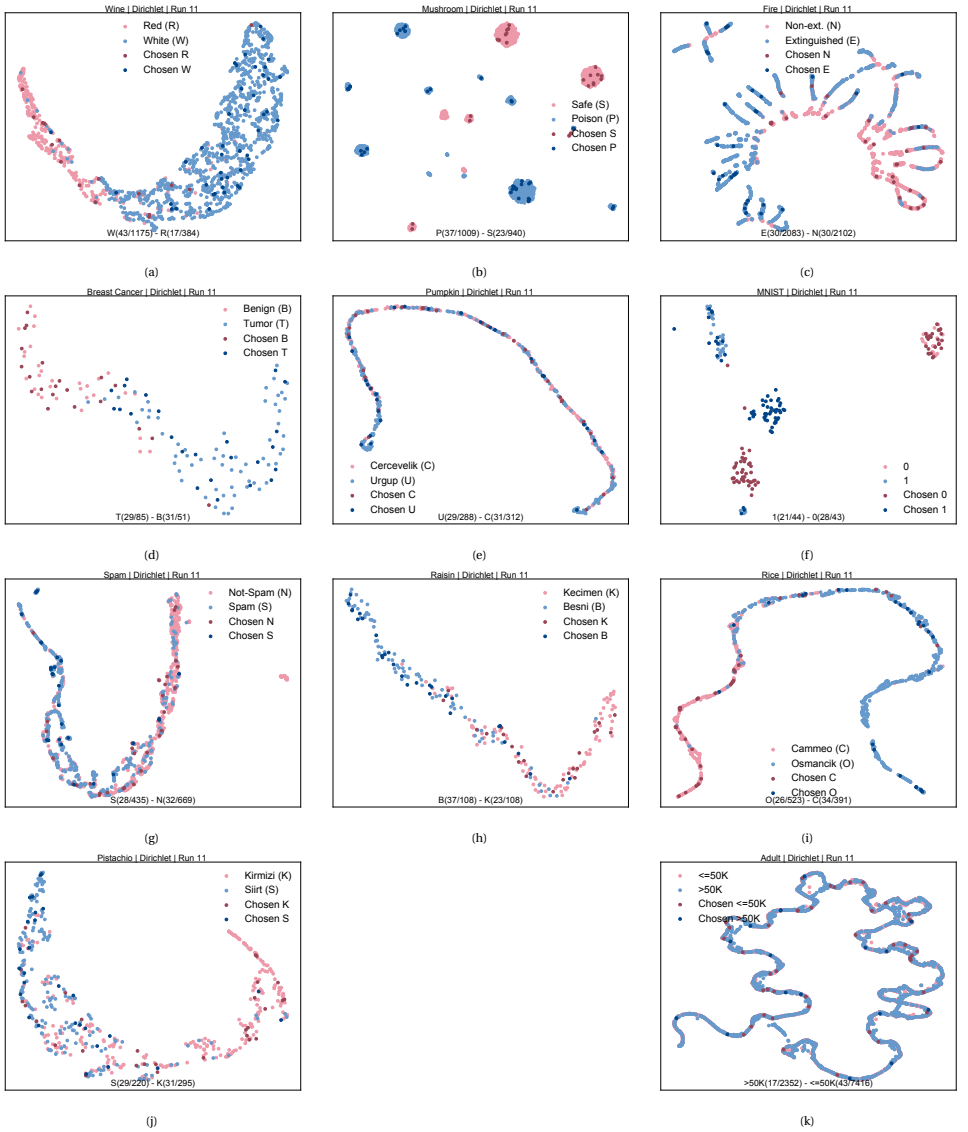


Figure 4.S5: **Impact of Dirichlet bias induction on the UMAP latent space.** Samples selected by Dirichlet bias highlighted on the respective latent UMAP space of the labeled train set for each of the 11 datasets: (a) wine, (b) mushroom, (c) fire, (d) breast cancer, (e) pumpkin, (f) MNIST, (g) spam, (h) raisin, (i) rice, (j) pistachio, and (k) adult. Results are shown for run 11 (arbitrarily chosen).

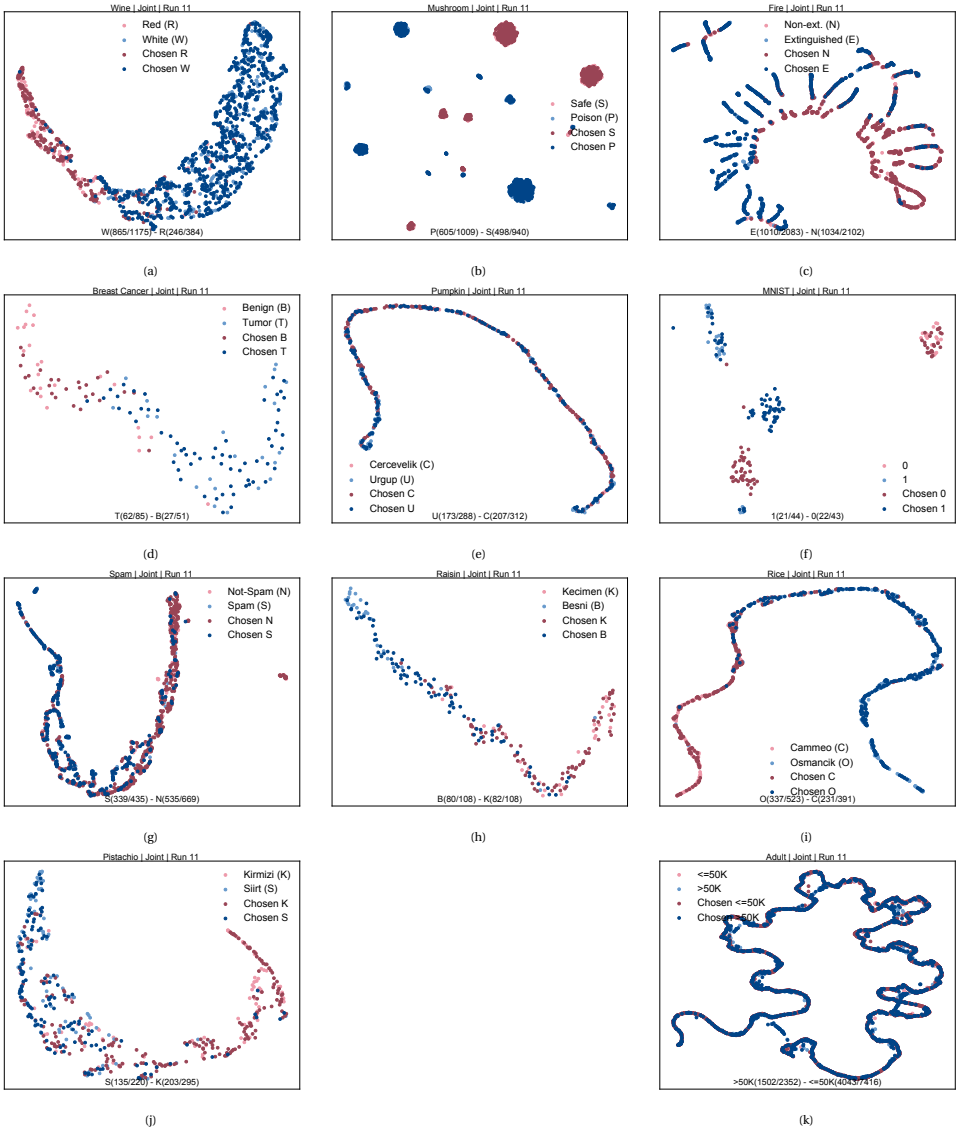


Figure 4.S6: **Impact of joint bias on the UMAP latent space.** Samples selected by joint bias, highlighted on the respective latent UMAP space of the labeled train set for each of the 11 datasets: (a) wine, (b) mushroom, (c) fire, (d) breast cancer, (e) pumpkin, (f) MNIST, (g) spam, (h) raisin, (i) rice, (j) pistachio, and (k) adult. Results are shown for run 11 (arbitrarily chosen).

SEMI-SUPERVISED METHODS ON OTHER BIAS INDUCTION TECHNIQUES

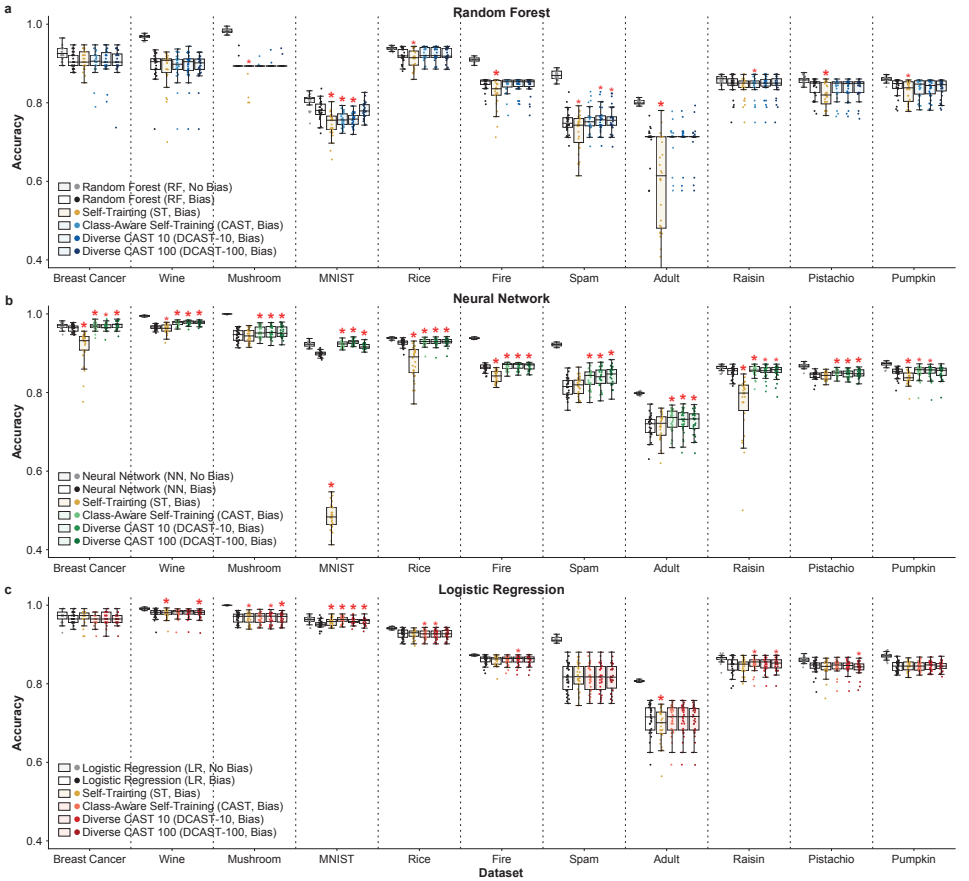


Figure 4.S7: **Performance of semi-supervised bias mitigation upon random subsampling.** Accuracy of supervised and semi-supervised learning methods with **(a)** RF, **(b)** NN, and **(c)** LR models across 11 datasets. Results for 30 runs: each training on a different split of the train set into labeled and unlabeled sets, all evaluated on the same original test set. Models included (top to bottom): supervised RF/NN/LR models trained on the original (No Bias) or biased (Bias) labeled set; and semi-supervised RF/NN/LR models, using conventional self-training (ST) on the biased labeled train set plus the unlabeled test set, or (D)CAST on the biased labeled train set plus the unlabeled train set. Red asterisks (*) denote statistically significant changes in accuracy over 30 runs for each semi-supervised approach compared to supervised learning on the biased labeled set, using one-sided Wilcoxon signed-rank tests (larger asterisks indicate $p < 0.01$, smaller asterisks $0.01 < p < 0.05$).

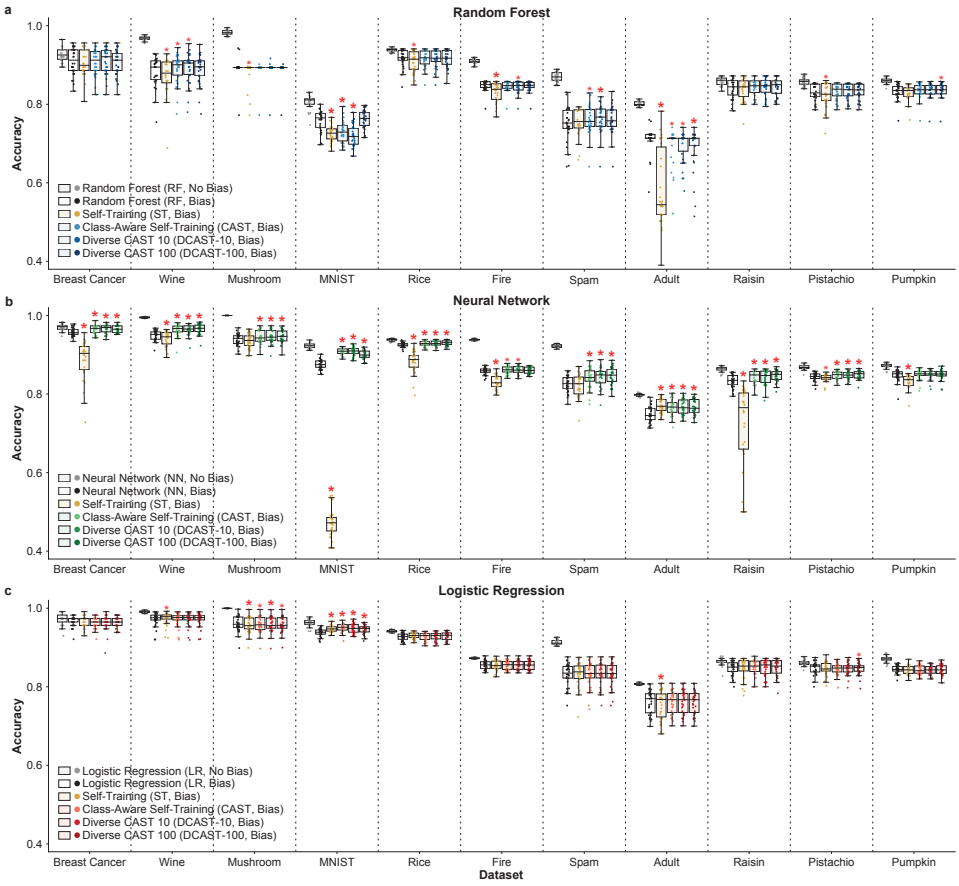


Figure 4.S8: **Performance of semi-supervised bias mitigation under Dirichlet bias.** Accuracy of supervised and semi-supervised learning methods with **(a)** RF, **(b)** NN, and **(c)** LR models across 11 datasets. Results for 30 runs: each training on a different split of the train set into labeled and unlabeled sets, all evaluated on the same original test set. Models included (top to bottom): supervised RF/NN/LR models trained on the original (No Bias) or biased (Bias) labeled set; and semi-supervised RF/NN/LR models, using conventional self-training (ST) on the biased labeled train set plus the unlabeled test set, or (D)CAST on the biased labeled train set plus the unlabeled train set. Red asterisks (*) denote statistically significant changes in accuracy over 30 runs for each semi-supervised approach compared to supervised learning on the biased labeled set, using one-sided Wilcoxon signed-rank tests (larger asterisks indicate $p < 0.01$, smaller asterisks $0.01 < p < 0.05$).

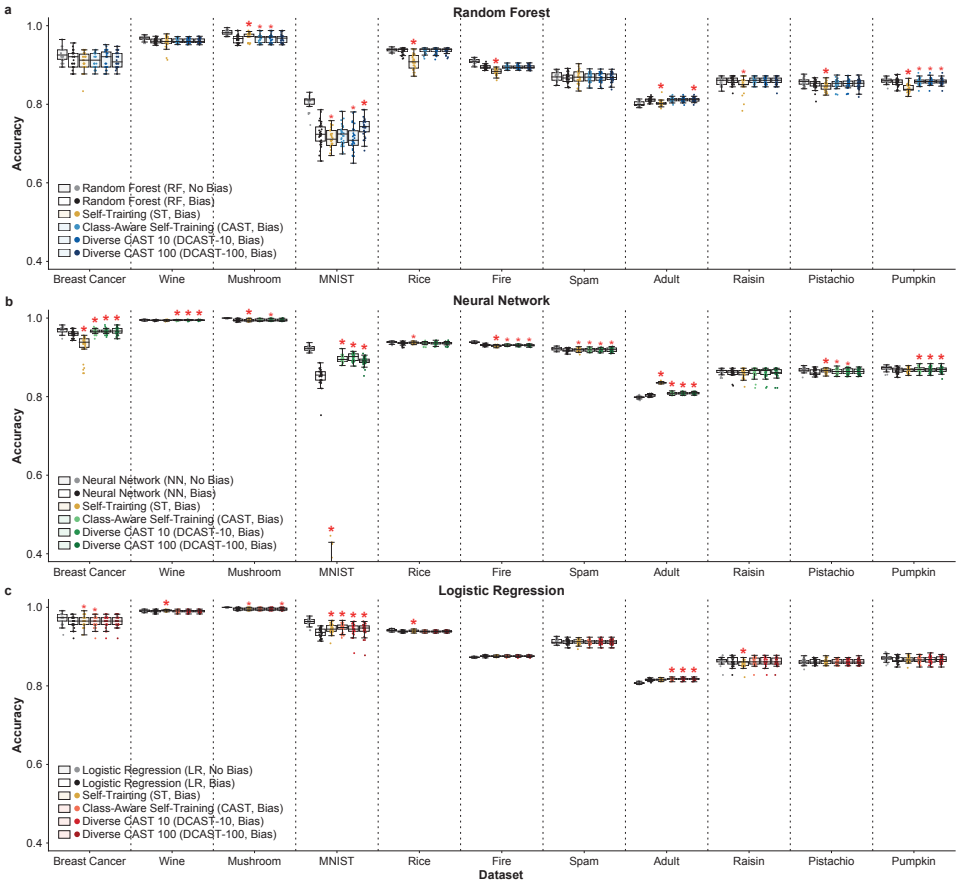


Figure 4.S9: **Performance of semi-supervised bias induction under joint bias.** Accuracy of supervised and semi-supervised learning methods with **(a)** RF, **(b)** NN, and **(c)** LR models across 11 datasets. Results for 30 runs: each training on a different split of the train set into labeled and unlabeled sets, all evaluated on the same original test set. Models included (top to bottom): supervised RF/NN/LR models trained on the original (No Bias) or biased (Bias) labeled set; and semi-supervised RF/NN/LR models, using conventional self-training (ST) on the biased labeled train set plus the unlabeled test set, or (D)CAST on the biased labeled train set plus the unlabeled train set. Red asterisks (*) denote statistically significant changes in accuracy over 30 runs for each semi-supervised approach compared to supervised learning on the biased labeled set, using one-sided Wilcoxon signed-rank tests (larger asterisks indicate $p < 0.01$, smaller asterisks $0.01 < p < 0.05$).

