Computational models for clinical drug response prediction aligning transcriptomic data of patients and pre-clinical models

Mourragui, S.M.C.

DOI
10.4233/uuid:9e4a11a1-e7cb-4c56-b69c-a1ee0a502f0f

Publication date
2023

Document Version
Final published version

Citation (APA)
Mourragui, S. M. C. (2023). *Computational models for clinical drug response prediction: aligning transcriptomic data of patients and pre-clinical models*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:9e4a11a1-e7cb-4c56-b69c-a1ee0a502f0f

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

COMPUTATIONAL MODELS FOR CLINICAL DRUG RESPONSE PREDICTION

ALIGNING TRANSCRIPTOMIC DATA OF PATIENTS AND PRE-CLINICAL MODELS

COMPUTATIONAL MODELS FOR CLINICAL DRUG RESPONSE PREDICTION

ALIGNING TRANSCRIPTOMIC DATA OF PATIENTS AND PRE-CLINICAL MODELS

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 13 February 2023 at 15:00 o'clock

by

Soufiane Marc Charles MOURRAGUI

Ingenieur Civil des Mines de Paris,
Paris Sciences et Lettres, Paris, France.
Born in Rouen, France.

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus	chairman
Prof. dr. L.F.A. Wessels	Delft University of Technology and Netherlands Cancer Institute, promotor
Prof. dr. ir. M.J.T. Reinders	Delft University of Technology, promotor
Prof. dr. M. Loog	Delft University of Technology and University of Copenhagen, promotor
Independent Members:	
Prof. dr. E. Voest	Utrecht University and Netherlands Cancer Institute
Prof .dr. E. Barillot	Institut Curie, France
Dr .ir. J. de Ridder	Utrecht Medical Center
Prof. dr .ir. G. Jongbloed	Delft University of Technology
Reserved Member:	
Prof. dr. ir. B.P.F. Lelieveldt	Delft University of Technology and Leiden University Medical Center.

Prof. dr. M.A. van de Wiel has contributed greatly to the preparation of this dissertation.

Keywords: cell lines, pre-clinical models, translational, cancer, transfer learning, machine learning, gene expression, predictive models, drug response, single cell

Printed by: ProefschriftMaken

Front & Back: Stefanie van den Herik.

ISBN 978-94-6384-414-7

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

Tout bon raisonnement offense.

Le Rouge et le Noir, Stendhal

CONTENTS

SUMMARY

Extensive efforts in cancer research over the past decades have markedly improved diagnosis and treatments, leading to better outcomes for cancer patients. Paradoxically, however, these discoveries have begun to shed light on a level of complexity that rules out the emergence of a universal cancer treatment. As any tumor is now known to be essentially a unique disease, clinicians and researchers are moving towards a new paradigm, termed "precision medicine", which consists of designing bespoke lines of treatment for each patient.

This paradigm-shift has been fueled by international consortia that have characterized large collections of tumors, thereby providing a vast reference for cancer heterogeneity. Two main strategies have been employed: sequencing of tumor biopsies directly extracted from patients or studying pre-clinical models, i.e., tumor cells cultured in artificial environments. While the first strategy generates clinically faithful data, the second strategy is flexible and cost-effective, and allows for the study of effects of various drugs at different concentrations.

Based on the large amount of data generated from pre-clinical models, computer scientists have developed various machine learning algorithms to model drug response based on these data. However, these models do not take into account the complexity of human tumors and the differences between model systems and human tumors, and are therefore not directly applicable in a clinical setting. In this thesis, we aim at bridging this gap. Specifically, we develop algorithms to integrate and align data generated from the two aforementioned strategies with a goal to predict drug response in patients from datasets generated using pre-clinical models.

SAMENVATTING

Uitgebreide inspanningen op het gebied van kankeronderzoek in de afgelopen decennia hebben de diagnose en behandelingen aanzienlijk verbeterd, wat heeft geleid tot betere vooruitzichten voor kankerpatiënten.

Echter, al deze bevindingen hebben ook licht geworpen op een niveau van complexiteit dat de opkomst van een generieke kankerbehandeling uitsluit.

Nu bekend is dat elke tumor in wezen een unieke ziekte is, gaan clinici en onderzoekers op weg naar een nieuw paradigma, “precision medicine”, dat ernaar streeft om voor elke patiënt een op maat gemaakte behandeling te ontwikkelen.

Deze paradigmaverschuiving is aangewakkerd door internationale consortia die grote verzamelingen tumoren karakteriseerden, en zodoende een enorme referentie vormen voor de heterogeniteit van kanker.

Er werden hiervoor twee strategieën gebruikt: 1) Sequentiebepaling van tumorbiopten die rechtstreeks van patiënten werden afgenomen, of 2) het bestuderen van preklinische modellen, d.w.z. tumorcellen gekweekt in een kunstmatige omgeving. Terwijl de eerste strategie meer realistische gegevens genereert, is de tweede strategie flexibel en kosteneffectief, wat het mogelijk maakt om de effecten van verschillende geneesmiddelen in verschillende concentraties te bestuderen.

Gezien de grote hoeveelheid gegevens die zijn gegenereerd uit preklinische modellen, hebben computerwetenschappers verschillende algoritmen voor machine learning ontwikkeld om de respons op geneesmiddelen te modelleren. Deze modellen houden echter geen rekening met de complexiteit van menselijke tumoren en zijn niet direct toepasbaar in een klinische setting. In dit proefschrift willen we deze kloof overbruggen. In het bijzonder ontwikkelen we algoritmen om gegevens die zijn gegenereerd met de twee bovengenoemde strategieën te integreren en op elkaar af te stemmen. Vervolgens gebruiken wij deze kennis om de respons op geneesmiddelen bij patiënten te voorspellen op basis van datasets die zijn gegenereerd met behulp van preklinische modellen.

RÉSUMÉ

La recherche contre le cancer a permis, ces dernières décennies, d'améliorer substantiellement diagnostiques et traitements, donnant ainsi de meilleures perspectives à un grand nombre de patients. Cependant, tous ces efforts de recherches ont mis en lumière un fort niveau de complexité, qui, paradoxalement, rend improbable l'émergence d'une thérapie unique. Ayant fait le constat que chaque cancer est unique, les médecins et les chercheurs se tournent désormais vers un nouveau paradigme, la médecine personnalisée, consistant à individualiser les stratégies thérapeutiques.

Dans le cadre de ce changement de paradigme, plusieurs consortiums internationaux ont séquencé un grand nombre de tumeurs, établissant une vaste ressource pour analyser l'hétérogénéité entre tumeurs. Ces consortiums suivent deux stratégies : le séquençage de biopsies directement prélevées de patients, ou l'étude de modèle dits pré-cliniques, c'est-à-dire la culture de cellules cancéreuses dans un environnement artificiel. Bien que la première stratégie soit plus réaliste, la seconde offre une grande flexibilité, permettant d'étudier l'effet de différents traitements.

Exploitant ces grandes bases de données pré-cliniques, de nombreux informaticiens ont développé des algorithmes d'apprentissage machine pour modéliser la réponse des tumeurs à différentes thérapies. Ces modèles, cependant, ne prennent pas en compte la complexité observée chez les patients et ne peuvent donc pas aisément être utilisés en clinique. Dans cette thèse, nous développons plusieurs algorithmes pour intégrer et aligner les données pré-cliniques et données patients, avec un objectif partagé : utiliser l'information disponible dans les bases de données pré-cliniques pour mieux prédire la réponse d'un patient à certains médicaments.

1

INTRODUCTION

Cancer is a major global health issue, with 19.3 million diagnosis and 10 million deaths for the year 2020, making it the leading cause of mortality for people under 70 years old in 57 countries [1]. Any progress in our understanding of this disease is therefore poised to have an impact for patients, be it by extending their lifespan or alleviating distressful symptoms. To reach this wishful goal, cancer researchers have made use of synthetic experimental models like cell-lines, animal models or organoids. We first give a brief presentation and overview of existing experimental cancer models, with a focus on cancer cell lines. We then present how these models have been used to study drug response prediction and highlight different modes of drug resistance in cancer patients. Finally, we focus on a specific area of research which employs machine learning methodologies to predict drug response from genomic data.

1.1. EXPERIMENTAL MODELS FOR CANCER RESEARCH

1.1.1. CANCER IS AN HETEROGENEOUS DISEASE

A healthy and functioning tissue consists of the aggregation of hundreds of thousands of cells, all working together in an orderly manner. To guarantee the integrity of the tissue, or **homeostasis**, each cell follows a set of instructions. These instructions are pre-recorded in the **DNA molecules** which form the genetic material of the cell. Following the central dogma of biology (Figure 1.1A), selective portions of the DNA molecules, called **genes**, are transcribed into **RNA** molecules. These molecules have diverse functions, but an important portion, called **mRNA**, are subsequently translated to **proteins** which are the main effectors of the cell.

This well-oiled machine can, however, go awry for various reasons. A first cause of disruption lies in the modification of the genetic material (Figure 1.1B). This can take several forms: base-pair modifications (**mutations**), repetitions or deletions of long stretches of DNA (**copy number aberrations**), fusion of normally-distant portions of DNA (**translocations**) or a local change in the base-pair order (**inversions**). These disruptions, called **somatic alterations**, are supplemented by modifications in the spatial layout of these long DNA

molecules inside the cell nucleus (Figure 1.1)C). The chromosomes are surrounded by a myriad of proteins and small molecules which together form the **chromatin** environment. For any gene, its expression is directly controlled by the surrounding chromatin environment ; modification thereof can thus have a drastic impact on gene expression and on the whole cell-wire. One example of such **epigenetic alteration** corresponds to the addition of a **methyl group** at a specific locus in one of the DNA molecules. Accumulation of such methyl groups in a gene, or in its vicinity, drives its expression down. Conversely, the absence of methyl groups leads to a higher expression.
Given the size of the genetic information (3 billions base-pairs), the possible alterations are endless, and although cells have sophisticated processes to avoid such disruptions, genetic and epigenetic alterations are not rare. Although they infrequently individually induce a profound change in cell behavior, the accumulations of hundreds, if not thousands, of such alterations within a few decades can drastically modify the function of a cell. Division and proliferation of this aberrant cell can compromise a tissue, thereby causing the growth of a malignant tumor. As a direct consequence, each cancer presents a highly specific set of alterations, which causes a large variability among patients.
Another layer of complexity stems from the evolutionary development of a tumor population. An alteration which provides a cell with a competitive advantage will rapidly expand into a **clone**, i.e., a group of cells harboring this very alteration. Emergence of subsequent alterations will divide this clone even further, leading to a high **intra-tumor heterogeneity**. Although a majority of sub-clones would eventually disappear under evolutionary pressure, we empirically observe that tumors are not mono-clonal and are made of a few different clones, each presenting a specific set of aberrations [2–4]. Understanding cancer development therefore requires charting the heterogeneity arising from these two levels of complexity. Even though common molecular, physiological and metabolic patterns are harbored by most tumors, i.e., the so-called **hallmarks of cancer** [5–7], cancer researchers are in need of versatile tools to precisely chart this complex landscape of oncogenic profiles.

1.1.2. CELL LINES ARE MODEL SYSTEM TO STUDY CANCER

A **cell line** is a human biopsy grown in an artificial environment – usually a Petri dish. Although they require a long protocol to be successfully cultured, cell lines have been shown to be a highly cost-effective way to study cancer. To this date, more than a thousand different cell lines have been established to study cancer, and some of them have been extensively studied (e.g. HeLa, RPE1, ...). This large variety of cell lines allows researchers to effectively chart the wide genomic landscape previously mentioned and compare various tumors in different conditions.
In the past decade, different consortia have been established to comprehensively characterize cancer cell lines: **GDSC** [8, 9], **CCLE** [10, 11] and **CTRP** [12, 13], to name the three largest ones. These three consortia have molecularly characterized over one thousand cancer cell lines by measuring gene expression levels (**RNA-seq**), the presence of certain mutations (**whole exome sequencing**), copy number profiles, methylation status at important loci (**methylation array**), as well as the quantification of specific proteins (**RPPA**). More recently, techniques such as single cell sequencing (**scRNA-seq** and **scATAC-seq**) have also been used to further characterize cell lines at a more refined reso-

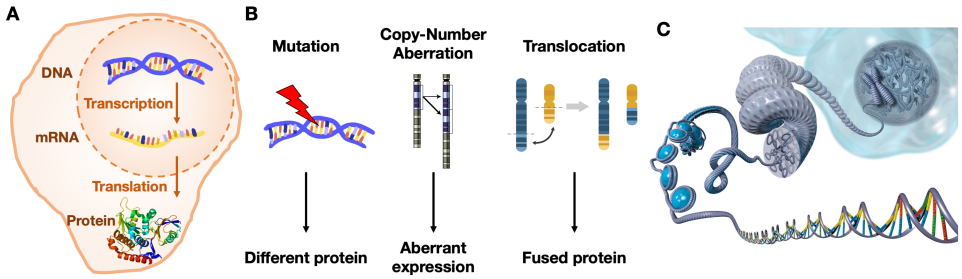


Figure 1.1 – **Genetic and epi-genetic alterations disrupt cell behaviors and cause tumor formation.** (**A**) Schematic of the central dogma. The genetic information (DNA), located in the nucleus (dark orange) is transcribed into smaller molecules, called RNA ; this process is called transcription, and governs gene expression. Some of these molecules, called mRNA, translocate to the cytoplasm (light orange) where they are translated into proteins. These proteins are the main cell effectors and are responsible for most physical functions. (**B**) Different genetic alterations can occur in a cell: mutations, copy-number aberrations (CNA) and translocations. These alterations can modify the translated protein which may result in different effects on the cell behavior. (**C**) The DNA material has a complex spatial organization and is surrounded by a complex biochemical environment. The spatial organization is compacted by the wrapping of chromosomes around small molecular complexes, called histones ; this wrapping has a direct influence on gene expression and disruption thereof can have a blasting impact.

lution. Alongside recent consortia which have characterized thousands of patients (e.g. TCGA or METABRIC), these efforts greatly help improve our understanding of cancer.

1.1.3. CELL LINES SUFFER FROM SEVERE LIMITATIONS

The protocol to establish a cancer cell line can be tedious, and not all tumor cells can be grown into a cell line. A first difficulty lies in the total absence of micro-environment, i.e., the population of non-cancerous cells present within and around a tumor, which plays a critical role in carcinogenesis. As a consequence, a successful cell line lineage must not be too much dependent on its surrounding micro-environment to thrive. Furthermore, as a cell line culture simply consists of an agar medium supplied with necessary growth factors, normal cells typically stop proliferating after a few replications. This creates a severe **sample selection bias** where only the most aggressive tumors are transformed into cell lines [14]. Indeed, these aggressive tumors are most likely to have broken down most dependencies with their surrounding stroma and are therefore more likely to happily grow in an agar plate.

Another key limitation arises from experimental artifacts caused by the way cell lines are cultured [15]. Some cell lines have been established several decades ago and their dissemination in various labs has caused them to genetically drift apart. As there is no universal experimental protocol and strategy to culture cell lines, it has been observed that cell lines originating from the same ancestral patients, harbor different mutations and copy-number profiles, directly affecting gene expression and important cell phenotypes such as drug response. As a consequence of this evolutionary differences, the key assumption that cell lines represent homogeneous mono-clonal populations is severely violated. This genetic drift furthermore hinders reproducibility, as two cultures of the same cell line can differ substantially.

Finally, and more fundamentally, cell lines are **very simplified model systems** which do not account for key aspects of tumor development and drug resistance. Especially, in the case of epithelial cancers, as hinted above, a tumor's mass critically depends on interactions with the surrounding stroma and micro-environment. For instance, cell lines completely lack vasculature, which obviates the need for **angiogenesis**, a key hallmark of cancer. Moreover, the complete absence of an immune system severely simplifies drug resistance and alleviates an important source of external pressure.

1.1.4. ADVANCED EXPERIMENTAL MODELS PARTIALLY ADDRESS THESE ISSUES

To account for these shortcomings, more complex experimental models have been developed. A first idea involves adding immune and stromal cells, alongside epithelial cancer cells, to reconstruct the tumor complexity. This **co-culturing** strategy has recently been exploited to study response to immuno-therapy by adding cytotoxic T-cells to traditional cell line cultures [16]. Other synthetic models, called **organoids**, resemble tumors by reconstructing the 3D tissue architecture of a tumor. To that end, pluripotent stem-cells are extracted from a patient and grown in a synthetic environment. Essential nutrients are provided, but no other cells are added. The cultured stem cells are expected to differentiate and reconstruct the tissue they were extracted from, thereby providing a useful and fertile ground for experimentation.
Moving closer to the human setting, researchers have also largely exploited animal models. Mammals evolutionary closely related to humans are ideal candidates as they can develop cancers which harbor strong genetic and morphological similarities with human tumors. Due to their small size and fast reproduction, mouse models were quickly selected and used for two main research purposes. A first aim is to understand the effect of certain genetic alterations on cancer formation. To do so, genetic alterations are introduced into a **mouse model** using genetic engineering. Mice can then be monitored over their lifespan to study the effect of the engineered alterations. A different strategy is employed when studying the response of pre-existing human tumors to different anti-cancer compounds. To reach this aim, researchers have developed **patient derived xenografts**, or PDXs, which are made of human tumors engrafted into immuno-deficient mice. This allows researchers to study the effect of a drug **in-vivo**.
Although these models represent a significant improvement over cell lines, they are not without flaws. A first problem comes from their cost-effectiveness : due to their higher complexity, the formation of organoids, mouse models and PDXs typically requires more time, experimental know-how and logistics. This in turn limits the scale of potential drug or CRISPR screens. Furthermore, some issues encountered with cell lines are also observed in PDX models [17] and certain cancer types have poor establishment rates in organoid models [18]. Finally, these models are perfect and none of them perfectly recapitulate the biology observed in humans.

1.2. DRUG RESPONSE

The large diversity among tumors has a direct clinical impact: when a patient enters the clinic, it is tedious to predict the most effective treatment regimen. We here intro-

Drug	Target or Mode of action	Cancer type
Chemotherapies		
Paclitaxel	Micro-tubule stabilization	Ovary, Esophagus, Breast, Lung, Cervix, Pancreas
Gemcitabine	Nucleoside analogs	Testis, Breast, Ovary, Lung, Pancreas, Bladder
Carboplatin	DNA cross-links	Ovary, Lung, Head and Neck, Brain
Cisplatin	DNA cross-links	Testis, Ovary, Cervix, Breast, Bladder, Head and Neck, Esophagus, Lung, Brain
5-Fluorouracil (5-FU)	Thymidylate synthase inhibition	Colon, Esophagus, Stomach, Pancreas, Breast, Cervix
Irinotecan	Topoisomerase inhibitor	Colon, Lung
Doxorubicin	Topoisomerase inhibitor	Breast, Bladder, Lymphoma, Leukemia
Etoposide	Topoisomerase inhibitor	Testis, Lung, Lymphoma, Leukemia, Brain, Ovary
Vinblastine	Micro-tubules disruption	Lymphoma, Lung, Bladder, Brain, Skin, Testis
Targeted therapies		
Trametinib	MEK1/MEK2	Skin
Ulixertinib	ERK1/ERK2	/
Vemurafenib	BRAF	Skin
Dabrafenib	BRAF	Skin, Lung
Erlotinib	EGFR	Lung, Pancreas
Gefitinib	EGFR	Lung
Olaparib	PARP (BRCA1/2 mutation)	Breast, Ovary, Pancreas
Trastuzumab	ERBB2 (Her-2)	Breast, Stomach
Afatinib	ERBB2 (Her-2)	Lung
Lapatinib	ERBB2 (Her-2)	Breast
Imatinib	BCR/ABL	Gastric, Leukemia
Nutlin-2	MDM2-TP53 complex	/

Table 1.1 – **Example of anti-cancer drugs routinely employed in clinical care.**

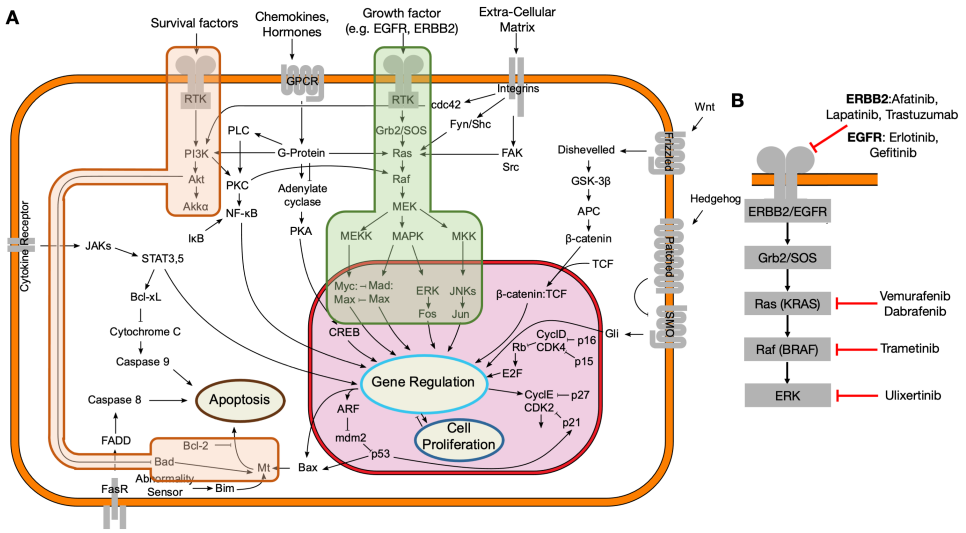


Figure 1.2 – **Tumor formation results from the disruption of molecular pathways**. (**A**) Flow-chart of the most important cancer-related pathways ; molecular information flows from the surface membrane (dark orange) to the nucleus where it can lead to the expression or the repression of down-stream genes. Each directed arrow corresponds to the functional activation of a daughter-protein by a parent protein, while a hammer-like arrow corresponds to a functional inactivation or a down-regulation. MAPK pathway is indicated a light green; PI3K pathway as light orange. (**B**) Several small molecules have been developed to inhibit the MAPK pathway by acting on observed disruptions like a mutation on EGFR, KRAS or BRAF, the amplification of ERBB2 or the deregulation of ERK.

duce the different mechanisms exploited by pre-malignant cells to grow into advanced malignancies and show how these can be therapeutically exploited. Finally, we present important concepts to measure and study drug response.

1.2.1. PROTEINS ARE ORGANIZED IN A NETWORK OF MOLECULAR PATHWAYS

As previously described, a cell is made of a large number of effectors, called **proteins**, which all function together to allow the cell to proliferate, replicate, communicate with surrounding cells, and among other tasks. The genetic material of human cells code for more than 20 000 proteins which are the products of more than a billion years of evolution. As a result, this vast web of proteins is organized in groups which have co-evolved to operate sophisticated tasks, and more importantly, to insure protection against potential disruptions such as genetic aberrations or viral infections. More precisely, the proteins are organized in chains, called **molecular pathways**, where proteins iteratively interact with one another: one parent protein biochemically communicates with another one, which in turns communicates with a different one, until a final molecule exerts its function. This organization results in a complex molecular network where a wide array of signals flows across the cell. A simplistic model of the cell highlighting key molecular pathways involved in cancer development can be found in Figure 1.2A.

A key example of signaling pathways are the **Ras pathways** [19], and specifically the **MAPK pathway** (Figure 1.2A, highlighted in green) [20, 21]. Each cell receives various sig-

nals from its surrounding environment. These signals indicate, for instance, whether the cell should proliferate, differentiate or die. Such signals take the form of small molecules that are exchanged between cells. Among these are **growth factors**, which play a key role in cancer development [22]. A growth factor transmits information by binding to a **growth factor receptor**, a protein from the family of **receptor tyrosine kinases** (RTK), located across the cell membrane. As a result, a complex biochemical reaction leads to the **phosphorylation** of certain peptides on the cytoplasmic tail of the receptor, which attracts two molecules, Grb2 ("Grab 2") and Sos, that form a biochemical complex. This complex activates another molecule, called **Ras**, which will then lead to a phosphorylation cascade: Ras → Raf → MEK → ERK1/ERK2. ERK1 and ERK2 finally translocate to the nucleus, where they activate a number of **transcription factors** responsible for chromatin remodeling and cell proliferation. Ras is also, among others, responsible for the activation of the **PI3K pathway** [23–25] (Figure 1.2A, highlighted in orange), which involves similar bio-chemical reactions which lead to the downstream inhibition of **apoptosis**, and therefore impedes the cell to activate its self-destruction signaling cascade.
Virtually all functions within a cell result from a similar **signal transduction** scheme. The specific biochemical reactions by which information is transmitted from one protein to another are very diverse: some, for instance, functionally activate a protein by phoshorylation [26], others by tagging proteins for degradation (**ubiquitylation**), and others modify the structure of a protein to enable or disable certain functional pathways. The spatial localization of a specific protein can also be crucial in the signal transduction. This is the case for the Ras pathways discussed above where the growth factor receptor brings together proteins which would have, in normal conditions, a limited chance to interact.
Finally, although we presented pathways with a linear mode of signal transduction, molecular pathways harbor non-linearities that are a necessary condition for a precise control of complex phenotypes. A first source of non-linearities stems from **feedback loops** which can either lead to self-inhibition of the pathway (negative feedback-loop) or an amplification (positive feedback loop). For instance, in the case of mitogenic signals transduced by the Ras pathways, ERK1 triggers the expression of a gene coding for the Sprouty protein [27], which prevents Grb2 and Sos to interact, thereby stopping Ras signaling at its very root. Positive feedback loops are observed in developmental biology, and are used to sustain decisions operated by a cell. Non-linearities can also emerge from protein interactions, especially when a message can only be passed after recruitment of two or more proteins. **Transcription factors** complexes provide a perfect example: gene expression initiation requires the recruitment of several molecules, potentially downstream of different pathways, yet a lack of one component would block the expression.

1.2.2. DISRUPTIONS OF MOLECULAR PATHWAYS LEAD TO CANCER GROWTH

As we discussed in Subsection 1.1.1, the complex structure of the cell can be disrupted by somatic alterations and epigenetic modifications. Whilst the latter affect individual molecular pathways by either enhancing or decreasing the transcription of particular proteins, the disruptions caused by somatic alterations are more diverse. A first source of disruptions comes from copy number amplifications, which act by significantly increasing the transcription of the encoded proteins. An important example is **ERBB2**, a growth

factor receptor important in many solid cancers [28]: the amplification of this gene increases the transcription of the protein which lives on the surface membrane. ERBB2 is a receptor tyrosine kinase upstream of the RAS pathway (Subsection 1.2.1) and is activated by **dimerization**. To trigger the signaling cascade, a ligand needs to bind to two proteins at the same time. When the concentration of ERBB2 increases, so does the probability of dimerization on the surface membrane. This in turn over-stimulates downstream pathways, including the pro-proliferation and anti-apoptotic MAPK and PI3K pathways.
Copy number deletions have the opposite effect on a cell. Deletions are exploited by cancer cells to shut down certain pathways, preferentially pro-apoptotic and anti-proliferative ones. **BRCA1** and **BRCA2** exemplify this vulnerability [29]. These two proteins insure the integrity of the genome by triggering alerts whenever replication errors are found, and specifically **double strand breaks**. These alerts can have two effects: either the cell repairs the defect, or the apoptosis pathway is triggered to avoid a subsequent accumulation of alterations. A lower expression of either BRCA1 and BRCA2 caused by copy number deletion would then temper the downstream protective pathways and allow replication errors to accumulate, increasing the chance of malignant tumor growths. BRCA1/2 are examples of genes which protect a cell against tumor formation and are called **tumor suppressor genes**. On the other hand, genes which are over-expressed and lead to cell proliferation are called **proto-oncogenes**.
Proto-oncogenes can also suffer from mutations which alter the form of the encoded protein. This is the case for instance in **EGFR**, another growth factor receptors upstream of the Ras pathways [30], and a member of the same family as ERBB2. Selective mutations on this gene code for a truncated protein which harbors a fundamental flaw: it no longer requires a ligand to activate the downstream pathways and is said to be **constitutively activated**. Instead of transmitting an exogenous signal, the truncated protein will sustain a continuous pro-mitogenic signal, causing the cell to rapidly replicate and grow. This alteration is frequently observed in lung, ovary and breast cancers. Other proto-oncogenes like **KRAS** (from the family coding for Ras) [31], **PI3K** [32–34] or **BRAF** (downstream of Ras) [35] can also suffer from similar mutations. The resulting revamped gene is referred to as an **oncogene**.
By altering the function of a protein, mutations can also impact tumor suppressor genes. The first important example of such action is **TP53**, also known as **p53** [36]. TP53 is the cornerstone of several anti-proliferative processes: it can trigger DNA damage repair, block the cell division should the genome integrity be compromised, and prompt the apoptosis pathway if the damage is too severe. Disruption of TP53 is therefore necessary for an aspiring cancer cell. This can be achieved by certain mutations of P53 that dwarf its function. Importantly, owing to its complex mode of action, one mutated allele of TP53 effectively reduces the action of the gene by a factor of 16, thereby granting a complete freedom to proliferate. Other important tumor suppressor genes suffer from equivalent mutations, like **RB1**, or BRCA1/BRCA2.
Finally, a rarer form of pathway alterations arises from translocations. This is for instance the case in a large number of leukemias where the tyrosine kinase ABL1 (chromosome 9) is found adjacent to the BCR gene (chromosome 22). This transposition results in an oncogenic fused gene called **BCR/ABL1** which modifies ABL1 to become constitutively active [37]. The result is a constant firing of mitogenic signals from the resulting pro-

tein with a lack of sensitivity to negative feedback, and as a consequence a proliferative phenotype potentially leading to malignancy.

1.2.3. EXPLOITING PATHWAY ALTERATIONS THERAPEUTICALLY

This acquisition of oncogenic alterations, which is consubstantial with the development of an aggressive and malignant phenotype, can however be exploited. A first approach consists of using **cytotoxic** drugs, called **chemotherapies**, which aim at damaging and killing cancer cells. This strategy capitalizes on two frequently observed cancer vulnerabilities: a lack of a proper **cell-cycle control**, and a **high proliferation** rate. Mitosis in normal cells is a highly regulated process where a diligent control of the cell integrity prevents replication, should any genetic aberration be found. Tumor cells however usually jettison this control pathway, thereby allowing genetic aberrations to go through cell cycle unnoticed. By inducing alterations, **DNA-damaging agents** (e.g. **Cisplatin**, **Carboplatin**, **Oxaliplatin** or **Temozolomide**) cause aberrations to accumulate in cancer cells, thus compromising their ability to function. Although these compounds offer some **cancer cell specificity** – normal cells have an intact repair apparatus and can cope with the damage – their complex and pleiotropic modes of action do affect non-cancerous cells leading to important side-effects and, sometimes, the formation of a secondary tumor [38]. Historically, this strategy has started to be widely used in the clinic in the second half of the 20th century and is today part of the **standard of care** for most solid tumors. A non-exhaustive list of standard-of-care chemotherapies is provided in Table 1.1.
A more recent strategy consists of targeting precise cancer vulnerabilities. As presented in Subsection 1.2.2, a cancer phenotype results from many genetic and epigenetic alterations disrupting various molecular pathways. Taking advantage of recent technological advances like **next-generation sequencing**, these alterations can be comprehensively charted [39]. This provides clinicians with a clear picture of what sets tumor cells apart from healthy cells and hints at therapeutic strategies, called **targeted therapies**, to exploit these differences. Let's take the example of the MAPK pathway (Figure 1.2A, highlighted in green). This proliferation pathway is hijacked by tumor cells to accelerate their evolution and presents alterations in many cancers: EGFR mutation in lung cancer, KRAS mutation in colon cancer, ERBB2 amplification in breast cancer, BRAF mutation in skin melanoma. In the case of EGFR and ERBB2 aberrations, drugs have been developed to replace their ligands without prompting any phosphorylation cascade: small molecules like **Erlotinib**/**Gefitinib** (EGFR) or **Afatinib**/**Lapatinib** (ERBB2), and monoclonal antibodies like **Trastuzumab** (ERBB2). Mutations activating BRAF are specifically targeted by drugs like **Vemurafenib** or **Dabrafenib**. Downstream of BRAF, MEK can also be directly inhibited by **MEK-inhibitors** like **Trametinib**. Finally, ERK can also be inhibited, for instance by **Ulixertinib**. All these drugs exploit a specific known vulnerability of the cancer cells to reduce the activity of the MAPK pathway (Figure 1.2B). As these cancer cells have often built their whole expansion strategy on the deregulation of this pathway, a phenomenon known as **oncogene addiction**, inhibition thereof often leads to a good initial response. Therapeutic strategies have also been developed for targeting other pathways: **Olaparib** to exploit BRCA1/2 deficiencies, **Alpelisib** for PI3K pathway deficiencies and **Imatinib** for hematopoietic tumors harboring a BCR/ABL fusion. A non-exhaustive list of standard-of-care targeted therapies is provided in Table 1.1.

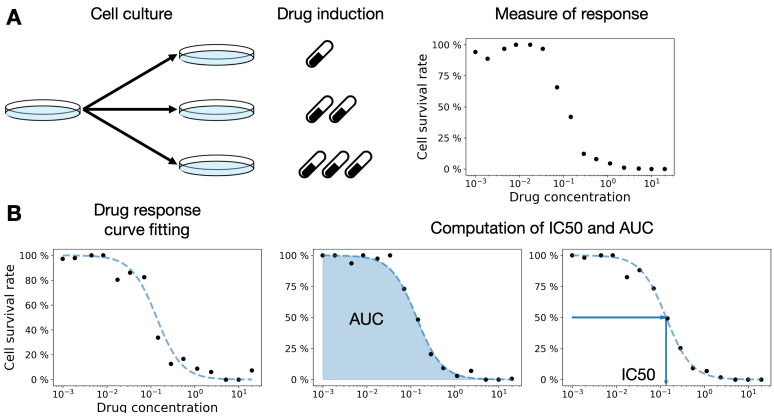


Figure 1.3 – **High-throughput drug screening allows fast evaluation of anti-cancer drugs**. (**A**) A given anti-cancer drug can be studied by subjecting one cell line to a wide array of concentrations. The proportion of surviving cells can then be extracted for each concentration, forming a drug-response curve. (**B**) A drug-response curve is specific to a pair made of one cell line and one drug. To aggregate the different drug response curve values, we first fit a sigmoid function. This sigmoid function can then be exploited to compute the Area Under the Curve (AUC), which corresponds to the integral of the drug response curve. Another metric, called IC_{50}, consists of measuring the concentration of the drug which kills half of the cell population.

Although promising, these targeted therapies are limited by several factors. First, designing a compound to target a specific vulnerability is a challenging task: the compound must fit in very specific areas of the targeted molecule, while not altering the function of other proteins which are often evolutionary related. Second, a drug can be limited by its **pharmacokinetics**, i.e. by its propensity to dilute in the plasma, to be metabolized, and ultimately to access the cancer cells and exert its action. Drug pharmacokinetics depend on the drug itself, but also on the patient's characteristics, which creates another source of important variability. Finally, **toxicity** can impede a functioning drug to be used in the clinic. All the therapeutic strategies listed above come at the cost of severe toxic effects on the healthy cells, which can harm the patient more than the pathology itself. These limitations make the drug development process complex and cause an important **attrition rate**: although the space of possible compounds is very large, the number of drugs with proven clinical benefit is remarkably low.

1.2.4. MEASURING AND STUDYING DRUG RESPONSE USING MODEL SYSTEMS

Studying whether a cancer patient will be sensitive or resistant to a particular treatment is the product of several genetic, epigenetic, physiological, metabolic and even psychological factors. Owing to their high versatility (Subsection 1.1.2), cell lines offer an ideal structure to study the impact of genetic and epigenetic variability on the response of a tumor cell population to a particular drug regimen. Exploiting recent technological advances in screening technology, a cell line can be efficiently and cost-effectively subjected to a drug at different concentrations (Figure 1.3A). For each concentration, the proportion of surviving cells (called **viability**) can be evaluated. Comparing this survival rate to the drug concentration (log-scale) yields a **drug response curve**. This drug re-

Loss function	Definition	Examples of algorithms
Least-squares	$\ell(x,y) = (x-y)^2$	ElasticNet, neural networks, Kernel Ridge Regression
Hinge loss	$\ell(x,y) = \max(0, 1-xy)$	SVM
Logistic	$\ell(x,y) = \log(1+e^{-xy})$	Logistic regression

Table 1.2 – **Loss functions frequently used in supervised learning algorithms.**

sponse curve can be processed *in-silico* to derive useful metrics indicative of resistance and sensitivity (Figure 1.3B). After fitting a **sigmoid**-like curve to the measurements [40], two key metrics can be derived:

- **Area Under the response Curve** (**AUC**), corresponding to the area under the sigmoid (i.e. its integral). It provides an average of the cell line viability across drug concentrations. It is usually compared to the AUC of a flat sigmoid, corresponding to a non-responder.

- **Half-inhibitory concentration** (**IC_{50}**): IC_{50} corresponds to the concentrations which kills half of the cell population.

Both these measures provides an indication of **potency**: low IC_{50} and AUC indicate that a small concentration kills a large proportion of the cell population. Conversely, a large IC_{50} and AUC imply a resistance to the drug. These two measures are highly correlated. More complex and advanced model systems (Subsection 1.1.3) can also be employed in such drug-screening endeavors. However, their higher complexity prohibits screening at the scale possible in cell lines. The measure resulting from such screens differ also from the ones computed on cell lines: since mouse models are 3-dimensional, the viability measure employed is based on the measured tumor volume. Comparing this tumor volume between untreated and treated PDXs provide a powerful response read-out.

1.3. MACHINE LEARNING

Once a large compendium of drug response data has been garnered, a natural step is to look for ways to relate it to the genomic characterization of the model systems under study. A powerful approach, called **machine learning**, consists of finding relationships and patterns within the genomic data, or between the genomic data and the drug response, by means of statistical inference. We present here a brief overview of these methods and refer the reader to *The Elements of Statistical Learning* [41] and *Pattern Recognition and Machine Learning* [42] for a thorough and authoritative presentation.
In our case, we consider to have $n > 0$ samples and $p > 0$ features. A feature can be, for instance, a gene in the case of gene expression, a genomic region in the case of copy-number, a SNP in the case of mutations or a CpG island for methylation. The genomic data of interest is stored in a matrix $X \in \mathbb{R}^{n \times p}$, called **design matrix**, where each row corresponds to the characterization of a single sample. The drug response measurements are stored in a vector $Y \in \mathbb{R}^n$ with a pairing between the rows of Y and X (Figure 1.4A).

Activation function	Definition
ReLU	$\sigma(x) = \max(0, x)$
Sigmoid	$\sigma(x) = (1 + e^{-x})^{-1}$
Hyperbolic tangent	$\tanh x$

Table 1.3 – **Activation/Linked functions frequently used in neural networks and generalized linear models.**

Kernel	Definition
Linear	$K(x,y) = x^T y$
Polynomial	$K(x,y) = (x^T y + b)^a, a \in \mathbb{N}, b \in \mathbb{R}.$
Gaussian	$K(x,y) = \exp\left(-\gamma \|x-y\|^2\right) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right), \gamma, \sigma > 0$
Laplacian	$K(x,y) = \exp\left(-\gamma \|x-y\|\right) = \exp\left(-\frac{\|x-y\|}{\sigma}\right), \gamma, \sigma > 0$
Matérn	$K(x,y) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\sqrt{2\nu}\frac{\|x-y\|}{\sigma}\right)^{\nu} K_\nu\left(\sqrt{2\nu}\frac{\|x-y\|}{\sigma}\right), \nu, \sigma > 0$

Table 1.4 – **Kernels, or similarity functions, frequently used in machine learning.** Γ denotes the Gamma function, K_ν the modified Bessel function of the second kind of order ν. Please note that, for both the Laplacian and the Gaussian kernels, two equivalent definition are given; both are used in this thesis.

1.3.1. SUPERVISED LEARNING ALLOWS DRUG RESPONSE MODELLING

EMPIRICAL RISK MINIMIZATION (ERM)

A first category of machine learning techniques, called **supervised learning**, consists of finding a function which relates X to Y. Formally, given a set of functions $\mathscr{F} \subset \{f : \mathbb{R}^p \mapsto \mathbb{R}\}$ which map the genomic profiles to a real number, a supervised learning algorithm finds the function $f^* \in \mathscr{F}$, which is a solution of the **empirical risk minimization** (ERM) problem, defined as

$$f^* = \arg\min_{f \in \mathscr{F}} \frac{1}{n}\sum_{i=1}^{n} \ell\left(f(X_i), Y_i\right) + R(f). \tag{1.1}$$

In Equation (1.1), ℓ, called the **loss function** 1.2, penalizes predictions $f(X_i)$ deviating from the actual values Y_i. R, called the **regularization term**, allows to add further constrains to the function f. Different classes of supervised learning algorithms have been developed, and these differ by the class of functions $\mathscr{F}$ they consider, the loss function ℓ, the regularizer R, and the optimization scheme used to solve Equation (1.1).

LINEAR MODELS

A first category, called **linear models** (Figure 1.4B), corresponds to affine projections of the genomic data matrix X; formally, $\mathscr{F}_{linear} = \{x \mapsto w^T x + b \mid w \in \mathbb{R}^p, b \in \mathbb{R}\}$. Various regularizers have been developed for these models; the most famous of them being arguably the **ElasticNet** regularization defined as $R(w) = \lambda\alpha\|w\|_1 + \lambda(1-\alpha)\|w\|_2^2$, with $\lambda > 0$ called the regularization parameter and $\alpha \in [0,1]$ called the ℓ_1-ratio. The ElasticNet optimization therefore boils down to find w^* and b^* such that:

$$w^*, b^* = \arg\min_{w \in \mathbb{R}^p, b \in \mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}\left(w^T X_i + b, Y_i\right)^2 + \lambda\alpha\|w\|_1 + \lambda(1-\alpha)\|w\|_2^2. \tag{1.2}$$

These linear models, albeit fast, robust and often competitive, can potentially suffer from their simplicity. Biological systems are notorious for being non-linear (Subsection 1.2.1), and predicting a certain phenotype from linear models is poised to be restrictive. Various extensions of linear models have been proposed in the statistics and machine learning literature and we present two independent extensions which are widely used in the computational biology community: deep **neural networks** and **kernel methods**.

DEEP NEURAL NETWORKS

A natural way to add a non-linearity is to combine a linear classifier with a non-linear function $\sigma : \mathbb{R} \mapsto \mathbb{R}$. This non-linear function, called **linked function** in statistics and **activation function** in deep learning (Table 1.3), is an intrinsic property of the model which can reflect some prior-knowledge. The functions considered by these approaches, called **generalized linear models**, are of the type $\sigma\left(w^T x + b\right)$ with $w \in \mathbb{R}^p$ and $b \in \mathbb{R}$. The regularizer usually consists of a mixture of ℓ_1 and ℓ_2 norms, *à la* ElasticNet.
In order to add some more complexity to the model, another idea consists of combining non-linear blocks. For instance, the input $x \in \mathbb{R}^p$ could feed into d_1 different generalized linear models, yielding d_1 weights $W_1 = \left[w_{1,1}, .., w_{1,d_1}\right] \in \mathbb{R}^{p \times d_1}$ and bias term $B_1 = \left[B_1, .., B_{d_1}\right]$. Each of these d_1 single models is called a **neuron**, and the d_1 neurons forms a **layer**. These d_1 neurons can finally be aggregated in a final linear model, called **output neuron**, with parameters $W_2 \in \mathbb{R}^{d_1}$ and $b_2 \in \mathbb{R}$. The corresponding function $f(\cdot; W_1, b_1, b_2, W_2)$ is therefore formally defined as:

$$\forall x \in \mathbb{R}^p, \quad f(x; W_1, b_1, b_2, W_2) = \sum_{k=1}^{d_1} \left[W_{2,k} \sigma \left(\sum_{j=1}^{p} W_{k,j} x_j + b_{1,k} \right) + b_{2,k} \right]. \tag{1.3}$$

The function defined in Equation (1.3) is a function with **one hidden-layer**. Following the same idea, additional layers can be added, creating a **deep neural network** (Figure 1.4C-D). The regularizer employed in neural networks usually consists of a sum of ℓ_2 norms of the neurons weights and is called **weight decay**. The ERM is usually solved by stochastic gradient descent, exploiting the chain rule – a process called **back-propagation**. Additional regularization methods have been developed in the deep learning fields, such as **dropout**, **early stopping** or **batch normalization**, but these act on the optimization strategy rather than the regularization function.

NON-PARAMETRIC KERNEL METHODS

The deep neural networks consist in a bottom-up approach, where non-linearities are combined. A complementary class of approaches, called **kernel methods**, consists of first expanding the number of features by integrating a large number of potential non-linearities, and then performing a linear approach on this expanded space. This method relies on the notion of **positive-definite kernels** (p.d.), which are functions $K : \mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R}$ which satisfy

$$\forall k > 0, \forall x_1, .., x_k \in \mathbb{R}^p, \forall \lambda_1, .., \lambda_k \in \mathbb{R}, \quad \sum_{1 \leq i,j \leq k} \lambda_i \lambda_j K\left(x_i, x_j\right) \geq 0. \tag{1.4}$$

A few examples of kernels are given in Table 1.4. Positive kernels are especially interesting due to two theorems. The first one, called **Moore–Aronszajn theorem**, states that a

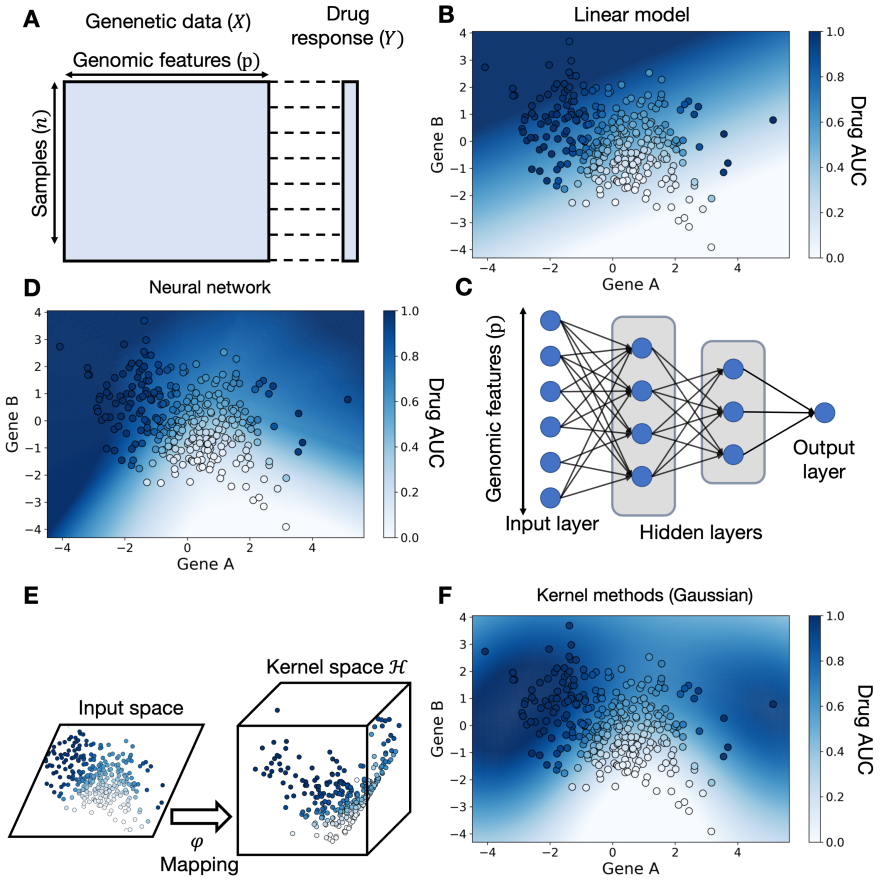


Figure 1.4 – **Supervised learning methods**. (**A**) A supervised machine learning task finds a function f which approximates the **labels** Y by means of the input matrix X. X contains for each sample (row) its molecular characterization, e.g. its gene expression profile. (**B**) Heatmap of a synthetic 2D example where drug response is predicted by a linear models, overlayed by the points used for training the model. (**C**) A neural network is organized in **layers**, each one containing **neurons**. The information,e.g. gene expression profiles, flows from left to right, and follows a serie of non-linear transformation. The end neuron, called **output neuron**, approximates the labels. (**D**) Same data as in (**B**) modeled by a neural network. (**E**) Kernel methods implicitly map the gene expression profiles into a richer **kernel space** $\mathcal{H}$ where linear methods are performed. (**D**) Same data as in (**B**), modeled by a kernel ridge regression.

p.d. kernel can be seen as a linear method in a more complex space, which is an intrinsic property of the kernel ((Figure 1.4E). Formally, for any p.d. kernel, there exists a Hilbert space $\mathcal{H}$ and a mapping φ from the feature space $\mathbb{R}^p$ to $\mathcal{H}$ such that

$$\forall x, y \in \mathbb{R}^p, \quad K(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}. \tag{1.5}$$

The functions considered in the ERM are of the types $x \mapsto \langle \omega, \varphi(x) \rangle_{\mathcal{H}}$ with $\omega \in \mathcal{H}$ and corresponds to a linear projection in the new complex space; the regularization corresponds to $\lambda \|\omega\|^2_{\mathcal{H}}$ with $\lambda > 0$ hyper-parameter. However, the space $\mathcal{H}$ can be arbitrarily

complex – potentially infinite dimensional – and therefore computationally intractable. A second theorem, called the **Representer theorem**, solves this issue by showing that the solution of the ERM is a combination of the training samples after mapping in the complex space $\mathscr{H}$. Formally, if we denote by ω^* the ERM optimum, then

$$\exists \alpha \in \mathbb{R}^n, \quad \omega^* = \sum_{i=1}^{n} \alpha_i \varphi(X_i). \tag{1.6}$$

Combining the Representer theorem with the ERM formulation provides a computationally tractable way to optimize the problem, known as the **kernel trick**. Two loss functions are usually used for supervised kernel methods: the **hinge loss**, yielding **support vector machines** (SVM) and the least-squares loss, yielding **Kernel Ridge Regression** (KRR) (Figure 1.4F).

1.3.2. UNSUPERVISED LEARNING EXPLOITS LARGE UNLABELLED DATASETS

Obtaining drug response data, or any other label of interest, can be very expensive. As a result, unlabelled genomic datasets far outnumber the amount of labelled data, and various algorithms have been designed to extract and exploit patterns from these rich resources. This class of algorithms, usually referred to as **unsupervised learning**, comes in different flavors. A first goal of these methods consists of finding **clusters** in the data, i.e. groups of samples which share the same neighborhood while being distant from the rest of the dataset. Various algorithm have been developed for this purpose, like **K-mean**, **hierarchical clustering**, and more recently **Louvain** and **Leiden clustering**.
A second goal consists of mapping the high-dimensional input space to a lower-dimensional **latent space**, or **embedding space**, which captures most of the signal. This concentrated signal can then be exploited in several ways, e.g. for visualization, analysis of latent directions, or to reduce complexity of supervised learning task when only a subset of the dataset is labelled. These **dimensionality reduction** methods, sometimes also referred to as **manifold learning** approaches, consists, at least implicitly, of two main elements: an **encoder** function f_E^* and a **decoder** function f_D^*. The encoder maps the input space to the embedding space ; the decoder takes as input the embedding and maps it back to the input space. The combined function $f_D^* \circ f_E^*$ can be compared to the actual sample, and these two functions are the solution of an optimization scheme that minimizes this discrepancy over two sets of function $\mathscr{F}_E$ and $\mathscr{F}_D$:

$$f_E^*, f_D^*, \quad = \quad \underset{f_e \in \mathscr{F}_E, V \in \mathscr{F}_D}{\operatorname{argmin}} \sum_{i=1}^{n} D\left(X_i, f_D \circ f_E(X_i)\right). \tag{1.7}$$

The choice of a distance measure D (or pseudo-distance) and the two sets of functions $\mathscr{F}_E$ and $\mathscr{F}_D$ divide the dimensionality reduction methods in different categories.

1. A first choice consists of restricting $\mathscr{F}_E$ and $\mathscr{F}_D$ to linear projections. When, furthermore, the Frobenius norm is considered for D, then we obtain a popular algorithm called **principal component analysis** (PCA). Another choice consists of taking only non-negative projection matrices, resulting in **non-negative matrix factorization**.

2. A second choice consists of modelling both f_E and f_D as neural networks. The resulting architecture is called an **auto-encoder** and allows for a non-linear embedding of the input. An elegant refinement of these auto-encoder architectures, called **variational auto-encoders** (VAE), are also widely used in probabilistic modelling.

3. Finally, linear dimensionality reduction can be performed after kernel transformation. Similar to Kernel Ridge Regression, these non-linear dimensionality reduction methods consists of first mapping the input data to a kernel space $\mathscr{H}$, and then performing PCA in this potentially infinite-dimensional space. This approach, called **Kernel PCA**, also exploits the kernel trick presented above.

1.3.3. DIFFERENCES IN DATASET COMPOSITION CAN BE CORRECTED USING TRANSFER LEARNING

Machine learning algorithms are designed to extract relationships from a **training data**. However powerful, these algorithms are often specific to the data they have been trained on. Several cases have been reported showing poor **generalization** of such algorithms to data points not encountered during training. **Transfer learning** aims at circumventing this issue by correcting for differences between a **source** dataset, used for training, and a **target** dataset of interest. We refer the reader to [43, 44] for comprehensive reviews on transfer learning.

DIFFERENCES IN PROBABILITY DISTRIBUTION AND ERM

In Subsection 1.3.1, we presented the Empirical Risk Minimization used in supervised learning. Specifically, the first term in Equation (1.1) corresponds to the empirical approximation of the **risk** function denotes $\mathscr{R}$. Formally, in our setting, the risk associated with a function $f \in \mathscr{F}$ is defined as:

$$\mathscr{R}\left(f, \nu_{(X,Y)}\right) \quad = \quad \mathbb{E}_{X,Y\sim\nu_{(X,Y)}}\left[\ell\left(Y, f\left(X\right)\right)\right], \tag{1.8}$$

where $\nu_{(X,Y)}$ denotes the joint probability distribution of the inputs and the labels. The function f^* defined in Subsection 1.3.1 minimizes the quantity defined in Equation 1.8 over $f \in \mathscr{F}$ (e.g. linear functions, neural networks, kernel methods, etc.).
In a typical transfer learning scenario, the predictor f^* is trained on a source data following a probability distribution $\nu^S_{(X,Y)}$, but applied on a target data following a different probability distribution $\nu^T_{(X,Y)} \neq \nu^S_{(X,Y)}$. This causes $\mathscr{R}(f, \nu^T_{(X,Y)})$ to differ from $\mathscr{R}(f, \nu^S_{(X,Y)})$ for most (if not all) functions f. As a direct consequence, the optimal function f^* which minimizes $\mathscr{R}(\cdot, \nu^S_{(X,Y)})$ is not guaranteed to be the optimal function for the target.
A first approach to solve this issue consists of observing that:

$$\mathbb{E}_{\nu^T_{(X,Y)}}\left[\ell\left(Y, f\left(X\right)\right)\right] = \mathbb{E}_{\nu^S_{(X,Y)}}\left[\ell\left(Y, f\left(X\right)\right)\frac{\nu^T_{(X,Y)}\left(X,Y\right)}{\nu^S_{(X,Y)}\left(X,Y\right)}\right]. \tag{1.9}$$

This re-formulation, highlights a key problem: the joint probability distribution of the target should be known for all source samples. Since the target data is usually only par-

tially labelled, getting a good estimate of $\nu^T_{(X,Y)}(X,Y)$ for all source points is computationally challenging without further model assumptions.

DIFFERENT TRANSFER LEARNING ASSUMPTIONS

The two probability distributions $\nu^S_{(X,Y)}$ and $\nu^T_{(X,Y)}$ can differ in many different ways, leading to different methodologies. Specifically, using the Bayes rule, we have $\nu^S_{(X,Y)} = \nu^S_{Y|X}\nu^S_X = \nu^S_{X|Y}\nu^S_Y$ and $\nu^T_{(X,Y)} = \nu^T_{Y|X}\nu^T_X = \nu^T_{X|Y}\nu^T_Y$. These equalities allow the division of transfer learning methods in three categories:

- If $\nu^S_{Y|X} = \nu^T_{Y|X}$, the methods only differ by their prior distributions ; this setting is called **covariate shift**.
- If $\nu^S_{X|Y} = \nu^T_{X|Y}$, the methods differ by the prior distribution of their labels ; this setting is called **label shift**.
- If $\nu^S_X = \nu^T_Y$, the methods only differ by their posterior distributions.

This choice of assumption is further guided by the availability of data. The target data can either contain labels (**inductive transfer learning**) or be completely unlabelled (**transductive transfer learning**). In some cases, called **unsupervised transfer learning**, both source and target labels do not contain any labels.

1.4. CONTRIBUTION AND THESIS PLAN

Exploiting machine learning algorithms in various shapes and forms, several studies have attempted to predict drug response in cell lines [45–51]. However, cell lines models differ strongly from human tumors (Subsection 1.1.3) and these drug response models are poised to suffer from a poor transferability when applied in a clinical setting (Subsection 1.3.3). Capitalizing on promising early studies [52, 53], we set out to use transfer learning techniques (Subsection 1.3.3) to derive biomarkers predictive of drug response in the clinic. In this thesis, we make the following contributions:

1. Noting a large imbalance in predictive power towards gene expression [46, 54, 55], we develop **Percolate**, a methodology which derives DNA-based biomarkers by exploiting the common signal with gene expression (Chapter 2). To account for the peculiarities of each omic-data type, Percolate extends correlation-based alignment methodologies [56, 57] by means of Exponential-family PCA.

2. We then turn our focus to gene-expression based biomarkers. To systematically correct gene expression profiles for differences between cell lines and tumors, we develop a linear transfer learning technique, **PRECISE**, based on a comparison of principal components (Chapter 3). Technically, PRECISE extends previous subspace-based transfer learning approaches [58–60] by offering a simpler analytical expression for the geodesic paths on the Grassmann manifold based on the notion of Principal Vectors.