4.5.2. Supplementary Tables

Table 4.S1: **Class balance of original and biased labeled train sets, as well as biased selection ratio, for 11 datasets over 30 train runs.** The "class balance" columns indicate the ratio between the number of samples in the first class (Class 0) and the total number of samples in the original and the biased labeled sets. The "Selection ratio" columns refer to the ratio between the number of samples in the biased labeled set and the number of samples in the original labeled set per class. The columns "Avg." and "SD" contain the average and standard deviation of the values over 30 runs, respectively. For MNIST, only the statistics for the first 2 classes are reported. "CB" refers to class balance and "SR" refers to selection rate by bias.

Dataset	Bias	Original CB	Biased CB		SR (class 0)		SR (class 1)	
			Avg.	SD	Avg.	SD	Avg	SD
Adult	**Dirichlet**	0.241	0.249	0.058	0.006	0.001	0.006	0.000
Adult	**Hierarchy (0.9)**		0.500	0.000	0.013	0.000	0.004	0.000
Adult	**Joint**		0.271	0.004	0.653	0.010	0.557	0.006
Adult	**Random**		0.500	0.000	0.013	0.000	0.004	0.000
Breast Cancer	**Dirichlet**	0.625	0.478	0.063	0.337	0.045	0.614	0.075
Breast Cancer	**Hierarchy (0.9)**		0.505	0.011	0.353	0.000	0.576	0.024
Breast Cancer	**Joint**		0.671	0.029	0.688	0.040	0.563	0.067
Breast Cancer	**Random**		0.500	0.000	0.353	0.000	0.588	0.000
Fire	**Dirichlet**	0.498	0.533	0.061	0.015	0.002	0.013	0.002
Fire	**Hierarchy (0.9)**		0.500	0.000	0.014	0.000	0.014	0.000
Fire	**Joint**		0.500	0.008	0.503	0.015	0.498	0.014
Fire	**Random**		0.500	0.000	0.014	0.000	0.014	0.000
MNIST	**Dirichlet**	0.102	0.514	0.067	0.704	0.121	0.678	0.110
MNIST	**Hierarchy (0.9)**		0.496	0.019	0.641	0.034	0.665	0.035
MNIST	**Joint**		0.443	0.054	0.463	0.087	0.595	0.096
MNIST	**Random**		0.500	0.000	0.682	0.000	0.698	0.000
Mushroom	**Dirichlet**	0.518	0.550	0.069	0.033	0.004	0.029	0.004
Mushroom	**Hierarchy (0.9)**		0.500	0.000	0.030	0.000	0.032	0.000
Mushroom	**Joint**		0.544	0.011	0.597	0.015	0.536	0.014
Mushroom	**Random**		0.500	0.000	0.030	0.000	0.032	0.000
Pistachio	**Dirichlet**	0.427	0.456	0.065	0.124	0.018	0.111	0.013
Pistachio	**Hierarchy (0.9)**		0.500	0.000	0.136	0.000	0.102	0.000
Pistachio	**Joint**		0.411	0.018	0.656	0.057	0.701	0.032
Pistachio	**Random**		0.500	0.000	0.136	0.000	0.102	0.000
Pumpkin	**Dirichlet**	0.480	0.504	0.085	0.105	0.018	0.095	0.016
Pumpkin	**Hierarchy (0.9)**		0.500	0.000	0.104	0.000	0.096	0.000
Pumpkin	**Joint**		0.441	0.016	0.580	0.040	0.679	0.028
Pumpkin	**Random**		0.500	0.000	0.104	0.000	0.096	0.000
Raisin	**Dirichlet**	0.500	0.638	0.060	0.355	0.034	0.201	0.034
Raisin	**Hierarchy (0.9)**		0.500	0.000	0.278	0.000	0.278	0.000
Raisin	**Joint**		0.482	0.024	0.689	0.042	0.741	0.045
Raisin	**Random**		0.500	0.000	0.278	0.000	0.278	0.000
Rice	**Dirichlet**	0.572	0.522	0.066	0.060	0.008	0.073	0.010
Rice	**Hierarchy (0.9)**		0.500	0.000	0.057	0.000	0.077	0.000
Rice	**Joint**		0.604	0.009	0.630	0.022	0.553	0.023
Rice	**Random**		0.500	0.000	0.057	0.000	0.077	0.000
Spam	**Dirichlet**	0.394	0.524	0.050	0.072	0.007	0.043	0.004
Spam	**Hierarchy (0.9)**		0.500	0.000	0.069	0.000	0.045	0.000
Spam	**Joint**		0.388	0.007	0.780	0.020	0.800	0.012
Spam	**Random**		0.500	0.000	0.069	0.000	0.045	0.000
Wine	**Dirichlet**	0.754	0.801	0.055	0.041	0.003	0.031	0.009
Wine	**Hierarchy (0.9)**		0.500	0.000	0.026	0.000	0.078	0.000
Wine	**Joint**		0.786	0.007	0.725	0.043	0.606	0.048
Wine	**Random**		0.500	0.000	0.026	0.000	0.078	0.000

Table 4.S2: **Dataset statistics.** Statistics of the datasets used to evaluate bias induction and bias mitigation strategies.

Dataset	Number of samples	Feature types	Number of features	Classes	Balance
Breast cancer	569	Continuous	30	Malignant (0), Benign (1)	0: 37%, 1: 63%
Wine	6497	Continuous	11	Red (0), White (1)	0: 25%, 1: 75%
Mushroom	8124	Categorical	22	Poisonous (0), Edible (1)	0: 48%, 1: 52%
MNIST	1797	Continuous	64	10 classes: 0 to 9	0, ..., 9: 10%
Fire	17442	Mixed (1C)	6	Non-extinction(0), Extinction (1)	0: 50%, 1: 50%
Spam	4601	Continuous	57	Safe (0), Spam (1)	0: 61%, 1: 39%
Adult	48842	Mixed (7C)	13	Earns <50k (0), Earns >50k (1)	0: 76%, 1: 24%
Rice	3810	Continuous	7	Cammeo (0), Osmancik (1)	0: 43%, 1: 57%
Raisin	900	Continuous	7	Kecimen (0), Besni (1)	0: 50%, 1: 50%
Pistachio	2148	Continuous	17	Kirmizi (0), Siit (1)	0: 57%, 1: 43%
Pumpkin	2500	Continuous	13	Cercevelik (0), Urgup (1)	0: 52%, 1: 48%

Table 4.S3: **Parameter values of models trained using supervised learning and bias mitigation methods.** Model types: LR, logistic regression (scikit-learn implementation); RF, random forest (LightGBM implementation); and NN, neural network (Keras implementation).

Method	Hyperparameter	Brief description	Value
LR	penalty	Regularization type	L2
	C	Inverse of reg. strength	5.0
	max_iter	Maximum iteration	100
RF	subsample	Subsample ratio of training samples	0.9
	subsample_freq	Frequency of subsample	1
	min_child_weight	Minimum sum of instance weight (Hessian) needed in a child	0.01
	reg_lambda	L2 regularization term	5
	num_leaves	Maximum tree leaves	31
	max_depth	Maximum tree depth	-1
	n_estimators	Number of decision trees	100
NN	activation	Activation function of layers	RELU
	optimize	Optimization strategy	Adam
	loss	Loss function to optimize	Cross entropy

5

METRIC-DST: MITIGATING SELECTION BIAS THROUGH DIVERSITY-GUIDED SEMI SUPERVISED METRIC LEARNING

Yasin I. TEPELI

Mathijs DE WOLF

Joana P. GONÇALVES

This chapter is published in: arXiv (2024), doi: doi.org/10.48550/arXiv.2411.18442 (submitted)
Supplementary material is also available online at: https://doi.org/10.48550/arXiv.2411.18442

Selection bias poses a critical challenge for fairness in machine learning, as models trained on data that is less representative of the population might exhibit undesirable behavior for underrepresented profiles. Semi-supervised learning strategies like self-training can mitigate selection bias by incorporating unlabeled data into model training to gain further insight into the distribution of the population. However, conventional self-training seeks to include high-confidence data samples, which may reinforce existing model bias and compromise effectiveness. We propose Metric-DST, a diversity-guided self-training strategy that leverages metric learning and its implicit embedding space to counter confidence-based bias through the inclusion of more diverse samples. Metric-DST learned more robust models in the presence of selection bias for generated and real-world datasets with induced bias, as well as a molecular biology prediction task with intrinsic bias. The Metric-DST learning strategy offers a flexible and widely applicable solution to mitigate selection bias and enhance fairness of machine learning models.

5

5.1. INTRODUCTION

Machine learning (ML) algorithms enabling predictive modeling and data-driven decision-making have contributed important advances across disciplines. The increasing pervasiveness of ML in society also raises awareness about its potential impact on people's lives and the need to ensure fairness in predictions made by ML models. Selection bias is one of the most common sources of unfairness in ML, where the training data is not representative of the underlying population, with some groups or profiles appearing more prominently while others might be excluded [1–5].

Mitigating selection bias is crucial to ensure fairness, accuracy, and reliability of machine learning models. Several approaches have been proposed to address this issue, including data preprocessing techniques [6–8], reweighting methods [9–15], and algorithmic fairness measures [16, 17]. Most of these methods are proposed under the umbrella term of domain adaptation (DA), which adjusts models to account for distribution shifts between source and target prediction domains. Available DA approaches typically focus on adapting models to specific test sets, which can limit the generalizability of the models beyond the train and test data.

Semi-supervised learning has gained traction to address bias by leveraging abundant unlabeled data that might offer further insight into the true underlying distribution of the data but cannot be directly used in supervised learning. A common framework for semi-supervised learning relies on self-training that iterates between (i) building a model with supervised learning and (ii) using the model both to predict pseudo-labels for unlabeled samples and to select a subset to incorporate into the learning during the subsequent iteration. Conventional self-training selects pseudo-labeled samples based on model confidence, often focusing on the most confident predictions [18, 19], which can reinforce the bias in the data by incorporating samples similar to others already in the biased labeled set [20–22].

To counteract this confirmation bias, the DCAST [21] semi-supervised strategy gradually includes diverse pseudo-labeled samples above a relaxed confidence threshold. Diversity is achieved by choosing samples from distinct clusters, identified based on sample distances or dissimilarities. The preferred DCAST approach leverages distances within a learned class-informed latent space, rather than the original feature space, to lessen the influence of uninformative features. This can be especially important for high-dimensional data, however the approach cannot be combined with classifiers lacking such latent representations. Additionally, DCAST presumes that the different clusters in the latent space capture diverse sets of samples, which can be suboptimal if the data cannot be meaningfully clustered.

We introduce Metric-DST, a self-training framework relying on metric learning to enable more general selection bias mitigation for diversity-aware prediction models. Metric learning offers a suitable alternative to obtain a class-informed latent space [23] by optimizing a transformation of the original feature space to a lower dimensionality in a class-contrastive manner. Metric-DST uses this mechanism to learn a bounded latent space where distances between samples reflect both dissimilarity and class membership, and then generates random locations within the space to select diverse samples that are predicted by a companion classifier above a relaxed confidence threshold. Metric-DST exploits sample diversity during model learning to improve generalizability, and can be used with virtually any type of classifier.

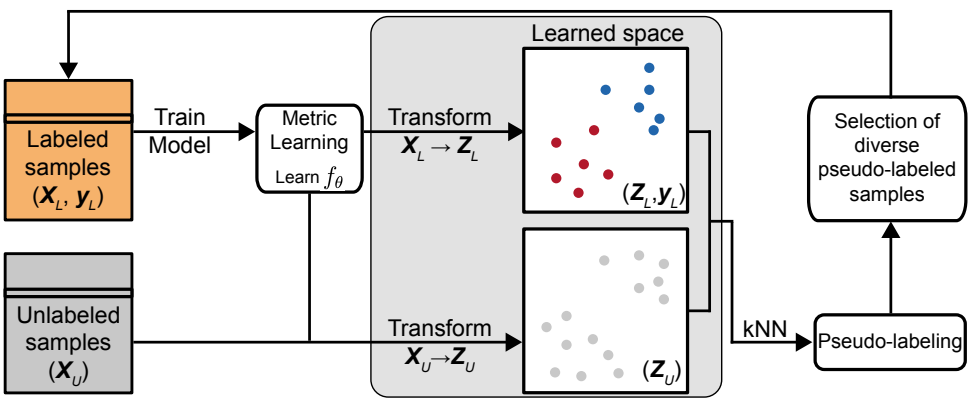


Figure 5.1: **Overview of the Metric-DST methodology.** A Metric-DST iteration encompasses 1) training a metric learning model on labeled data that can be used to transform both labeled and unlabeled samples into an embedding space, 2) obtaining predicted pseudo-labels and model confidence values for unlabeled samples using k-nearest neighbors (kNN) on the embedding space representations, 3) selecting diverse pseudo-labeled samples distributed across the learned embedding space and adding them to the labeled set for the subsequent iteration.

5.2. RESULTS AND DISCUSSION

5.2.1. MODEL LEARNING UNDER SELECTION BIAS WITH METRIC-DST

The aim of the proposed semi-supervised Metric-DST framework is to learn a prediction model with improved robustness to selection bias by leveraging available unlabeled data for additional representativeness of the underlying population distribution (Fig. 5.1). Using self-training, unlabeled samples are gradually pseudo-labeled and selected to be incorporated into model learning. Since conventional self-training is prone to reinforcing data bias, Metric-DST seeks to counter such behavior through the selection of diverse pseudo-labeled samples. To achieve this, Metric-DST exploits a metric learning model formulation to generate class-informative representations of samples in a bounded latent space. Briefly, at each self-training iteration, Metric-DST first learns a transformation function or model f_θ from the labeled samples $\boldsymbol{X}_L$ and respective labels $\boldsymbol{y}_L$ using metric learning with a contrastive loss to optimize class separation in the learned latent space (Fig. 5.1, Methods). The learned transformation f_θ is used to obtain embeddings or representations $\boldsymbol{Z}_U$ of the unlabeled samples $\boldsymbol{X}_U$ in the new space. Then, the learned representations are used by Metric-DST in two ways: (i) to make predictions and thus assign pseudo-labels to unlabeled samples, using a simple weighted k nearest neighbors classifier; and (ii) to select $p/2$ diverse pseudo-labeled samples per class as randomly generated points in the latent space whose nearest pseudo-labeled sample satisfies a relaxed confidence threshold μ.

We evaluated the bias mitigation ability of the proposed diversity-guided Metric-DST method against two approaches: Metric-ST, a similarly semi-supervised variant relying on conventional self-training without diversity; and Supervised, vanilla supervised learning. Generally, our goal was to investigate if Metric-DST could build models with improved robustness to selection bias and if the diversity strategy was effective in that regard. All three strategies used metric learning with an identical neural network architecture, in combination with weighted kNN for prediction. Different bias scenarios were also considered across generated and real-world benchmark binary classification datasets, as well as a molecular biology challenge inherently affected by selection bias called synthetic lethality prediction. Each of the three learning methods was assessed for each bias scenario across 10 different train/test splits (Methods).

5.2.2. METRIC-DST MITIGATES BIAS INDUCED TO GENERATED AND REAL-WORLD DATASETS

We first evaluated Metric-DST, and the Metric-ST and Supervised baselines, on binary classification tasks using artificially generated and real-world benchmark datasets with induced selection bias. Briefly, for each train/test split, the train set comprising 90% of the data was further randomly split into labeled (30%) and unlabeled (70%) subsets. The Supervised approach trained using labeled data alone, while Metric-(D)ST trained using both labeled and unlabeled data. For experiments using bias, selection bias was induced only to the labeled subset, enabling us to

assess if the trained model could generalize beyond the biased training data and also leverage the unlabeled data to do so. For comparison, we also trained separate models without bias induction and using a random selection of samples (as many as used in the biased selection).

Moons dataset and delta bias. The generated moons dataset contained 2000 data points in 2 dimensions, distributed over two classes, with the class-specific point clouds forming interleaving moon shapes (Fig. 5.2a). We induced selection bias using a technique termed delta bias to obtain a set of either 100 or 200 class-balanced samples in the vicinity of user-defined points Δ_0 and Δ_1 for classes 0 and 1, respectively. We also used two combinations of Δ points: identical for both classes, $\Delta_0 = \Delta_1 = (0,0)$; and different per class, with $\{\Delta_0 = (1,0.5), \Delta_1 = (0,0)\}$. The effect of delta bias was confirmed by visualizing the samples in the 2D space. We observed that the biased selection excluded relevant regions of the point clouds, which could shift the decision boundary of a classifier (Fig. 5.2a).

Selection bias had a noticeable impact on models built using supervised learning, where training on a biased selection generally resulted in lower performances compared to training on the original data (Fig. 5.2b, blue vs. grey), with differences in median AUROC between 0.02 and 0.28. The effect of supervised training on a biased selection was also larger than that of training on a random selection with the same number of samples (Fig. 5.2b, blue vs. purple), enabling us to disentangle the influence of bias and sample size. We further noticed that the decrease in supervised learning performance was stronger using selection bias with distinct Δ points per class, leading to median AUROC values of 0.69 and 0.84 for 100 and 200 samples, compared to 0.88 and 0.95 using identical Δ points (Fig. 5.2b, blue). The Metric-ST variant without diversity was unable to overcome the induced selection bias, leading to large variances accompanied by decreases in performance compared to supervised learning across all four bias settings (Fig. 5.2b, yellow vs. blue), three of which were statistically significant (p-values < 0.03). In contrast, the diversity-guided Metric-DST method showed a significant improvement in performance with 100 samples and identical Δ points (median AUROC: supervised 0.88, Metric-DST 0.93, p-value: 0.037) and no significant performance differences but smaller variances in performance for the three remaining bias settings compared to supervised learning (Fig. 5.2b, green vs. blue).