3. In Chapter 4, we posit that linearity is too strong an assumption for our clinical drug response prediction task. We employ the framework offered by kernel methods to introduce non-linearities in PRECISE and thereby derived **TRANSACT**. Using patient data from The Cancer Genome Atlas (TCGA) and the Hartwig Medical function (HWG), we validate TRANSACT on a wide range of drugs. We propose an interpretability scheme of TRANSACT which allows us to provide insights in the mechanisms associated with the resistance to Paclitaxel and Gemcitabine.

4. Finally, noting that neither cancer cell lines nor epithelial tumor cells are made of homogeneous mono-clonal populations, we developed **Sobolev Alignment** to extend the previous comparisons to scRNA-seq data (Chapter 5). Sobolev Alignment uses recent advances in deep generative modeling (VAE) and approximate kernel methods (Nyström approximation) to compare complex non-linear patterns of gene expression. We applied Sobolev Alignment to lung cancer cell lines and showed the conservation of immune-related pathways in cancer cell lines.

2

DESIGNING DNA-BASED PREDICTORS OF DRUG RESPONSE USING THE SIGNAL JOINT WITH GENE EXPRESSION

Soufiane MOURRAGUI, Marco LOOG, Mirrelijn VAN NEE, Mark A. VAN DE WIEL, Marcel J.T. REINDERS, Lodewyk F.A. WESSELS

Parts of this chapter have been deposited on Biorxiv. Extension of this chapter has been accepted for presentation at RECOMB 2023 and is currently under review at Genone Research.

2.1. INTRODUCTION

Over the course of their lifespan, human cells accumulate molecular alterations that result in the modification of behavior [61]. When aggregated at the tissue level, these alterations can compromise tissue homeostasis, in turn clinically impacting a patient [6]. Understanding the combined effect of these alterations is key to designing bespoke lines of treatment [62, 63]. These molecular alterations occur at different genomic levels and are recorded using different technologies, collectively referred to as 'omics' technologies. Each of these omic measurements offers only partial information regarding the compromised tissue. Aggregating different omic measurements, an analysis known as multi-omics integration, is therefore necessary to generate a comprehensive picture of the molecular features underlying a cancerous lesion [64, 65].

Owing to their high versatility, cell lines offer a cost-effective model system for drug response modelling [66]. Specifically, large scale consortia have industriously subjected a large number of cell lines to hundreds of different compounds, yielding valuable drug response measurements [9, 11, 67]. A key challenge resides in combining these response measurements with multi-omics data to study mechanisms of resistance and sensitivity[68]. Existing approaches focus on combining all omics data types and can be ordered based on the stage of the analysis at which the integration is performed[69]. At one extreme, early integration approaches [70, 71] first aggregate all features from all data types to process them all simultaneously. At the other extreme, late integration approaches first compute a representation of each data type individually, and subsequently combine these representations [56, 57, 72]. Several other methods can be positioned along this ordering, and differ by the analysis stage during which the grouping of data types is performed[73]. Although promising and encouraging, these methods do not take into account the quality of the data types and do not explicitly model their topology [54], i.e., how the data types relate to each other regarding information content and capacity to predict drug response. In particular, it has been observed that, although it has traditionally been the least clinically actionable data type, gene expression consistently prevails over other data types [55] and provides similar performance as early-integration approaches [46], obviating the need for complex integration strategies.

In order to maintain the predictive power of gene expression data, while exploiting the robustness of the most actionable data types, we present Percolate, an unsupervised multi-omics integration framework. Percolate sets itself apart from other integration approaches as it aims to eliminate gene expression measurements from the final predictor, rather than integrating it with all other data types. This is achieved by extracting the joint signal between gene expression and the other data types in an iterative fashion. First the joint signal between gene expression and Data type 1 (e.g. mutations) is extracted. Then the remaining signal (not shared with Data type 1) is employed to extract the joint signal of gene expression with Data type 2 (e.g. copy number data). This procedure is repeated for all omics data types. In this way, the gene expression signal is "percolated" down the other omics data types, ideally extracting all predictive signal from the gene expression data. We first show that comparing gene expression to other data types individually recovers a known topology of multi-omics data. We then show that the information shared between individual omic data types and gene expression increases drug response predictive performance for the individual omic data types. Finally,

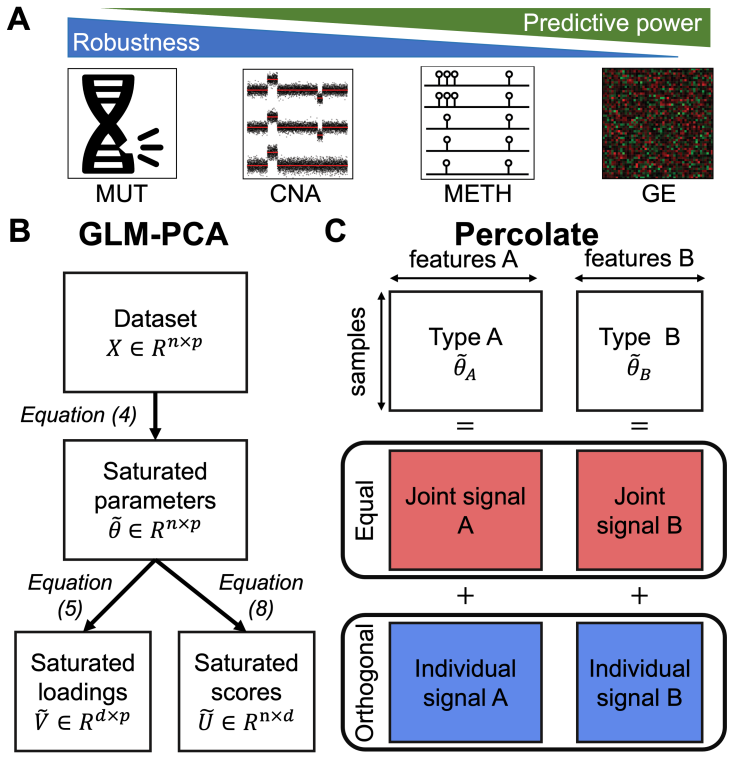


Figure 2.1 – **Dissecting multi-omics topology using Percolate bridges the gap between predictive and robust data types.** (**A**) Trade-off between robust data types (MUT, CNA) and predictive types (METH, GE). (**B**) Workflow of our implementation of GLM-PCA, which relies on the projection of saturated parameters. (**C**) Workflow of Percolate, which extends JIVE to non-Gaussian settings by comparing the low-rank structures of saturated parameter matrices.

reconstructing the joint signal solely from mutation, copy-number and methylation, we show that the signatures derived from "percolating" gene expression down these data types recapitulate the drug response predictive performance of these data types. Here we make the following technical contributions:

- We employ a popular framework, called JIVE [56, 57], which we extend to non-Gaussian noise models.
- We exploit an extension of PCA, called GLM-PCA, for which we present a versatile implementation.
- We develop an out-of-sample extension for JIVE, and specifically for the case when only one of the two data types is available.

2.2. METHODS

2.2.1. TRADE-OFF BETWEEN ROBUST AND PREDICTIVE TYPES

We consider four data types: mutations (MUT), copy number aberrations (CNA), methylation (METH) and gene expression (GE). MUT and CNA directly measure genetic aberrations and therefore rely on DNA measurements. Due to several biological and technological factors, these measurements are highly robust and suffer from limited technical artefacts. On the other end of the spectrum, GE measures RNA abundance, a process known for exhibiting large biological variability and prone to technical artefacts. Between these two extremes, methylation offers an intermediate level of robustness. However, when it comes to drug response prediction, the order is reversed: GE offers, on average, a better predictive performance than METH, and significantly outperforms MUT and CNA [45, 46, 66]. This leads to a trade-off between robustness and predictive ability (Figure 2.1A) with MUT and CNA being the most robust and least predictive and GE being the most predictive and least robust, with METH rating at the intermediate level in terms of robustness and predictive capacity.

2.2.2. EXPONENTIAL FAMILY DISTRIBUTION

Our integrated approach is inspired by AJIVE [57], a computational approach which takes as input two paired datasets and computes a joint and a data-specific signals. AJIVE is an extension of the JIVE model [56], which we selected, among other extensions [74, 75], for its computational tractability and its mathematical formulation which is amenable to the derivation we propose. JIVE, AJIVE, and derivations thereof, critically rely on Principal Component Analysis (PCA) which assumes a Gaussian noise model on the data [76, 77]. To extend this framework to non-Gaussian settings, we make use of a generalized formulation that can deal with a wider class of parametric distribution models, i.e., the so-called exponential families [78].

Let $\mathcal{X} \subset \mathbb{R}^p$, we say that a random vector $Z \in \mathcal{X}$ follows an **exponential family distribution** if its probability density function f can be written as

$$\forall z \in \mathcal{X}, \quad f(z|\theta) = h(z) \exp\left(\eta(\theta)^T T(z) - A(\theta)\right), \tag{2.1}$$

where $T : \mathcal{X} \to \mathbb{R}^q$ ($q > 0$) is called the **sufficient statistics**, $\theta \in \mathbb{R}^q$ the **exponential parameter**, $\eta : \mathbb{R}^q \to \in \mathbb{R}^q$ the **natural parametrization**, $A : \mathbb{R}^q \to \mathbb{R}$ the **log-partition function** and $h : \mathcal{X} \to \mathbb{R}^+$ the base measure. The exponential family encompasses a broad set of distributions (Table A.1), including the Gaussian distribution with unit variance, the Poisson, the Bernoulli, the Beta or the Gamma distributions. Practically, the functions A, T and η are modelling choices which can be tuned for any specific application.

2.2.3. SATURATED MODEL PARAMETERS

We consider a dataset $X \in \mathbb{R}^{n \times p}$ with n samples and p features which we model using an exponential family defined by $\mathcal{E} = (T, A, \eta)$ (Subsection 2.2.2). This choice of a distribution is usually motivated by prior knowledge on the data, e.g. Bernoulli for binary data. If we assume that each element in the data matrix X follows this exponential family distribution, we define the **saturated parameters** of X given this exponential family, $\widetilde{\theta}(X; \mathcal{E})$, as the matrix of $n \times p$ natural parameters which minimise the following negative

Data type	Family distribution
Copy-number aberration (CNA)	Log-Normal or Gamma
Gene expression (GE)	Negative-Binomial
Methylation (METH)	Beta
Mutation (MUT)	Bernoulli

Table 2.1 – **Exponential family distributions**. Gaussian distribution is assumed to have unit variance. The dispersion parameter r is fixed for the Negative Binomial.

log-likelihood:

$$\begin{aligned} \widetilde{\theta}(X;\mathscr{E}) &= \underset{\theta\in\mathbb{R}^{n\times q}}{\operatorname{argmin}} \quad \mathscr{L}(\theta;X,\mathscr{E}) \\ \text{where} \quad \mathscr{L}(\theta;X,\mathscr{E}) &\widehat{=} \sum_{i=1}^{n}\sum_{j=1}^{p} A\left(\theta_{i,j}\right) - \eta\left(\theta_{i,j}\right)^{T} T\left(X_{i,j}\right). \end{aligned} \tag{2.2}$$

The saturated parameters correspond to single-sample maximum likelihood estimates. In the particular case where A is differentiable with an invertible differential, assumption which holds for the usual distributions, let's define g as $g^{-1} = \left(\frac{dA}{d\eta}\right)^{-1} \circ T$. Equation (2.2) admits the following solution

$$\widetilde{\theta}(X;\mathscr{E}) = g^{-1}(X). \tag{2.3}$$

Note that g is defined using the derivative of A with regards to η, not θ. Equation 2.3 shows that the saturated parameters correspond to a dual representation of the data motivated by prior knowledge on the data-distribution. We will exploit this representation *à la PCA* to find the main sources of variations.

2.2.4. GENERALIZED LINEAR MODEL PRINCIPAL COMPONENT ANALYSIS (GLM-PCA)

Principal Component Analysis (PCA) admits three equivalent definitions: maximisation of projected variance, minimisation of reconstruction error and maximisation of a Gaussian likelihood with unit-variance. In **GLM-PCA**, also referred to as **Exponential PCA** in the literature, this Gaussian likelihood is replaced by an exponential family distribution, as was first proposed by *Collins et al* [79]. The approach from *Collins et al* computes the saturated parameters using an SVD-like decomposition, which yields three different matrices. Refinements of this idea which solves a similar optimisation problem, have been proposed in the literature [80, 81] and offer competitive routines for the computation of these three matrices. Another take on this problem, which relies on the projection of saturated parameters, has recently been developed by *Landgraf et al* [82]. This approach offers the advantage of a simpler single-matrix optimisation instead of concomitantly optimising on three. Furthermore, the out-of-sample extension relies on a matrix multiplication and is thus computationally fast. We therefore turned to this implementation. Formally, given an exponential family characterized by $\mathscr{E} = (T, A, \nu)$ and a number of

principal components d, GLM-PCA solves the following optimisation problem:

$$\widetilde{V}(X;\mathscr{E},d),\ \widetilde{\mu}(X;\mathscr{E},d) \quad = \quad \underset{\substack{V\in\mathbb{R}^{d\times p},\mu\in\mathbb{R}^p \\ VV^T=I_d}}{\operatorname{argmin}} \quad \mathscr{L}\left(\left(\widetilde{\theta}(X;\mathscr{E})-1_n\mu^T\right)V^TV+1_n\mu^T\ ;\ X,\mathscr{E}\right), \tag{2.4}$$

with $\mathscr{L}$ defined in Equation (2.2) and 1_n a vector of size n made of ones. The optimisation problem presented in Equation (2.4) corresponds to the minimisation of the negative log-likelihood after projection of the centered saturated parameters on a d-dimensional subspace. The solution $\widetilde{V}$ is the orthogonal matrix which best reconstructs the data-likelihood. We refer the reader to Chapter **??** for further details on the distributions we used.

The solution of Equation (2.4) is an optimisation problem with a Stiefel-manifold constraint, which we solved by using recent advances in auto-differentiation [83] and optimisation on Riemmannian manifolds [84]. We modelled the functions A, T and the negative log-likelihood using PyTorch; stochastic gradient descent (SGD) on the Stiefeld-manifold was performed using McTorch. Such a formulation allows to employ a large variety of exponential family distribution without the need for heavy and potentially cumbersome Lagrangian computations. Our optimisation scheme relies on four hyper-parameters: number of factors (or principal components), learning rate, number of epochs and batch size. To determine them, we compute the Akaike Information Criterion (AIC) of the complete data for various values of d and different hyper-parameters [85]. For a GLM-PCA model with d PCs, the AIC corresponds to the sum of the data log-likelihood and the number of model parameters, which we estimate as the dimensionality of the Stiefel manifold $\left\{V\in\mathbb{R}^{d\times p}|VV^T=I_d\right\}$, equal to $pd-d(d+1)/2$. Among all trained models, we select the one which harbors the smallest AIC.

2.2.5. COMPARISON OF GLM-PCA DIRECTIONS BY PERCOLATE

We consider two datasets $X_A\in\mathbb{R}^{n\times p_A}$ and $X_B\in\mathbb{R}^{n\times p_B}$ with paired samples (rows) but potentially different features. We first perform GLM-PCA independently on X_A and X_B using two different exponential family distributions, yielding d_A and d_B factors, respectively denoted as $\widetilde{V}_A$ and $\widetilde{V}_B$. We furthermore denote by $\widetilde{\theta}_A$ and $\widetilde{\theta}_B$ the saturated parameters of datasets A and B respectively, and $\widetilde{\mu}_A$ and $\widetilde{\mu}_B$ the intercept terms. To compute the sample scores resulting from these two GLM-PCAs, we perform an SVD decomposition on $\left(\widetilde{\theta}_A-1_n\widetilde{\mu}_A^T\right)\widetilde{V}_A^T\widetilde{V}_A=\widetilde{U}_A\Sigma_AW_A^T$ and $\left(\widetilde{\theta}_B-1_n\widetilde{\mu}_B^T\right)\widetilde{V}_B^T\widetilde{V}_B=\widetilde{U}_B\Sigma_BW_B^T$. By construction, these SVDs are of respective ranks d_A and d_B.

To compare the two sets of samples scores, $\widetilde{U}_A$ and $\widetilde{U}_B$, we aggregate them in a matrix M which we decompose by SVD:

$$M \quad = \quad \left[\widetilde{U}_A,\widetilde{U}_B\right] \quad = \quad U_M\Sigma_MV_M^T \quad . \tag{2.5}$$

The top left-singular vectors correspond to sample scores which are highly correlated between $\widetilde{U}_A$ and $\widetilde{U}_B$, since these two matrices are consisting, by construction, of uncorrelated factors. Following the same intuition as in AJIVE, these can be understood as the **joint signal**: restricting to the top r_J components ($r_J<\min(d_A,d_B)$), we obtain the matrix $\widetilde{U}_J\in\mathbb{R}^{n\times r_J}$ corresponding to the joint view. We furthermore denote by $\Sigma_J\in\mathbb{R}^{r_J\times r_J}$

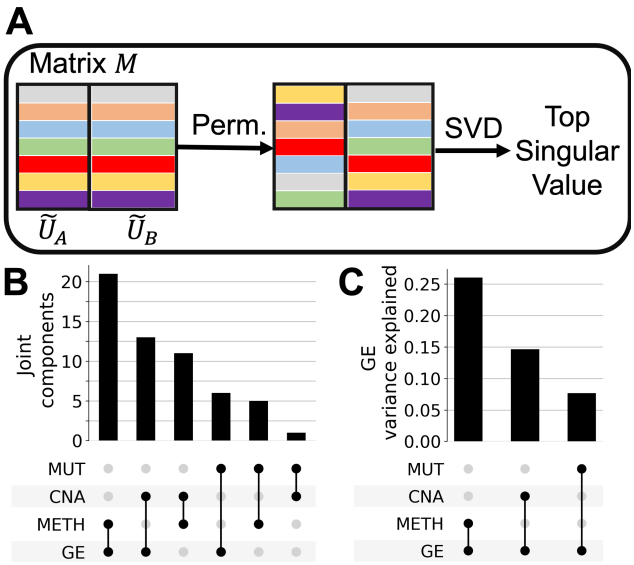


Figure 2.2 – **Assessing the number of joint components.** (**A**) Schematic of the sample-level permutations we perform to estimate the number of joint components. (**B**) Venn-diagram of the number of joint components obtained using the permutation scheme. (**C**) Ratio of variance explained for the GE saturated parameters matrix after projection on the joint components.

the diagonal square matrix containing the top r_J singular values of M. Finally, we define the **individual signal** of A, denoted as $\widetilde{U}_I^A$, as the signal from $\widetilde{U}_A$ not present in $\widetilde{U}_A$. We define $\widetilde{U}_I^B$ similarly:

$$\begin{aligned} \widetilde{U}_I^A &= \left(I_n - \widetilde{U}_J \widetilde{U}_J^T\right) \widetilde{U}_A \\ \widetilde{U}_I^B &= \left(I_n - \widetilde{U}_J \widetilde{U}_J^T\right) \widetilde{U}_B \end{aligned}. \tag{2.6}$$

The Percolate workflow is summarised in Figure 2.1B-C.
In order to set the number of joint components r_J, we employ a sample-level permutation scheme. We first independently permute the rows of $\widetilde{U}_A$ and $\widetilde{U}_B$, which we then aggregate as in Equation (2.5) to obtain the singular values. We perform 100 such permutations independently and retrieve the first singular value for each. Finally, we set r_J as the number of elements in Σ_M over one standard deviation from the mean of the permuted singular values (Figure 2.2A).

2.2.6. PROJECTOR OF JOINT SIGNAL

AJIVE does not provide an out-of-sample extension, and we here propose a derivation thereof by rewriting the matrix U_J as a function of the saturated parameters. If we decompose the matrix V_M as $V_M = \left[V_{M,A}^T \, V_{M,B}^T\right]^T$ such that $V_{M,A}^T$ contains the first d_A

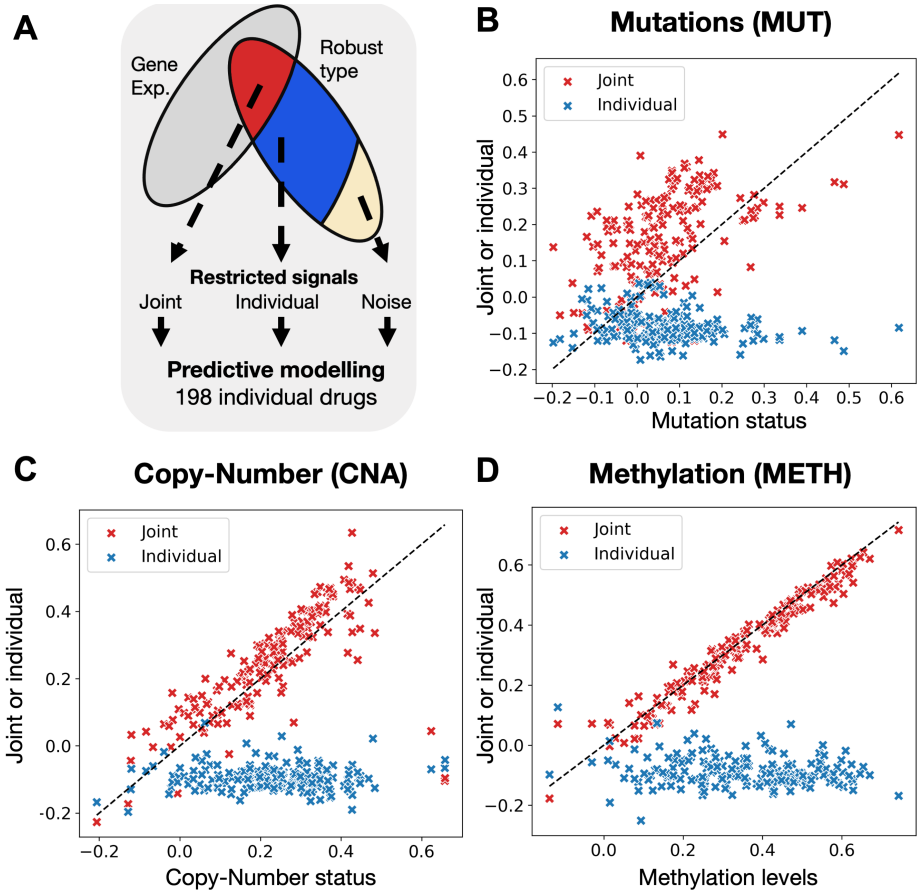


Figure 2.3 – **The joint signal between robust and gene expression contains most of the predictive signal.** (**A**) Workflow of our approach. (**B**) Predictive performance for MUT when using Percolate between MUT and GE. Each point corresponds to a single drug, with the x-axis corresponding to the predictive performance obtained using the original mutation data, and the y-axis by either the joint (red) or the individual (blue) signals. (**C**) Predictive performance for CNA, similarly displayed as in **B**. (**D**) Predictive performance for METH, similarly displayed as in **B**.

columns of V_M^T and $V_{M,B}^T$ the last d_B ones, we obtain (Section A.4):

$$\begin{aligned} \widetilde{U}_J &= \widetilde{U}_{J,A} + \widetilde{U}_{J,B} \\ \text{with } &\begin{cases} \widetilde{U}_{J,A} = (\widetilde{\theta}_A - 1_n \widetilde{\mu}_A^T)\, \widetilde{V}_A^T \widetilde{V}_A W_A \Sigma_A^{-1} V_{M,A} \Sigma_J^{-1}. \\ \widetilde{U}_{J,B} = (\widetilde{\theta}_B - 1_n \widetilde{\mu}_B^T)\, \widetilde{V}_B^T \widetilde{V}_B W_B \Sigma_B^{-1} V_{M,B} \Sigma_J^{-1} \end{cases} \end{aligned} \tag{2.7}$$

The formulation of $\widetilde{U}_J$ presented in Equation (2.7) highlights the additive contribution of both dataset to the joint signal. At test time, both views are therefore required to estimate the joint signal. To tackle the issue of missing data-view, we propose a nearest-neighbor imputation of the unknown joint-term. Let's consider, without loss of generality, that only the view A is available. The joint signal has been computed using the two data matrices X_A and X_B, yielding $\widetilde{U}_{J,A}$ and $\widetilde{U}_{J,B}$. The second term contains r_J terms, and we train r_J corresponding k-Nearest-Neighbors (kNN) regressors. The test dataset $Y_A \in \mathbb{R}^{m \times p_A}$ can be projected on the joint signal by replacing the saturated parameter $\widetilde{\theta_A}$ in Equation 2.7 with the saturated parameter of the test data. We then estimate the second term by means of the r_J kNN regression models. Adding these two terms yields an estimate of the joint signal.

2.2.7. DRUG RESPONSE PREDICTION

We assess the predictive performance of a dataset by employing ElasticNet [86], which has been shown, inspite of its relative simplicity, to outperform more complex non-linear models when it comes to drug response prediction [45, 66, 87]. For a given dataset, we perform nested cross-validation as follows. First, datasets are stratified into 10 groups of equal size. For each group (10%), we employ a 3-fold cross-validation grid search on the remaining 90% to determine the optimal ElasticNet hyper-parameters (ℓ_1-ratio and penalization). We then fit this optimal ElasticNet model on the 90% to predict the class labels on the 10%. Repeating this procedure, we obtain one cross-validated estimate per sample and we define the **predictive performance** as the Pearson correlation between these estimates and the actual values.

2.2.8. DATA DOWNLOAD, MODELLING AND PROCESSING

We consider four data types in our analysis (Table 2.1) which we modelled using different exponential family distributions (Sections A.1, A.2 and A.3). The GDSC data was accessed on January 2020 from CellModel Passport [9]. For GE, MUT and CNA, we restricted to protein coding genes known to be frequently mutated in cancer, referred to as the **mini-cancer genome** [88]. GE was corrected for library size using TMM normalization [89] and mutations were restricted to non-silent.

2.3. RESULTS

2.3.1. THE BREAKDOWN OF THE JOINT SIGNALS HIGHLIGHTS THE TOPOLOGY OF MULTI-OMICS DATA

To compare data types, we employ Percolate using the distributions defined in Table 2.1, and a number of PCs set using the procedure presented in Subsection 2.2.4 (Figure A.2).

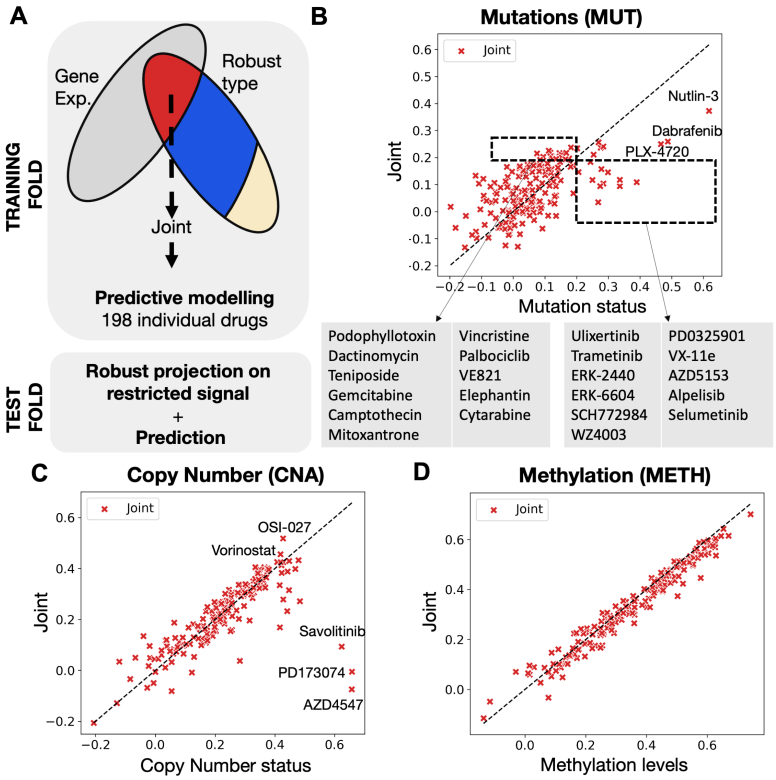


Figure 2.4 – **Robust-type-based signatures created from Percolate recapitulate drug response.** (**A**) Schematic of the cross validation experiment. (**B**) Results for MUT with a special zoom on drugs predictive for joint but not robust (left) and for robust but not join (right). (**C**) Results for CNA. (**D**) Results for METH.

For each comparison, setting the number of joint components is a crucial step, as it defines the threshold between the joint and individual signals. For that purpose, we used a sample level permutation test (Figure 2.2A, Subsection 2.2.5).
We observe that GE shares 21 joint components with METH, 13 with CNA and only 6 with MUT, which is coherent with the gradient put forward in Figure 2.1. We furthermore observe that MUT is consistently the data type with the least number of joint components (Figure 2.2B), highlighting the weakness of the signal coming from MUT data, corroborating previous measured topologies of multi-omics data [54]. To measure the strength of the underlying joint signals, we computed the proportion of GE variance explained by the joint directions (Figure 2.2C), computed as the ratio between the joint signal variance and the variance of the GE's saturated parameters matrix. We observe that the joint signal between GE and METH explains 26% of GE variance, while this figures drops to 14% and 7% for CNA and MUT, respectively. These observations highlight the existence of a joint signal, of which the predictive performance can be interrogated.