Overall, on the moons dataset, Metric-DST delivered models with increased robustness to induced delta bias compared to conventional self-training (Metric-ST). The proposed diversity-guided approach also performed comparably or better than supervised learning.

Higher-dimensional two-cluster datasets and hierarchy bias. We complemented the generated data using 8 balanced binary classification datasets of 2000 samples spread over two clusters per class. The datasets spanned four dimensionalities or numbers of features (16, 32, 64, and 128), paired with an additional setting

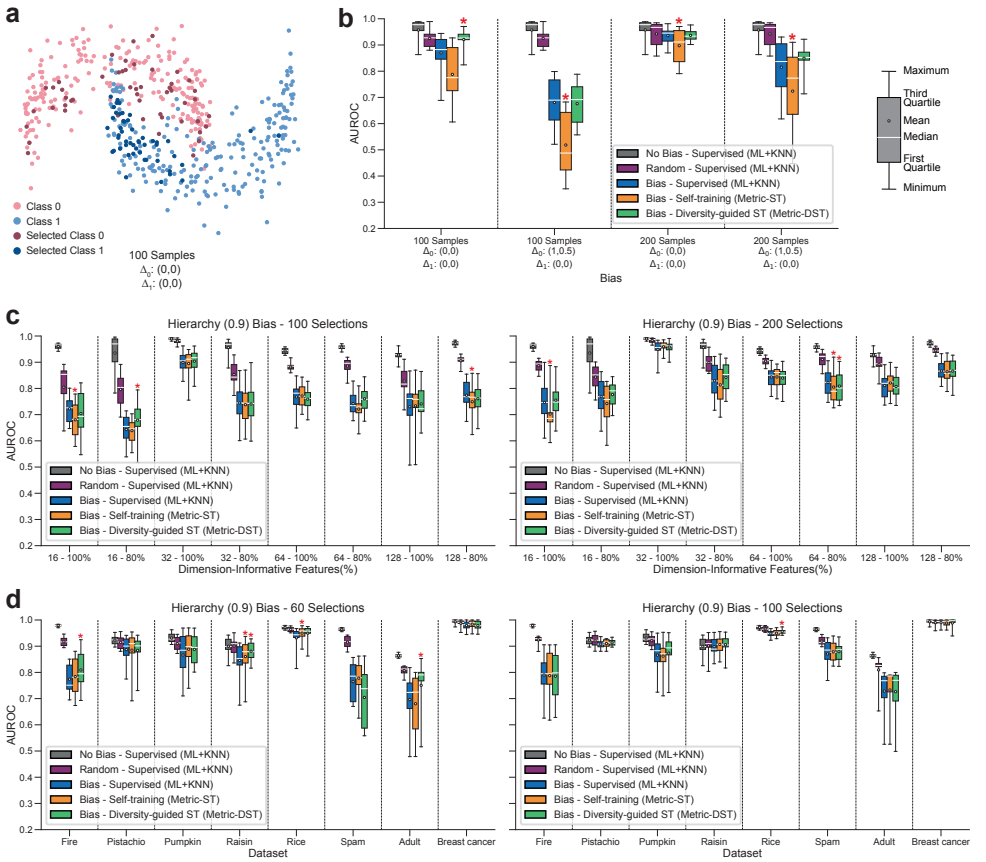


Figure 5.2: **Mitigation of selection bias induced to generated and real-world benchmark data.** **(a)** Samples selected by delta bias ($\Delta_0 = \Delta_1 = (0,0)$ for classes 0 and 1) highlighted on a scatter plot of the artificially generated 2D moons dataset. Performance (AUROC) of supervised and semi-supervised Metric-(D)ST methods using metric learning and kNN on: **(b)** generated 2D moons dataset of 2000 samples with four delta bias induction settings, selecting 100 or 200 samples with $\Delta_0 = \Delta_1 = (0,0)$ and $\{\Delta_0 = (1,0.5), \Delta_1 = (0,0)\}$, **(c)** generated higher-dimensional datasets of 2000 samples and 16, 32, 64, and 128 features with hierarchy bias induction (ratio $b = 0.9$) selecting 100 or 200 samples, and **(d)** eight real-worlddatasets with hierarchy bias induction ($b = 0.9$) targeting the selection of 60 and 100 samples. Results of 10-fold cross-validation, with all methods evaluated using the same folds (train/test splits) and the same divisions of the train sets into labeled and unlabeled subsets. Methods included: supervised model trained on the complete labeled set (No Bias), on a biased selection (Bias), or on randomly selected samples (Random, same number as the biased selection); and semi-supervised models, using conventional self-training (Metric-ST) or diversity-guided self-training (Metric-DST) on the biased labeled train set plus the unlabeled train set. The red asterisks stand for significant difference (p-value<0.05) between the performances of the method with asterisk and the biased supervised method based on a two-sided Wilcoxon signed-rank test.

determining whether 100% or 80% of those features were informative for the classification task. We selected a biased subset of 100 or 200 samples from each dataset using hierarchy bias with bias ratio $b = 0.9$ [21], which favored samples from one specific cluster identified de novo per class (Methods, Supplementary Fig. 5.S1).

Training on a random selection of 100 or 200 samples caused a decrease in the performance of supervised learning across 7 of the 8 datasets compared to training without bias (Fig. 5.2c, purple vs. grey). The biased selection using hierarchy bias

led to a further decrease in supervised model performance beyond the impact of random selection and respective reduction in sample size, with a change in median AUROC between 0.06 and 0.17 for 100 samples and between 0.03 and 0.14 for 200 samples (Fig. 5.2c, blue vs. purple). The Metric-ST method relying on conventional self-training was comparable or worse than supervised learning concerning robustness to induced hierarchy bias, and led to significant decreases in performance for 2 out of the 8 datasets for both 100 and 200 selected samples (Fig. 5.2c, yellow vs. blue, p-values < 0.03). Metric-DST was mostly comparable to supervised learning, with only two significant differences: a performance increase for 100 samples with 16 dimensions of which 80% informative (Fig. 5.2c, green vs. blue, p-value < 0.05), and a performance drop for the 200 sample selection of the 64-dimensional dataset with 80% informative features (p-value 0.004). We also observed non-significant increases in median AUROC for 100 samples with 64 dimensions of which 80% informative (medians 0.73 vs 0.76) and for 200 samples with 16 and 32 dimensions of which 80% informative (medians 0.77 vs 0.79 for 16 dimensions and 0.83 vs 0.85 for 32 dimensions).

On the generated higher-dimensional datasets, Metric-DST displayed superior robustness to induced hierarchy selection bias compared to the Metric-ST approach. Mostly Metric-DST was able to protect the supervised learning performance, with occasional very modest improvements.

Real-world benchmark datasets with hierarchy bias. Event though artificially generated data and bias induction may offer some sense of control over the conditions of the experiments, there is still a multiplicity of factors to consider, and it is unlikely that the generated datasets capture the complexity and exhibit the behavior of real-world datasets. For this reason, we also evaluated the mitigation of selection bias on 8 real-world binary classification tasks using public datasets. We induced hierarchy selection bias with ratio $b = 0.9$, targeting selections of 60 and 100 samples due to the limited size of some datasets (Methods, Supplementary Fig. 5.S2).

Training on the biased sample selection led to an overall decrease in the performance of supervised learning models compared to training on the original data or a random selection (Fig. 5.2d). The effect of the induced hierarchy bias was however less pronounced using the larger 100 sample selection, and did not significantly affect model performance for datasets like *Raisin* and *Breast cancer* for which the sample count corresponded to a substantial portion of the data ($\simeq$153 samples in the labeled training set before bias induction).

Using the 60 sample selection, Metric-ST improved performance in two datasets, *Raisin* and *Rice* (p-values 0.006 and 0.020). Metric-DST resulted in significantly improvements for three datasets, *Fire* (p-value 0.049), *Raisin* (p-value 0.020), and *Adult* (p-value 0.002). Additionally, Metric-DST increased performance in the *Fire* dataset as well, but the change was not significant (p-value 0.064). While Metric-ST showed potential, Metric-DST demonstrated a greater overall impact. Using the 100 sample selection, neither semi-supervised Metric-(D)ST approach delivered

significant performance improvements consistently across datasets: only on one instance Metric-DST improved significantly over supervised learning on the biased data, on the *Rice* dataset (p-value 0.020). It is worth noting that the larger biased selection of 100 samples did not affect the original performance as much, leaving limited room for improvement for semi-supervised learning methods. Some datasets, especially *Breast cancer*, could also potentially harbor easily separable classes, a dynamic that may cause biased selections to still capture the original decision boundary, thereby rendering semi-supervised methods less effective. Overall, Metric-DST showed improved robustness to selection bias compared to Metric-ST, and the ability to preserve or improve performance compared to supervised learning across all datasets.

5.2.3. METRIC-DST MITIGATES SELECTION BIAS FOR SYNTHETIC LETHALITY PREDICTION

The evaluation with induced biases on generated and real-world benchmark datasets enabled us to assess the effectiveness of the learning methods in cases where the biases in the data are unknown or difficult to characterize. However, artificially induced biases also have their limitations, and the insights gained from such experiments might not translate well to real-world prediction tasks inherently affected by complex selection biases. To cover this scenario, we finally evaluated Metric-DST on a molecular biology challenge, called synthetic lethality (SL) prediction, where the set of labeled samples available for training is known to be biased.

We performed three experiments to evaluate Metric-DST on SL prediction, which were designed to control the extent of the difference in selection bias between paired train and test sets (Methods).

Randomized split for similar train/test selection bias. We assessed the supervised and semi-supervised learning methods on SL prediction for each of five distinct cancer types under similar selection bias between train and test sets. The supervised model showed noteworthy median AUPRC performances for the BRCA and LUAD cancer types (0.854 and 0.837, respectively). Metric-ST and Metric-DST both led to marginal, non-significant improvements in median AUPRC performance compared to supervised learning for LUAD (0.843 and 0.851), and Metric-DST also for BRCA (0.859) (Fig. 5.3a, green vs. blue). Possibly due to the ample sample sizes (Supplementary Table 5.S1-5.S3) and high starting performances of BRCA (1443 SL, 1010 non-SL pairs) and LUAD (594 SL, 5509 non-SL pairs), the use of additional pseudo-labeled data yielded inconsequential performance gains.

We noticed improvements of Metric-ST and Metric-DST over supervised learning in median AUPRC for the cancer types with more limited numbers of labeled samples, including CESC, OV, and SKCM. However, owing to relatively large variances, the only significant improvement was seen with Metric-DST for CESC (Fig. 5.3a, green

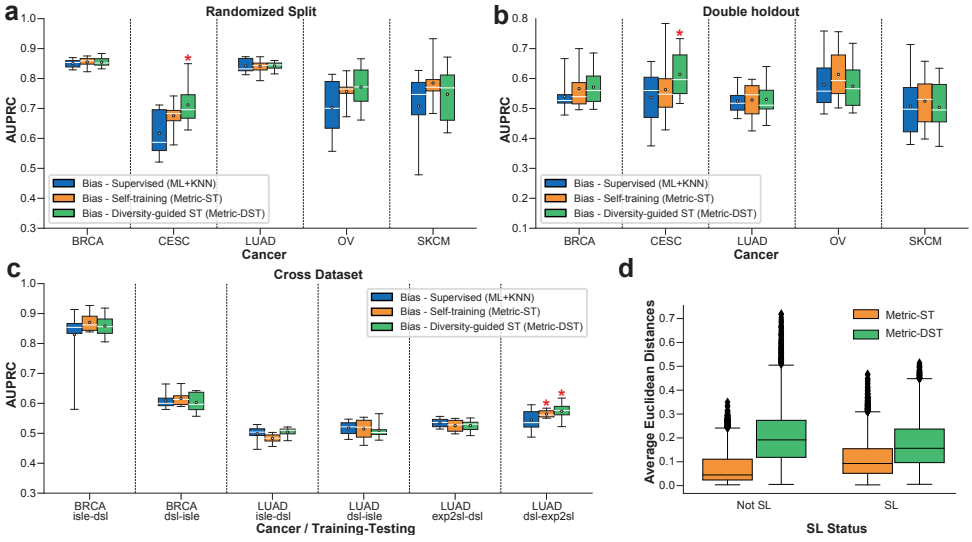


Figure 5.3: **Mitigation of intrinsic selection bias for synthetic lethality prediction.** Prediction performance (AUPRC) of synthetic lethality prediction models trained and tested per cancer type using supervised learning or the semi-supervised Metric-ST and Metric-DST methods for 10 train/test splits. Three types of splits were used to control the degree of similarity in selection bias between the train and test sets: **(a)** *Randomized split*, **(b)** *Double holdout*, **(c)** *Cross dataset*. For **(a)**, **(b)**, and **(c)**, boxplots include all points (no outlier detection), and the white circles denote the mean values. **(d)** Average Euclidean distances between pseudo-labeled samples selected by Metric-ST and Metric-DST per class, with diamonds denoting outliers. The red asterisks denote significant differences in performance (p-value < 0.05) between the method with an asterisk and the biased supervised method based on a two-sided Wilcoxon signed-rank test.

vs. blue, p-value 0.014). Additionally, both Metric-DST and Metric-ST seemed superior to supervised learning for CESC and OV in median AUPRCs (Metric-DST CESC: 0.696, OV: 0.772; Metric-ST CESC: 0.683, OV: 0.765; supervised CESC: 0.587, OV: 0.701). We also saw a moderate non-significant improvement in median AUPRC of the Metric-(D)ST methods over supervised learning for the SKCM dataset (Metric-DST 0.769, Metric-ST 0.771, supervised 0.747).

In summary, the application of Metric-DST looked cautiously promising in the context of a randomized split, preserving similar biases between train and test sets, for cancer types with more limited sample sizes (CESC, OV, and SKCM).

Double holdout for distinct train/test selection bias. We also assessed Metric-DST with paired train and test sets yielding different biases, adopting a double holdout technique where gene overlap between test and train sets was entirely prevented. This restrictive split resulted in a diminished train set size, reaching its lowest for the CESC dataset with only 90 samples.

Relative to the randomized split experiment, supervised learning using double holdout resulted in lower median AUPRCs (*Randomized split* vs. *Double holdout* in BRCA, OV, CESC, SKCM, and LUAD: 0.853 vs. 0.527, 0.701 vs. 0.558, 0.587 vs. 0.560, 0.747 vs. 0.497, and 0.837 vs. 0.517, respectively) (Fig. 5.3b). This was expected due to the restrictions imposed by the double holdout to ensure zero overlap in

individual genes, in addition to zero overlap in gene pairs between train and test sets. Although some performance differences could be observed between the Metric-(D)ST methods and supervised learning for the BRCA, LUAD, OV, and SKCM datasets, none of them reached statistical significance. Metric-ST showed higher median AUPRC than Metric-DST and supervised learning for LUAD, OV, and SKCM, while Metric-DST did better in this regard for BRCA and CESC. The only significant improvement in AUPRC performance was recorded for CESC, with Metric-DST outperforming the supervised model (median AUPRC 0.60 vs. 0.56, p-value 0.010). It is important to note that the semi-supervised methods did not cause significant decreases in performance relative to supervised learning.

Multiple factors might explain the lack of effectiveness of Metric-DST for some cancer types. For instance, the restrictions imposed by the double holdout procedure may have caused too extreme differences in biases between the train and test sets, due to the absence of shared genes. An additional contributing factor could be the reduction in train set size, exemplified by the CESC dataset (Supplementary Table 5.S4). The impact of these constraints also resulted in a large performance decrease for the baseline supervised model (Fig. 5.3a-b), making the recovery more difficult for the semi-supervised techniques which rely heavily on an initial successful model.

Cross dataset split with naturally occurring selection bias. To evaluate bias mitigation with naturally occurring differences in selection bias between train and test sets, we set up the data splits to train using SL labeled samples from one study and test on SL labeled samples from another study, encompassing six permutations across three studies (ISLE, dSL, and EXP2SL).

For BRCA, when trained on ISLE and tested on dSL, both Metric-ST and Metric-DST induced an increase in the minimum AUPRC performance by over 0.2, but overall there were no significant differences in performance between the two semi-supervised methods and supervised learning (Fig. 5.3c). For LUAD, the Metric-(D)ST methods resulted in significant performance improvements only for the setting that trained on dSL and tested on EXP2SL significant differences (median AUPRC: Supervised 0.536; Metric-ST 0.561 with p-value 0.049; Metric-DST 0.576 with p-value 0.014). The remaining study combinations did not reveal significant changes either, but we observed small decreases in median AUPRC for Metric-ST trained on ISLE and tested on dSL, as well as for Metric-DST trained on dSL and tested on ISLE.

Taking all experiments on synthetic lethality prediction into account, it is important to highlight that the two semi-supervised Metric-(D)ST methods significantly outperformed supervised learning on three scenarios, while never performing significantly worse. Instances where Metric-ST and Metric-DST yielded no clear impact might be attributed to multiple factors, including the inherent complexity of the problem with baseline supervised learning performances hovering around 0.5, or extreme disparities between the train and test sets.

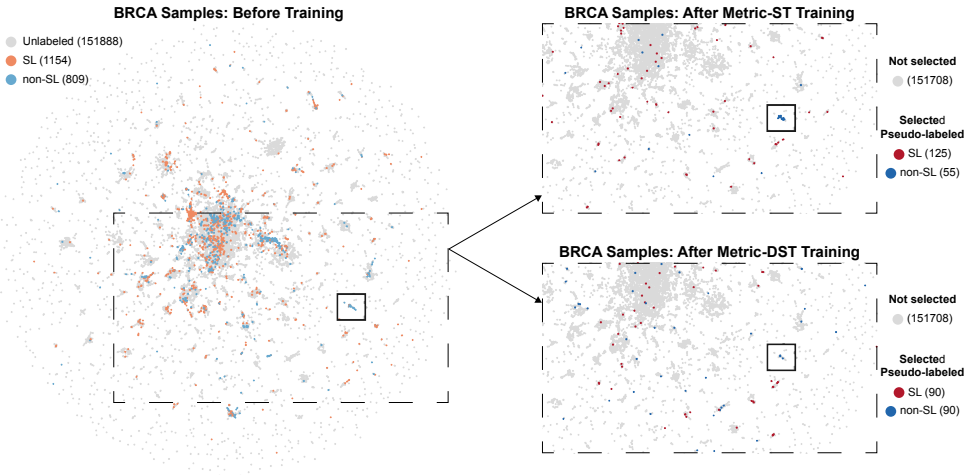


Figure 5.4: **UMAP projections of the SL dataset for BRCA.** On the left, the training samples are highlighted before the training. The top right plot shows the pseudo-labeled samples selected by Metric-ST and the pseudo-labeled samples selected by Metric-ST during the training. The number of samples of each class is stated in parentheses. The highlighted box highlights a cluster dominated by gene pairs containing the gene CDH1.