2.3.2. ROBUST SIGNAL PREDICTIVE OF DRUG RESPONSE IS CONCENTRATED IN THE JOINT PART

We then investigated the relevance of the joint and individual signals when it comes to drug response prediction. Considering one robust data type at a time (MUT, CNA or METH), we first decomposed the original robust data type into a signal joint with GE and an individual signal specific to the robust data type. We then computed, for 195 drugs (Methods), the predictive performance for these two signals and compared it to the original robust robust data (Figure 2.3A, Subsection 2.2.7). To ensure a proper comparison between joint, individual and cell-view, the cross-validation was performed using the same folds for all datasets.
We first analyzed the results obtained between MUT and GE data (Figure 2.3B). We observe that for most drugs, the predictive performance of the joint signal exceeds the predictive performance of the original robust signal, except for a number of drugs of which the response is quite well predicted based on MUT only. This set includes the drugs Nutlin-3, Dabrafenib, and PLX-4720. In contrast, the individual signal shows no predictive performance (Pearson correlation below 0) for most drugs, indicating an absence of drug response related signal in the individual portion. We then turned to CNA where the choice of distribution was unclear, with, to the best of our knowledge, no clear precedent on how to model such data. Due to the observed behavior of CNA data, we opted for two possible distributions: Log-normal and Gamma distributions (Table A.1). We observe that the joint signal computed using a Gamma-distribution yields better performances than the log-normal model (Figure A.3A-B). When using a Gamma distribution, a conclusion similar to the MUT data can be reached with the majority of drugs predicted well with the joint signal except three drug, AZD4547, PD173074 and Savolitinib (Figure 2.3C). This advocates for using the Gamma distribution for analyzing CNA data and shows that the joint signal presents an increased performance while the individual signal is not predictive. Finally, we studied the drug response performance obtained after decomposing METH using GE (Figure 2.3D). We observe that the joint signal presents a similar predictive performance as the original methylation data. The individual signal is, again, not predictive. These results highlight the potential of restricting predictors to

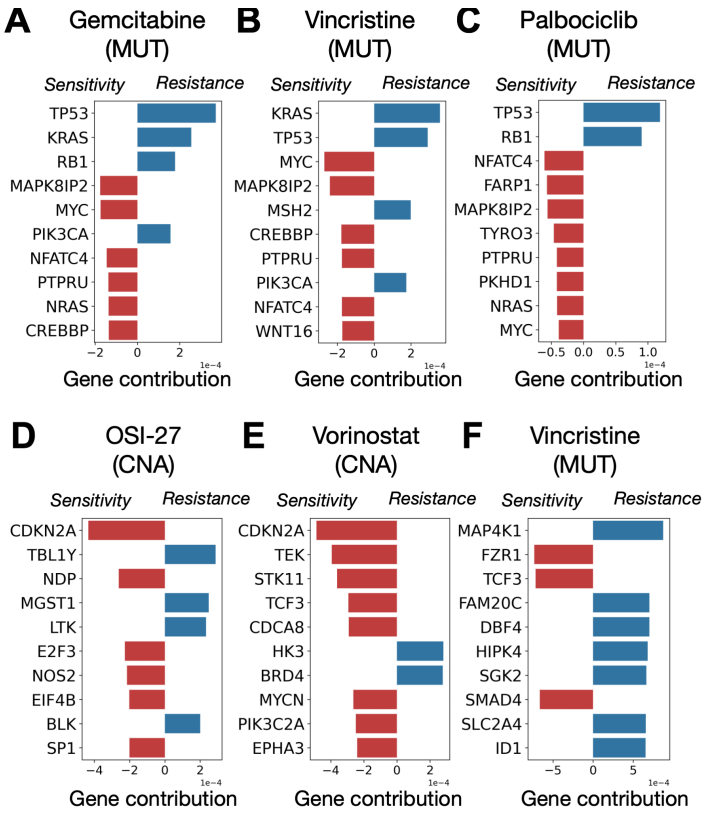


Figure 2.5 – **Study of joint signals contributing to improved performance.** For each drug, we report the top 10 largest gene regression coefficients from the joint signal, in absolute values. We first analysed the joint biomarkers created from MUT data for Gemcitabine (**A**), Vincristine (**B**) and Palbociclib (**C**). We then turned to CNA-based signatures for OSI-27 (**D**), Vorinostat (**E**) and Vincristine (**F**).

the joint signal for robust data types.

2.3.3. OUT-OF-SAMPLE EXTENSION RECAPITULATES THE PREDICTIVE PERFORMANCE OF ROBUST SIGNAL

In order to compute the joint signal between one robust data type and GE, one needs to have access to both modalities. However, the purpose is to become independent of non-robust GE measurements. In order to study whether the joint signal could be estimated without access to gene expression, when the predictor is applied to a test case, we exploited our out-of-sample extension (Subsection 2.2.6). We employed this algorithm to compute the drug response predictive performance of the joint signal estimated using the robust data alone (Figure 2.4A). Dividing the data in ten independent folds, we performed a cross-validation estimation as follows. For each train-test division of the data, we trained a Percolate instance on the 90% of the data, the training set containing GE and the robust data type. The resulting joint information was then used to train an ElasticNet model to predict drug response. The remaining 10% (test data) were then used to first estimate the joint signal, solely based on the robust data (Subsection 2.2.6). This joint signal was then used as input into the ElasticNet model to predict the response on this test set. Finally, we computed the predictive performance as indicated in Subsection 2.2.7.

When analyzing results for MUT (Figure 2.4B), we first observe a clear drop in performance for the joint signal compared to the previous results (Figure 2.3B). This suggests that the GE portion of the joint signal (Equation 2.7) contains a significant portion of predictive signal, which is less well captured by our out-of-sample extension. Nonetheless, we observe that 11 drugs show a predictive performance above 0.2 for joint but not for the robust data. In contrast, 11 drugs show the opposite effect, including seven which target the MAPK pathway – MEK (Trametinib, PD0325901, Selumetinib) and ERK (ERK2440, ERK6604, Ulixertinib, SCH772984). BRAF inhibitors Dabrafenib and PLX-4720 also show a drop in performance. This suggests that constitutive activation of the MAPK pathway is not recapitulated by the joint signal. Nonetheless, the joint signal generated by Percolate helps increase performance for several poorly predictive drugs and is therefore of interest to study various response mechanisms. We then turned to CNA (Figure 2.4C) and observe a modest decrease in predictive performance compared to the performance on the original CNA profiles. Three drugs show a spectacular drop as the response can not be predicted by the joint signal – Savolitinib (cMET), PD173074 (FGFR) and AZD4547 (FGFR). In contrast, three drugs show improved performance for the joint signal – OSI-027 (mTOR), Navitoclax (HDAC) and Vincristine (tubulin). Finally, we repeated the experiment for METH (Figure 2.4D) and observe that predictive performances of the joint signal is remarkbly comparable to the predictive performance on the original METH data, with most drugs falling showing less than 2% relative performance difference (Figure A.4C). Taken together, these results show that the joint signal recapitulates the drug response performance abilities of DNA-based measurements.

2.3.4. STUDY OF GENES CONTRIBUTING TO THE JOINT SIGNALS

We then set out to study the underlying mechanisms associated with the predictors derived from the robust data types (Subsection 2.3.3) which also lead to improved perfor-

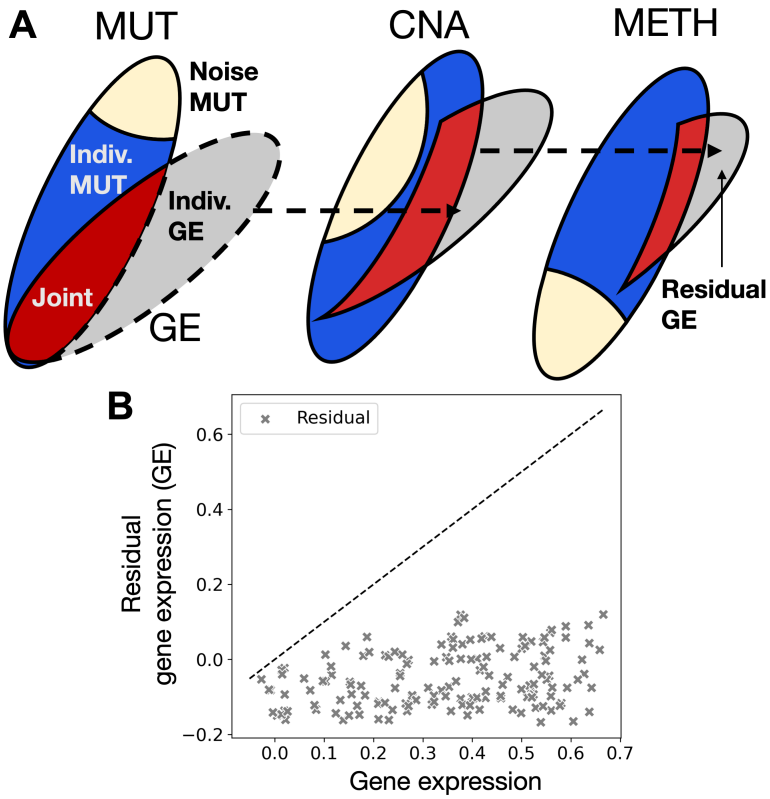


Figure 2.6 – **The signal joint with DNA-based measurements deprives gene expression from any predictive power.** (**A**) Schematic of our iterative procedure to remove from GE any signal joint with robust data type. (**B**) Predictive performance of the resulting residual gene expression compared to the predictive performance of the complete gene expression.

mance. For a given drug, we trained an ElasticNet model on the joint signal, yielding one regression coefficient per joint component. Using the relationship from Equation 2.7, we obtain a regression coefficient for each gene. A positive coefficient indicates that larger values of the saturated parameters, caused by a mutation or amplification of the supporting gene, are associated with resistance. In contrast, a negative coefficient indicates that larger values of the saturated parameters are associated with sensitivity.
For MUT, we studied the mode of action of three drugs for which the joint signal performs well (Figure 2.4B): Gemcitabine (Figure 2.5A), Vincristine (Figure 2.5B) and Palbociclib (Figure 2.5C). We observe that TP53 mutation status is associated with resistance to three drugs, concordant with earlier observations showing that TP53 mutant are more resistant to chemotherapy [90]. Resistance to Gemcitabine and Vincristine is also associated with KRAS and PI3KCA mutations, known for their proliferative potential [91, 92]. Interestingly, mutations in MYC and MAPK8IP2 are associated with sensitivity to these three drugs. Three other drugs show a drop in predictive performance on the joint signal as compared to the original signal: Nutlin-3, Dabrafenib and PLX-4720 (Figure 2.4B). We observe that the known targets of these drugs exhibit a large coefficient: TP53 for Nultin-3 (known resistance biomarker) and BRAF for Dabrafenib and PLX-4720 (Figure A.5). These three drugs highlight a limitation of our approach: GLM-PCA generates scores which aggregates the contributions of several genes. Highly-specific drugs, like Nutlin-3 (Mdm2-inhibitor) or BRAF/MEK-inhibitors not only target a specific protein, but mutations in the target are excellent response predictors. Such cases do not benefit from the GLM-PCA aggregation as a single feature alone is predictive.
Next we turned to CNA where three drugs: OSI-27 (Figure 2.5D), Vorinostat (Figure 2.5E) and Vincristine (Figure 2.5F), which all showed increased performance when the joint signal is employed as compared to the original CNA data. For both OSI-27 (mTORC1) and Vorinostat (HDAC), we observe that amplification of CDKN2A (p16) is associated with sensitivity. P16 acts as a tumor-suppressor by slowing down the early progression of the cell-cycle and its loss is here associated with resistance for these two drugs. Finally, Vincristine's predictor shows that MAP4K1's amplification as a predictor of resistance. Such result is coherent with what we observed for MUT (Figure 2.5B) where mutations on KRAS were associated with resistance.

2.3.5. ITERATIVE APPLICATION OF PERCOLATE DEPRIVES GENE EXPRESSION FROM PREDICTIVE POWER

Finally, we questioned whether some signal predictive of drug response is still present in gene expression. To this end, we studied the GE signal after it has been stripped of all the signal it shares with MUT, METH or CNA. To remove all signal associated with robust data types from GE, we used Percolate iteratively on GE, starting with the *least* predictive data type (MUT), followed by CNA and ending with the *most* predictive data type (METH) (Figure 2.6A). Specifically, we first "percolate" GE through MUT to obtain an individual GE signal (not shared with MUT), which is then percolated through CNA to obtain a second GE individual signal, which is then finally percolated through METH, resulting in the individual GE signal we denote as *residual gene expression*. We finally assessed the predictive performance of this residual gene expression and compared it to the predictive performance of the original GE (Figure 2.6B, Subsection 2.2.7). We observe

that no drug reaches a Pearson correlation above 0.16, indicative of a complete lack of predictive performance in the residual GE. This shows that removing the signal joint with DNA-based measurements deprives gene expression from any predictive ability.

2.4. DISCUSSION

Designing multi-omics predictors of drug response has highlighted the existence of a trade-off between robust and predictive data types. To study this trade-off, we developed Percolate, a method which decomposes a pair of data types into a joint and an individual signal. After showing that the strength of the joint signal recapitulates the known topology between data types, we showed that the joint signal contains more predictive power than any robust data type alone. Exploiting our out-of-sample extension, we showed that the joint signal, computed from robust data types alone, recapitulates most of the predictive performance of each original robust signal. Finally, we showed that the gene expression signal predictive of drug response is fully captured by robust data types through Percolate.

Technically, Percolate extends JIVE in two different ways. First, by using GLM-PCA instead of PCA, we tailor the dimensionality reduction step to the specific data under consideration. Second, we developed an out-of-sample extension which allows to estimate the joint signal, even in the absence of one data-modality. For our analysis, we made use of standard distributions from the exponential family: Negative Binomial, Gamma, Beta or Bernoulli. Our implementation of GLM-PCA is versatile and any exponential family distribution can be employed in our framework, provided it can be auto-differentiated by PyTorch. Employing more complex distribution, like the inverse-gamma for copy-number is a fruitful avenue to improve on our methodology.

3

PRECISE: A DOMAIN ADAPTATION APPROACH TO TRANSFER PREDICTORS OF DRUG RESPONSE FROM PRE-CLINICAL MODELS TO TUMORS

Soufiane MOURRAGUI, Marco LOOG, Mark A. VAN DE WIEL, Marcel J.T. REINDERS, Lodewyk F.A. WESSELS

Parts of this chapter have been published in Bioinformatics as proceedings of ISMB/ECCB 2019.

3.1. INTRODUCTION

Cancer is a heterogeneous disease that arises due to the accumulation of somatic genomic alterations. These alterations show high levels of variability between tumors resulting in heterogeneous responses to treatments. Precision medicine attempts to improve response rates by taking this heterogeneity into account and tailoring treatment to the specific molecular make-up of a given tumor. This requires the identification of biomarkers to identify the set of patients that will benefit from a given treatment while sparing those that will not benefit the unnecessary side-effects. However, as there are limited patient response data for a wide range of drugs, pre-clinical modes such as cell lines and patient-derived xenografts (PDXs) have been employed to generate large-scale data sets that enable the development of personalized treatment strategies based on the data-driven identification of biomarkers of response. More specifically, hundreds of pre-clinical models have not only been extensively molecularly characterized, but, more importantly, their response to hundreds of drugs have also been recorded. This has resulted in large public resources containing data derived from cell lines (GDSC1000, [9]) and PDX models (NIBR PDXE, [93]).

These pre-clinical resources can be employed to build predictors of drug response which are then transferred to the human setting, allowing stratification of patients for drugs the patients have not yet been exposed to. Geeleher *et al.* applied this approach by simply correcting for a batch effect between the cell line and tumor data sets and then directly transferring the cell line predictor to the human setting ([52, 53]). This already yielded some promising results: it recovered well-established biomarkers such as the association between Lapatinib sensitivity and ERBB2 amplifications. However, when directly transferring a predictor from the source domain (cell lines) to the target domain (human tumors) one assumes that the source and target data originate from the same distribution. While the differences between pre-clinical models and human tumors have been studied extensively ([15, 17, 94]), the most obvious differences include the absence of an immune system in both cell lines and PDXs and the absence of a tumor micro-environment and vasculature in cell lines. One can therefore not assume similarity between the source and target distributions.

Transfer Learning aims at addressing this issue (see ([43, 44]) for a general review). Transfer learning methods can be assigned to different categories depending on the availability of source and target labels and on the specific relation between these source and target datasets. Since we have a very small number of labeled tumor samples, but a wealth of labeled pre-clinical models, our approach falls into the category referred to as *transductive*[1]. Since the features (i.e. the genes) are the same in the source and target domains, our problem requires a *domain adaptation* strategy, sometimes also referred to as *homogeneous domain adaptation*.

As previously mentioned, the marginal distributions of pre-clinical models and tumors are expected to be different. However, we assume that drug response is, for a large part, determined by biological phenomena that are conserved between pre-clinical models and human tumors. Therefore there should exist a set of features (genes) for which the conditional distribution of drug response given these features is comparable

[1] While this terminology is not widely used in the community, we follow the categorization employed in ([43]).

across cell lines, PDXs and human tumors. Different methodologies have been proposed to find such a common space and these can be divided in two main categories ([95]). For approaches in the first category, called *data-centric*, a common subspace can be found directly from both the pre-clinical models and the tumors by aligning the marginal distributions. This can be done by, for instance, using the Maximum Mean Discrepancy either exactly or employing semi-definite-programming ([96, 97]), or by using approximations based on multiple-kernel learning ([98]) or empirical kernel maps ([99]). Approaches in the second category, called *subspace-centric*, perform the domain adaptation correction by first reducing the dimensionality, and then aligning the low-rank representations ([58, 59, 100]). In the first category, the marginal distributions are directly aligned, suggesting that the empirical distribution would sufficiently accurately reflect the real behavior of the source and target samples. If, for instance, the source dataset consists of a proportion of ER positive samples that is very different from the target dataset, such a direction would be discarded as it is too dissimilar between the source and target. Clearly, this would be undesirable, as it represents a very important variable in breast cancer. The second category does find the directions of important variations, and then compares these directions between source and target. Hence, these approaches do not directly compare the distribution and are less subject to the sample size issue and to sample selection bias. We chose to employ the latter approach.

We present PRECISE (**P**atient **R**esponse **E**stimation **C**orrected by **I**nterpolation of **S**ubspace **E**mbeddings), a methodology based on domain adaptation that trains a regression model on processes that human tumors share with pre-clinical models. Fig. 3.1 shows the general workflow of PRECISE. We first independently extract factors from the cell lines, PDXs and human tumors by means of linear dimensionality reduction. We then use a linear transformation that geometrically matches the factors from one of the pre-clinical models to the human tumor factors ([59]). Subsequently we extract the common factors (*principal vectors (PVs)*) defined as the directions that are the least influenced by the linear transformation (Fig. 3.1A). After selection of the most similar principal vectors, we compute new feature spaces (based on this selection) by interpolating between the source domain (cell line or PDX principal vectors) and the target domain (human tumor principal vectors). The feature spaces resulting from this interpolation allow a balance to be struck between the chosen model system and the tumors ([58, 100]).[2] From the set of interpolated spaces, the *consensus representation* is obtained by optimizing the match between the marginal distributions of the chosen pre-clinical model and the human tumor data projected on these interpolated features. These consensus features are finally used to train a regression model using data from the pre-clinical model of choice. We use this regression model to predict tumor drug response (Fig. 3.1B). As these features are shared between the pre-clinical models and the human tumors, the regression model is expected to generalize to human tumors. We finally use known biomarker-drug associations (from independent data sources, e.g. mutation status, copy number) as positive controls to validate the predictions of the model in human tumors (Fig. 3.1C).

This work contains the following novel contributions. First, we introduce a scalable and flexible methodology to find the common factors between pre-clinical models and

[2] Our method, although using the notion of canonical angle, differs markedly from Canonical Correlation Analysis. Indeed, in our case, samples are not paired and no cross-correlation can be computed.

human tumors. Second, we use this methodology to quantify the transcriptional commonality in biological processes between cell lines, PDXs and human tumors, and we show that these common factors are biologically relevant. Third, we show how these common factors can be used in regression pipelines to predict drug response in human tumors and that we recover well-known biomarker-drug associations. Finally, we derive an equivalent, faster and more interpretable way to compute the geodesic flow kernel, a widely used domain adaptation method in computer vision. Our approach builds up on the work of ([58, 100]) but extend these approaches by first removing irrelevant non-transferable information automatically, and second by finding consensus features within the interpolation scheme to counter the bias towards source features induced by Ridge regression.

3.2. MATERIAL AND METHODS

3.2.1. NOTES ON TRANSCRIPTOMICS DATA

We here present the datasets employed in this study. Further notes on pre-processing can be found in Subsection B.1.2.

THE CELL LINE DATASET

We used the GDSC1000 dataset to train predictors on the cell lines. GDSC1000 contains IC_{50}-values for a wide range of drugs. Amongst these drugs, we restricted ourselves to drugs that are either cytotoxic chemotherapies or targeted therapies and have shown an effect on at least one cancer type. This resulted in a set of 45 drugs employed in this study (Subsection B.1.1). The gene expression profiles of 1,031 cell lines are available in total, including 51 breast cancer and 40 skin melanoma lines. Gene expression data are available in the form of FPKM and read counts.

THE PDX DATASET

We used the Novartis PDXE dataset which includes the gene expression profiles of 399 PDXs, including 42 breast cancer PDXs and 32 skin melanoma PDXs. Transcriptomics data are available in the form of FPKM.

THE HUMAN TUMOR DATASET

We extracted gene expression profiles for human tumors from TCGA. Specifically, we employed gene expression profiles for 1,222 breast cancers and 472 skin melanoma cancers. Both FPKM and read counts are available. Mutation and copy number aberrations have been downloaded from the cBioPortal ([101]). Translocation data has been downloaded from TumorFusions ([102]).

3.2.2. THE COSINE SIMILARITY MATRIX

Transcriptomics data are high-dimensional with $p \sim 19.000$ features (genes) and since these genes are highly correlated, only some combinations of genes are informative. A simple – yet robust ([103]) – way to find these combinations is to use a linear dimensionality reduction method, like PCA, that breaks the data matrix down in d_f factors independently for the source (cell lines or PDXs) and target (human tumors) such that

$$\forall i \in \{s,t\}, \quad \mathbf{X}_i = \mathbf{S}_i \mathbf{P}_i \quad \text{with} \quad \mathbf{P}_i \mathbf{P}_i^T = \mathbf{I}_{d_f}, \tag{3.1}$$

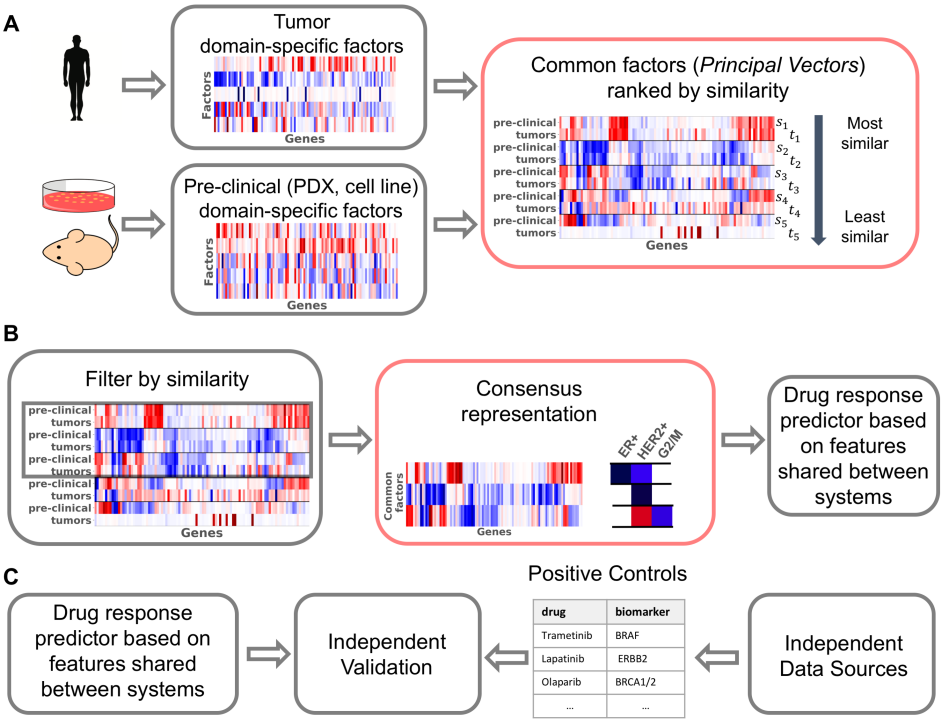


Figure 3.1 – **Overview of PRECISE and its validation**. **(A)** Human tumor and pre-clinical data are first processed independently to find the most important domain-specific factors (using, for instance PCA). These factors are then compared, aligned and ordered by similarity, yielding *principal vectors (PVs)*. The first PVs are pairs of vectors that are geometrically very similar and capture strong commonality between human tumors and pre-clinical models, the PVs at the bottom represent dissimilarities between human tumors and pre-clinical models. **(B)** A cut-off in similarity enables the retention of processes that are common. After interpolation between these most similar pre-clinical and tumor PVs, a *consensus representation* is computed by balancing the influence of human tumor and pre-clinical PVs. We performed a gene set enrichment analysis on these features to assess that they were clinically relevant. A tumor-aware regression model is finally trained by projecting pre-clinical and human tumor transcriptomics data on this consensus representation. **(C)** In order to validate our model, we use positive controls from independent data sources such as copy number or mutation data. These positive controls are established biomarker-drug associations. We compare the predictions of our model to predictions obtained based on these independent established biomarkers. Red boxes highlight our contributions.

where s and t refer to the source and target, respectively; $\mathbf{X}$ represents the (n x p) transcriptomics dataset where each row represents a sample and each column a gene; $\mathbf{I}_{d_f}$ is the identity matrix and $\mathbf{P} \in \mathbb{R}^{d_f \times p}$ contains the *factors* in the rows (i.e. the principal components). Since these factors are computed independently for the source and the target, we refer to them as *domain-specific factors*. Here we only consider PCA (Principal Components Analysis) since it is widely adopted by the community, and for its direct link to variance that acts as a first-order approximation in the comparison of distributions. Our method is, however, flexible and any linear dimensionality reduction method can be used.

Once domain-specific factors have been independently computed for both the source and target, a simple way to map the source factors to the target factors is to use the *subspace alignment* approach suggested by ([59]). This approach finds a linear combination ($\mathbf{M}^*$) of source factors that reconstructs the target factors as closely as possible:

$$\mathbf{M}^* = \underset{\mathbf{M} \in \mathbb{R}^{d_f \times d_f}}{\operatorname{argmin}} \left|\left|\mathbf{P}_s^T \mathbf{M} - \mathbf{P}_t^T\right|\right|_F = \mathbf{P}_s \mathbf{P}_t^T \,, \tag{3.2}$$

which is the least squares solution under orthogonality constraints from Equation (3.1). This optimal transformation consists of the inner product between the source and target factors and therefore quantifies the similarity between the factors. We will therefore refer to it as the *cosine similarity matrix*. It is also referred to as Bregman matrix divergence in the literature.

3.2.3. COMMON SIGNAL EXTRACTION BY TRANSFORMATION ANALYSIS

As we will show in Subsection 3.3.1, matrix $\mathbf{M}^*$ is far from diagonal, indicating that there is not a one-to-one correspondence between the source- and target-specific factors. Moreover, using $\mathbf{M}^*$ to map the source-projected data onto the target domain-specific factors would only remove source-specific variation, leaving target-specific factors and the associated variation untouched.

To understand this transformation further, we performed a SVD, i.e. $\mathbf{P}_s \mathbf{P}_t^T = \mathbf{U} \Gamma \mathbf{V}^T$, where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal of size d_f and Γ is a diagonal matrix. $\mathbf{U}$ and $\mathbf{V}$ define orthogonal transformations on the source and target domain-specific factors, respectively, and create a new basis for the source and target domain-specific factors:

$$s_k = \left(\mathbf{P}_s^T \mathbf{U}\right)_{.,k} \text{ and } t_k = \left(\mathbf{P}_t^T \mathbf{V}\right)_{.,k} \quad \text{for all } k \in \{1,..,d_f\}\,. \tag{3.3}$$

These define the *principal vectors* (PVs) ([104]) that have the following equivalent definition:

$$\begin{aligned} \forall k \in \{1,..d_f\}, \quad s_k, t_k = & \underset{s \in \text{span}(\mathbf{P}_s),\ t \in \text{span}(\mathbf{P}_t)}{\operatorname{argmax}} s^T t \\ \text{s.t} \quad \forall i < k, & \begin{cases} s_i \perp s \\ t_i \perp t \end{cases} \text{ and } s^T s = t^T t = 1 \end{aligned} \,. \tag{3.4}$$

$s_1,..,s_{d_f}$ define the same span as the source-specific factors – and so do $t_1,..,t_{d_f}$ with the target-specific factors. PVs thus retain the same information as the original domain-specific factors, but their cosine similarity matrix (Γ) is diagonal. The PVs $\{(s_1,t_1),..,(s_{d_f},t_{d_f})\}$ are derived from the source and target domain-specific factors and the pairs are sorted in

Algorithm 1 PRECISE

Require: source data $\mathbf{X}_s$, target data $\mathbf{X}_t$, number of *domain-specific factors* d_f, number of *principal vector* d_{pv}.
$\mathbf{P}_s \leftarrow d_f$ source *domain-specific factors* (e.g. Principal Components)
$\mathbf{P}_t \leftarrow d_f$ target *domain-specific factors* (e.g. Principal Components)
$\mathbf{U}, \mathbf{V}, \Gamma \leftarrow$ SVD of $\mathbf{P}_s \mathbf{P}_t^T = \mathbf{U}\Gamma\mathbf{V}^T$
$\mathbf{Q}_s \leftarrow \left(\mathbf{P}_s^T \mathbf{U}\right)^T$
$\mathbf{Q}_t \leftarrow \left(\mathbf{P}_t^T \mathbf{V}\right)^T$
$\Phi = \Phi_{\mathbf{Q}_s, \mathbf{Q}_t}$ as specified in Equation (3.7)
for $i \leftarrow 1$ to d_{pv} **do**
  $\mathbf{S}_i \leftarrow [\Phi_i(0), \Phi_i(0.01), .., \Phi_i(1)]^T$
  $\mathbf{X}_{s,i}^{proj} = \mathbf{X}_s \mathbf{S}_i$
  $\mathbf{X}_{t,i}^{proj} = \mathbf{X}_t \mathbf{S}_i$
  $\tau_i \leftarrow$ time of optimal matching between columns of $\mathbf{X}_{s,i}^{proj}$ and $\mathbf{X}_{t,i}^{proj}$
end for
$\mathbf{F} \leftarrow \left[\Phi_1(\tau_1), \Phi_2(\tau_2), .., \Phi_{d_{pv}}\left(\tau_{d_{pv}}\right)\right]^T$
$X_s^{proj} \leftarrow X_s \mathbf{F}$
$X_t^{proj} \leftarrow X_t \mathbf{F}$
Train a regression model on X_s^{proj}
Apply it on the projected target data X_t^{proj}

decreasing order based on their similarity. The top PVs are very similar between source and target while the bottom pairs are very dissimilar. For this reason we restricted the analysis to the top d_{pv} principal vectors. In (Equation 3.4), PVs have been defined as unitary vectors that maximise the inner product. The similarities therefore range between 0 and 1, and can thus be interpreted as the cosines of *principal angles* defined as

$$\forall k \in \{1, .. d_f\}, \quad \theta_k = \arccos\left(s_k^T t_k\right). \tag{3.5}$$

We define $\mathbf{Q}_s$ and $\mathbf{Q_t}$ as the matrix with the ordered principal vectors of the source and the target, respectively, with the factors in the rows.

3.2.4. FACTOR-LEVEL GENE SET ENRICHMENT ANALYSIS

In order to associate the principal vectors and the consensus representation with biological processes, we use Gene Set Enrichment Analysis ([105]). For each factor (i.e. a principal vector or a consensus factor), we projected the tumor data onto it, yielding one score per tumor sample. These sample scores where then used in the GSEA package as continuous phenotypes. We employed sample-level permutation to assess significance based on 1000 permutations. We used two curated gene sets from the MSigDB package: the canonical pathways (cp) and the chemical and genetic perturbations (cgp).

3.2.5. BUILDING A ROBUST REGRESSION MODEL

Given the common factors, we can create a drug response predictor based on these pairs of PVs. There are different ways to use these pairs of PVs. We could restrict ourselves to either the source or target PVs, but it would only support one of the two domains. Alternatively, we could use both source and target PVs. However, this would also be sub-optimal for the following reason. Source PVs are computed using the source data and maximize the explained variance of the source. Hence the source data projected on the source PVs is likely to have higher variance than the source data projected on the target PVs, since target PVs have not been optimized for the source data. If we were to apply penalized regression on the source data projected on both the source and target PVs, it would preferably select the source PVs. This would in turn lead to a loss of generalizability as source-specific information is weighed more heavily that target specific information.

One way to circumvent this issue is to construct a new feature space using 'intermediate' features based on interpolation between the spaces spanned by the the source and target PVs. For instance, in the plane that joins s_1 and t_1, the rotations from the former to the latter vector could contain a better representation. The intermediate features are expected to be domain invariant as they represent a trade-off between source and target domains and can thus be used in a regression model.

There is an infinite number of parameterizations for the intermediate features that join the source to the target PVs. As suggested in ([100], [58]), we consider the geodesic flow representing the shortest path on the Grassmannian manifold. We derive (see Subsection B.2.2 for the complete proof) a parameterization of the geodesic as a function of the PVs. Let's define Π and Ξ as

$$\forall \tau \in [0,1], \quad \begin{cases} \Pi(\tau) &= \operatorname{diag}\left(\frac{\sin((1-\tau)\theta_i)}{\sin(\theta_i)}\right)_i \\ \Xi(\tau) &= \operatorname{diag}\left(\frac{\sin(\tau\theta_i)}{\sin(\theta_i)}\right)_i \end{cases}, \tag{3.6}$$

where $\operatorname{diag}(\cdot)$ is the diagonal matrix. The intermediate representations are then defined by the geodesic path, that corresponds to a rotation for each pair of PVs:

$$\Phi : \tau \in [0,1] \mapsto \mathbf{Q}_s^T \Pi(\tau) + \mathbf{Q}_t^T \Xi(\tau) \ . \tag{3.7}$$

This geodesic path contains, for each pair of PVs, the features forming a rotating arc between the source and the target PVs. This formulation of the geodesic flow has the advantage of being based on the PVs, and not the domain-specific factors, in contrast to the formulation used in ([100]) and ([58]). Non-similar PVs can be removed prior to interpolation.

However, we show in Subsections B.2.3 and B.2.4 that projecting on all these features is equivalent, even in the infinite case, to projecting onto both the source and the target principal vectors, with the undesirable consequences described above. It is therefore preferable to create, for each pair of PVs, a single interpolated feature that strikes the right balance between the information contained in the source and target spaces. Consequently, a regression model trained on source data projected on the interpolated feature would generalize better on the target space. To construct these interpolated, or *consensus features*, we use the Kolmogorov-Smirnov (KS) statistic as a measure of sim-

ilarity between the source and target data, both projected on a candidate interpolated feature.

Specifically, by denoting d_{pv} as the number of selected PVs, for $i \in \{1,..,d_{pv}\}$ and $\tau \in [0,1]$ the feature at position τ between the i^{th} pair is defined as $\Phi_i(\tau) = (\Phi(\tau))_{.,i}$. For each pair of PVs, we then select the position τ_i that minimizes the KS statistic between the distributions of the source and target data projected on this feature. Let denote by D the KS statistic between the two projected datasets. We thus define τ_i as:

$$\tau_i = \min_{\tau \in [0,1]} D\Big(\mathbf{X}_s \Phi_i(\tau), \mathbf{X}_t \Phi_i(\tau)\Big). \tag{3.8}$$

This optimization is performed using a uniformly spaced grid search in interval $[0,1]$ with step size 0.01, moving between the source and the target.

This process is repeated for each of the top d_{pv} PVs, resulting in an optimal interpolation position for each. These positions are then plugged back into the geodesic curve to yield the domain-invariant feature representation $\mathbf{F}$ defined as:

$$\mathbf{F} = \Big[\Phi_1(\tau_1), \Phi_2(\tau_2), .., \Phi_{d_{pv}}\Big(\tau_{d_{pv}}\Big)\Big]^T. \tag{3.9}$$

The source data can now be projected on these features and the resulting data set can be used for training a regression model that can be more reliably transferred to the target (human tumor) data.

3.2.6. NOTES ON IMPLEMENTATION

Once the number of principal components and principal vectors have been set (see Subsection B.5 for an example), the only hyper-parameter that needs to be optimized is the shrinkage coefficient (λ) in the regression model. We employed a nested 10-fold cross validation for this purpose. Specifically, for each of the outer cross validation folds, we employed an inner 10-fold cross-validation on 90% of the data (the outer training fold) to estimate the optimal λ. To this end, in each of the 10 inner folds, we estimated the common subspace, projected the inner training and test fold on the subspace, trained a predictor on the projected inner training fold and determined the performance on the projected inner test fold as a function of λ. After completing these steps for all 10 inner folds, we determined the optimal λ across these results. Then we trained a model with the optimal λ on the outer training fold and applied the predictor to the remaining 10% of the data (outer test fold). We then employed the Pearson correlation between the predicted and actual values on the outer test folds as a metric of predictive performance. Note that every sample in an outer test folds is never employed to perform either domain adaptation nor in constructing the response predictor in that same fold.

The methodology presented in this section is available as a Python 3.7 package available on our GitHub page. The domain adaptation step has been fully coded by ourselves and the regression and cross-validation uses scikit-learn 0.19.2 ([106]).

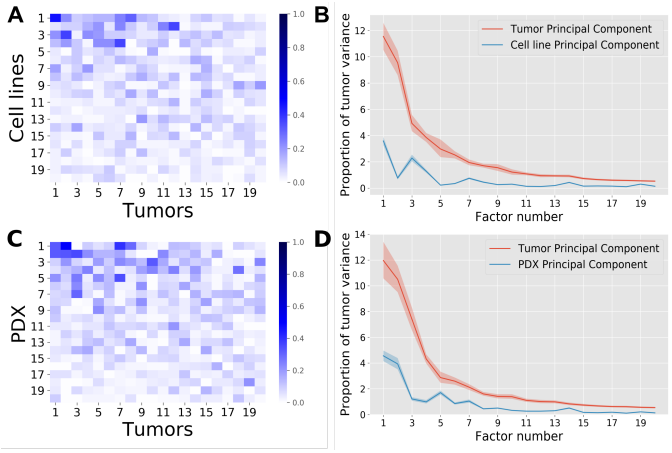


Figure 3.2 – Comparison of domain-specific factors between source and target with the source being cell lines **(A,B)** or PDXs **(C,D)**. **(A,C)** The absolute cosines similarity, i.e. the absolute values of the scalar products between source and target PVs. High similarities are found between some factors but no clear 1-1 correspondence is visible (absence of high values on the diagonal). **(B,D)** The ratio of tumor variance explained. Human tumor data were projected on the source and target PVs, the variance of the projected data on each direction was computed and divided by the total human tumor variance. The shaded regions represent the 98% confidence intervals obtained by bootstrapping the human tumor samples. Overall, the first five cell line factors each explain more than 1% of the human tumor variance each, while for PDXs this is achieved for the first seven factors. The non-monotonic behavior of the PDX and cell line principal component curves shows that the human tumor variance is not supported by the same directions than the pre-clinical models, which necessitates domain adaptation. PDXs show slightly higher similarity to human tumors than cell lines.

3.3. RESULTS

3.3.1. PRE-CLINICAL MODELS AND HUMAN TUMORS SHOW LIMITED SIMILARITY

The cosine similarity matrix $\mathbf{M}^*$ presented in Subsect. 3.2.2 gives an indication of the similarity between the source and the target principal components. A clear correspondence between factors would yield a diagonal matrix, allowing a single target principal component to be assigned to a single source principal component. Using data presented in Subsect. 3.2.1, Fig. 3.2A and Fig. 3.2C instead, show that this is clearly not the case as each source principal component shows similarity to a number of target principal components. Roughly speaking, for cell lines (Fig. 3.2A), the top four source factors show high similarity with the top ten target factors. The similarity between PDXs and human tumor principal components is generally higher and holds for a larger set of factors (Fig. 3.2C). This is to be expected since PDXs are believed to show higher resemblance to human tumors than cell lines.

When tumor data is projected on the cell line principal components, the explained variance accounts for around 30% of the variance explained when mapping the data on the human tumor principal components (Fig. 3.2B). For PDXs, on the other hand, this amounts to 40% (Fig. 3.2D), again indicating that PDXs resemble tumors more closely than cell lines. The bootstrap confidence intervals obtained by bootstrapping the tumor samples show that the obtained variance proportions differ significantly (Fig. 3.2B and Fig. 3.2D). However, for both model systems, the explained variance is relatively small, indicating that the data for both model systems are not drawn from the same probability distribution as the human tumors, underscoring the need for a proper alignment of the datasets prior to transferring a predictor from the pre-clinical modes to the human tumors.

In order to show that some gene-level structure is shared between these systems, we permuted the order of the genes in the source data only. We then computed the cosine similarities and target explained variance as before (Subsect. B.3.2). Neither the cosine similarity values, nor the variance explained were as high as for the original unshuffled data, suggesting that model systems and tumor cells do share some feature-level structure. We also compared the target data to samples drawn uniformly from a gaussian with a random covariance matrix in order to study whether the similarity between source and target data is significant. As shown in (Subsect. B.3.3), this also yields values three to four orders of magnitude lower than observed. This all shows the existence of a shared signal between source and target.

3.3.2. PRINCIPAL VECTORS CAPTURE COMMON BIOLOGICAL PROCESSES

The source and target principal components are employed to compute the 'common factors' or PVs for both the source and target (see Subsect. 3.2.3). Fig. 3.3A shows that the source and target PVs exhibit a perfect one-to-one correspondence. This is not unexpected since, by construction, this cosine similarity matrix is the central diagonal matrix (Γ) in the SVD of the optimal transformation between source and target (Equation 3.2). The source and target PVs are directions that respectively support the source and the target variance and are ranked by their pairwise similarity. When 20 principal compo-

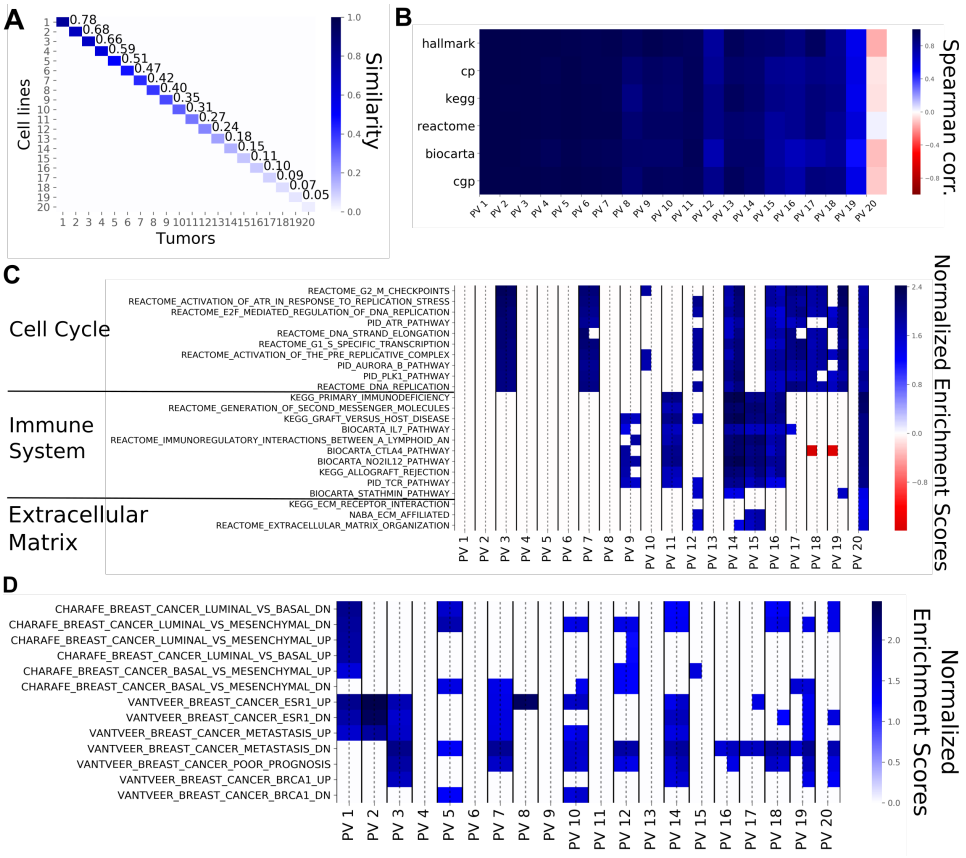


Figure 3.3 – Principal vectors (PVs) computed from breast cell lines and breast tumors from 20 principal components. **(A)** The Cosine Similarity matrix for cell line and tumor PVs. The values on the diagonal show the similarities within the corresponding pairs of PVs. Similarity starts at 78% and goes down to 2% for the last pair (not shown). The off-diagonal values are almost zero, showing that pairs of PVs of unequal rank are orthogonal to one another. **(B)** The Spearman correlations between the Normalized Enrichment Scores (NES) of source and target PVs for the different gene sets employed. The top principal vectors show similar enrichments while the bottom ones show little similarity, even negative correlation. This shows that top principal vectors represent the same biological phenomena. **(C)** The NES based on the Canonical Pathways for each PV pair with the NES for the source PV on the left and the NES for the target PV on the right (separated by a dashed line). Non-significant gene sets are represented as white cells. For this figure panel, we selected the ten gene sets that were most highly enriched in the first five PVs, the ten gene sets that showed the highest enrichment in the bottom PVs as well as all the gene sets related to extra-cellular matrix. The top PVs are exclusively enriched in pathways related to cell cycle. Immune system-related pathways are enriched in the middle and bottom PVs and PVs at the bottom tend to show enrichment for the target PVs only. **(D)** The NES for each PV as displayed in **(C)**, for the CHARAFE and VANTVEER gene sets. The top principal vectors are significantly enriched in sets associated with breast cancer subtypes.

nents are computed between breast cell lines and breast tumors, the similarity between the principal vectors ranges from 0.78 to 0.02, indicating that the low ranking pairs are almost orthogonal. Although 0.78 could seem like a low value for the top similarity, such a value remains significantly large for a 19,000-dimensional space. Fig. B.5A depicts the top 20 PVs obtained for breast PDXs and breast tumors and shows that the top similarity coefficients are higher than for cell lines, as expected.

To determine how PVs are related to genes, we calculated the contribution of each gene to the PVs. We subsequently employed these contributions in a gene set enrichment analysis (see Subsect. 3.2.4) to compute the association between the pathways in a given data base and the PVs. This resulted in a vector of pathway scores for every PV. We then computed the pathway similarity between a pair of PVs as the correlation of the pathway scores. Fig. 3.3B shows these correlations for the top 20 pairs of cell line and human tumor PVs for six different pathway databases. Correlations are close to one for the first principal vectors, indicating high pathway similarities, whereas the pathway similarity decreases (even becoming negative) for the lower ranked PVs. Taken together this shows that the principal vectors capture shared pathway information between the model systems and the human tumors.

When zooming in on the individual pathway similarities, we observe roughly two types of behaviors (Fig. 3.3C and Fig. 3.3D). First, some gene sets show significant enrichment for the top PVs as well as lower ranked PVs. These gene sets are related to breast cancer subtypes, cell cycle and DNA replication, i.e. gene sets one would expect to be enriched in both cell lines and human tumors. Second, some gene sets show enrichment for the more dissimilar PVs. Most of these gene sets are related to the response of the immune system and the extra-cellular matrix, entities which are not fully present in the cell lines. Fig. B.5C shows the results of the gene set enrichment analysis for breast PDXs and human tumors. We observe roughly the same behavior as for the cell lines and human tumors, especially regarding the gene sets enriched in the top PVs and the enrichment of immune related sets. However, we do observe that the extra-cellular matrix shows enrichment in higher ranked PVs (PV 5 and 7) which is in line with what one would expect in a PDX model. Taken together, this indicates that the PVs that are most similar between pre-clinical models and human tumors provide a mechanism to capture the information shared between model systems are tumors, while discarding processes that behave differently.

3.3.3. THE CONSENSUS REPRESENTATION YIELDS REDUCED BUT COMPETITIVE PERFORMANCE

Using PRECISE, we have derived a consensus feature representation which is both biologically informative and shared between the source and target. This representation can be used in a regression model trained on the source drug response data. Since ([45]), demonstrated that regularized linear models such as Ridge regression or ElasticNet ([86]) yield state-of-the-art performance for drug response prediction and since it is widely used, we will be employing Ridge regression.

We computed the predictive performance of PRECISE for 84 drug and tumor type combinations (Subsect. B.1.1). For a given drug and tumor type combination, we used PRECISE with all cell lines – i.e. the 1.001 cell lines across the 31 tissue types – as source

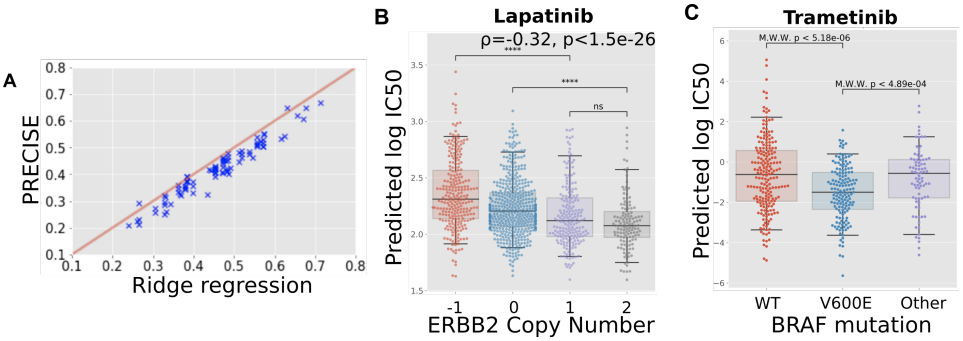


Figure 3.4 – Predictive performance assessment. **(A)** Scatterplot of the performance of PRECISE and Ridge regression. Predictive performance for each approach was computed as the Pearson correlation between the measured and the 10-fold cross-validated predicted IC_{50}s. PRECISE and Ridge regression performance is strongly correlated, with PRECISE showing a slight drop in performance. **(B)** Predicted drug response (predicted IC_{50}) for breast tumors based on a predictor of Lapatinib response trained on all the cell lines – i.e. across all tissue types – employing gene expression data only. ERBB2 copy number status in tumors correlates significantly with predicted IC_{50} values, validating the predictions, as tumors with ERBB2 amplifications are known to be more sensitive to Lapatinib. **(C)** Predicted drug response (predicted IC_{50}) for skin melanoma tumors based on a predictor of Tramatinib response trained on all the cell lines – i.e. across all tissue types – employing gene expression data only. The PRECISE predictions are validated by the established fact that $BRAF^{V600E}$-mutated tumors show a significantly higher sensitivity while the tumors bearing other mutations in BRAF do not show responses that differ from wild-type tumors.

and the corresponding tumor type as target. For example, to predict response to Vemurafenib in melanoma, we used all the cell lines from the GDSC panel, regardless of tumor type, as source, and melanoma tumors from the TCGA as target. We used 70 principal components since the variance explained shows a plateau after 70 principal components (Subsect. B.5.1). For the selection of the top PVs, we compared the obtained similarity values of the source and target PVs to the similarities obtained with random data and put the cutoff at the top 40 PVs (Subsect. B.5.2). We then computed the interpolated feature space employing the KS statistic and employed this feature space in the subsequent Ridge regression models.

As shown in Fig. 3.4A, PRECISE achieves predictive performances that are reduced but comparable with a Ridge regression model trained on the raw cell line gene expression data. The Pearson correlation between Ridge regression and PRECISE performance is $r = 0.97$ with a median relative reduction in Pearson correlation of 0.039. As it is the aim of the consensus representation to focus on the commonalities between the cell lines and the human tumors, it is to be expected that such a representation will not fully capture the variation in the cell lines as it is also adapted to the variation in the human tumors. Hence, a small drop in performance is not unexpected, as long as it results in improved predictions on the human tumor data, which we will demonstrate in the next section.

3.3.4. DOMAIN-INVARIANT REGRESSION MODELS RECOVER BIOMARKER-DRUG ASSOCIATIONS

To validate our predictor on human tumor data, we follow ([53]) by comparing the prediction in human tumors with the performance of independent, known biomarkers. Since we exclusively use gene expression to stratify patients in terms of their response to a certain therapy, we can use known biomarkers derived from other data sources – e.g. mutations or copy number changes – to create an independent response stratification of the human tumors against which we can compare our approach. In order to predict the IC_{50} in human tumor, we trained PRECISE using all the cell lines, irrespective of their tissue of origin, and the corresponding tumor type.

The first known and clinically employed association between a biomarker and response to a drug that we tested is the association between the presence of an ERBB2 amplification and response to Lapatinib in breast cancer. Since Lapatinib specifically targets the ERBB2 growth factor, breast tumors that overexpress this growth factor, and are therefore addicted to this signal, respond well to this therapy. An accurate stratification of the breast tumors by PRECISE would therefore show a negative association with the level of ERBB2 amplification in the tumors, as the tumors predicted to be most sensitive (lowest IC_{50}) would show the highest level of ERBB2 copy number amplification. This is exactly what we observe in Fig. 3.4B, where the predicted IC_{50} and the observed ERBB2 copy number amplification level show a Spearman correlation of $\rho = -0.32$. In addition, the copy number loss and copy number neutral categories show statistically significant differences in predicted IC_{50} compared to the samples harboring a copy number gain.

The second known association that we investigated is the association between the presence of a $BRAF^{V600E}$ mutation and response to MEK inhibitors in skin melanomas, while other BRAF mutations have not exhibited this association. As shown in Fig. 3.4C, the IC_{50}s predicted for Trametinib by PRECISE show a significant difference between tumors bearing a $BRAF^{V600E}$ mutation and tumors that are wild type for this gene. In addition, the PRECISE predictions also show a significant difference in predicted Trametinib response between tumors bearing a $BRAF^{V600E}$ mutation and tumors bearing other mutations in this gene.

Other known associations we tested against are shown in Fig. B.9: Dabrafenib sensitivity predicted by $BRAF^{V600E}$ mutations(Fig. B.9A), Vemurafenib sensitivity predicted by $BRAF^{V600E}$ (Fig. B.9B) mutations, Imatinib sensitivity predicted by BCR/ABL translocations in Acute Myeloid Leukemia (Fig. B.9C), Olaparib sensitivity predicted by BRCA1 deletion in breast cancer (Fig. B.9D) and Talazoparib sensitivity predicted by BRCA1 deletion in breast cancer (Fig. B.9E). These results have been computed using all the cell lines as source. When using solely cell lines from the tissue under consideration, the same results are observed, with a larger difference in predicted sensitivity for $BRAF^{V600E}$ mutated tumors for Trametinib (Fig. B.9G), Dabrafenib (Fig. B.9H) and Vemurafenib (Fig. B.9I), and a very clear separation of BCR/ABL translocated tumors from the rest under Imatinib treatment (Fig. B.9J).

Finally, we compared results from PRECISE with two baseline approaches. We replicated results from ([53]) by employing Ridge regression on either the raw or ComBat corrected gene expression data. Correcting for batch effect with ComBat provides a baseline comparison for our methodology. Lapatinib sensitivity is predicted as well as by PRE-

CISE (Fig. B.10A) with or without the ComBat pre-processing step. For Dabrafenib, the association with $BRAF^{V600E}$ is recovered, with and without ComBat. However, while the strength of the association obtained with ComBat is comparable to the association recovered with PRECISE, the strength diminishes without ComBat pre-processing (Fig. B.10C). In contrast to PRECISE, neither Ridge regression nor Ridge regression in combination with ComBat were able to recover the association between the $BRAF^{V600E}$ mutation and Trametinib.

In summary, PRECISE is capable of retrieving all tested associations between known biomarkers and drug response, while current state-of-the-art approaches fail to recover all these associations.

3.4. DISCUSSION

Using high throughput sequencing and screening technologies, scientists have leveraged the versatility of pre-clinical models over the past decade to create powerful predictors of drug response. However, due to the intrinsic differences between cell lines, PDXs and real human tumors, these predictors can not be expected to directly translate to the human setting. We have quantified the overlap in terms of the transcriptomics signal between these pre-clinical models and human tumors. We then introduced PRECISE, a domain adaptation framework that finds shared mechanisms between pre-clinical models and human tumors that display the same behavior across these systems.

PRECISE generates PVs, pairs of factors that capture the common variance between pre-clinical models and human tumors. The top pairs of PVs are most similar, and thus recapitulate molecular behavior shared between the systems. The least similar PVs, can be discarded since they correspond to mechanisms not shared across systems.

These vectors depend on the choice of linear dimensionality reduction methods employed. We employed PCA, but other methods have been proposed in the literature (e.g. [70, 107]), all having different qualities (e.g. being biologically meaningful, filtering out noise, etc.). The versatility of our method enables the use of any dimensionality reduction scheme, as long it finds an informative linear subspace.

Interpolation between the source and target principal vectors gives rise to features that balance the contribution of pre-clinical models and tumors. We showed (Sect. B.2) that interpolating between the principal vectors is equivalent to employing the Geodesic Flow Kernel approach that relies on the geodesics on the Grassmannian manifold ([58]) and has already yielded state-of-the-art performance in Computer Vision.

Recently, other ways to interpolate between the source and the target domains have been proposed, such as ([108]) that use a spline instead of the geodesic. We devised a simple, yet effective interpolation scheme between the source and target PVs where we employed the similarity based on the Kolmogorov-Smirnov statistic between the source and target data projected on the interpolated space to arrive at a consensus representation. This representation strikes the right balance between the pre-clinical models and the human tumors.