METRIC-DST PROMOTES DIVERSITY IN SELECTED PSEUDO-LABELED SAMPLES

To verify if the diversity approach of Metric-DST was able to select more diverse samples, we analyzed the Euclidean pairwise distances between pseudo-labeled samples assigned to the same class label in the learned embedding space, using the BRCA *Randomized split* as an example (Fig. 5.3d). The distances were larger on average for pseudo-samples selected by Metric-DST, confirming a more heterogeneous sample selection compared to Metric-ST.

We also examined the distribution of selected pseudo-labeled samples in a UMAP projection of the labeled and unlabeled samples original feature space onto two dimensions (Fig. 5.4). The projection showed no clear linear separation between the labeled samples of the two classes, SL and non-SL, reflecting the complexity of the prediction task and its underlying decision boundary. In addition, most clusters apparent in the UMAP embedding contained no labeled samples that could be used for supervised training, which further illustrates the lack of representation and the extent of the selection bias problem in synthetic lethality.

More detailed analysis revealed that Metric-ST incorporated a total of 55 and 125 pseudo-labeled samples, respectively assigned non-SL and SL labels. Of the 55 non-SL pseudo-labeled samples, 29 were in a cluster dominated by the gene CDH1 (Fig. 5.4). This cluster originally contained 507 labeled and unlabeled samples, of which 501 contained the gene CDH1. The fact that the method focused heavily on one cluster demonstrates the main drawback of using conventional self-training and relying on model confidence alone to mitigate the effect of selection bias. In contrast, Metric-DST was able to select a more varied set of 90 pseudo-labeled samples, of which only 6 originated from the “CHD1” cluster.

Together, these findings highlight the ability of Metric-DST to promote diversity

while incorporating unlabeled samples into the learning of a prediction model.

5.3. CONCLUSION

In this work, we proposed Metric-DST, a semi-supervised framework coupled with metric learning to build prediction models with improved robustness to sample selection bias. Metric-DST relies on self-training to incorporate unlabeled samples for additional representation and insight into the underlying distribution of the population. Crucially, Metric-DST introduces a strategy to counter confirmation bias of conventional self-training by learning from a more diverse set of samples. Diversity is introduced via metric learning of a class-contrastive representation, which facilitates the pseudo-labeling and identification of dissimilar unlabeled samples to include in the training.

Evaluation using artificially generated and real-world datasets with induced selection bias suggested the potential of self-training to enhance model generalizability, yet also its susceptibility to exacerbate data bias. The proposed diversity-guided approach, Metric-DST, showed greater resilience than conventional self-training, albeit with modest performance improvements. Application to synthetic lethality prediction showed that semi-supervised metric learning could augment performance in scenarios where train and test sets yielded similar or distinct naturally occurring selection biases. It was reassuring that Metric-DST was able to preserve the performance obtained with supervised learning or deliver more robust models in all contexts, and especially under challenging conditions, such as with limited numbers of training samples or weak baseline models. Ultimately, the effectiveness of Metric-DST is contingent upon factors such as the performance of the underlying base model, the type and extent of the data bias, and the ratio of features to samples, among others. Future work warrants a deeper exploration of the potential of the Metric-(D)ST learning framework, including refinement of neural network architectures and loss functions. Leveraging metric learning as a means of diversifying pseudo-sample selection in combination with various classifiers could further expand the scope of the model. We also envision further addressing the limitations of the existing pseudo-labeled sample selection approach, which could be extended to ensure a more comprehensive representation of the embedding space by excluding unpopulated regions.

5.4. EXPERIMENTAL PROCEDURES

5.4.1. METRIC-DST

Metric-DST is a semi-supervised ML framework based on metric learning to obtain an embedding function or transformation that is informative for a classification task of interest with increased robustness to selection bias. Learning is accomplished via self-training, where the transformation is gradually refined by incorporating a diverse selection of newly pseudo-labeled unlabeled examples into the training

process. The learned transformation serves the dual purpose of predicting pseudo-labels and assessing sample diversity to counter the data bias.

Each self-training iteration involves three steps: (1) learn a metric embedding function from the labeled data such that the latent representation of a sample also yields pertinent information about class separation, (2) pseudo-label unlabeled samples based on the learned transformation so they can be considered as candidates for selection and training, (3) select a diverse subset of pseudo-labeled samples and include them in the labeled train set for the next iteration.

LEARNING OF A METRIC EMBEDDING FUNCTION

At iteration t, Metric-DST first learns a transformation function or model $f_\theta^{(t)}$ based on the labeled samples in matrix $\boldsymbol{X}_L^{(t)}$ and the corresponding binary labels $\boldsymbol{y}_L^{(t)}$ using metric learning. The general goal is to learn a transformation of an individual sample vector $\boldsymbol{x}$ to a latent embedding representation $\boldsymbol{z} = f_\theta^{(t)}(\boldsymbol{x})$, guided by class assignments and inter-sample distances, such that samples of the same class are closer together and samples from different classes are distanced further apart in the learned embedding space. Various model architectures could be used for the transformation, in this case we used a feed-forward neural network with a single hidden layer. The model is optimized based on the contrastive loss function designed to minimize intra-class distances and maximize inter-class distances of samples in the embedding space (Eq. 5.1).

$$\mathscr{L}_{contrastive} = \sum_{(i,j)\in P} \mathbb{1}_{y_i=y_j} \max\{0, d_{i,j} - m_{pos}\} + \mathbb{1}_{y_i\neq y_j} \max\{0, m_{neg} - d_{i,j}\} \tag{5.1}$$

Here, $d_{i,j}$ denotes the Euclidean distance between samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in the embedding space, thus $d_{i,j} = d(f_\theta^{(t)}(\boldsymbol{x}_i), f_\theta^{(t)}(\boldsymbol{x}_j))$. Symbol P represents the set of all sample pairs within a training batch, and the indicator function $\mathbb{1}_{condition}$ takes value 1 if the condition holds or 0 otherwise. The positive and negative margins, m_{pos} and m_{neg}, are used to prevent the algorithm from forcing samples with the same labels to overlap completely or samples with different labels to be separated infinitely. Specifically, the distance between samples with the same labels only increases the loss when it exceeds the positive margin, and the distance between samples with different labels stops contributing to the loss once the distance exceeds the negative margin.

Once the transformation has been learned from the labeled samples $\boldsymbol{X}_L^{(t)}$, it can be applied to obtain embedding representations for unlabeled samples in $\boldsymbol{X}_U^{(t)}$ as well. We denote the embedding matrix containing the representations of all samples, labeled and unlabeled, by $\boldsymbol{Z}^{(t)}$.

PSEUDO-LABELING OF UNLABELED SAMPLES THROUGH METRIC EMBEDDING

The transformation model $f_\theta^{(t)}$ learned from the labeled data cannot be directly used to make predictions and thus assign pseudo-labels to unlabeled samples. To

classify the unlabeled samples, Metric-DST applies a weighted version of k nearest neighbors (kNN) to the embedding matrix $\boldsymbol{Z}^{(t)}$ with the learned representations $\boldsymbol{z}^{(t)} = f_\theta^{(t)}(\boldsymbol{x})$ of all samples. For a given unlabeled sample i with representation $\boldsymbol{z}_i^{(t)} \in \boldsymbol{Z}^{(t)}$, Metric-DST identifies the set $N_i^{(t)}$ of its k closest labeled samples in $\boldsymbol{Z}^{(t)}$. The prediction class probability $\bar{y}_i \in [0,1]$ for sample $\boldsymbol{x}_i$ is then calculated as a weighted average of the probabilities of the k neighbors, as given by Eq. 5.2. The calculation factors in the distance of each neighbor representation to $\boldsymbol{z}_i$, so that closer neighbors contribute more than farther ones.

$$\bar{y}_i = \frac{\sum_{n \in N_i^{(t)}} y_n \times (1 - d_{i,n}) + (1 - y_n) \times d_{i,n}}{|k|} \tag{5.2}$$

The probability $\bar{y}_i$ represents the confidence of the model, where values close to 1 and 0 indicate high confidence in predicting class 1 and class 0, respectively. The final class label $\hat{y}_i$ is obtained by thresholding the probability value $\bar{y}_i$ as per Eq. 5.3.

$$\hat{y}_i = \begin{cases} 1, & \text{if } \bar{y}_i > 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{5.3}$$

5

SELECTION OF DIVERSE PSEUDO-LABELED SAMPLES

After assigning pseudo-labels, Metric-DST selects which newly pseudo-labeled samples to include in the labeled set for the subsequent training iteration.

Conventional self-training (ST) typically chooses the p newly pseudo-labeled samples with the highest confidence [19], where p is a user-defined parameter. The reliance on confidence alone promotes confirmation bias, where the model is likely to follow and strengthen the selection bias present in the labeled data. Additionally, ST is not class-aware in that it does not consider that the model may not be similarly confident about prediction of different classes, which could further lead to unwanted biases such as class imbalance.

To address both issues, Metric-DST performs diversity-guided self-training (DST), which introduces sample diversity and class balancing into the selection of pseudo-labeled samples using the learned metric embedding. Diversity is achieved through randomness in the choice of each pseudo-labeled sample as follows. First, Metric-DST creates a candidate point in learned embedding space $\boldsymbol{Z}^{(t)}$ as a tuple of randomly generated coordinates in the range $[0,1]$. Then, the pseudo-labeled sample closest to the candidate point is identified based on the Euclidean distance (Eq. 5.2). The selected pseudo-labeled sample is designated for inclusion in the labeled train set for the subsequent iteration if the confidence on its prediction surpasses a predefined relaxed threshold μ. Class balance is achieved by selecting $p/2$ positive and $p/2$ negative pseudo-labeled samples sequentially using the aforementioned procedure for each self-training iteration. If Metric-DST fails to secure a sufficient number of pseudo-labeled samples within $50 \times p$ attempts for any one self-training iteration, undersampling of the majority class is employed to

obtain a class balanced set of pseudo-labeled samples.

5.4.2. EVALUATION OF METRIC-DST

We evaluated the Metric-DST semi-supervised model learning strategy proposed to mitigate sample selection bias against two baselines: Metric-ST, also a semi-supervised approach based on metric learning to train models using both labeled and unlabeled data, but paired with conventional self-training and thus missing the class-awareness and diversity elements of Metric-DST; and Supervised, referring to the traditional supervised metric learning technique to train models from labeled data alone. We used the same neural network model architecture as a basis with all learning strategies, consisting of an 8-dimensional hidden layer and a 2-dimensional output layer. Unless otherwise specified, the batch training size was set to 64, and the confidence threshold μ was set to 0.9. We further relied on weighted kNN with $k = 5$ to make predictions based on the metric embedding of a sample. Finally, we assessed the bias mitigation ability of Metric-DST across a range of binary classification tasks and selection bias scenarios, ranging from artificially generated and real-world benchmark data with induced selection bias to an important prediction task in molecular biology intrinsically affected by selection bias.

DATASETS AND SELECTION BIAS

Generated 2-dimensional moons dataset and induced delta bias. We generated the simplest "moons" dataset as a binary class-balanced set of 2000 samples or points in a 2-dimensional space, such that the samples of the two classes formed interleaving half circles (or moons), using the *make_moons* function from scikit-learn [24]. Selection bias was induced by choosing an equal number of samples from each class, while favoring samples closer to a point in space with user-defined coordinates Δ_i for each class i. We refer to this type of bias as delta bias, where we set the selection probability of each sample $\boldsymbol{x}$ according to its distance to the point $\Delta_{class(\boldsymbol{x})}$ associated with the corresponding class label $class(\boldsymbol{x})$, and then selected samples without replacement based on their normalized selection probabilities. The selection probability of a sample $\boldsymbol{x}$ was defined to decrease exponentially with the Manhattan distance to $\Delta_{class(\boldsymbol{x})}$, multiplied by a factor of 2 denoting bias strength: $P_{\boldsymbol{x}} = e^{-2\times\left(|x_1-\Delta_{class(\boldsymbol{x}),1}|+|x_2-\Delta_{class(\boldsymbol{x}),2}|\right)}$, where x_1 and x_2 are the coordinates of $\boldsymbol{x}$ and $\Delta_{class(\boldsymbol{x}),1}$ and $\Delta_{class(\boldsymbol{x}),2}$ are the coordinates of $\Delta_{class(\boldsymbol{x})}$ in the 2D space, respectively. Four different biased selections of the moon dataset were generated, two of 100 samples and two of 200 samples, combined with $\Delta_0 = (0,0)$ and $\Delta_1 = (0,0)$ or $\Delta_0 = (1,0.5)$ and $\Delta_1 = (0,0)$.

Generated higher-dimensional datasets and induced hierarchy bias. We created 8 n-dimensional datasets, each containing 2000 samples with binary class-balanced labels and forming two sample clusters per class, using the *make_classification* function from scikit-learn [24]. Each n-dimensional dataset was generated with f of the n dimensions independent and informative for the prediction task, and the

remaining $n-f$ dimensions as linear combinations of the f informative features. Briefly, the procedure for the informative dimensions creates a f-dimensional hypercube with sides measuring 3 units, then generates clusters of samples distributed around the vertices of the hypercube (within 1 standard deviation), and finally assigns two randomly chosen clusters to each class. The $n-f$ additional dimensions are generated by linearly combining randomly selected informative features. We generated 8 datasets spanning four dimensionality values (16, 32, 64, and 128), combined with 80% or 100% of informative features. We induced selection bias using hierarchy bias [21], a multivariate technique which identifies clusters of samples and makes a biased selection of k samples per class, where a bias ratio parameter b is used to skew the representation of samples selected from one specific cluster relative to the others. To achieve this, hierarchy bias performs agglomerative hierarchical clustering until it obtains one cluster with at least k samples, and then selects $k \times b$ samples uniformly at random from such cluster plus $k \times (1-b)$ samples uniformly at random from the remaining data. For the experiments with generated high-dimensional datasets, we used a challenging hierarchy bias with ratio $b = 0.9$ to create two biased selections of 100 and 200 class-balanced samples.

Real-world binary classification benchmark datasets and induced hierarchy bias. We used 8 publicly available binary classification benchmark datasets of varying dimensions, feature types, and complexity: 5 from the UCI Data Repository [25] (breast cancer, adult, spam, raisin, rice) and 3 from other sources including pistachio [26], fire [27], and pumpkin [28]. To induce selection bias, we again used hierarchy bias [21] with bias ratio $b = 0.9$ to create two biased selections of 60 and 100 class-balanced samples per dataset. The numbers of selected samples were chosen to be feasible and consistently applied across all real-world benchmark datasets.

Synthetic lethality dataset and inherent selection bias. To assess the bias mitigation ability of Metric-DST on a real-world prediction task inherently affected by selection bias, we focused on the molecular biology challenge of synthetic lethality prediction. Synthetic lethality refers to a relationship between two genes, relevant for cancer therapy [29, 30], whereby the loss-of-function of both genes leads to cell death but loss-of-function of either gene independently is not lethal[31]. Computational prediction of synthetic lethality (SL) gene pairs is key to generate promising candidates for the discovery of new SL relationships. However, the existing labeled gene pairs used for training SL prediction models suffer from extensive selection bias [32], as they are often limited to specific disease-related genes, gene families, or pathways [33–37].

Following recent work on supervised SL prediction models ELISL [38], we represented each sample or gene pair by a 128-dimensional vector expressing a relationship between the embedding representation vectors of the two genes, based on amino acid sequence. This formulation was introduced to reflect the functional

similarity of a pair of genes, and emerged as the most successful predictor of SL in ELISL models. We used SL labeled samples from 5 different cancers [38]: breast (BRCA), lung (LUAD), ovarian (OV), skin (SKCM), and cervix (CESC) (Supplementary Table 5.S1). In addition to the labeled SL gene pairs, we used a set of unlabeled samples comprising pairwise combinations of 572 genes involved in cancer and DNA repair pathways, excluding any samples already present in the labeled set [38] (Supplementary Table 5.S1). We did not use bias induction techniques with SL data, since the goal of this particular use case was to assess the behavior of the different model learning strategies in the presence of naturally occurring selection bias. We leveraged such bias for evaluation as described below.

Training and evaluation of prediction models

Generated and real-world binary classification tasks. We trained and evaluated all models using 10-fold cross-validation (CV), stratified by class. The CV procedure generated a split into train set (90%) and test set (10%) for each fold, with the train set further split randomly into labeled (30%) and unlabeled (70%) subsets. Supervised metric learning models were trained per fold on the corresponding labeled train subset, as well as biased and random selections of it. Metric-DST and Metric-ST were used to learn models per fold from the corresponding labeled train subset, as well as its biased and random selections, together with the unlabeled train subset. For the Metric-(D)ST methods, the number p of selected pseudo-labeled samples was set as the greatest even integer smaller than or equal to $\sqrt{n}$, with n referring to the number of labeled samples available for training. We induced selection bias to the labeled train subset using either delta or hierarchy bias, depending on the dataset, as previously described. Each trained model was evaluated on the unbiased test set of the corresponding fold for which it was learned using the area under the receiver-operating characteristic curve (AUROC) as performance metric. The same folds and train set splits were used across all experiments. We tested the significance of performance differences between the supervised model learned from biased data and Metric-(D)ST using two-sided Wilcoxon signed rank tests and a p-value threshold of 0.05.

Synthetic lethality prediction. We evaluated Metric-DST, Metric-ST, and supervised metric learning for SL prediction with three experiments, each involving 10 runs of model training and evaluation based on different train/test splits. We largely followed an experimental setup previously proposed and refined to assess robustness to selection bias in SL prediction [32, 38].