We subsequently projected the data on this consensus representation and trained a regression model which takes the distribution of the tumor gene expression data into account. This work considered Ridge and ElasticNet, but our approach is versatile and can be employed in combination with any classification or regression approach.

We showed that a Ridge regression model based on the consensus representation achieves slightly reduced performance compared to state-of-the-art approaches applied directly to the raw cell line gene expression data. This is to be expected, as the consensus representation filters out cell line specific information while capturing more relevant tumor variation, hence enabling efficient transfer to the tumor samples.

We finally compared our predictions to the performance of known biomarkers such as $BRAF^{V600E}$ in skin cancer or ERBB2 amplification in breast cancer and show that our method can reliably recover the associations between these biomarkers and their companion drugs. We show that response to Lapatinib can be predicted better when all cell lines are used for domain adaptation. On the other hand, using all cell lines reduced the power to predict response to Vemurafenib, although the resulting association with $BRAF^{V600E}$ mutation status remained significant. This might be due to the ubiquity of the ERBB2 amplification in several tumor types, in contrast to the $BRAF^{V600E}$ mutation that is specific to particular tumor types.

We restricted our study to domain adaptation based on transcriptomics data only, as it has been shown to be the most predictive data type ([45]). However, dissimilar behavior between pre-clinical models and human tumors might also be present in other molecular data types. A multi-omic drug response predictor should also correct for these differences, which will require a multi-omic domain adaptation approach which accommodates the unique data characteristics of each molecular data type.

Other methods have recently been proposed to tackle the problem of transferring pre-clinical predictors to human tumors. In ([109]), the authors create a correlation network for each omics data type and jointly map these networks onto a protein-protein interaction network. They then select the cliques that are conserved across omics-layers. In ([110]), the authors present an elegant framework for fold-change prediction in humans based on data from mouse models. Using fold-change data from both humans and mouse, a linear model is fitted at the gene-level. This linear model is then used to predict fold changes in human tumors for novel conditions. The problem of translating from model systems to human has broad applications and we envision that it will be a very active area of future research.

4

PREDICTING PATIENT RESPONSE WITH MODELS TRAINED ON CELL LINES AND PATIENT DERIVED XENOGRAFTS BY NON-LINEAR TRANSFER LEARNING.

4.1. INTRODUCTION

The accumulation of somatic alterations on the genome and epigenome transforms healthy cells into malignant tumor cells. Although these alterations are required for tumor growth, they also confer vulnerabilities on tumor cells. Some well-known examples of such genetic vulnerabilities are the amplification of ERBB2 in breast cancer [111], the $BRAF^{V600E}$ mutation in skin melanoma [112] or the BCR/ABL fusion in leukemia [113]. These vulnerabilities have been successfully exploited clinically by directing drugs against them. However, for the vast majority of cancer patients, no clear biomarkers exist. Hence, expanding our arsenal of accurate biomarkers would pave the way for personalized medicine, by identifying the most effective drug for each patient [114].

In order to discover such biomarkers, pre-clinical models have been used extensively in the past decades, either in the form of cell lines, patient-derived xenografts (PDX) or organoids. This was partially fueled by the relative ease with which these model systems can be subjected to drug screening. This has led to break-through discoveries with broad clinical impact [115]. However, Paul Valery's statement, "*what is simple is always wrong; what is not, is unusable*"" [116], also applies to these model systems. Specifically, their simplicity also confers weaknesses: the lack of a micro-environment in cell lines, and the

Parts of this chapter have been published in the Proceedings of the National Academy of Sciences (PNAS).

absence of an immune system in cell lines, PDXs and organoids. These shortcomings are further amplified by culture artefacts [15**?**] that lead to a reduced clinical significance of these models [94, 117] and a high attrition rate in drug development [118].
Computational approaches that correct for these differences are therefore needed to improve the identification of truly predictive biomarkers [119]. In the particular case of cancer, approaches that identify biomarkers are divided into two distinct categories. In the first category, mechanistic models are developed on pre-clinical models and subsequently "humanized" to focus on the similarities between pre-clinical models and human tumors [109]. The second category approaches the problem in a statistical fashion. Using molecular profiles and drug screens from large-scale panels of pre-clinical models [9, 45, 93], cell line drug response predictors are inferred [66, 120]. The resulting models are then applied to predict the sensitivity of patients to certain drugs. Although already promising, these approaches either do not take into account the fundamental differences between pre-clinical models and human tumors [121], or only model these differences as a technical batch effect [52, 53, 122]. Recently, transfer learning and multi-task learning approaches have been developed to explicitly take these differences into account, either partially using tumor responses during training [123, 124], or solely based on pre-clinical labels while employing linear approaches to correct for differences between pre-clinical models and human tumors (Chapter 3).
We present **TRANSACT** (**T**umor **R**esponse **A**ssessment by **N**on-linear **S**ubspace **A**lignment of **C**ell-lines and **T**umors), a versatile framework for subspace-based transfer learning [44, 58, 60, 97, 125] which enables the transfer of drug response predictors trained on a source domain (e.g. cell lines and PDXs) to a target domain (e.g. human tumors). TRANSACT employs the powerful and robust mathematical framework of Kernel methods [47, 48, 50, 68, 126, 127] to capture both linear and non-linear molecular processes expressed in both the source and target domains. In doing so, we obviate the need for cell-line pre-selection [128–130] and limit the loss of statistical power. While TRANSACT cannot compensate for inherent deficiencies in model systems, it identifies and exploits the space where model systems do represent human tumors accurately. First, we demonstrate that, compared to existing methods [53, 122, 131], modeling non-linearities using TRANSACT improves drug response prediction in PDXs after training on cell line responses only. We fix the hyperparameter controlling the degree of non-linearity on the PDX data and then employ TRANSACT to transfer predictors of drug response trained on cell lines to two human tumor datasets: primary tumors from TCGA and metastatic lesions from the Hartwig Medical Foundation (HMF). Specifically, the median performance of TRANSACT exceeds that of competing approaches in 7 of 13 challenges on TCGA and the HMF set. Importantly, this performance improvement is attained without any training on data from the human tumors. We finally employ the interpretability of our approach to identify genes and pathways associated with drug response. We provide a thorough mathematical derivation of our algorithm in which we propose a principled way to compare kernel principal components based on loadings by extending the framework of principal vectors to the non-linear kernel PCA setting. We generated a completely reproducible pipeline and a fully open-source software package.

4.2. RESULTS

4.2.1. TRANSACT: GENERATING NON-LINEAR MANIFOLD REPRESENTATIONS TO TRANSFER PREDICTORS OF RESPONSE FROM PRE-CLINICAL MODELS TO TUMORS

TRANSACT compares genomic signals contained in the source (e.g., pre-clinical models) and target (e.g., human tumors) datasets, and outputs a consensus space – a representation of processes that are present in both datasets. The nature of this representation depends on the similarity function, K, that characterizes the relationships between samples (Subsection 4.4.4). Depending on the similarity function employed, various types of non-linear relationships can be represented in the consensus space. For instance, in the case of a Gaussian similarity function, these non-linearities include constant, linear, second and higher-order interaction terms (Subsection 4.4.9).

In a first step, TRANSACT computes processes active in pre-clinical models and human tumors, referred to as *Non-Linear Principal Components* (NLPCs) (Figure 4.1A, Figure C.1A-B). These NLPCs correspond to non-linear combinations of gene activities that capture the variation in source and target sets (Figure C.2A). However, these two sets of processes typically display limited similarity, simply because pre-clinical models are not perfect models of human tumors (Figure C.1C). In order to capture the biological signal common to both pre-clinical models and tumors, we align the two sets of NLPCs using the notion of *Principal Vectors* (PVs) (Figure 4.1B).These PVs are pairs of non-linear processes ¬– one pre-clinical and one tumor process – ranked by decreasing similarity (Figure 4.1B). The top PVs correspond to highly similar processes, while bottom PVs are essentially different processes. We first filter out PVs with low similarity (below 0.5) in order to discard information specific to either pre-clinical models or tumors (Figure 4.1C). Since the remaining PVs represent pairs of highly correlated processes, we perform, within each PV pair, an interpolation between the pre-clinical and the tumor processes (Figure 4.1C). We then select one intermediate vector that best balances the contribution of each dataset (Figure 4.1C, Figure C.1E). These intermediate processes are called Consensus Features and correspond to biological processes that are 1) important in both pre-clinical and tumor signals, and 2) geometrically filtered to ensure that the signal is not specific to either one of the datasets. We then project pre-clinical and tumor samples on the Consensus Features (Figure 4.1D, Figure C.1F, Subsection 4.4.7). Finally, we use the projected scores as input in a predictive model of drug response trained using pre-clinical response data (Figure C.1G).

We theoretically show that, in the case of a linear similarity function, TRANSACT reduces to PRECISE [131] (Subsection C.9) and is fundamentally different from approaches such as Canonical Correlation Analysis (CCA)[132] (Subsection C.10).

4.2.2. NON-LINEARITIES IMPROVE RESPONSE PREDICTION OF PREDICTORS TRANSFERRED FROM CELL LINES TO PATIENT-DERIVED XENOGRAFTS (PDXS)

When it comes to predicting drug response in one model system, it is known that inducing non-linearities can lead to improved performance[50], although linear methods remain competitive [45, 87].We investigated whether the introduction of non-linearities in

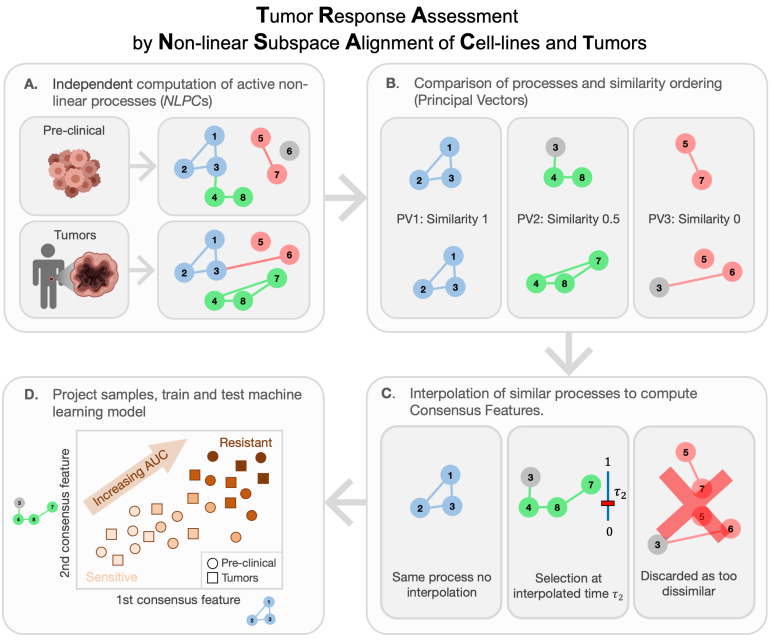


Figure 4.1 – **TRANSACT generates a non-linear representation to transfer predictors of drug response from pre-clinical models to tumors**. **A** In the first step, we use a non-linear dimensionality reduction method to find biological processes active in pre-clinical models and in tumors. This step is performed independently in pre-clinical models and tumors and gives two sets of non-linear processes called NLPCs. Here we consider 8 genes and 3 NLPCs for both pre-clinical models and tumors. A colored circle means that the corresponding gene contributes to the NLPC, while a grey circle is a gene that does not contribute. A colored connection is the interaction, or product, of two genes that contribute to the NLPC. For instance, the red pre-clinical NLPC represents the expression of Genes 5 and 7, and the product of the expression of Genes 5 and 7. **B** These two sets of processes are compared and ranked by similarity. For that purpose, we compute Principal Vectors (PV), which are pairs of processes, one from pre-clinical, one from tumor, ordered by decreasing similarity. Here the first PV is conserved between pre-clinical models and tumors, the second shows a 50% similarity while the last PV corresponds to two distinct processes. **C** We first discard the PVs with low similarity, e.g., PV 3. We then aggregate each PV pair into one Consensus Feature (CF) by finding an intermediate feature that balances the effect of the pre-clinical and tumor dataset. **D** Pre-clinical and tumor data are finally projected on each consensus feature, yielding a sample-by-CFs matrix. These Consensus Feature scores represent the activity of biological processes essentially important for both pre-clinical models and tumors. These scores can then be used in any machine learning model to predict drug response.

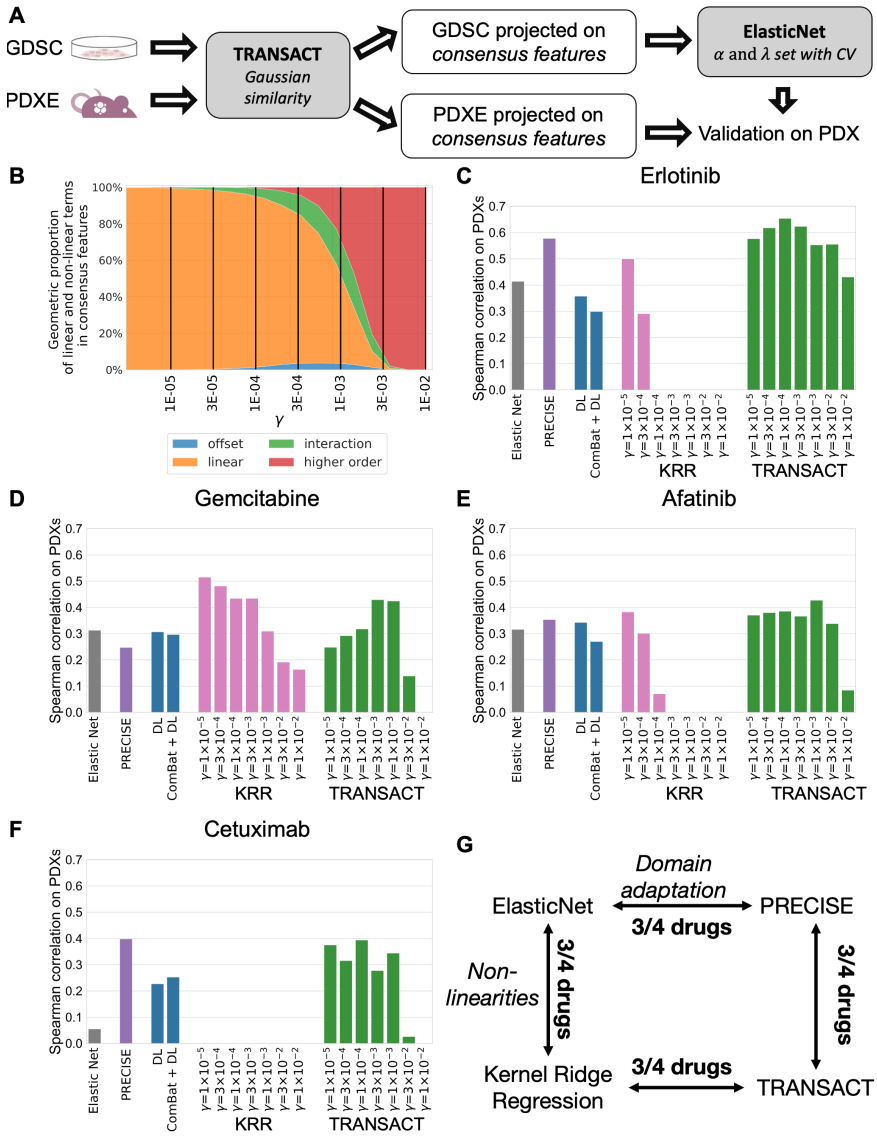


Figure 4.2 – **Impact of modeling non-linearities for drug response prediction transfer from cell lines to PDXs.**. **A** Main workflow of the prediction on PDXs. Using cell lines and PDX gene expression, we compute consensus features and project each dataset onto these. We then train a predictor of drug response (ElasticNet regression) on the projected scores of cell lines and the cell line response, measured as the Area Under the drug response Curve (AUC). Finally, we use this regression model to predict drug response on PDXs and correlate the predicted AUC to the known best average response. **B** Proportion of non-linearities induced by the Gaussian similarity function as a function of the scaling factor γ. For different values of γ, we compute the average contribution over all consensus features of offset, linear, interaction and higher order features (Subsection 4.4.9). Offset is here to be understood as the exponential of the squared depth and does not correspond to a constant term. We finally evaluate the response prediction on PDX models for different values of γ, and for five competing approaches: ElasticNet, Deep Learning (DL), ComBat+DL, PRECISE and a Kernel Ridge Regression (KRR) with same non-linearities as TRANSACT. We report results for Erlotinib (**C**), Gemcitabine (**D**), Afatinib (**E**) and Cetuximab (**F**) which show the Spearman correlation between predicted AUC and Best Average Response on the PDX models. (**G**) Diagram summarizing the effect of non-linearities and domain adaptation in our predictor.

the computation of sample similarities resulted in improved response prediction of predictors trained on cell lines (source domain) and transferred to PDXs (target domain). Since gene expression is known to have predictive power comparable to other omics datasets combined[45, 46, 54, 66], we restricted our analysis to the expression of 1780 genes known to be related to cancer [88]. Using TRANSACT, we computed consensus features for cell lines (1049 cell lines from 26 different tissues) and all PDXs (399 samples from 5+ different tissues) (Subsection 4.4.7). We projected the gene expression data of all cell lines and all PDXs onto these consensus features. We employed ElasticNet to train models of drug response. We employed the projected cell line expression data as input and the drug response, quantified as the area under the drug response curve (AUC), as target output (Subsection 4.4.3). We applied this trained predictor on the projected PDX expression data and compared the predicted response to the measured best average response by Spearman Correlation (Figure 4.2A). We made use of the standard Gaussian similarity function (Subsection 4.4.4) to vary the level of non-linearity introduced. This similarity function is characterized by a single scaling factor γ, whose size is directly proportional to the proportion of non-linearity introduced (Figure 4.2B). We studied the predictive performance in PDXs for seven different values of γ, ranging from a set of consensus features with an almost purely linear ($\gamma = 1 \times 10^{-5}$ to an almost purely non-linear composition ($\gamma = 1 \times 10^{-2}$). We compared the performance of *TRANSACT* to three approaches that do not perform domain adaptation: *ElasticNet* [86], a deep learning regression model (Subsection 4.4.10), referred to as *Deep Learning* (*DL*), and a Kernel Ridge Regression model (*KRR*) with same non-linear kernel settings as *TRANSACT*. We further compared it to two state-of-the-art domain adaptation approaches: ComBat batch correction followed by a deep learning regression (ComBat+DL)[122] and PRECISE[131], a linear domain adaptation approach. All models were trained to predict response to four different drugs (Erlotinib, Cetuximab, Gemcitabine, Afatinib) for which we had response data available for both PDX models and cell lines (Figure 4.2C-F). For *ComBat+DL* and *DL*, we report the median performance obtained over 50 independent random initializations (Subsection 4.4.10). Three other drugs were also studied: Paclitaxel, Ruxolitinib and Trametinib, however, these show no significant association between predicted and actual response in PDXs for any of the tested methods.

The studied methods can be divided along two axes: linear vs non-linear and domain-adapted vs non-adapted (Figure 4.2G) and we evaluated the performances along these axes for the four drugs. For *KRR* and *TRANSACT* we performed the comparisons for the values of gamma that gave the best performance. We observe that non-linear methods (*KRR* and *TRANSACT*) prevail over linear approaches (*ElasticNet* and *PRECISE*) for three of the four drugs in each separate comparison (Figure 4.2G). Furthermore, domain adapted approaches (*PRECISE* and *TRANSACT*) prevail over non-domain adapted approaches (*ElasticNet* and *KRR*) for three of the four drugs in each comparison (Figure 4.2G).

When considering Deep Learning-based approaches, we observe, in general, a clear improvement for approaches that employ domain-adaptation (*PRECISE* and *TRANSACT*) over those that either do not (*DL*), or use a naïve correction (*ComBat+DL*), confirming our earlier observation, namely the necessity to correct the input signal when moving from the source to the target domain. Moreover, this also suggests that the correction

required to transfer from cell lines to PDXs is more complicated than correcting for a technical batch effect as performed by ComBat. When comparing *TRANSACT* to *DL*, a non-linear and non-adapted method, we observe better performance for TRANSACT for all 4 drugs.
We note that for *KRR*, additional non-linearity tends to reduce performance. In contrast, the introduction of additional non-linearity in TRANSACT increases performance. Specifically, we observe for several drugs that the predictive performance increases with the scaling factor until a maximal performance is reached ($\gamma = 10^{-4}$ for Erlotinib, Cetuximab and Afatinib and $\gamma = 10^{-3}$ for Gemcitabine), after which the predictive performance drops. As we only have three drugs in common between the PDX and human cohorts, we decided to fix the scaling factor to the average of these two values ($\gamma^* = 5Ö10^{-4}$) and employ the associated consensus space to transfer the predictors of response to the tumor samples. For the drugs in common, we applied the predictors with drug-specific values of γ optimized on the PDX models to the TCGA and HMF cohorts. We only did so for the drugs where the drug specific value of γ differed from γ^*, i.e. not for Afatinib. For Gemcitabine, we observe a small increase in performance (0.01 in AUC) for TCGA and no difference for HMF, while for Cetuximab the prediction result still failed to reach significance. As a further check of the selected value of γ^*, we analyzed the properties of the consensus space obtained using γ^*. We observe a concentration of the offset contribution in the top consensus features and an increasing proportion of non-linear terms contribution to lower order features (Figure C.7C). The UMAP[133] projection of the consensus features shows a clear co-clustering of cell lines and PDXs of the same tissue (Figure C.7D).

4.2.3. CONSENSUS FEATURES BETWEEN CELL LINES (GDSC) AND HUMAN TUMORS CONSERVE PRIMARY TUMOR INFORMATION.

With the scaling factor (γ) calibrated on PDX models, we moved to the clinical setting to investigate domain adaptation between cell lines and two different human tumor datasets: primary tumors from TCGA and metastatic lesions from the HMF. We selected 30 consensus features in the GDSC-TCGA analysis (Figure C.9) and 20 in the GDSC-HMF analysis (Figure C.10). We arrived at these numbers by first selecting NLPCs based on the inflection point of the cumulative eigenvalues, and subsequently only retaining PVs with a similarity above 0.5. We observe that the consensus features computed between GDSC and TCGA (Figure 4.3A) and between GDSC and HMF (Figure 4.3B) show a concentration of offset and linearities in the top consensus features, and overall the same proportion of non-linearities as in the GDSC-to-PDXE analysis (Figure 4.3C).
In order to visualize the structure retained in the consensus space, we embedded our consensus scores into a 2D space using UMAP[133]. We observed that primary tumors cluster together based on their tissue type (Figure 4.3D). Most cell lines cluster with the tumors from a similar tissue of origin. However, some groups of cell lines cluster together but away from the tumors with the same tissue of origin, as observed in previous studies[128, 129]. For example, there is a cell line cluster consisting of PNS (Peripheral Nervous System) and bone cell lines that co-clusters with CNS (Central Nervous System) tumors. To quantify the degree of co-clustering of cell lines and tumors, we compared distances between tumors and cell lines from similar and non-similar tissues, and ob-

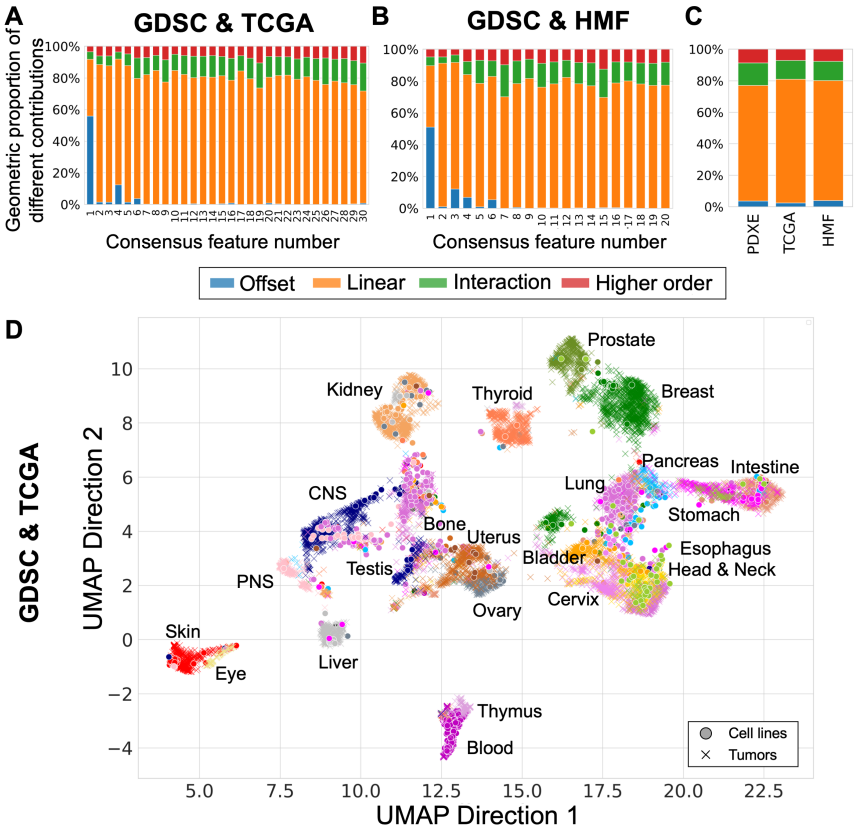


Figure 4.3 – **Pan-cancer consensus features between cell lines and tumors conserve tissue type information.** We used cell lines (GDSC) as source data and compute two sets of consensus features with two different target datasets: primary tumors (TCGA, **A** and **D**) and metastatic lesions (HMF, **B** and **E**). (**A**) Proportion of linear and non-linear contributions to each of the 30 GDSC-to-TCGA consensus features. (**B**) Proportion of linear and non-linear contributions to each of the 20 GDSC-to-HMF consensus features. (**C**) Comparison of global contributions in the three analyses, i.e. GDSC-to-PDXE, GDSC-to-TCGA and GDSC-to-HMF. (**D**) UMAP plot of primary tumors (TCGA, 21 tissues) and cell lines (GDSC, 22 tissues) projected on the consensus features, using the same parameters as selected in Figure 2. The full legend for Panels D and E is depicted in Figure C.11B.

served, as expected, a higher similarity between tumors and cell lines from the same tissue (Figure C.11D). Metastatic lesions show a weaker clustering based on the primary tumor's tissue of origin (Figure C.11A and E). This is not unexpected, as the expression profiles are derived from biopsy sites distant from the primary tissue. Of particular interest, we observe the existence of a hematopoietic cell-line cluster that co-clusters with metastatic samples from various biopsy sites. Most of these tumor samples (7 out of 12 samples) are lymph node metastases and most likely display a hematopoietic expression profile due to blood infiltration in the samples (Figure C.11C).

4.2.4. CONSENSUS FEATURES INCREASE TRANSFER OF RESPONSE PREDICTORS FROM CELL LINES TO PRIMARY TUMORS AND METASTATIC LESIONS

To further validate our approach, we transferred response predictors from cell lines to the TCGA and HMF collections of human tumors. First, we projected the GDSC and TCGA expression data onto the GDSC-TCGA consensus features. Then we trained, for each drug, a regression model using solely the cell line response data (measured as AUC). These drug-specific regression models were then used to predict response on the projected TCGA data, for patients that received the target drug as monotherapy or in combination with other standard-of-care therapies (Subsections **??**). Finally, we compared the predicted patient responses to the known categorical clinical responses using a one-sided Mann-Whitney test and computed the corresponding effect size. We trained models for seventeen different drugs (Table 1). We compared the performance of TRANSACT to the performance obtained by four competing approaches (*ElasticNet*, *DL*, *ComBat+DL* and *PRECISE*) (Table 1, Figure 4.4A, Subsection 4.4.10). For the Deep Learning approaches (*DL* and *ComBat+DL*), we selected the architecture and hyper-parameters for each drug by 5-fold cross-validation on GDSC. We subsequently trained 50 models with different and independent initializations and reported the median performance obtained on TCGA.

ElasticNet and *PRECISE* obtain significant associations (bold entries in Table 1 and Table 2) for three and six drugs, respectively, but neither approach ever outperforms (i.e. achieves a larger AUC) all other approaches. *DL* and *ComBat+DL* achieve significant associations for eight and five drugs, respectively – however, both approaches outperform all others (red, bold entries in Table 1 and Table 2) for only three and one drug, respectively. In contrast, *TRANSACT* achieves significant associations for seven drugs and obtains a larger AUC than all other approaches for five drugs.

For both deep learning approaches, we observe an important dependency on the network initialization (Figure 4.4A). More importantly, we observe no correlation between the training error achieved on the source domain (cell lines) and the prediction accuracy on the target domain (human tumors), making it impossible to select a proper initialization solely based on the source domain performance (Figure C.12A, Figure C.13A). Results obtained with *TRANSACT*, on the contrary, do not depend on a random initialization.

For the HMF data, we repeated the steps above, while employing the GDSC-HMF consensus features as well as the HMF and GDSC expression and response data. We trained models for six drugs (Table 2, Figure 4.4B). Across all approaches, we observe a signif-

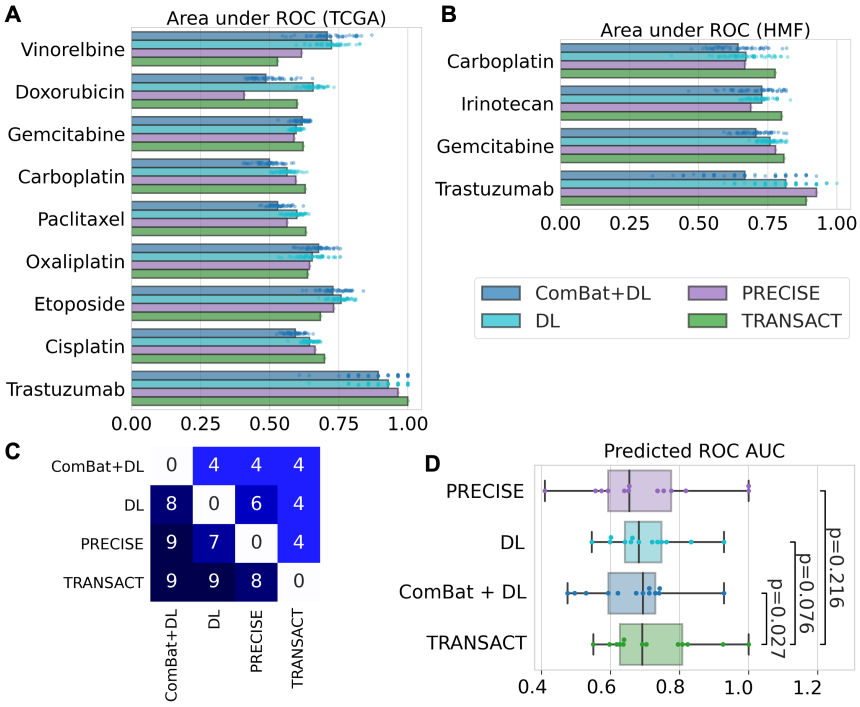


Figure 4.4 – **Consensus features improve response prediction in patients.**(**A**) We used consensus features computed between the GDSC and each tumor dataset to train a predictor for 17 drugs on TCGA and 6 drugs on HMF, using only GDSC drug response. We then predicted the continuous response in patients for each drug and compared this predicted value to the observed clinical response using a one-sided Mann-Whitney test (Table 1, Table 2). We performed the same prediction tasks using three state-of-the-art approaches (ComBat + Deep Learning, Deep Learning and PRECISE) and we summarized the results in a boxplot for TCGA (**A**) and HMF (**B**), restricting to the drugs where at least one method reaches a significant prediction. For ComBat+DL and DL, we display the results obtained using 50 different random initializations. (**C**) Table of performance comparing each method; to read as "method on y-axis obtained higher AUC than method on x-axis", e.g. "PRECISE obtained higher AUC than ComBat+DL for 9 drugs". (**D**) Comparison of ROC AUCs obtained with each of the four methods for the 13 drugs with significant prediction. P-values computed using a one-sided Wilcoxon paired rank-sum test.

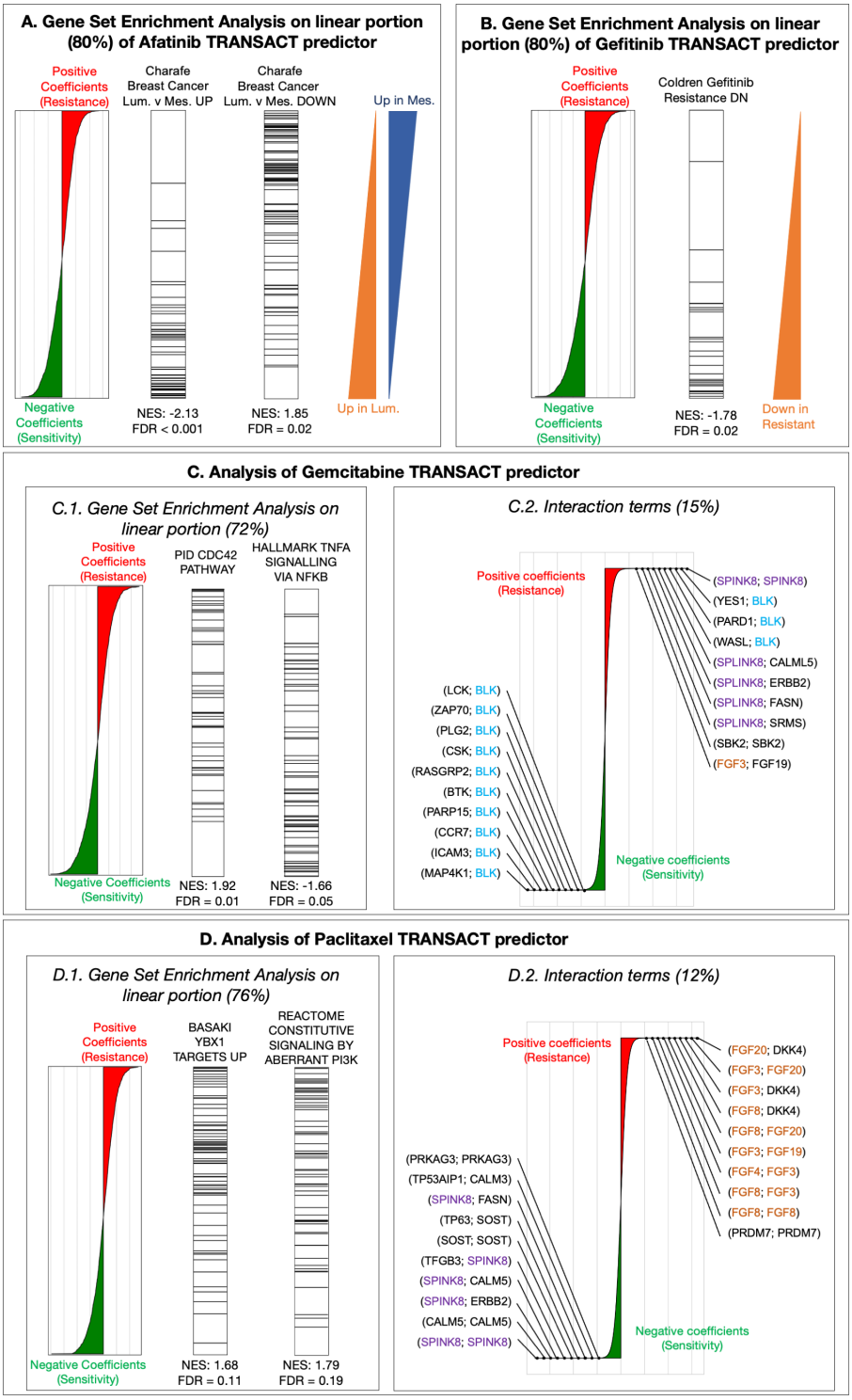
icant association between the predicted AUC and clinical responses for four of the six drugs (Table 2, Figure C.14). *PRECISE* reaches significance for two drugs, whereas *ElasticNet*, *DL* and *Combat+DL* reach significance for a single drug. *TRANSACT* outperforms *PRECISE* on three drugs, and *ComBat+DL*, *Deep Learning* and *ElasticNet* on four drugs. *TRANSACT* achieves a borderline non-significant association for Paclitaxel but achieves an effect size of 0.7. In contrast, all competing approaches fail to achieve any association with effect sizes around 0.5, corresponding to random chance. Also for the HMF cohort, deep learning approaches show a strong dependency on parameter initialization (Figure 4B) and a lack of correlation between source and target domain performance (Figure C.12B, Figure C.13B).

In summary, across the 23 drug prediction challenges on the TCGA and HMF cohorts, 13 lead to a significant prediction for at least one method. Amongst these, *TRANSACT* performs best, reaching significance in 11 challenges, followed by *Deep Learning* and *ComBat+DL* reaching significance in 9 and 6 challenges, respectively. *TRANSACT* yields larger AUCs than *Deep Learning*, *ComBat+DL* and *PRECISE* for 9, 9 and 8 drugs respectively (Figure 4.4C). It should be noted that only 3 of these comparisons are significant based on the bootstrap CIs of the AUC. Nevertheless, when comparing methods in a paired fashion across the 13 drugs where at least one method reaches significance, we observe that AUCs obtained by *TRANSACT* are significantly larger than the AUCs obtained by *ComBat+DL* ($p = 0.021$, one-sided paired Wilcoxon rank-sum test) (Figure 4.4D). When comparing *TRANSACT* to *Deep Learning* or *PRECISE*, we observe larger obtained median AUCs, but these differences are not significant ($p = 0.07$ and $p = 0.21$, respectively).

4.2.5. INTERPRETABILITY OF CONSENSUS FEATURES CONFIRMS KNOWN MECHANISMS FOR TARGETED THERAPIES AND UNVEILS POTENTIAL BIOMARKERS OF SENSITIVITY FOR CYTOTOXIC DRUGS

Figure 4.5 – **Interpretability of TRANSACT consensus features recapitulates modes of action for Afatinib and Gefitinib, and highlights mechanisms of sensitivity and resistance to Gemcitabine and Paclitaxel.** (**A**) For the Afatinib (HER-2 protein kinase inhibitor) response model trained on GDSC samples projected on GDSC-to-TCGA consensus features, we show the contribution of each gene to the predictor (left): positive (negative) weights in the predictor indicate that high (low) expression of the genes leads to resistance (sensitivity) represented by larger (smaller) AUCs. We performed a PreRanked gene set enrichment analysis (GSEA) on these weights and highlight genes associated to two significantly enriched gene sets: "Charafe [...] Luminal vs Mesenchymal UP" (center-left) and "Charafe [...] Luminal vs Mesenchymal DOWN" (center-right). The two triangles on the right illustrate these two gene-set distributions. (**B**) Interpretation of the linear part of the Gefitinib (EGFR inhibitor) model. We display gene weights ordered by contribution (left) and the ranks of genes known to be down regulated in Gefitinib resistant tumors (right). (**C**) Interpretation of Gemcitabine predictor. (**C**.1.) We display the positions of genes associated to two significantly enriched gene sets: "CDC42 pathway", associated to positive weights (resistance), and "TNFγ-signalling", associated to negative weights (sensitivity). (**C**.2.) Interaction terms make 15% of Gemcitabine predictor. We show the distribution of their weights annotated with the ten largest weights (resistance) and ten smallest weights (sensitivity). (**D**) Interpretation of Paclitaxel predictor. (**D**.1.) We display the positions of genes associated to two significantly enriched gene sets, both to resistance: "Basaki YBX1 target UP" and "constitutive signaling by PI3K-aberrant signaling". (**D**.2.) Interaction terms for Paclitaxel, that make 12% of the predictor, alongside the top-10 resistant and top-10 sensitive interactions.

Finally, we made use of the interpretability of our approach to mechanistically validate our predictors (Subsection 4.4.9). We first validated targeted therapies with documented modes of action. We started with the TRANSACT predictor of response for Afatinib, a


A. Gene Set Enrichment Analysis on linear portion (80%) of Afatinib TRANSACT predictor
Positive Coefficients (Resistance)
Charafe Breast Cancer Lum. v Mes. UP
Charafe Breast Cancer Lum. v Mes. DOWN
Up in Mes.
Negative Coefficients (Sensitivity)
NES: -2.13
FDR < 0.001
NES: 1.85
FDR = 0.02
Up in Lum.
B. Gene Set Enrichment Analysis on linear portion (80%) of Gefitinib TRANSACT predictor
Positive Coefficients (Resistance)
Coldren Gefitinib Resistance DN
Negative Coefficients (Sensitivity)
NES: -1.78
FDR = 0.02
Down in Resistant
C. Analysis of Gemcitabine TRANSACT predictor
C.1. Gene Set Enrichment Analysis on linear portion (72%)
Positive Coefficients (Resistance)
PID CDC42 PATHWAY
HALLMARK TNFA SIGNALLING VIA NFKB
Negative Coefficients (Sensitivity)
NES: 1.92
FDR = 0.01
NES: -1.66
FDR = 0.05
C.2. Interaction terms (15%)
Positive coefficients (Resistance)
(SPINK8; SPINK8)
(YES1; BLK)
(PARD1; BLK)
(WASL; BLK)
(SPLINK8; CALML5)
(SPLINK8; ERBB2)
(SPLINK8; FASN)
(SPLINK8; SRMS)
(SBK2; SBK2)
(FGF3; FGF19)
(LCK; BLK)
(ZAP70; BLK)
(PLG2; BLK)
(CSK; BLK)
(RASGRP2; BLK)
(BTK; BLK)
(PARP15; BLK)
(CCR7; BLK)
(ICAM3; BLK)
(MAP4K1; BLK)
Negative coefficients (Sensitivity)
D. Analysis of Paclitaxel TRANSACT predictor
D.1. Gene Set Enrichment Analysis on linear portion (76%)
Positive Coefficients (Resistance)
BASAKI YBX1 TARGETS UP
REACTOME CONSTITUTIVE SIGNALING BY ABERRANT PI3K
Negative Coefficients (Sensitivity)
NES: 1.68
FDR = 0.11
NES: 1.79
FDR = 0.19
D.2. Interaction terms (12%)
Positive coefficients (Resistance)
(FGF20; DKK4)
(FGF3; FGF20)
(FGF3; DKK4)
(FGF8; DKK4)
(FGF8; FGF20)
(FGF3; FGF19)
(FGF4; FGF3)
(FGF8; FGF3)
(FGF8; FGF8)
(PRDM7; PRDM7)
(PRKAG3; PRKAG3)
(TP53AIP1; CALM3)
(SPINK8; FASN)
(TP63; SOST)
(SOST; SOST)
(TFGB3; SPINK8)
(SPINK8; CALM5)
(SPINK8; ERBB2)
(CALM5; CALM5)
(SPINK8; SPINK8)
Negative coefficients (Sensitivity)

small molecule inhibitor of the EGFR family, which includes HER2 (Figure 4.5A). We performed a gene set enrichment analysis of the linear terms that constitutes 80% of the predictor. Most enriched gene sets are related to breast cancer subtypes as defined by *Charafe and colleagues*[134] where, contrary to the definition based on the intrinsic breast cancer subtypes, the Luminal subtype contains both ER+ and HER2+ tumors. The top ranked gene set amongst the genes associated with sensitivity (genes with a negative coefficient in the predictor) are genes associated with the "Luminal" subtypes (FDR < 0.001). Conversely, genes associated with resistance (genes with a positive coefficient in the predictor) show enrichment for the "Mesenchymal" molecular signatures, shared by basal and mesenchymal subtypes. This corresponds with HER2 negative samples, which is in line with our expectation as absence of the drug target would indicate lack of response. Similarly, in the TRANSACT response predictor for Gefitinib (EGFR inhibitor) the genes constituting the linear portion and associated with sensitivity (negative predictor coefficients) show an enrichment for genes downregulated in Gefitinib resistant tumors (Figure 4.5B). Interestingly, two gene sets related to breast cancer subtypes also show a significant enrichment in the negative coefficients of the predictor, linked to sensitivity: "Luminal vs Basal Down" (NES=-1.94, FDR<0.001) and "Luminal vs Mesenchymal Up" (NES=-1.85, FDR=0.004). The first gene set contains EGFR, the target of Gefitinib, implying that high levels of EGFR are, as one would expect, associated with sensitivity. The association of the second set with Gefitinib response is supported by the fact that mesenchymal tumors have been shown to be resistant to EGFR inhibition [135]. Further support is provided by the presence of two genes in the leading edge of the enrichment: ERBB2 ¬¬and PTPN6 (SHP-1) (Dataset S4). ERBB2 is a member of the EGFR family which heterodimerizes with EGFR resulting in activation of the EGFR pathway. Such cells tend to be sensitive to inhibition of the pathway, i.e. to Gefitinib treatment. PTPN6, on the other hand, inhibits the PI3K pathway [136, 137], activation of which is a known resistance mechanism to Gefitinib[138]. Therefore, high levels of PTPN6 prevents the pathway from being activated to circumvent Gefitinib treatment effects.

Cytotoxic drugs such as Gemcitabine or Paclitaxel have complex modes of actions involving different pathways, crosstalk between which remains challenging to understand. Since the predictions of these two drugs showed a significant association in both PDXs and patients, we set out to interpret the mechanisms of sensitivity or resistance inferred by our predictor. In Gemcitabine (Dataset S5), we observe that over-expression of the CDC42 pathway is a significant marker of resistance (FDR = 0.012, Figure 4.5C.1.) together with pathways linked to microtubule formation and cell migration (Figure C.15), both known to be promoted by CDC42[139]. Together, these enriched pathways highlight CDC42 over-expression as a potential mechanism of Gemcitabine resistance, which suggests the use of CDC42 inhibitors [140, 141] for Gemcitabine-resistant tumors. Another interesting finding is the significant enrichment of TNFα signaling in the genes associated with sensitivity (FDR=0.046) (Figure 4.5C.1.). A clinical trial has shown that co-administration of TNF with gemcitabine improves patient survival and further inhibits tumor growth[142], lending additional credence to this finding. Last, we observe a concentration of sensitive interactions involving BLK, a pro-apoptotic Src-proto-oncogene involved in B-cell signaling and differentiation (Figure 4.5C.2.). Since hematopoietic cell lines respond better to Gemcitabine, these interactions can either act as a tissue-type

marker or could potentially represent a sensitive pathway.
Finally, we looked for enriched pathways in the Paclitaxel predictor (Figure 4.5D.1., Dataset S6) and observed three potential mechanisms of resistance. We first observe that in the linear terms, the genes associated with resistance are significantly enriched in genes linked to silencing of YBX1[143] (FDR=0.106), a gene associated with proliferation in certain tumor types[144]. . In ovarian cancer, YBX1 has been shown to regulate ABCB1 expression levels, a gene related to Paclitaxel resistance [145–148]. Our pan-cancer analysis therefore further supports the role of drug transporters in Paclitaxel resistance. Second, we observe a significant enrichment in genes associated with resistance for PI3K activation (FDR=0.18), which is corroborated by the observed activation of PI3K/AKT/mTOR signaling pathway in Paclitaxel-resistant cancer cells[149, 150]. Moreover, a recent investigation suggests that PI3K catalytic subunits can regulate ABCB1 expression[151]. Finally, when it comes to the non-linear part, we observe a concentration of fibroblast growth factors interactions in the non-linearities associated with resistance, in particular FGF3, FGF20 and FGF8 and FGF4 (Figure 4.5D.2., Dataset S6). This behavior, although suggested by previous studies[152, 153], is all the more interesting as cell lines do not contain any micro-environment that would elicit such resistance.

4.3. DISCUSSION

We introduced an approach to integrate pre-clinical and clinical data in a fully unsupervised way. Our approach geometrically aligns sample-to-sample similarity matrices and extracts directions of important variations for both datasets, without requiring any sample-level pairing. By performing a geometrical alignment instead of a direct distribution comparison, our approach limits the effect of a potential sample selection bias. This geometrical alignment is implicitly performed in a space induced by our similarity function, which enables the integration of various assumptions regarding non-linearities in the system. Although we restricted ourselves to a single Gaussian similarity function for all drugs, designing similarity functions that incorporate a wide range of prior knowledge, specifically tailored for each drug, is a potentially promising avenue. Learning the similarity matrix, e.g. using multiple kernel learning[154] or deep learning methods such as variational auto-encoders[155], may also help increase performance.
TRANSACT compares directions computed using Kernel PCA, but our approach can be extended to other basis expansion methods by modulating the way the coefficients α^s and α^t are computed. More generally, our method is versatile, generalizable, and can be applied beyond the scope of our study, e.g., to integrate single cell sequencing data.
We showed that the consensus features can be used to build translatable predictors of drug response. Although we do not require a strong covariate shift assumption as in a previous study[156], we do assume the functions modeling the response from these consensus features follow the same monotonicity in pre-clinical models and human tumors. This assumption, albeit reasonable, may be questioned.
In this study, we limited ourselves to gene expression. Making use of other genomics levels – e.g., mutations, copy number, methylation, chromatin accessibility – may help refine the prediction by providing additional signal. The integration of our approach with multi-omics integration strategies[72, 73] may offer a solution to the translation of multi-omics signatures.

Finally, we evaluated the predictors on a variety of drug prediction tasks in human data that are quantitatively far greater, and mechanistically more diverse than prior work. We were able to predict response in patients that received a particular treatment, either as monotherapy or in combination with other therapies, even though the cell line predictors were trained on monotherapies only. We convincingly demonstrate that response predictions are now substantially better than random guessing for a number of therapies of high clinical importance, such as platinum-based chemotherapies, gemcitabine and paclitaxel. In addition, we include results of a new dataset from HMF which provides independent validation of performance on TCGA data. Intriguingly, none of the methods were able to predict the human responses to cyclophosphamide. However, no effect in vitro has been observed for this pro-drug, which might be considered as a negative control for the approaches.

Although our results are encouraging, we recognize that the drug response prediction models we present here are still far from clinical applicability. For example, one would never withhold a standard-of-care therapy based on the accuracy with which the presented predictors can identify non-responsive patients. However, a more likely scenario where such predictors could become useful sooner, is in providing guidance in drug repurposing for patients that have become refractory to all standard-of-care treatments. To reach accuracies that are acceptable for clinical application, we anticipate that large-scale model system drug (combination) screens could provide the required training samples sizes.

The recent advent of immuno-therapies calls for methods with the ability to predict the clinical response from model systems. This requires model systems capable of mimicking the action of the immune system and screening technologies able to measure the response for large panels. We believe that our approach can be extended to such problems once data is made available.

4.4. METHODS

4.4.1. PUBLIC DATA DOWNLOAD AND PRE-PROCESSING

GDSC DATASET, DOWNLOAD AND PROCESSING

We made use of the GDSC1000 cell line panel[9], which contains complete molecular profiles for 1,049 cell lines (Figure C.3). Gene expression is provided in the form of both read counts and FPKM. For both settings, we normalized the dataset for library-size and potential sampling artifacts using TMM[157] and log-transformed the adjusted read counts[158, 159]. Finally, we performed a gene-level mean-centering and standardization. When comparing GDSC to PDXE, we employed the FPKM data; in the two other analysis (GDSC to TCGA and GDSC to HMF), we made use of the read count data. In this way, FPKM and read count were never used at the same time.

NOVARTIS PDXE DATASET, DOWNLOAD AND PROCESSING

We made use of NIBR PDXE dataset for patient-derived xenografts[93], which contains the gene expression profiles of 399 PDXs (Figure C.4). Gene expression is provided in the form of FPKM. We normalized the dataset using TMM [157] and log-transformed the adjusted read counts[158, 159]. Finally, we performed a gene-level mean-centering and standardization.

TCGA DATASET, DOWNLOAD AND PROCESSING

We made use of the TCGA dataset for analyzing human biopsies[160], which comprises 10,347 human tumors (Figure C.5). Gene expression is provided in the forms of both read counts and FPKM and we used the same pre-processing pipeline as for GDSC. Response data have been obtained from Ding et al[161]. Following *Ding et al*, for each drug, we consider the response to patients who were administered a particular drug either as monotherapy or in combination with other drugs.

4.4.2. HARTWIG MEDICAL FOUNDATION DATASET (HMF) DOWNLOAD AND PROCESSING

We validated our approach on a cohort of 1,049 patients provided by the Hartwig Medical Foundation – referred to as HMF (Figure C.6A). Gene expression was measured for each metastatic sample prior to indicated drug regimen. We used MultiQC for quality control[162], salmon v1.0.0 for alignment to reference transcriptome [163], and finally edgeR for gene-level quantification[89]. Comparison with results obtained using STAR[164] and featureCounts[165] shows high degree of concordance (Figure C.6D) and we used this comparison to refine our filtering. Read counts were then processed using the same pipeline as in GDSC and TCGA.

Drug response was measured in 802 unique metastatic samples using the RECIST criteria. Response was measured differently for each patient (Figure C.6B) with most patients having one single measure of response around 10 to 15 weeks after treatment started (Figure C.6C). Since we are interested in the response of the drug given the molecular characterization measured, we considered for each patient the first response after treatment. Since most drugs are administered in combination, we considered, for each drug, all the patients that received it, whether in combination with other drugs or as monotherapy. For instance, in the case of Gemcitabine, we predicted drug response for all patients that received Gemcitabine as part of their treatment strategies.

4.4.3. MEASURE OF DRUG RESPONSE

In our different analysis, we rely on drug response measurements, either to train a predictor (GDSC), or to validate it (PDXE, TCGA and HMF). For cell lines (GDSC), drug response is measured as *Area Under the drug response Curve*, referred to as AUC. For PDX, drug response is measured as *Best Average Response*, which corresponds to the relative variation of tumor volume induced by a certain treatment. For both AUC and Best Average Response, large values are associated with resistance. For TCGA and HMF, clinical responses have been measured using the RECIST criteria[166]. Based on various metrics, patients get assigned to one of the following four groups: *Complete Response* (CR), *Partial Response* (PR), *Stable Disease* (SD) and *Progressive Disease* (PD). Following the division used in previous works[122, 123], we divide TCGA patients in two categories: *Responders* (CR and PR) and *Non Responders* (SD and PD). For HMF, we discriminate between each possible couple: PR vs PD, PR vs SD and SD vs PD. Since only a couple of patients showed a complete response, we did not consider these patients.

4.4.4. MATHEMATICAL NOTATION

We denote by p the number of genes. We consider one source dataset $\mathscr{X}_s = \{x_1^s, .., x_{n_s}^s\} \subset \mathbb{R}^p$ and one target dataset $\mathscr{X}_t = \{x_1^t, .., x_{n_t}^t\} \subset \mathbb{R}^p$ with corresponding source and target data matrices $X_s \in \mathbb{R}^{n_s \times p}$ and $X_t \in \mathbb{R}^{n_t \times p}$.
We consider a similarity function K – also called kernel function – that assigns to two samples a scalar value that is large for similar samples and small for dissimilar samples. In this work, we assume the kernel to be positive definite (Proposition C.2), and specifically use the following two kernels:

- Linear kernel: $K^{linear}(x, y) = x^T y$.
- Gaussian kernel, or Radial Basis Function: $K_\gamma^{rbf}(x, y) = \exp(-\gamma \|x - y\|^2)$, with $\gamma > 0$.

We denote by K_s the matrix of similarity between source samples, K_t between target samples and $K_s t$ the matrix of similarity between source and target (Definition C.2.2). These similarity matrices are then mean-centered (Definition C.3.1), yielding matrices $\widetilde{K}_s, \widetilde{K}_t$ and $\widetilde{K}_{st}$.

4.4.5. KERNEL PCA BY EIGEN-DECOMPOSITION OF CENTERED KERNEL MATRIX FOR CAPTURING DIRECTIONS OF PRINCIPAL VARIANCE

Using the so-called Kernel Trick (Proposition C.2) the similarity matrices previously presented can be seen as sample-covariance matrices and therefore decomposed to compute principal components inside the embedded space, a procedure known as *Kernel PCA* [167]. We perform Kernel PCA on source and target data independently to compute d_s and d_t principal components respectively. Kernel PCA on the source dataset consists of an eigen-decomposition of the matrix $\widetilde{K}_s$, yielding $\alpha^s \in \mathbb{R}^{d_s \times n_s}$, while Kernel PCA on the target dataset decomposes $\widetilde{K}_t$, yielding $\alpha^t \in \mathbb{R}^{d_t \times n_t}$ (Definition C.4.1).