The *Randomized split* experiment assessed SL prediction performance without explicitly evaluating bias effects: the labeled gene pairs were randomly split into 20% train and 80% test data per run, with both subsets then expected to exhibit similar biases (Supplementary Table 5.S2 for the distribution of classes). The two other experiments evaluated the ability of the model learning strategies to mitigate selection bias in training data. The *Double holdout* split was set up to promote

distinct biases between train and test data by distributing the labeled gene pairs into disjoint train/test sets per run, but this time also enforcing zero overlap of individual genes in addition to no overlap in gene pairs (More details in Supplementary Methods). For the *Cross dataset* experiment, we took advantage of the fact that different SL studies focus on distinct sets of genes and thus naturally yield varying selection bias. We therefore split the labeled gene pairs based on the three SL studies from which they were obtained: ISLE [39], dSL [40], EXP2SL[41]. Considering only cancer types and studies with a sufficient number of samples, models were trained using labeled pairs from one study and tested on labeled pairs from another study. Any gene pairs overlapping between the train and test sets, due to their inclusion in multiple studies, were removed from the train set.

For all three experiments, train and test sets were class-balanced at the start of each run by randomly undersampling the majority class, and 20% of the train set was used as a validation set for early stopping (Supplementary Table 5.S3-5.S5 for the number of samples in each experiment). Each model was trained until the validation loss did not decrease for five consecutive rounds of self-training, with the final performance evaluated on the test set. We measured performance using the area under the precision-recall curve (AUPRC) score, given that SL prediction places a greater emphasis on detecting positive SL pairs and negative pairs (non-SL) cannot be confidently identified or validated. The AUPRC score is suitable for measuring performance in this scenario, as it does not take correctly predicted negatives into account. We assessed the significance of performance differences in SL experiments using two-sided Wilcoxon signed ranked tests and a p-value significance threshold of 0.05.

The hyperparameters of Metric-(D)ST, namely the confidence threshold μ and number of pseudo-labeled samples p to select per iteration, could be set judiciously for the application to other datasets using controlled bias induction. Since the effect of these hyperparameters could be more challenging to predict for the synthetic lethality dataset with inherent selection bias, we performed grid search to identify the hyperparameter values leading to the lowest validation loss per run for each experiment (Supplementary Table 5.S6-5.S8). The final performance was obtained on the test set using the model with the selected hyperparameter values.

5.4.3. RESOURCE AVAILABILITY

DATA AND CODE AVAILABILITY

The data used in this article were obtained from publicly available sources, detailed in the Experimental procedures section. The raw data necessary to reproduce the experiments are accessible via Figshare at 10.6084/m9.figshare.27720726.v2. An implementation of the dataset generation, bias induction, and Metric-DST method in Python has been made available under an open source license at github.com/joanagoncalveslab/Metric-DST.

REFERENCES

[1] D. Wu *et al.* “Correcting sample selection bias for image classification”. In: *2008 3rd International Conference on Intelligent System and Knowledge Engineering*. Vol. 1. 2008, pp. 1214–1220.

[2] C. Persello and L. Bruzzone. “Active and Semisupervised Learning for the Classification of Remote Sensing Images”. In: *IEEE Transactions on Geoscience and Remote Sensing* 52.11 (2014), pp. 6937–6956.

[3] J. W. Richards *et al.* “ACTIVE LEARNING TO OVERCOME SAMPLE SELECTION BIAS: APPLICATION TO PHOTOMETRIC VARIABLE STAR CLASSIFICATION”. In: *The Astrophysical Journal* 744.2 (Dec. 2011), p. 192.

[4] F. Shen *et al.* “Reject inference in credit scoring using a three-way decision and safe semi-supervised support vector machine”. In: *Information Sciences* 606 (Aug. 2022), pp. 614–627.

[5] M. Melucci. “Impact of Query Sample Selection Bias on Information Retrieval System Ranking”. In: *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. 2016, pp. 341–350.

[6] J. Blitzer, R. McDonald, and F. Pereira. “Domain Adaptation with Structural Correspondence Learning”. In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. EMNLP ’06. Sydney, Australia: Association for Computational Linguistics, 2006, pp. 120–128. ISBN: 1932432736.

[7] B. Fernando *et al.* “Unsupervised Visual Domain Adaptation Using Subspace Alignment”. In: *2013 IEEE International Conference on Computer Vision*. 2013, pp. 2960–2967.

[8] W. M. Kouw *et al.* “Feature-Level Domain Adaptation”. In: *J. Mach. Learn. Res.* 17.1 (Jan. 2016), pp. 5943–5974. ISSN: 1532-4435.

[9] B. Zadrozny. “Learning and Evaluating Classifiers under Sample Selection Bias”. In: *Proceedings of the Twenty-First International Conference on Machine Learning*. ICML ’04. Banff, Alberta, Canada: Association for Computing Machinery, 2004, p. 114. ISBN: 1581138385.

[10] C.-H. Chang and J.-H. Lin. “Decision Support and Profit Prediction for Online Auction Sellers”. In: *Proceedings of the 1st ACM SIGKDD Workshop on Knowledge Discovery from Uncertain Data*. U ’09. Paris, France: Association for Computing Machinery, 2009, pp. 1–8. ISBN: 9781605586755.

[11] C.-W. Seah, I. W.-H. Tsang, and Y.-S. Ong. “Healing Sample Selection Bias by Source Classifier Selection”. In: *2011 IEEE 11th International Conference on Data Mining*. 2011, pp. 577–586.

[12] M. Sugiyama, M. Yamada, and M. C. du Plessis. “Learning under nonstationarity: covariate shift and class-balance change”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 5.6 (Aug. 2013), pp. 465–477.

[13] T. D. Nguyen, M. Christoffel, and M. Sugiyama. “Continuous Target Shift Adaptation in Supervised Learning”. In: *Asian Conference on Machine Learning*. Ed. by G. Holmes and T.-Y. Liu. Vol. 45. Proceedings of Machine Learning Research. Hong Kong: PMLR, Nov. 2016, pp. 285–300.

[14] J. Huang *et al.* “Correcting Sample Selection Bias by Unlabeled Data”. In: *Proceedings of the 19th International Conference on Neural Information Processing Systems*. NIPS’06. Canada: MIT Press, 2006, pp. 601–608.

[15] W. Du and X. Wu. “Fair and Robust Classification Under Sample Selection Bias”. In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. CIKM ’21. Virtual Event, Queensland, Australia: Association for Computing Machinery, 2021, pp. 2999–3003. ISBN: 9781450384469.

[16] A. Liu and B. Ziebart. “Robust classification under sample selection bias”. In: *Advances in Neural Information Processing Systems* 1 (Jan. 2014), pp. 37–45.

[17] W. M. Kouw and M. Loog. “Robust domain-adaptive discriminant analysis”. In: *Pattern Recognition Letters* 148 (Aug. 2021), pp. 107–113.

[18] J. E. van Engelen and H. H. Hoos. “A survey on semi-supervised learning”. In: *Machine Learning* 109.2 (Nov. 2019), pp. 373–440.

[19] D.-H. Lee. “Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks”. In: *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)* (July 2013).

[20] A. Radhakrishnan *et al. Enhancing Self-Training Methods*. 2023.

[21] Y. I. Tepeli and J. P. Gonçalves. *DCAST: Diverse Class-Aware Self-Training Mitigates Selection Bias for Fairer Learning*. 2024. arXiv: 2409.20126 [cs.LG].

[22] E. Arazo *et al. Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning*. 2019.

[23] S. Chopra, R. Hadsell, and Y. LeCun. “Learning a similarity metric discriminatively, with application to face verification”. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1. 2005, 539–546 vol. 1.

[24] F. Pedregosa *et al.* “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[25] D. Dua and C. Graff. *UCI Machine Learning Repository*. 2017.

[26] I. A. Ozkan, M. Koklu, and R. Saraçoğlu. “Classification of Pistachio Species Using Improved K-NN Classifier”. In: *Progress in Nutrition* 23.2 (July 2021), e2021044. ISSN: 1129-8723.

[27] M. Koklu and Y. S. Taspinar. “Determining the Extinguishing Status of Fuel Flames With Sound Wave by Machine Learning Methods”. In: *IEEE Access* 9 (2021), pp. 86207–86216.

[28] M. Koklu, S. Sarigil, and O. Ozbek. “The use of machine learning methods in classification of pumpkin seeds (Cucurbita pepo L.)” In: *Genetic Resources and Crop Evolution* 68.7 (June 2021), pp. 2713–2726.

[29] P. C. Fong *et al.* “Inhibition of Poly(ADP-Ribose) Polymerase in Tumors fromBRCAMutation Carriers”. In: *New England Journal of Medicine* 361.2 (July 2009), pp. 123–134. ISSN: 1533-4406.

[30] L. Hutchinson. “PARP inhibitor olaparib is safe and effective in patients with BRCA1 and BRCA2 mutations”. In: *Nature Reviews Clinical Oncology* 7.10 (Sept. 2010), p. 549.

[31] D. A. Chan and A. J. Giaccia. “Harnessing synthetic lethal interactions in anticancer drug discovery”. In: *Nature Reviews Drug Discovery* 10.5 (Apr. 2011), pp. 351–64.

[32] C. Seale, Y. Tepeli, and J. P. Gonçalves. “Overcoming selection bias in synthetic lethality prediction”. In: *Bioinformatics* 38.18 (July 2022). Ed. by K. Borgwardt, pp. 4360–4368. ISSN: 1367-4811.

[33] C. Jacquemont *et al.* “Non-specific chemical inhibition of the Fanconi anemia pathway sensitizes cancer cells to cisplatin”. In: *Molecular Cancer* 11.1 (2012), p. 26.

[34] C. M. Toledo *et al.* “Genome-wide CRISPR-Cas9 Screens Reveal Loss of Redundancy between PKMYT1 and WEE1 in Glioblastoma Stem-like Cells”. In: *Cell Reports* 13.11 (Dec. 2015), pp. 2425–39.

[35] D. Etemadmoghadam *et al.* “Synthetic lethality between CCNE1 amplification and loss of BRCA1”. In: *Proceedings of the National Academy of Sciences* 110.48 (Nov. 2013), pp. 19489–94.

[36] C. G. Hubert *et al.* “Genome-wide RNAi screens in human brain tumor isolates reveal a novel viability requirement for PHF5A”. In: *Genes & Development* 27.9 (May 2013), pp. 1032–45.

[37] D. Kranz and M. Boutros. “A synthetic lethal screen identifies FAT1 as an antagonist of caspase-8 in extrinsic apoptosis”. In: *The EMBO Journal* 33 (Jan. 2014), pp. 181–97.

[38] Y. I. Tepeli, C. Seale, and J. P. Gonçalves. “ELISL: early–late integrated synthetic lethality prediction in cancer”. In: *Bioinformatics* 40.1 (Dec. 2023). Ed. by J. Wren. ISSN: 1367-4811.

[39] J. S. Lee *et al.* “Harnessing synthetic lethality to predict the response to cancer treatment”. In: *Nature Communications* 9.1 (June 2018), p. 2546.

[40] S. Das *et al.* “DiscoverSL: an R package for multi-omic data driven prediction of synthetic lethality in cancers”. In: *Bioinformatics* 35.4 (July 2018). Ed. by R. Schwartz, pp. 701–702.

[41] F. Wan *et al.* “EXP2SL: A Machine Learning Framework for Cell-Line-Specific Synthetic Lethality Prediction”. In: *Frontiers in Pharmacology* 11 (Feb. 2020).

5.5. SUPPLEMENTARY MATERIALS

5.5.1. SUPPLEMENTARY FIGURES

Figure 5.S1: **Impact of hierarchy bias on the UMAP latent space for a generated higher dimensional dataset.** 100 Samples selected by hierarchy (0.9) bias highlighted on the latent UMAP space of the labeled train set for artificially generated higher dimensional dataset with 16 dimensions and 80% informative features. Results are shown for run 1 (arbitrarily chosen).

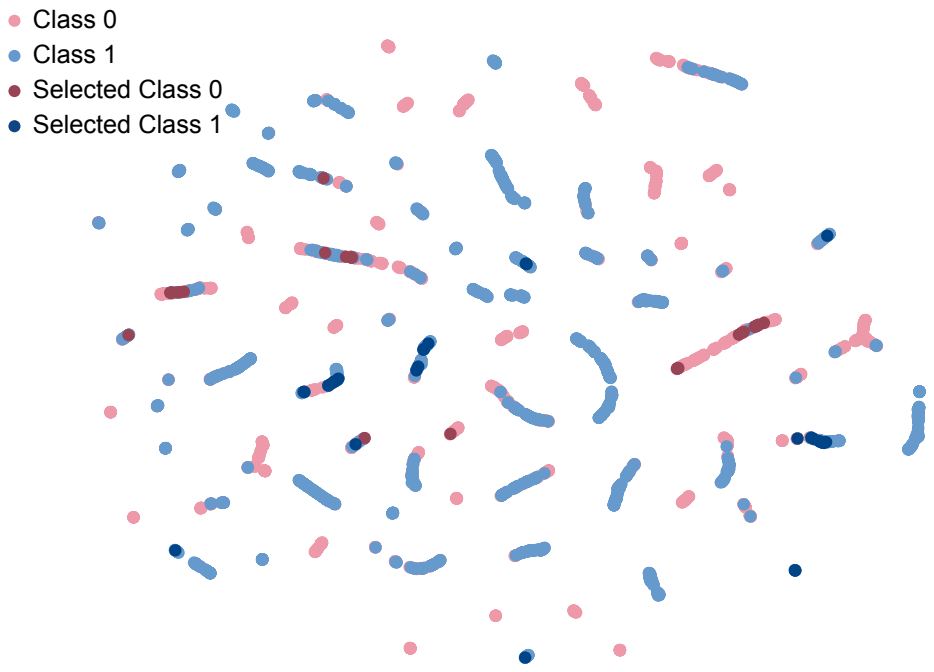


Figure 5.S2: **Impact of hierarchy bias on the UMAP latent space for a real world fire dataset.** 60 Samples selected by hierarchy (0.9) bias highlighted on the latent UMAP space of the labeled train set for fire. Results are shown for run 7 (arbitrarily chosen).

5.5.2. SUPPLEMENTARY TABLES

Table 5.S1: Numbers of synthetic lethality labeled and unlabeled samples or gene pairs per cancer type.

Cancer	Total	SL	non-SL	Unlabeled
BRCA	2453	1443	1010	151888
OV	805	253	552	151972
CESC	4900	144	4756	150964
SKCM	18407	107	18300	151545
LUAD	6103	594	5509	150944

Table 5.S2: Final distribution of classes in ST in the Randomized split experiments. The percentage of final train sets that are reported are averaged over 10 runs.

Cancer	Share majority class (%)	Share minority class (%)
BRCA	55 ± 2	45 ± 2
OV	69 ± 3	31 ± 3
SKCM	70 ± 5	30 ± 5
CESC	68 ± 6	32 ± 6
LUAD	56 ± 2	44 ± 2

Table 5.S3: Sizes of individual runs for the randomized split experiment after splitting the data and balancing for each set.

Run	Cancer	Train	SL	non SL	Validation	SL	non SL	Test	SL	non SL
1	BRCA	1294	647	647	324	162	162	402	201	201
2	BRCA	1294	647	647	324	162	162	402	201	201
3	BRCA	1294	647	647	324	162	162	402	201	201
4	BRCA	1294	647	647	324	162	162	402	201	201
5	BRCA	1294	647	647	324	162	162	402	201	201
6	BRCA	1294	647	647	324	162	162	402	201	201
7	BRCA	1294	647	647	324	162	162	402	201	201
8	BRCA	1294	647	647	324	162	162	402	201	201
9	BRCA	1294	647	647	324	162	162	402	201	201
10	BRCA	1294	647	647	324	162	162	402	201	201
1	OV	324	162	162	80	40	40	102	51	51
2	OV	324	162	162	80	40	40	102	51	51
3	OV	324	162	162	80	40	40	102	51	51
4	OV	324	162	162	80	40	40	102	51	51
5	OV	324	162	162	80	40	40	102	51	51
6	OV	324	162	162	80	40	40	102	51	51
7	OV	324	162	162	80	40	40	102	51	51
8	OV	324	162	162	80	40	40	102	51	51
9	OV	324	162	162	80	40	40	102	51	51
10	OV	324	162	162	80	40	40	102	51	51
1	CESC	184	92	92	46	23	23	58	29	29
2	CESC	184	92	92	46	23	23	58	29	29
3	CESC	184	92	92	46	23	23	58	29	29
4	CESC	184	92	92	46	23	23	58	29	29
5	CESC	184	92	92	46	23	23	58	29	29
6	CESC	184	92	92	46	23	23	58	29	29
7	CESC	184	92	92	46	23	23	58	29	29
8	CESC	184	92	92	46	23	23	58	29	29
9	CESC	184	92	92	46	23	23	58	29	29
10	CESC	184	92	92	46	23	23	58	29	29
1	SKCM	138	69	69	34	17	17	42	21	21
2	SKCM	138	69	69	34	17	17	42	21	21
3	SKCM	138	69	69	34	17	17	42	21	21
4	SKCM	138	69	69	34	17	17	42	21	21
5	SKCM	138	69	69	34	17	17	42	21	21
6	SKCM	138	69	69	34	17	17	42	21	21
7	SKCM	138	69	69	34	17	17	42	21	21
8	SKCM	138	69	69	34	17	17	42	21	21
9	SKCM	138	69	69	34	17	17	42	21	21
10	SKCM	138	69	69	34	17	17	42	21	21
1	LUAD	760	380	380	190	95	95	238	119	119
2	LUAD	760	380	380	190	95	95	238	119	119
3	LUAD	760	380	380	190	95	95	238	119	119
4	LUAD	760	380	380	190	95	95	238	119	119
5	LUAD	760	380	380	190	95	95	238	119	119
6	LUAD	760	380	380	190	95	95	238	119	119
7	LUAD	760	380	380	190	95	95	238	119	119
8	LUAD	760	380	380	190	95	95	238	119	119
9	LUAD	760	380	380	190	95	95	238	119	119
10	LUAD	760	380	380	190	95	95	238	119	119

Table 5.S4: Sizes of individual runs for the double holdout experiments

Run	Cancer	Train	SL	non SL	Validation	SL	non SL	Test	SL	non SL
1	BRCA	520	260	260	130	65	65	214	107	107
2	BRCA	532	266	266	132	66	66	210	105	105
3	BRCA	460	230	230	114	57	57	216	108	108
4	BRCA	524	262	262	130	65	65	208	104	104
5	BRCA	558	279	279	140	70	70	202	101	101
6	BRCA	494	247	247	124	62	62	202	101	101
7	BRCA	460	230	230	116	58	58	256	128	128
8	BRCA	520	260	260	130	65	65	224	112	112
9	BRCA	534	267	267	134	67	67	202	101	101
10	BRCA	538	269	269	134	67	67	206	103	103
1	OV	142	71	71	36	18	18	42	21	21
2	OV	142	71	71	36	18	18	50	25	25
3	OV	136	68	68	34	17	17	40	20	20
4	OV	144	72	72	36	18	18	44	22	22
5	OV	148	74	74	36	18	18	44	22	22
6	OV	144	72	72	36	18	18	50	25	25
7	OV	136	68	68	34	17	17	46	23	23
8	OV	152	76	76	38	19	19	44	22	22
9	OV	140	70	70	36	18	18	46	23	23
10	OV	140	70	70	34	17	17	44	22	22
1	CESC	90	45	45	22	11	11	28	14	14
2	CESC	90	45	45	22	11	11	30	15	15
3	CESC	90	45	45	22	11	11	28	14	14
4	CESC	92	46	46	22	11	11	30	15	15
5	CESC	90	45	45	22	11	11	28	14	14
6	CESC	88	44	44	22	11	11	30	15	15
7	CESC	92	46	46	22	11	11	30	15	15
8	CESC	88	44	44	22	11	11	28	14	14
9	CESC	92	46	46	22	11	11	28	14	14
10	CESC	90	45	45	22	11	11	24	12	12
1	SKCM	120	60	60	30	15	15	22	11	11
2	SKCM	120	60	60	30	15	15	22	11	11
3	SKCM	120	60	60	30	15	15	22	11	11
4	SKCM	120	60	60	30	15	15	22	11	11
5	SKCM	120	60	60	30	15	15	22	11	11
6	SKCM	120	60	60	30	15	15	22	11	11
7	SKCM	120	60	60	30	15	15	22	11	11
8	SKCM	120	60	60	30	15	15	22	11	11
9	SKCM	120	60	60	30	15	15	22	11	11
10	SKCM	120	60	60	30	15	15	22	11	11
1	LUAD	322	161	161	80	40	40	106	53	53
2	LUAD	334	167	167	84	42	42	104	52	52
3	LUAD	322	161	161	80	40	40	100	50	50
4	LUAD	334	167	167	84	42	42	106	53	53
5	LUAD	334	167	167	84	42	42	104	52	52
6	LUAD	380	190	190	94	47	47	120	60	60
7	LUAD	340	170	170	84	42	42	106	53	53
8	LUAD	348	174	174	86	43	43	108	54	54
9	LUAD	322	161	161	80	40	40	100	50	50
10	LUAD	330	165	165	82	41	41	106	53	53

Table 5.S5: Sizes of datasets for the Multiple SL label sources experiments.

Training study	Test study	Cancer	Train	SL	non-SL	Test	SL	non-SL	Unlabeled
ISLE	dSL	BRCA	1509	573	935	960	885	75	151882
dSL	ISLE	BRCA	893	854	39	1575	590	985	151882
ISLE	dSL	LUAD	4897	168	4729	711	372	339	150944
dSL	ISLE	LUAD	711	372	339	4897	168	4729	150944
EXP2SL	dSL	LUAD	2676	307	2369	711	372	339	150944
dSL	EXP2SL	LUAD	711	372	339	2676	307	2369	150944

Table 5.S6: Selected parameters for the randomized split experiments. For μ, the confidence threshold, the values 0.80, 0.85, 0.90 and 0.95 were tested. For p, number of pseudo-labeled samples to add in each iteration of self-training, the values 10, 20 and 50 were tested.

Cancer	p	μ
BRCA	0.90	20
OV	0.85	20
CESC	0.90	10
SKCM	0.90	10
LUAD	0.90	10

Table 5.S7: Selected parameters for the double holdout experiments. For μ, the confidence threshold, the values 0.70, 0.75, 0.80, 0.85, 0.90 and 0.95 were tested. For p, number of pseudo-labeled samples to add in each iteration of self-training, the values 6, 10 and 20 were tested.

Cancer	p	μ
BRCA	0.85	6
OV	0.80	6
CESC	0.75	6
SKCM	0.75	20
LUAD	0.90	10

Table 5.S8: Selected parameters for the Multiple SL label sources experiments. For μ, the confidence threshold, the values 0.75, 0.80, 0.85, 0.90 and 0.95 were tested. For p, number of pseudo-labeled samples to add in each iteration of self-training, the values 4, 6, 10 and 20 were tested.

Training study	Test study	Cancer	p	μ
ISLE	dSL	BRCA	6	0.85
dSL	ISLE	BRCA	4	0.95
ISLE	dSL	LUAD	10	0.80
dSL	ISLE	LUAD	10	0.85
EXP2SL	dSL	LUAD	6	0.85
dSL	EXP2SL	LUAD	10	0.80

5.5.3. SUPPLEMENTARY METHODS

DETAILS OF THE DOUBLE HOLDOUT EXPERIMENT

To assess the performance of the proposed methods when the train and test set follow different biases, we performed an experiment where the gene pairs in the train and test sets did not have any genes in common. By decoupling the genes in the train set from the test set, we constructed an experiment where the two sets do not originate from the same distribution and do not follow the same sample selection bias. In this experiment, we could evaluate the ability of the methods to transfer knowledge learned on one distribution to data with a different bias. For BRCA, CESC, LUAD, and OV, we divided the set of all individual genes instead of pairs into two sets: a training and a test gene set. Then all pairwise combinations of genes with available SL labels were generated within each set while trying to protect the ratio of samples between the training and test set to 4:1. We generated 10 different runs where in each run, the gene sets were selected randomly. This separation ensured that there was no overlap between the two sets of gene pairs. In contrast, for the SKCM dataset, since the gene *MYC* was dominant and only 60 samples did not contain the *MYC* gene, we constructed the test set always from these pairs without *MYC* gene. Then, for the training set, we used all pairs except those 60 and any other pair that had any gene overlap with these 60 samples.

6

DISCUSSION

Throughout the years, understanding cancer and effectively treating cancer patients has been a top priority in the medical field. As much as our understanding of cancer has improved, and significant progress has been made in technology and treatments, we also realize there is still a lot to uncover about cancer and its treatment. Cancer treatment originally focused more on surgeries, and cancer research depended more on prior biological knowledge and experiments in the lab to discover additional characteristics of cancer that could be exploited by new therapies. With recent advances in computational tools and molecular data measurement technologies, many more targetable vulnerabilities of cancer can be uncovered in a data-driven manner, without the need for prior knowledge. Especially with the fast progress of machine learning (ML) and the ability to mathematically find relations and patterns in complex high-dimensional molecular data, identifying treatment opportunities with computational tools has gained momentum in cancer research [1]. In this thesis, we focused on improved solutions for the computational prediction of candidate cancer treatment targets and, more generally, for building more generalizable machine learning models in the presence of selection bias. In this chapter, we discuss our conclusions, as well as the remaining gaps in the literature along with future research directions to address them. Furthermore, we briefly offer our views on the prospects of machine learning for the identification of effective cancer treatments as well as on fundamental issues related to bias in machine learning.

6.1. EFFECTIVE CANCER TREATMENTS

Targeted therapy has led to improved cancer treatment outcomes and patient prognoses over the years. However, effectively treating cancer remains an ongoing challenge. Due to the inherent heterogeneity of cancer, the development of resistance, and other complicating factors, even targeted therapies can sometimes prove ineffective. Effective targeted therapy first requires the precise identification of cohorts that could benefit from specific treatments and the determination of the most suitable therapeutic approach, whether directly or indirectly targeting cancer drivers. With the advances in high-throughput technologies that generate vast amounts of molecular data, as well as more complex computational tools,

we can now uncover deeper insights, such as further stratifications within cancer cohorts and potential genetic interactions that may effectively target cancer drivers. Nevertheless, there are still many hidden insights that can be uncovered.

6.1.1. UNCOVERING GENETIC INTERACTIONS

SYNTHETIC LETHALITY BETWEEN GENES

For certain cancer driver genes, even when their mechanisms are well understood, there are currently no effective methods to target them due to challenges such as inaccessible binding sites or inactivity of the driver gene. In these scenarios, genetic interactions like synthetic lethality offer promising alternatives. Since many cancer driver genes are associated with cellular proliferation and survival, their alteration and mutation transform normal cells into cancer cells. As this happens, cancer cells may simultaneously become dependent on other genes and pathways for survival.

Although numerous computational tools exist to identify synthetic lethal relationships, they often fall short in predictions for certain cancer types or are susceptible to bias, which often restricts their ability to identify interactions solely to specific genes or gene families. To address this, **in Chapter 2**, we introduced ELISL, a framework that integrates biological information about genes and gene pairs from multiple sources to create an ensemble ML model capable of predicting the synthetic lethality status of gene pairs. To build a robust and reliable model, ELISL makes use of both context-specific data sources specific to each cancer tissue type such as omics from cancer patients and cancer cell lines, and context-free data sources such as protein sequence similarity and protein-protein interaction networks. While context-specific data sources already used by other methods help uncover cancer-specific relations and reflect the variation between patients, they are also limited as they can be sparse for some cancer types. Context-free data sources, especially protein aminoacid sequences, contain information about the functional properties of the proteins, which could be informative of related gene function and genetic interactions. Moreover, aminoacid sequences are well-characterized and available for all proteins, thus preventing the sparsity problem. Additionally, ELISL relies on feature-based ML using forest ensembles to build models with increased generalizability in the presence of selection bias prevalent in SL training data. ELISL showed superior performance compared to models based on matrix factorization [2, 3] or network graphs [4, 5], which were more prone to follow the bias in the data. Furthermore, we showed that ELISL could be used in a cross-cancer setting where a model trained on a cancer type was used to predict SL relations for another cancer type. Lastly, we showed that mutations in the top SL pairs predicted by ELISL for breast cancer associated with longer patient survival, suggesting they could harbor therapeutic potential.

The Success of Amino Acid Sequence Information. The results from ELISL show the importance of amino acid sequences, which are believed to contain essential information about protein functions [6], in understanding genetic interactions.

Since functional relatedness is a key factor in determining genetic interactions, the information encoded within amino acid sequences is extremely valuable. However, amino acid sequence alone might not provide all necessary information about genetic interactions. For instance, protein-protein interactions could be informative as well, although they are typically based on physical and experimentally curated interactions that also do not cover the entire collection of known proteins and interactions between them.

Another important question that arises from the success of using amino acid sequences is to what extent does the computational method used to extract information contribute to this success? We know that the methods extracting embeddings from aminoacid sequences are well-established and they can provide valuable insight for various problems [6]. For example, deep learning models can uncover sequence features that correlate with functional relatedness and other biological properties. However, these methods are not always entirely accurate and they may introduce biases or miss critical information. Therefore, while we suggest that information extracted by deep learning models from amino acid sequences can be valuable across various problems, we should not ignore how much these methods specifically impact the problem. For example, what is the gain of using complex embedding extraction methods compared to one-hot binary encoding in various downstream tasks? This issue is beyond identifying synthetic lethality (SL) interactions, it also affects other fields where amino acid sequences and embedding methods are utilized. For example, applications in protein structure prediction, function annotation, or drug-target interaction prediction may face similar issues. Thus, we need to ensure that these methods are reliable and effective in downstream tasks across different biological problems.

Challenges with Context-Specific Data Sources. Data sources collected from patients or cancer cell lines have profoundly influenced cancer research across a wide range of applications. However, these types of data are still collected by humans from a relatively narrow group of patients or cell lines, constrained by various limitations such as time, cost, and technology, as well as the quality versus quantity trade-off in sample collection. Consequently, for different cancer types originating from various tissues, varying number of available samples limit the effectiveness of the computational methods that heavily depend on the data. Additionally, for some cancer types, data collection may lack quality, resulting in numerous missing data points. Typically, missing data or technical errors in a few samples have a diminishing impact as more data is collected. However, for cancer types with a small number of samples, these issues remain critical. Thus, it is essential that future data collection efforts are standardized and that samples are carefully evaluated. When using existing datasets, it may be wise to exclude certain data points to avoid introducing confusion while preserving meaningful biological variability. However, it is not straightforward to distinguish between technical errors and genuine biological deviations. While some methods, such as outlier detection, can help in deciding which samples to exclude, without detailed metadata that includes technical and

biological variables, the likelihood of error in selection remains high.

IS SYNTHETIC LETHALITY ONLY BETWEEN PAIRS OF GENES?

Current computational approaches for predicting synthetic lethality (SL) mainly identify interactions between pairs of genes. However, the mechanisms underlying SL are often far more complex and typically involve entire biological pathways rather than just individual genes. A prominent example is the well-known synthetic lethal interaction between BRCA1/2 mutations and PARP which arises from a dependency between a pathway and a gene [7]. Inhibition of PARP leads to an accumulation of double-stranded breaks due to not being able to repair single-stranded breaks. Simultaneously, mutations in the BRCA1/2 genes result in a deficiency in the homologous recombination repair pathway, rendering the cells less capable of repairing double-stranded breaks. Consequently, the inhibition of PARP in cells with defective homologous recombination repair leads to an outstanding accumulation of DNA damage and, ultimately, cell death. This interaction illustrates that synthetic lethality is not simply a relationship between two genes but rather a result of broader pathway vulnerabilities and compensatory mechanisms that become critical when one pathway is disrupted.

Furthermore, synthetic lethal relationships are not necessarily limited to interactions between just two entities. For example, studies have shown that the BRCA-PARP synthetic lethal interaction can be further enhanced by the additional inhibition of RAD52, a protein involved in an alternative DNA repair pathway [8]. Thus, targeting multiple genes or pathways together could enhance therapeutic efficacy, particularly as cancer cells frequently develop resistance to established treatments. Expanding the concept of synthetic lethality to include interactions involving multiple genes or pathways could help find promising strategies for overcoming resistance and improving the effectiveness of cancer therapies.

Most existing computational models are limited to the gene-gene pair concept, limiting our understanding of more extensive genetic relationships that involve gene families, gene sets, or even entire pathways. Capturing these broader interactions computationally is a complex process. Traditional gene-based approaches often fail to account for the multi-layered, network-driven nature of cellular processes that drive SL. To bridge this gap, the integration of knowledge graphs could offer a powerful alternative.

A knowledge graph is a structured representation of knowledge where entities (such as genes, proteins, and pathways) are connected through edges that denote relationships, such as functional associations, co-expression, or other regulatory interactions [9]. Knowledge graphs can capture the intricate network of interactions that underlie cellular function and can incorporate diverse data types, including genomics, transcriptomics, proteomics, and metabolomics. By embedding these graphs within SL prediction frameworks, it becomes possible to explore more complex relationships beyond simple gene pairs. This approach could enable the identification of SL relationships among gene families or pathways, offering

a more holistic understanding of synthetic lethality and potentially revealing new therapeutic opportunities that target these broader biological contexts. However, these network-driven approaches should be used with caution, for example only as a supplement to a robust unbiased method, as their previous performance suggest that methods based on previously known SL interaction graphs are likely to follow the bias in the data [10].

UNCOVERING OTHER GENETIC INTERACTIONS

While synthetic lethality has garnered the most attention in the field of cancer therapeutics, there are other genetic interactions such as synthetic rescue that occur when the lethal effects of a mutation in one gene are alleviated by mutations or alterations in another gene [11]. This concept can be used to reactivate tumor suppression components. For example, if a tumor suppressor gene is inactivated due to mutations, identifying a synthetic rescue interaction could provide a way to restore the function in the cells. Currently, besides a couple of studies [12, 13], synthetic rescue interaction is largely overlooked in computational studies compared to synthetic lethality predictions.

The primary challenge in identifying synthetic rescue and other genetic interactions lies in the complexity of their relationships, which often go beyond simple binary relations. Current computational methods primarily frame the problem as a binary classification task, where the model predicts whether a specific relationship, such as synthetic lethality, exists between a pair of genes. This binary approach limits our ability to explore the full spectrum of genetic interactions. Instead, there is a need for developing multi-class classification models that can predict not only the existence of a genetic interaction but also classify the type of interaction whether it is synthetic lethality, synthetic rescue, or another form of genetic interplay.

6.1.2. STRATIFYING COHORTS DEEPER

Stratifying cancer patient groups is an effective technique to find more refined cancer treatments. However, cohorts identified by stratification may still contain subgroups that develop resistance to current treatments. Additionally, as more data becomes available, we can see that new patients who are assigned to a subcohort may not respond well to the targeted therapy designed for that specific subcohort. Over time, each level of subcohorts can become inadequate. For example, initial stratification by tissue of origin becomes progressively insufficient. Or patient cohorts stratified by cancer driver genes also become a limitation. Consequently, more refined stratification is required that incorporates various biomarkers such as specific mutation types in particular genes, combinations of mutations, alterations in passenger genes, and other epigenetic markers.

In Chapter 3, we introduced OncoStratifier, a framework designed to further stratify oncogene-addicted cohorts by identifying drugs that induce the cohort to split into two distinct groups with different responses to treatment. Unlike other stratification

problems where ML models can be effectively applied, oncogene-addicted cohorts often consist of a small number of samples where it is harder to use ML methods that typically require large datasets. Therefore, we utilized drug response data from cancer cell lines—one of the more reliable data sources—to identify subcohorts of oncogene-addicted cohorts with differential responses to a given treatment. To measure the stratification potential within a cohort, we used Shannon entropy [14], which quantifies the ambiguity in drug response among the group. Simultaneously, we ensured that the same drug did not produce a similar stratifying effect in the WT cohorts, thereby identifying exclusive stratifiers for specific oncogenes.

Our strategy for identifying stratifiers exclusive to a specific oncogene is based on two key considerations. First, to accurately analyze an oncogene that is driving cancer, the ideal scenario would involve a cohort where that oncogene is the sole driver, meaning no other mutated cancer drivers are present. However, such cohorts are either non-existent or too small to support robust analysis. As a result, both oncogene-addicted and wild-type (WT) cohorts may include samples where cancer is also driven by another gene. This overlap introduces a complication: if the response of a drug response appears stratified within an oncogene-addicted cohort, it may be due to the presence of another cancer driver within a subset of that cohort.

For instance, if a drug specifically targets a secondary oncogene and this secondary oncogene is mutated in half of the cohort primarily driven by the first oncogene, the response to the drug will appear stratified. To exclude such cases, we also examine whether the response to the drug is stratified in the cohort where the primary oncogene is wild-type assuming that the mutational status of first and secondary oncogenes are not identical across the cohort. If the response is stratified in this cohort as well, it suggests that the drug is likely targeting another cancer driver, and the observed stratification is not related to the primary oncogene of interest. This additional check ensures that the identified stratifiers are genuinely associated with the primary oncogene and are not confounded by other underlying genetic factors.

Moreover, for some of the stratified cohorts, we identified mutational biomarkers that could transition the study to clinical settings involving cancer patients. This approach is particularly valuable for well-known cases of oncogene addiction that currently lack effective treatment or targeting options, such as KRAS mutations in colorectal cancer [15].

WHEN TO STOP STRATIFYING

The process of stratifying cancer cohorts leads to questions about the depth of stratification needed. The biggest challenge with deeper stratification is that subcohorts become increasingly smaller with each level of stratification. As cohort sizes shrink, the statistical power to detect meaningful differences also diminishes. Additionally, the evolution of cancer cells, which can gain resistance to therapies, is not readily characterized and thus does not provide sufficient data for further stratification. This lack of data makes it hard to achieve clinically meaningful stratification.

While stratifying deeper to identify more precise subcohorts, we should also consider the limitations due to sample size and data availability. Excessive stratification can lead to subcohorts that are clinically not valuable since they might focus on irrelevant patterns to cancer such as age, sex, race, or other protected attributes. Patient stratification research is in need of a theoretical framework to guide these kind of stratification decisions, focusing on the clinical utility and the potential to develop effective, targeted therapies.

When and How to Utilize Machine Learning Models in Stratification

In our study, we chose not to use ML models for stratification purposes. Firstly, supervised learning models require pre-existing annotations for each sample, and in stratification problems, there is usually no ground truth or established annotations available. This lack of annotated data makes supervised learning models less commonly utilized in the literature for such applications. Another reason is that supervised learning models, which require labeled training data, typically need a large number of sample points to accurately recognize patterns and establish correct relationships between sample characteristics and their corresponding labels. With a small number of samples, either no solution is found that maps the features to the corresponding labels, or the models find a suboptimal local solution that may be valid for the small sample set but is not a globally optimal solution that would work for all samples. This means that the performance of supervised learning models is influenced by the quantity of labeled data available. More samples generally lead to better performance and more robust models [16]. Oncogene-addicted cohorts often have very small sample sizes, frequently less than ten, which makes it difficult to apply supervised learning effectively in these scenarios.

On the other hand, unsupervised learning methods like clustering aim to discover inherent structures or patterns within data without the need for labeled outputs. Unlike supervised learning, which requires mapping features to specific labels, clustering focuses mostly on the similarities and differences among samples based on their feature representations. Since the aim is not uncovering one specific type of pattern, these models can group samples solely based on intrinsic properties, which may reflect different patterns. While having more data can improve the stability and resolution of the patterns identified in both learning types, some clustering methods can still be effective with a smaller number of samples if those samples possess informative features that capture the essential characteristics of the data. The effectiveness of clustering methods relies more on the distinct separation between intended clusters rather than the sheer quantity of samples [17]. For instance, methods like k-means are less affected by sample size as they partition the data into clusters based on feature similarities regardless of the number of samples [17]. In contrast, methods based on the number of neighbors, such as DBSCAN, can still be slightly affected by sample size because they require enough neighbors to form meaningful clusters [17]. Therefore, although not all, but some of the clustering methods may be less sensitive to the number of samples, unlike the supervised learning methods. However, it is important to note that with the advancement of

clustering methods that use complex deep learning structures, number of samples may still be as important as supervised learning due to the high number of parameters to learn.

However, although clustering can find meaningful patterns with fewer samples, the patterns used to cluster the data may not have clinical or practical utility, as they are not guided by specific labels or outcomes as in supervised learning. Therefore, the identified groups may lack therapeutic value because they do not correspond to any available treatments or actionable targets, rendering the stratification clinically irrelevant. For example, a subcohort might be defined by a unique combination of mutations that is not related to any known therapeutic pathways. This limitation diminishes the practicality of the stratification, as it does not result in actionable clinical interventions.

Therefore, unsupervised ML methods require additional analysis post-stratification to determine the relevance and clinical applicability of the identified subgroups. First, the subcohorts need to be characterized by specific genomic, molecular, or phenotypic characteristics, if such characteristics exist. Then, subsequent analysis is necessary to identify any targetable factors within these groups and propose potential treatments. This multi-step process of post-analysis and validation can be time-consuming and resource-intensive, complicating the integration of unsupervised learning into clinical stratification workflows. Without detailed follow-up, there is a high risk that the identified subcohorts may not translate into meaningful clinical outcomes.

In summary, the applicability of machine learning (ML) methods for patient stratification, whether supervised or unsupervised, should be critically investigated before application. Supervised ML models should be employed only when there is a sufficient sample size for robust model training and unsupervised methods should be used only when preliminary evidence suggests that stratification could lead to actionable therapeutic insights due to extensive post-stratification validation. This preliminary evidence can be defined in different ways, such as a subcohort starting to show resistance to an already accepted therapy while retaining its previously identified biomarker. Stratification can also be tailored to the applicability of an existing drug that has passed toxicity tests if the drug can be repurposed for a different subcohort than originally intended.

With current problems in both supervised and unsupervised learning, integrating clinical relevance with computational techniques can be the only way to make ML valuable for cancer stratification and precision medicine. This may involve utilizing features previously linked with clinical relevance to remove potentially noisy features, using clinical knowledge as constraints to guide unsupervised stratification, or employing them in post-stratification validation. Additionally, domain experts can be engaged to guide stratification by providing very limited input to correct model errors, similar to reinforcement learning, [18] or supplying further information where the model struggles to make any decisions, as in active learning techniques.

6.2. Selection Bias in Machine Learning Models

Previous studies on SL prediction [10], as well as our work in Chapter 2 (ELISL), have shown that only certain gene pairs, consisting of specific genes or genes from particular gene sets, are annotated with SL status. This results in a dataset that is biased towards specific genes. Since most supervised learning methods rely on these pre-annotated SL pairs for training, the models they produce become also biased. This is a classic example of selection bias, where samples in a dataset are not selected randomly or independently of any factor. As a result, training models on such biased datasets that do not represent the true data distribution, can lead to reduced generalizability and models that overlook or mispredict cases that are originating from a distribution that is different from the training data distribution. The recognition and examination of the impact of selection bias on ML models remain limited in bioinformatics, as traditional evaluation methods, such as randomly splitting a dataset into train and test sets, are often used. This approach may not provide a fair evaluation of an ML method because, even if the split is random, both the train and test sets will share similar characteristics and inherent biases already present in the original dataset.

In Chapter 4, to evaluate the generalizability of ML models learned using a biased train set for a prediction task of interest, we introduced "hierarchy bias", a cluster-based method designed to artificially induce bias to any dataset by selecting specific samples. Hierarchy bias induces selection bias separately for each class by first finding clusters using hierarchical clustering, and then influencing the number of samples selected from a specific cluster relative to the others. Unlike other bias induction techniques [19, 20], this method allows users to control the total and biased number of samples selected, and can introduce different biases to different classes, making it more effective in altering the original ideal decision boundary compared to previous approaches. Machine learning (ML) models trained on datasets with induced hierarchy bias consistently showed a decline in generalizability compared to models trained on the same datasets with other existing bias induction techniques. This indicates that hierarchy bias can effectively introduce selection bias that alters the correct decision boundary. Therefore, when an independent validation dataset is unavailable for measuring generalizability, we can artificially induce bias and generate train and test sets with different distributions, even though they are derived from the same dataset. This approach allows us to measure model performance in terms of generalizability by providing train and test sets with discrepancy in distribution and different decision boundaries.

In Chapter 4, we proposed DCAST, a framework aimed at mitigating selection bias by incorporating diverse unlabeled (not annotated) samples into the learning of ML models. The DCAST strategy employs a semi-supervised technique called self-training (ST) [21], which iteratively adds to the train set new unlabeled samples whose labels are predicted (pseudo-labels) with high confidence by the latest model trained on the already available (pseudo-)labeled data. For models learned from biased data, focus on high-confidence predictions is likely to yield samples more similar to the biased data, making self-training prone to further strengthening the existing bias. To

address this, DCAST selects pseudo-labeled samples by identifying distinct clusters of unlabeled samples for each class and then selecting one sample from each cluster based on a relaxed confidence threshold, assuming that different clusters contain samples with varying characteristics. The DCAST framework consistently delivered models with improved generalizability compared to conventional self-training and other domain adaptation methods. Evaluation across multiple real-life datasets with different characteristics demonstrated that DCAST is applicable across diverse problems, including binary and multi-class classification. However, for effective selection of diverse pseudo-labeled samples using high-dimensional data, DCAST ideally requires ML models to have a latent space informed by class labels (supervised embedding). This requirement is not feasible with all models, such as logistic regression, which thus limits the applicability of the method.

In Chapter 5, we introduced Metric-DST, a framework similar to DCAST that also incorporates diverse unlabeled samples using self-training, but relies on metric learning to obtain supervised embeddings that can be exploited for prediction by any ML model of interest. Metric learning finds a function that converts original features into a lower-dimensional representation optimized to separate samples of different classes. Metric-DST first finds a bounded latent space representation of the samples using metric learning and then selects pseudo-labeled samples nearby random coordinates in this space to promote diversity. Any other ML model can then be employed within this new space to determine pseudo-labels for unlabeled samples, thereby making the method agnostic to the type of ML model used for predictions. Metric-DST showed potential for mitigating selection bias across simple toy datasets, real-world ML benchmark datasets, and synthetic lethality prediction. While Metric-DST may be more model-agnostic than DCAST, it does introduce an intermediate step to learn a supervised embedding representation that is not as well integrated as when this is intrinsically part of the prediction model. This also means that Metric-DST functions more like a black-box compared to DCAST, limiting user control and impact.

While both methods are potential solutions for mitigating selection bias within semi-supervised settings, some limitations remain.

Fundamental Assumptions for Incorporating Unlabeled Samples. All self-training methods, including DCAST and Metric-DST, begin by training with a labeled set and assume that the initial model can accurately predict the labels of unlabeled samples. If the initial performance of the model is not substantially better than random guessing, self-training methods may fail to assign correct labels to unlabeled samples, rendering them ineffective. Based on this, we can also argue that the starting performance of a model trained on biased labeled data will most likely issue biased predictions and thus have modest generalizability. However, in the presence of selection bias, even if the unbiased prediction performance (generalizability) were to be only slightly better than random chance, predictions that are close to the decision boundary, thus with low confidence, are more likely to be incorrect while predictions with a high confidence that are further away from the decision

boundaries are less likely to be incorrect unless the change in decision boundary is extreme such as finding a linear boundary instead of a nonlinear one. Selection bias usually causes slight shifts and rotations in the decision boundary, which predominantly affect samples close to the decision boundary that originally have low prediction confidence. Consequently, high-confidence predictions are least affected by changes in the decision boundary, especially in cases where there is a shift without rotation and the support remains the same. Therefore, considering high-confidence predictions while promoting diversity makes diverse selection feasible.

Semi-supervised learning, including self-training, relies heavily on the smoothness and clustering assumptions, besides other assumptions due to label propagation [22]. Furthermore, one of our proposed methods, DCAST, is also built heavily on the cluster assumption to identify diverse unlabeled samples. Smoothness assumes that samples close to each other share the same labels, while clustering assumes that data is composed of clusters and that samples within the same cluster are likely to share the same label. Although self-training uses supervised ML model predictions rather than a purely similarity-based method such as k-nearest neighbors, these assumptions still apply in self-training, as samples with similar characteristics from the same cluster are likely to receive similar predictions. However, a fundamental issue is that not all features in a dataset are necessarily relevant to the prediction task of interest. For example, in predicting patient response to therapy based on gene expression profiles, there may be more than 20,000 genes in the feature space, but only two genes might be relevant to the specific therapy response. As a result, many samples (patients) might appear similar when considering all 20,000 genes but not when considering only the two relevant genes or vice versa. Therefore, this was an important aspect we considered when designing DCAST and Metric-DST methods and their procedures for assigning pseudo-labels and promoting diversity. For example, DCAST selects one sample from each identified cluster, and if this process was done considering the full original feature space we might add samples from different clusters that looked diverse in the full feature space but quite similar when only the informative features were considered. This motivated the use of class informative supervised embeddings instead: lower-dimensional sample representations learned using supervised ML methods that incorporate class information to ensure that the dimensions of the new latent space and any distances based on them are more relevant to the prediction problem.

However, while supervised embeddings can be extracted from random forests or neural network-based models, simpler models like logistic regression do not provide a class-informative latent space. For these types of algorithms, suitable alternatives include using only a selection of class informative features from the original feature space, or relying on an approach such as Metric-DST to learn a class informative embedding based on which predictions can be made using an independent ML model.

Problem of Endless Hyperparameter Space. As ML models become more complex, they are associated with an increasing number of hyperparameters that require

extensive tuning. Normally hyperparameter values are determined by the user, but with automated algorithms the best possible hyperparameters for the specific task can be found by looking at the performance on the left out validation set. When these models are combined with other models in a pipeline or integrated into another framework, the number of hyperparameters further increases since additional parameters that control the integration of models to pipelines or frameworks also need to be configured. As the number of hyperparameters grows, it becomes nearly impossible for humans to manually identify the optimal combination. Although automated hyperparameter optimization is widely used, when working with biased datasets the automatic tuning of hyperparameters may overfit or overoptimize the model for the specific dataset in question to the point that it might fail when used on other datasets with slightly different characteristics. Therefore, we often focused on what type of model and architecture were fundamentally required more than trying to find the most optimal hyperparameters for performance.

6.2.1. THE POTENTIAL OF DOMAIN ADAPTATION METHODS IN GENERALIZABILITY

Domain adaptation methods are designed to address discrepancies between the source (training) and target (test) sets, either independently or during model training [23]. Thus, they typically focus exclusively on specific source and target set pairs, requiring the retraining of the model whenever the target set to be tested changes. Thus, they are mainly used when the objective is to improve model performance for a specific target set, such as predicting the outcome for a new patient from a different hospital than those represented in the source set. However, their ability to make models generalizable across all types of target sets without retraining is not yet well-explored.

Similar to semi-supervised ML methods, domain adaptation can be applied using unlabeled samples that are different than the samples in target set. Technically it only requires the models to treat the unlabeled set as the target set. These methods do not require labels for the target set, which allows them to be applied to an unlabeled set instead. The idea is to align the distributions of the source set and the unlabeled set to develop a model that can predict any target sample and evaluate the model on different target sets. While it is technically feasible to use domain adaptation methods with a large, readily available set of unlabeled samples that better approximates the true underlying distribution of the population, their effectiveness can be limited by the complexity of the decision boundary and the distribution of the unlabeled data. This limitation arises because most domain adaptation methods assume that the correct decision boundary, or its support within the target set, is represented in the source set. However, this assumption may not hold when using a large and diverse unlabeled set that requires a much more complex decision boundary than that represented by the biased source set. Therefore, we should also evaluate domain adaptation methods in semi-supervised settings where the primary objective is to improve generalizability rather than to adapt a model to a specific target set.

6.2.2. Potential of Multiomics in Mitigating Bias

Selection bias arises from the way samples are chosen during data collection or annotation and is not necessarily related to the features of the samples used during training. However, one could argue that since these samples are selected in a biased manner based on at least one of their characteristics, the way these characteristics are represented in the feature space may also be important for mitigating selection bias if such characteristics are present among the features. Thus, using feature spaces that are independent from the bias source is important. Since the source of the selection bias is unknown for most applications, it may be better to use different feature spaces, while mitigating bias. In bioinformatics, samples can have different types of measurements at the molecular and cellular level, termed generally as omics, including gene expression, somatic mutations, protein expression, methylation, and copy number variation as well as other biological data sources. It is widely known that each of these data sources contains both common and exclusive information about samples, such as patients in cancer-related studies. Previous literature has shown that the performance of ML models can be improved by using multiple omics. Following this reasoning, we can also use multiomics data that reflect different characteristics together to mitigate bias.

Self-training conventionally is designed and used for single omics to predict the pseudo-labels of unlabeled samples, thus not considering different aspects of using multi-omics. Co-training [24] is a modified version of self-training where multiple subdatasets that are different portions of the features or samples of the original dataset are created and utilized while incorporating unlabeled samples. In co-training, two or more classifiers are trained on different subdatasets of the same data. These classifiers then iteratively label the unlabeled samples for each other. The success of co-training relies on the assumption that each view is conditionally independent given the class label, and that each view is sufficient for learning. Inspired by the co-training framework, we can modify our methods so that the pseudo-labels are not predicted by just one data source but by multiple sources or omics. One strategy would be to use multiple datasets/omics to train multiple models, with all of them contributing to the pseudo-label prediction. For example, we could consider adding a pseudo-labeled sample to the training set only if all ML models from all datasets agree on the pseudo-label of the sample or if a majority of the models agree on a label. In short, we would still use self-training procedure but in each iteration, we can use the ensemble of multiple ML models to decide which samples to incorporate. our main ML model for the self-training would be an ensemble ML model. While this technique is intuitive and may improve performance by finding more reliable predictions, it is unknown if it would contribute to the diversity approach we have used while selecting pseudo-labeled samples to be incorporated. However, if the feature space of one or more datasets is more representative, they may still affect how the sample space is used for diversity.

Another approach would be to split the self-training process for each dataset, omic, or view, and only allow information sharing between them during pseudo-labeling. For example, there could be a separate self-training process for each omic, but in

each iteration, the pseudo-labeled samples added for one dataset could be added to another dataset's training set. By doing this, we aim to increase the information sharing between all datasets. Then, in the end, an ensemble model could be built by combining the resulting models of all datasets.

However, both ideas require careful examination due to many unanswered questions. What should be done if different models disagree on pseudo-labels? How much should each model contribute to the ensemble model? Should a model itself agree with the pseudo-labeled sample sent by other models? Could a model's good initial performance be hindered by an unsuccessful model during information sharing? Therefore, if multiomics are utilized for mitigating bias, these questions need to be carefully considered and answered.

6.3. FINAL REMARKS

In this thesis, we introduced a method to uncover genetic interactions, specifically synthetic lethality, that could be used to discover new strategies to treat cancer patients more effectively. We also developed a method to identify cancer patient groups experiencing oncogene addiction without current treatment, and propose potential treatment possibilities for them relying on existing drugs. Additionally, we presented computational methods to study, replicate, and reduce sample selection bias in ML.

Throughout our methods, we applied several key concepts, such as using pretrained ML models to extract valuable information from biological data, employing ensemble ML models to predict genetic interactions with multiomics, and utilizing drug entropy information to identify further stratification of underexplored cohorts. We also used various techniques in semi-supervised settings with available unlabeled data to address selection bias and unfairness of computational methods. Together, the proposed contributions strive to make computational methods that suggest potential cancer treatments fairer and more unbiased.

REFERENCES

[1] B. Zhang, H. Shi, and H. Wang. “Machine learning and AI in cancer prognosis, prediction, and treatment selection: A critical approach”. en. In: *J. Multidiscip. Healthc.* 16 (June 2023), pp. 1779–1791.

[2] H. Liany, A. Jeyasekharan, and V. Rajan. “Predicting synthetic lethal interactions using heterogeneous data sources”. In: *Bioinformatics* 36.7 (2020), pp. 2209–2216.

[3] J. Huang *et al.* “Predicting synthetic lethal interactions in human cancers using graph regularized self-representative matrix factorization”. In: *BMC Bioinformatics* 20.S19 (Dec. 2019), p. 657.

[4] Y. Long *et al.* “Graph contextualized attention network for predicting synthetic lethality in human cancers”. In: *Bioinformatics* 16 (Feb. 2021), pp. 2432–40.

[5] R. Cai *et al.* “Dual-dropout graph convolutional network for predicting synthetic lethality in human cancers”. In: *Bioinformatics* 36.16 (Mar. 2020), pp. 4458–65.

[6] M. Heinzinger *et al.* “Modeling aspects of the language of life through transfer-learning protein sequences”. en. In: *BMC Bioinformatics* 20.1 (Dec. 2019), p. 723.

[7] T. Helleday. “The underlying mechanism for the PARP and BRCA synthetic lethality: clearing up the misunderstandings”. en. In: *Mol. Oncol.* 5.4 (Aug. 2011), pp. 387–393.

[8] K. Sullivan-Reed *et al.* “Simultaneous Targeting of PARP1 and RAD52 Triggers Dual Synthetic Lethality in BRCA-Deficient Tumor Cells”. In: *Cell Reports* 23.11 (June 2018), pp. 3127–3136. ISSN: 2211-1247.

[9] D. N. Nicholson and C. S. Greene. “Constructing knowledge graphs and their biomedical applications”. en. In: *Comput. Struct. Biotechnol. J.* 18 (June 2020), pp. 1414–1428.

[10] C. Seale, Y. Tepeli, and J. P. Gonçalves. “Overcoming selection bias in synthetic lethality prediction”. In: *Bioinformatics* 38.18 (July 2022). Ed. by K. Borgwardt, pp. 4360–4368. ISSN: 1367-4811.

[11] Y. Han *et al.* “Genetic interaction-based biomarkers identification for drug resistance and sensitivity in cancer cells”. en. In: *Mol. Ther. Nucleic Acids* 17 (Sept. 2019), pp. 688–700.

[12] A. D. Sahu *et al.* “Genome-wide prediction of synthetic rescue mediators of resistance to targeted and immunotherapy”. en. In: *Mol. Syst. Biol.* 15.3 (Mar. 2019), e8323.

[13] M. Liu *et al.* “Synthetic viability induces resistance to immune checkpoint inhibitors in cancer cells”. en. In: *Br. J. Cancer* 129.8 (Oct. 2023), pp. 1339–1349.

[14] C. E. Shannon. “A mathematical theory of communication”. In: *Bell Syst. Tech. J.* 27.3 (July 1948), pp. 379–423.

[15] L. Huang *et al.* “KRAS mutation: from undruggable to druggable in cancer”. en. In: *Signal Transduct. Target. Ther.* 6.1 (Nov. 2021), p. 386.

[16] D. Rajput, W.-J. Wang, and C.-C. Chen. “Evaluation of a decided sample size in machine learning applications”. In: *BMC Bioinformatics* 24.1 (Feb. 2023). ISSN: 1471-2105.

[17] E. S. Dalmaijer, C. L. Nord, and D. E. Astle. “Statistical power for cluster analysis”. In: *BMC Bioinformatics* 23.1 (May 2022). ISSN: 1471-2105.

[18] R. Sutton and A. Barto. “Reinforcement Learning: An Introduction”. In: *IEEE Transactions on Neural Networks* 9.5 (1998), pp. 1054–1054.

[19] J. Huang *et al.* “Correcting Sample Selection Bias by Unlabeled Data”. In: *Proceedings of the 19th International Conference on Neural Information Processing Systems*. NIPS’06. Canada: MIT Press, 2006, pp. 601–608.

[20] A. Liu and B. Ziebart. “Robust classification under sample selection bias”. In: *Advances in Neural Information Processing Systems* 1 (Jan. 2014), pp. 37–45.

[21] G. J. McLachlan. “Iterative Reclassification Procedure for Constructing an Asymptotically Optimal Rule of Allocation in Discriminant Analysis”. In: *Journal of the American Statistical Association* 70.350 (1975), pp. 365–369. eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1975.10479874.

[22] J. E. van Engelen and H. H. Hoos. “A survey on semi-supervised learning”. en. In: *Mach. Learn.* 109.2 (Feb. 2020), pp. 373–440.

[23] W. M. Kouw and M. Loog. “A Review of Domain Adaptation without Target Labels”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.3 (Mar. 2021), pp. 766–785.

[24] A. Blum and T. Mitchell. “Combining Labeled and Unlabeled Data with Co-Training”. In: *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*. COLT’ 98. Madison, Wisconsin, USA: Association for Computing Machinery, 1998, pp. 92–100. ISBN: 1581130570.

ACKNOWLEDGEMENTS

After four long years, my journey is coming to an end, and through all my ups and downs, many people have stood by me. Completing a PhD is not only physically demanding but also mentally challenging, and it would have been incredibly difficult without the support of those by my side. During these challenging yet enriching times, I have deepened existing relationships and met incredible people in various places. Now, it is time to thank them all.

First of all, I would like to thank my supervisors. **Joana**, thank you for all the support you have given me throughout these four years. You were always available to help whenever I needed it and consistently handled matters with kindness. As a junior lab member, I feel that we grew together over time. **Marcel**, thank you for creating such a unique and welcoming lab environment. You were always open to conversation, encouraged me and others, and treated each of us fairly. DBL will always be my home lab, even if I leave. **Francesco**, I had the opportunity to work with you briefly during my research visit, and you left a lasting impact on me both personally and in my research with your enthusiasm and interest in all the work I have been doing. You and your lab also became my second home in research.

Furthermore, I would like to thank **Jasmijn** and **Ahmed** for always being attentive, inquiring about my well-being, and providing valuable feedback and engaging in scientific discussions. I would also like to thank **Jana**, **Erik**, and **Thomas** for their kindness and scientific input. Additionally, I am grateful to **Marunka** for all the help and support she has provided, as well as to **Saskia** and **Ruud** for their endless support.

My PhD started during COVID, which initially caused me significant anxiety. However, from my first day, there were people who helped me overcome the great challenges of coming to a new country, living alone in isolation, and starting a new project in an unfamiliar environment. **Tamim**, I cannot count how many times I have turned to you for help when I was unable to make decisions—whether regarding personal issues or organizing a basketball tournament. You have always been like a brother to me in the lab, offering incredible friendship and sound advice. I will always wave at you and your family from my balcony. Another person I have consistently turned to during tough times is **Ramin**, one of the founders of C++; I cannot even express how much of an impact you have had on me personally. Whether it was dealing with personal or workplace issues or seeking help with C++, you have always been there. Although you forgot me for a while (sorry, who are you?), I am glad that I can now be part of your life again. It is always a joy to see you, **Yosra**, **Loki**, and **Odin**. Next is my four-year companion, **Colm**, who worked alongside me, facing the same challenges and celebrating the same victories. I have always admired your emotional resilience and strength (and, of course, your physical strength as well). I feel that we grew together in this lab, and without your en-

couragement, this PhD would have been much harder and I would have never gone to the gym. Lastly, **Stavros**, although you transferred to a different lab halfway through my PhD, you remain one of the most influential figures I have had in the lab. Without your proactive nature, scientific arguments, and boldness, my PhD would have been far less fruitful. Please do not hurt yourself any further with our gifts.

A special thanks also goes to my paranymphs. **Kirti**, even though you were not here at the beginning, once you started your PhD here, you became one of my closest friends. I could trust and rely on you for everything, and you were always there when I felt down. Thank you for teaching me crocheting, for watching LIB together, for discussing all the TV characters with me, for going to all kinds of movies, and bringing **Nishad** with you. **Sander**, my first master's student and PhD friend—although our personalities are completely different (according to the test)—I always felt that I could trust you and your judgment at any time. Thank you for introducing me to the Dutch sandwich (cheese and bread) and carrots. I look forward to tailgating you more with my bike.

My appreciation also extends to everyone in GoncalvesLab. **Sara**, although I have known you for only a year, we have shared so much that I consider you a very close friend. You have brought out the emotional side of our small office and encouraged (even forced) us to think differently. I look forward to working with you for a long time. **Ivan**, fate and luck brought us together through the subletting of the house. Thank you for always standing up (literally) when needed. I wish you a lot of pepper. **Roy**, the most senior member of our small office, I always enjoyed our discussions; please take care of that evolving flower on your desk.

I have not forgotten DBL (or PRB), at least for now: **Stephanie**, my trauma-bonding friend, thank you for opening your office to us, chairing all C++ sessions, and listening to my complaints. **Mo** and **Arman**, I see you both as an inseparable duo; thank you for always being with me, for teaching me all the intricate parts of Dutch, and for making me feel welcomed and happy. It is always wonderful to see both **Munja** and **Zaytuna**. **Lieke**, I am still saddened by not being able to share a 3 p.m. coffee with you or organize joint birthday parties, but I hope to visit you in New York at least once. **Mostafa**, I have enjoyed our friendship—whether staying together at retreats and conferences, during gaming nights, or while driving in your car. Good luck in Eindhoven. **Gabriel**, I truly have no idea what I would have done without you on that final red piste during skiing. Thank you for saving me and for being open when it comes to discussing emotions. **Paul**, thank you for taking all the jokes so well; I have always valued hearing your opinions on "serious" topics. **Swier**, the "Thor from the Sea," thank you for all the morning coffees, gym tips, and the incredible ski trip. **Amelia** and **Chirag**, thank you for all your advice and friendliness; you always made me feel included. I am sure we will have a great time in India and at your wedding. **Timo**, I will never forget your excitement over Jenga at the retreat. **Jasper**, thank you for caring about people's emotions and for your deep yet humorous jokes. **Bram**, although you joined the group later, I believe you are one of the key elements that hold everything together. Keep your energy and always keep smiling. **Lorenzo**, I cannot get over how you found the postman on a random street and still made the effort to bring an SSD and mouse for me from Italy. I would not have been able to use the gift card without you. **Gerard**, thank you for always being there whenever I had

a question. Stay tall. **Inez**, although you joined recently, I appreciate that you are always present for all the activities, making them better with your smiles and laughter. **Bianca**, thank you for introducing me to incredibly "interesting" movies. **Chengyao**, thank you for sharing many nice memories. **Niek**, thank you for slowing your pace when running with us at the retreat. **Madelon**, thank you for demonstrating how kindness can spread rapidly when you are around. **Giorgio**, thank you for always being spontaneous and for teaching me all the secrets of Italian food. **Paolo**, I still cannot get over the fact that you joined a group of racing bikes with an Ov-fiets. **Claudio**, and **Katharina**, thank you for all the nice chatting and karting!! **Daniyal**, thank you for introducing us to South Indian cuisine. **Atilla**, thank you for imagining me as a tall person. **Aurora** and **Johanna**, thank you for the incredible apres-skis. **Rickard**, it is always interesting to hear a variety of different languages from you! **Alejandro**, thank you for being my ski roommate.

I would also like to thank **Daan**, **Ellie**, **Alex**, **Soufiane**, **Osman**, **Tom**, **Ombretta**, and **Daniele**for their valuable discussions and pleasant conversations.

Although it was only for four months, I would also like to thank every member of **Iorio-Lab** for their friendship. **Lorenzo**, it was quite unbelievable how quickly we became very close and spent every day and weekend together. If only I had known how to manage long-distance relationships! **AleDG**, thank you for all the food suggestions and for taking us to games and holidays. **Irene**, thank you for being the best neighbor and for going on a holiday every weekend; stay brave and continue to explore the world. **Thanos**, thank you for the nice collaboration and having a constant happiness! **Flavio**, never lose your energy and please take care of your back. **Anania**, thank you for all the good memories and for discovering "good" pizza places almost every day in Milan. **Linda**, **Lucia**, **AleV**, **Raf**, **Ricciardo**, and **Ottavio**, thank you all for the table football games and for always making me feel welcome.

Arda and **Ugur**, my lifelong friends who are as close as brothers, you have made this journey far easier simply by being in my life. It is hard to recall an important moment without both of you. I believe we will eventually be together in the same place in the future. **Ece** and **Polen**, thank you for all the support and friendship over the years. **Irem** and **Omer**, with the longest friendships I have to date, thank you for always being there for me and for sharing in my happiness. **Ipek**, my weekends with you and Arda are always among my favorite times.

I would also like to thank **Hasan**, **Gizem**, **Riley**, **Elcin**, **Mark**, **Gulpembe**, **Suheda**, and **Furkan** for making these four years as wonderful as possible.

My special and deepest thanks go to my family. First, to my parents, **Sevilay** and **Yusuf**: I will always be grateful for every sacrifice you made for me to be here—not just for the PhD. This achievement is yours as much as it is mine. My brother, **Ilbey**, you have always been an example to me throughout my life. Whatever I do, I do it knowing that you are behind me. My sister, **Ozge**, thank you for bringing such a positive dynamic to our family, for all the cakes you have baked for me, and, of course, for all the coffee gossip sessions we have shared. **Yusuf Goktug**, **Bahar Gokce**, and **Kerem Gökalp**, you are all like a sun rising in the sky to give light to our family. I am, and will always be, proud of all of you. (Anne ve baba, benim burada olmam için yaptığınız bütün fedakârlıkları hayatım

boyunca hatırlayacağım. Bu ve her başarım, benim kadar sizin de başarınızdır. Abi, hayatım boyunca verdiğim kararlar için önümde bir örnek oldun. Hala hayatımda yaptığım her şeyi senin arkanızda olduğunu bilerek yapıyorum. Özge abla, ailemize getirdiğin enerji ve bana abla olduğun için teşekkür ederim. Bana yaptığın kekleri ve birlikte kahveli dedikodu sohbetlerini hiçbir zaman unutmayacağım. Göktuğ, Bahar ve Kerem; hepiniz dünyaya doğan güneş gibi ailemize ışık oldunuz. Hepinizle hayatımın sonuna kadar gurur duyacağım.) Dear **Ollie**, you brought so much action to our lives—whether by escaping the house, jumping from the window, or simply sleeping on our laps.

Finally, my beloved wife and lifelong best friend, **Pınar**, we embarked on the same PhD journey over four years ago, and as we have for the past six years, we have supported each other every day. Your love, patience, and kindness have helped me overcome every obstacle. I cannot imagine a day passing without you.

CURRICULUM VITÆ

Yasin İlkağan TEPELI

26-04-1995 Born in Kale, Antalya, Türkiye.

EDUCATION

2020–2024 Ph.D. in Bioinformatics Lab, Delft University of Technology, The Netherlands
Thesis: Computational Tools for Optimizing Targeted Cancer Treatments and Addressing Bias
Promoter: Prof. dr. M.J.T. Reinders & Dr. J.P. Gonçalves

2023–2024 Visiting Ph.D. in IorioLab, Fondazione Human Technopole, Italy
Research: Stratifying Oncogene Addicted Cohorts by Drug Response

2018–2020 M.Sc. in Computer Science and Engineering, Sabanci University, Turkey
Thesis: Discovering Cancer Patient Subgroups with Functional Graph Kernels

2013–2018 B.Sc. in Computer Science, Bilkent University, Turkey
Thesis: AeroCast - Forecasting Flight Prices for Fixed Travel Days

AWARDS & HONORS

2020 2020 Best 3rd Poster Award (HIBIT)

2018 M.Sc. 2210-National Scholarship (TUBITAK)

2013 B.Sc. Comprehensive 100% Education Scholarship (Bilkent)

LIST OF PUBLICATIONS

1. S.A.C.H. Goossens, **Y.I. Tepeli**, C. Seale, J.P. Gonçalves, "SNMF: Integrated Learning of Mutational Signatures and Prediction of DNA Repair Deficiencies", *bioRxiv* (2024)

2. **Y.I. Tepeli**, M. de Wolf, J.P. Gonçalves "Metric-DST: Mitigating Selection Bias Through Diversity-guided Semi Supervised Metric Learning", *arXiv* (2024)

3. **Y.I. Tepeli**, J.P. Gonçalves, "DCAST: Diverse Class-Aware Self-Training Mitigates Selection Bias for Fairer Learning", *arXiv* (2024)

4. S. Pillay, **Y.I. Tepeli**, P. van Lent, T. Abeel, "A Metagenomic Study of Antibiotic Resistance Across Diverse Soil Types and Geographical Locations", *bioRxiv* (2024)

5. **Y.I. Tepeli**, C. Seale, J.P. Gonçalves, "ELISL: early–late integrated synthetic lethality prediction in cancer", *Bioinformatics* (2024)

6. C. Seale, **Y.I. Tepeli**, J.P. Gonçalves, "Overcoming selection bias in synthetic lethality prediction", *Bioinformatics* (2022)

7. H.I. Kuru, **Y.I. Tepeli**, O. Tastan, "GEGE: predicting gene essentiality with graph embeddings", *Düzce Üniversitesi Bilim ve Teknoloji Dergisi* (2022)

8. **Y.I. Tepeli**, A.B. Ünal, F.M. Akdemir, O. Tastan, "PAMOGK: a pathway graph kernel-based multiomics approach for patient clustering", *Bioinformatics* (2020)