4.4.6. COMPARING AND ALIGNING PRE-CLINICAL AND TUMOR NON-LINEAR PRINCIPAL COMPONENTS

Similarly to the *cosine similarity matrix* in other related works [59, 131], we define the non-linear cosine similarity matrix $\mathbf{M}^K$ as the matrix that geometrically compares the source NLPCs to the target NLPCs (Definition C.6.1). This matrix can be computed as follow (Proposition C.6.2):

$$\mathbf{M}^K = \alpha^s \widetilde{K}_{st} {\alpha^t}^T = \alpha^s C_{n_s} K_{st} C_{n_t} {\alpha^t}^T. \tag{4.1}$$

In a first step of our domain adaptation approach, we use the matrix $\mathbf{M}^K$ to align NLPCs, yielding non-linear principal vectors $s_1, .., s_d$ for the source and $t_1, .., t_d$ for the target domains, with $d = \min(d_s, d_t)$ (Definition C.5.1). These principal vectors are pairs of vectors: one linear combination of source NLPCs and one linear combination of target NLPCs, ordered by decreasing similarity with the first pair being the most similar. The computation of these PVs relies on the Singular Value Decomposition [104] of $\mathbf{M}^K$, $\mathbf{M}^K = \beta^s \Sigma {\beta^t}^T$, that helps us define the source and target sample importance loadings ρ^s

and ρ^t as follows (Proposition C.6.5)

$$\rho^s = \beta^{s^T} \alpha^s \quad \text{and} \quad \rho^t = \beta^{t^T} \alpha^t. \tag{4.2}$$

4.4.7. INTERPOLATION BETWEEN KERNEL PRINCIPAL VECTORS FOR BALANCING EFFECT OF SOURCE AND TARGET

Each pair of principal vectors contains two vectors that are geometrically similar. Projection on them will create two highly correlated covariates that would not be optimal for subsequent statistical treatment. In order to compute one vector out of each pair, we interpolate between the source and the target PV within each pair (Definition C.7.2). For the k^{th} PV, the interpolation is modulated by a coefficient τ_k that ranges between 0, when the interpolation returns the source PV, and 1, when the interpolation returns the target PV. This interpolation between vectors within each PV pair relies on two functions $\Gamma(\tau) = [\Gamma_1(\tau_1),..,\Gamma_d(\tau_d)]^T$ and $\xi(\tau) = [\xi_1(\tau_1),..,\xi_d(\tau_d)]^T$ defined as (Definition C.7.1). For a set of d interpolation coefficients $[\tau_1,..,\tau_d]$, we compute the projection of source and target datasets $F(\tau) \in \mathbb{R}^{(n_s+n_t)\times d}$ as follows (Theorem C.7.6)

$$F(\tau) = \begin{bmatrix} K_s & K_{st} \\ K_{st}^T & K_t \end{bmatrix} \begin{bmatrix} C_{n_s} & 0 \\ 0 & C_{n_t} \end{bmatrix} \begin{bmatrix} \rho^{s^T} & 0 \\ 0 & \rho^{t^T} \end{bmatrix} \begin{bmatrix} \Gamma(\tau) \\ \xi(\tau) \end{bmatrix}. \tag{4.3}$$

Such an interpolation between PVs balances the effect of source and target datasets. We prove that, in the case of a linear kernel, our interpolation scheme is equivalent to the one from previous approaches[58, 60] (Subsection C.9).
Within each pair of PVs, we select one intermediate representation where the source and target projections match the most. For the k^{th} PV-pair, we compare the source and target projected data using a Kolmogorov-Smirnov statistic and select the interpolation coefficient τ_k^* where the statistic is minimal. We obtain a set of optimal interpolation coefficients $\tau^* \in [0,1]^d$ when, for each PV, source and target influence are balanced. We call the corresponding vector consensus features. These consensus features show the minimal difference between source and target domain, a theoretical necessary condition for domain adaptation[168].

4.4.8. PREDICTION USING ELASTICNET

In order to predict drug response, we use ElasticNet regression[86]. ElasticNet is a linear model that imposes two penalties on the coefficients to predict: an ℓ_1 penalty that leads to a sparse model and an ℓ_2 penalty that jointly shrinks correlated features. We chose ElasticNet first because it has repeatedly been shown in the drug response prediction literature to give equivalent, if not better, predictive performance compared to complex models [45, 66, 87]. Using a linear classifier limits the complexity and therefore makes the transfer more robust in practice.

4.4.9. TAYLOR EXPANSION OF THE SIMILARITY FUNCTION FOR INTERPRETABILITY OF THE MODEL

In the case of RBF, we perform a PCA in an infinite-dimensional feature space. Although this space cannot be analytically computed, it can be approximated using a Taylor ex-

pansion [169] (Subsection C.8). For the q-th consensus feature, we differentiate three kinds of contributions (Definition C.8.4):

- Offset $\mathscr{O}_q$: a Gaussian term that models the squared depth.
- Linear contributions $\left(\mathscr{L}_{q,j}\right)_{1\leq j\leq p}$: a linear term, resembling the expression of one gene.
- Interaction terms $\left(\mathscr{I}_{q,j,k}\right)_{1\leq j,k\leq p}$: an interaction term that expresses the product of two genes.

These contributions can be computed from sample importance loadings of consensus features (Proposition C.8.7). We consider the contributions' sum-of-squares as a geometrical proportion since these sum up to one (Definition C.8.8). In order to look for enrichment in a particular consensus feature, we look for enrichment of particular gene sets[105]. Specifically, for the linear contribution, we compute the loading of all linear terms corresponding each to one gene. We then use these gene scores to perform a Pre-Ranked gene set enrichment analysis with 1000 permutations and use a threshold of 20% for FDR. Since these loadings correspond to a Euclidean geometric proportion, we used a squared statistic to compare them.

4.4.10. COMPARISON TO COMPETING APPROACHES

We compare TRANSACT to four different approaches. The two first approaches consist in applying a regression model trained on a source dataset to a target dataset without any correction; we use one linear ElasticNet model (referred to as *ElasticNet*) and a non-linear neural network model (referred to as *Deep Learning*). In both cases, we perform a grid-search 5-fold cross-validation on cell lines to select the model with the best performance: on *ElasticNet* we vary the ℓ_1 ratio and the total regularization ; on *Deep Learning*, following the protocol from *Sakellaropoulos et al* [122], we use a hyperbolic-tangent activation function while varying the global network structure, the ℓ_2 penalty and the input and output dropout levels (Dataset S7).
The other two approaches first correct the signal and then train a regression model. The third approach (referred to as *ComBat+DL*) reproduces the approach from *Sakellaropoulos et al* [122] by first performing a ComBat technical batch effect correction between source and target, and then applying a neural network on the corrected signal, similar to *Deep Learning* (Dataset S8). The last competing approach, referred to as *PRECISE*, consists in using a linear similarity function followed by an ElasticNet model, which is equivalent to PRECISE (Subsection C.9).
For the two deep learning approaches, we first performed cross-validation on the source dataset (with or without correction) to select the hyper-parameters and the network structures with the largest predictive performance. We then re-initialize the network and train it on the complete GDSC dataset.
In all comparisons, ROC curves and areas were computed using the pROC package[170]. The 95% Confidence Intervals (CI) were computed using the "bootstrap" sub-method with 1000 samplings with stratification.

5

IDENTIFYING COMMONALITIES BETWEEN CELL LINES AND TUMORS AT THE SINGLE CELL LEVEL USING SOBOLEV ALIGNMENT OF DEEP GENERATIVE MODELS.

Soufiane MOURRAGUI, Joe C. SIEFERT, Marcel J.T. REINDERS, Marco LOOG, Lodewyk F.A. WESSELS

Parts of this paper have been deposited on Biorxiv and an extension of this chapter is currently under revision at Genome Biology.

5.1. INTRODUCTION

Synthetic model systems, like cell lines, offer a highly cost-effective and versatile way to study human biology. In the case of cancer, the possibility to study these model systems under a wide array of conditions renders them particularly attractive for drug screening [9, 11, 171]. When compared to human tumors, however, these benefits are overshadowed by intrinsic limitations such as the lack of a vasculature or micro-environment [15, 17, 94]. Consequently, understanding the biological differences between pre-clinical models and patients is key to improving the translation of findings from cell lines to clinical implementation [118, 172].

Several computational studies have already attempted to characterize the molecular differences between cell lines and patients. The first category of approaches consist of designing machine learning tools to capture the common information relevant for transferring biomarkers of drug response [53, 109, 123, 124, 131, 173]. A second category of approaches compare the genomic landscapes of cell lines and tumors in an unsupervised fashion[128, 129, 174]. The latter studies have notably highlighted the existence of key differences in molecular profiles and identified clear differences between certain cancer cell lines and their associated tumors.

These insightful studies are based on bulk RNA-seq data, where gene expression is averaged across thousands of cells. Single cell sequencing technologies, like single-cell RNA sequencing (scRNA-seq), provide a more fine-grained view by measuring gene expression profiles for each individual cell. Here, we set out to assess the similarities and differences in transcriptional profiles between cell lines and tumors at single cell resolution. Specifically, we focus on cell line cultures [175, 176] and tumors extracted from lung cancer patients [177], as non-small cell lung cancer (NSCLC) cell lines have been shown to markedly drift from their tissue of origin [129]. Using a panel of lung cancer tumors and a panel of cell lines, we first show that the differences between cell lines and tumors cannot be modeled as a classical batch effect. As existing methods for comparing cell lines to tumors were not designed for scRNA-seq data [131, 173], we developed a novel computational approach, **Sobolev Alignment** (SA), which mobilizes recent advances in large scale kernel-based machine learning [178–180] and deep-learning-based probabilistic modelling of scRNA-seq profiles [181–183]. We show that the application of SA to three simulated single cell datasets accurately captures the (constructed) shared biology between datasets. Using SA, we then set out to characterize the shared and distinct biology between NSCLC cell lines and tumors. Amongst others, we show the conservation of a wide array of immune-related processes. By exploiting the biological processes shared between cell lines and tumors, we demonstrate that our approach recapitulates known modes of action of four clinically approved drugs, Finally, we analyzed the perturbation triggered by a drug with an unknown mode of action.

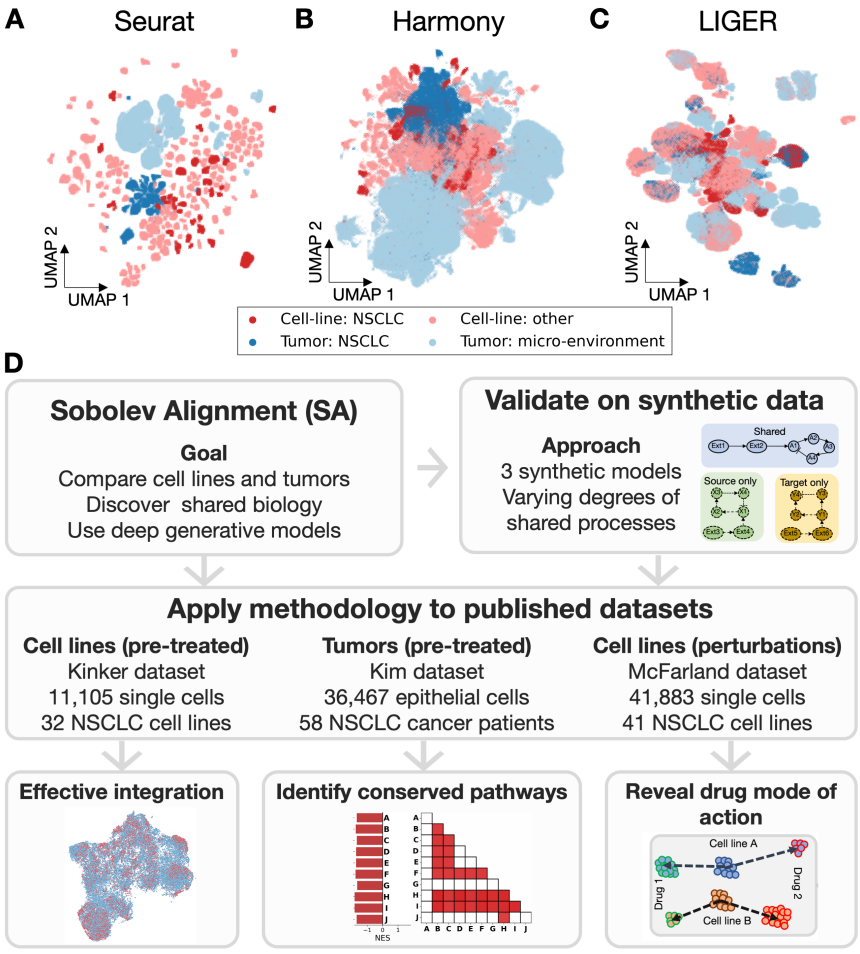


Figure 5.1 – **Transcriptional differences between cell lines and tumors obscure the traditional cell type division**. We aligned cells from 198 cell lines and tumor cells extracted from 58 NSCLC patients using three state-of-the-art batch-effect correction approaches: Seurat (**A**), Harmony (**B**) and LIGER (**C**). We here display the 2-dimensional UMAP visualization of the data after correction. Cell lines are divided in two cell-type categories: NSCLC (Non-Small Cell Lung Cancer) cell lines in dark red, and other types in light red; patient cells are divided between epithelial cells (coined NSCLC) and other cells (micro-environment) in dark and light blue respectively. (**D**) Workflow of our analysis.

5.2. RESULTS

5.2.1. DIVERGENCE BETWEEN GENE EXPRESSION PROFILES OF NON-SMALL CELL LUNG CANCER (NSCLC) CELL LINES AND TUMORS OBSCURES CELL-TYPE DEFINITION

Cancer cell lines established from biopsies of NSCLC patients are usually assumed to correspond to homogenous cell populations classified as differentiated epithelial cells. Following this assumption, integrating single cell profiles measured on NSCLC cell lines and tumors should yield a perfect co-clustering [132, 184–187]. This assumption is, however, challenged by recent studies which highlight significant differences in gene expression patterns [129, 173], mostly resulting from the lack of a micro-environment in cell lines. To assess this cell type drift and evaluate the ability of existing batch effect correction tools to correct it, we employed three state-of-the-art approaches [188, 189]: Seurat v3 [132], Harmony [190] and LIGER [187] (Methods). For cell lines, we use a panel of 53,512 single cells from 198 cell lines derived from 22 cancer types [175], including 11,105 single cells from 32 NSCLC cell lines (Figure D.1C), referred to as the *Kinker dataset*. For tumors, we use a panel of 208,506 cells from 58 NSCLC cancer patients at different disease stages [177], including 36,467 epithelial cells (Figure 5.1A-B), referred to as the *Kim dataset*. We characterized cells along two axes: 1) their origin, i.e., whether derived from cell lines or tumors, and 2) their type, i.e. whether the cells are epithelial NSCLC cells, micro-environment-related cells (tumors), or from a different cancer type (cell lines). We visualized the degree of co-clustering using UMAP [191]. For both Seurat (Figure 5.1A) and Harmony (Figure 5.1B), we observe no co-clustering of epithelial tumor cells and NSCLC cell line cells. LIGER offered a slightly better co-clustering (Figure 5.1C), with one cluster mixing cell-lines and tumors from the same cell type, although not perfectly. For all three methods, however, most NSCLC cancer cell lines are projected away from the epithelial tumor cells, indicating either a clear difference in gene expression profiles or an inability of current methods to correct this kind of batch effect. When considering at the tumor clustering, we furthermore observe a lack of mixing in epithelial tumor cells from different patients, contrasting with a good mixing in tumor micro-environment cells (Figure D.1D-F). This suggests profound differences between transcriptional profiles of cell lines and tumor cells, and the inability of standard single cell batch alignment pipelines to identify commonalities. To overcome this limitation, we set out to systematically identify and align biological processes present in cell lines and tumors (Figure 5.1D).

5.2.2. SOBOLEV ALIGNMENT COMPARES DEEP PROBABILISTIC MODELS BY KERNEL APPROXIMATION

We previously introduced two computational approaches to align bulk cell line and tumor gene expression profiles in an unsupervised manner: PRECISE [131] and TRANSACT [173], respectively based on PCA and kernel PCA. However, these two approaches, do not account for the specific properties of scRNA-seq data, such as zero-inflation or over-dispersion, which limit their applicability to single-cell data. To accommodate these properties, we replaced the (kernel-)PCA dimensionality reduction step with a Variational Auto-Encoder (VAE) [178] tailored to scRNA-seq data [182, 183]. A VAE con-

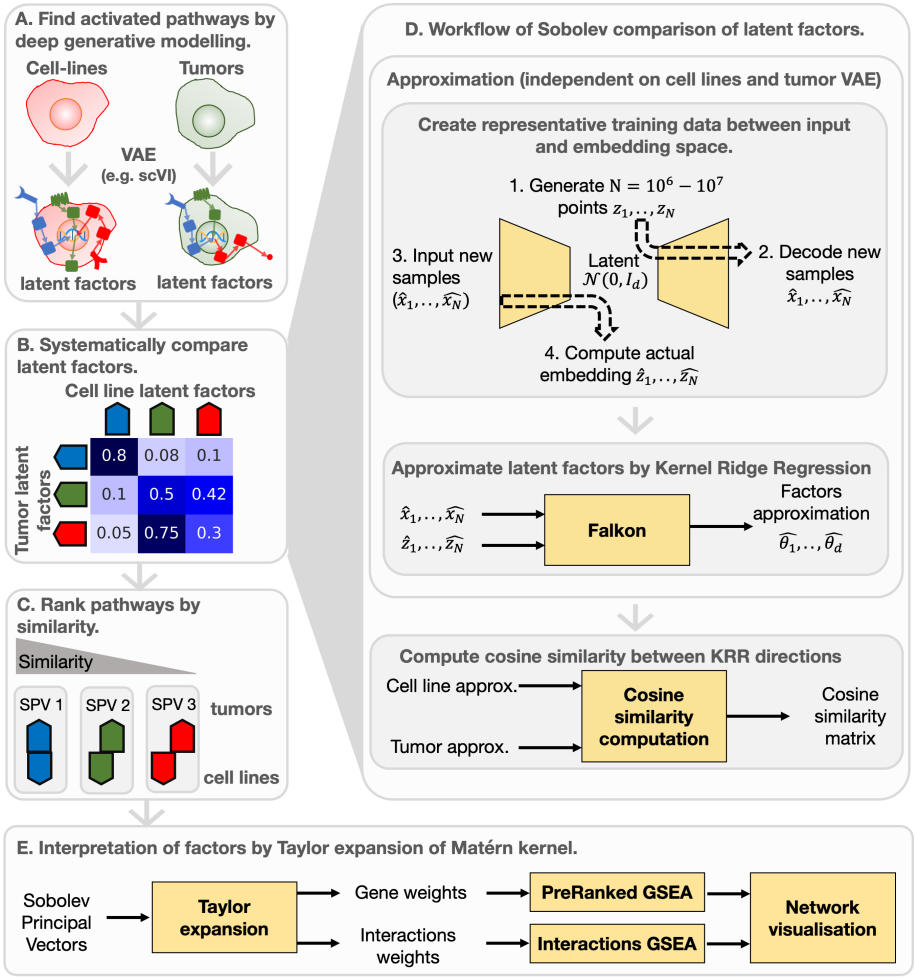


Figure 5.2 – **Sobolev Alignment systematically compares molecular processes active in cell lines and tumors**. (**A**) In the first step, we employ Variational Auto-Encoders (VAE) to aggregate single cell profiles into a handful of so-called latent factors. Each of these latent factors represents a complex non-linear gene combination. Leveraging a model tailored to scRNA-seq data, scVI, this decomposition also accounts for technical issues, such as dropout, dispersion, or library-size. We train two independent VAEs, one for each data stream, with different architectures (Methods, Figure D.2A), resulting in two sets of latent factors. (**B**) In order to relate these two sets of factors, Sobolev Alignment yields a cosine similarity matrix with values ranging from 0 to 1: 0 means that the biology supporting the two factors is completely different while 1 means that the genes supporting the latent factors are perfectly similar. (**C**) The cosine similarity matrix is finally decomposed to obtain Sobolev Principal Vectors (SPV) which are pairs of latent factors – one from cell lines (top), one from tumors (bottom) – ranked by gene-level similarity. Here, the first SPV corresponds to two highly similar processes, while subsequent SPV pairs contain less and less similar processes. (**D**) Technically, Sobolev Alignment leverages the generative nature of each VAE to build a large labeled training dataset. This dataset is then employed as labeled training data in a large-scale Kernel Ridge Regression, using Falkon, which approximates the encoder functions. We approximate cell line and tumor VAEs by two different sets of kernel machines, independently from one another but using the same Matérn kernel (Methods). We then compare these two sets of approximations, yielding the cosine similarity matrix (Section D.5). (**E**) In order to interpret the SPVs, we derived a scheme which relies on the Taylor expansion of the Gaussian kernel. This yields two sets of contributions: the contribution of genes and of interaction terms — which are the products of two genes. These weights are used in a Gene Set Enrichment Analysis (GSEA) framework to measure the enrichment of certain biological processes.

sists of two neural networks: one network which reduces the dimensionality of gene expression profiles to a much smaller number of latent factors, and a second which aims to reconstruct the original expression profiles from the latent factors. Akin to principal components, each latent factor represents a source of variation and captures combinations of correlated non-linear patterns in the studied dataset (Figure D.2A).
To capture processes present in either cell line or tumor data, we train two independent VAEs (Methods), one on cell lines and one on tumors (Figure 5.2A). This allows each model to capture the variability within either data set separately. (Figure D.2B). Next, based on the similarities between the two resulting sets of latent factors, cell line latent factors are matched with tumor latent factors based on a cosine similarity score (Figure 5.2B, Methods), where a high cosine similarity score indicates that two factors are overlapping and thus share underlying biology. Note that this cosine similarity is determined at the gene-level and takes into account how each gene, as well as each nonlinear combination of genes, influences the latent factor. The cosine similarity matrix between the two sets of latent factors is then used to generate pairs of matching factors (one from the cell lines and one from the tumors) that are ordered by decreasing similarity following the notion of Principal Vectors [104] (Figure 5.2C, Section D.6). Consequently, each resulting Sobolev Principal Vector (SPV) corresponds to a pair of processes – one from the cell line and one from the tumor data. By construction, the top SPVs correspond to non-linear combinations of genes common to cell lines and tumors, whereas the lower-ranked SPVs relate to cell line or tumor-specific biology. We show that the SPVs can be interpreted by exploiting a closed form solution for the Taylor expansion of the Gaussian kernel [169] (Figure 5.2E), which computes the contributions of genes and their interaction terms to each SPV (Methods). The gene-contributions are analyzed using standard PreRanked GSEA [105], while we derived an extension of GSEA for the interaction terms (Methods). Sobolev Alignment relies on several hyper-parameters, which we selected as indicated in Figure D.2D.
Our main technical contribution consists in overcoming two drawbacks of VAEs: the difficulty to interpret the resulting latent factors, and the absence of simple manipulation to obtain the contributions of the genes to the factors. Our strategy consists in approximating the VAE mapping using a Matérn kernel machine (Figure 5.2D), able to approximate a very wide class of functions, so-called Sobolev spaces, and mathematically easy to handle (Section D.3). Note that we cannot initially use this kernel machine to find the latent factors as they assume a gaussian noise model which is violated in scRNA-seq data. As the VAE is a generative model, we can train the kernel machine by generating a large data set ($\sim 10^6 - 10^7$ data points) from the trained scVI models (Figure 5.2D). Each latent factor is then approximated using Falkon [179, 180], a large-scale Kernel Ridge Regression tool based on the Nyström approximation (Methods).

5.2.3. SYNTHETIC DATA COMPARISON

In order to demonstrate the utility of SA, we applied it to synthetically constructed single cell datasets. We employed Dyngen [192] to generate the single cell data using three synthetic models. In each model we simulated processes associated with the source (cell line) and a target (tumor), respectively, and we varied the level of overlap (conservation) between the source and target.

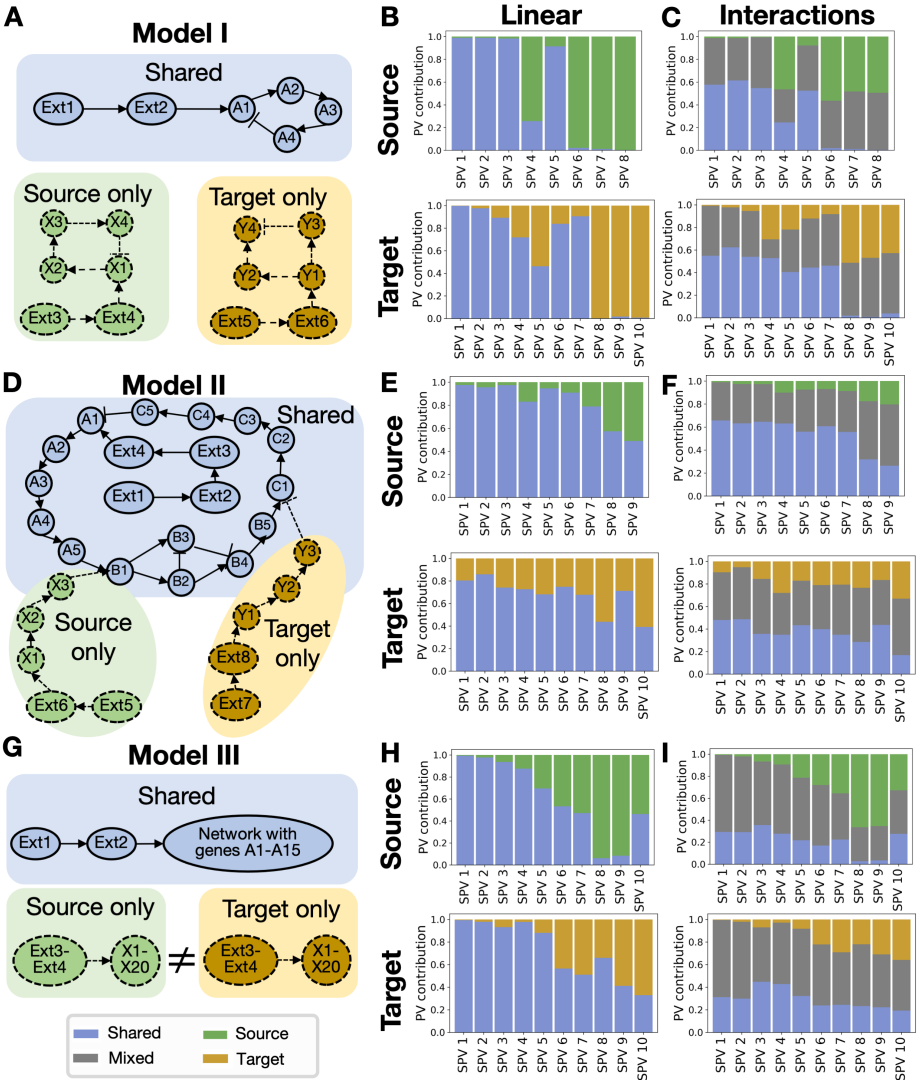


Figure 5.3 – **Sobolev Alignment faithfully discriminates shared from specific biological processes on three synthetic datasets**. We generated three models of increasing complexity to assess the capacity of Sobolev Alignment in discriminating common from data-specific signals. (**A**) Gene regulatory network of *model I*. A blue box indicates a gene shared between source and target, green a gene specific to source, and orange to target. (**B**) Proportion of gene weights for source (top) and target (bottom) SPVs. (**C**) Proportion of interaction weights for source (top) and target (bottom) PVs. (**D-F**) Results similarly displayed for *model II*. (**G-I**) Results similarly displayed for *model III*. (blue: common; grey: interactions with one common and one specific gene; green: source-specific; orange: target-specific).

For the first model (*model I*) the source and target both contain a conserved regulatory network (A1-A4), which is complemented with an independent network (X1-X4) in the source and an independent network (Y1-Y4) in the target (Figure 5.3A). Following *Cannoodt et al*, each network is initialized by "external" genes (Ext1-Ext6) (Figure 5.3A). Following Dyngen, source and target datasets furthermore differ by their kinetic parameters and the structure of housekeeping genes and transcription factors, which are used for generating the data, but excluded when applying SA. We ran the complete SA pipeline (Figure D.2A) and obtained 8 SPVs for the source and 10 for the target. To inspect the SPVs, we first look at their linear portion (Figure 5.3B), which account for 50% of the SPVs norm (Figure D.3B-C). The three top source and target SPVs exclusively consist of signal from the shared network. The subsequent two SPVs consist of a mix of shared and source/target-specific signal. Finally, the last three SPVs are made up solely of source/target specific genes. Next, we looked at the interaction genes (Figure 5.3C), which explain 30% of the variation in the SPVs (Figure D.3A-C) and observed a similar pattern. The first 3 SPVs contain a strong shared component and a strong mixed component (interactions between shared and source/target specific genes). In contrast, the last 3 SPVs contain a strong pure source/target-specific as well as mixed interactions.

Next, we constructed a more challenging dataset with a conserved component and source/target-specific external control of this shared component. Following an example proposed by *Cannoodt et al*, we designed a conserved cycling backbone pathway (Figure 5.3D). To this we added a linear pathway, specific to the source, providing an activating signal from an external gene (Ext1-5) to the shared component via B1. Similarly, specific to the target, we added a linear pathway inhibiting the shared component via C1 (Figure 5.3D). Analysis of the linear part (Figure 5.3E) and the interaction part of the SPVs (Figure 5.3F) showed a decreasing contribution of shared linear and interaction terms when going down the ranking of the SPVs. Specifically, shared linear terms account for 98% of source SPV1 and 80% of target SPV1, while these proportions decrease to 50% and 40%, respectively, for SPV9. Note that none of the SPVs solely contain shared signal or source/target-specific signal (as was the case for *model I*). This might be because, for *model II*, the source and target-specific genes have a direct influence on the conserved network. As a result, a significant amount of covariance is to be expected between the shared and source/target-specific genes. Nevertheless, SA does correctly identify the shared elements of the source and target in the top PVs while the source and target specific elements are represented in the lower ranked PVs.

Finally, *models I* and *II* harbor individual genes that are specific to either source or target. To further challenge our Sobolev Alignment, we created a *model III*, that has a shared component similar to *model II* (genes A1-A15), and a set of shared genes (X1-X20) but whose wiring is different between the source and target (Figure 5.3G). When applying SA, we observed that the linear part of the first three SPVs mostly consist of shared genes (Figure 5.3H) and that the non-linear part of these SPVs is dominated by either pure shared or mixed interactions (Figure 5.3I). Also here, the last three SPVs have a limited contribution of shared genes or interaction terms consisting of shared genes. The results from these three models demonstrate that Sobolev Alignment is able to capture the shared information between two datasets in the top ranked SPV components.

5.2.4. SOBOLEV ALIGNMENT EFFECTIVELY INTEGRATES CELL LINES AND TUMORS

We applied our Sobolev Alignment to compare treatment naïve NSCLC cell lines (*Kinker dataset*) and epithelial tumor cells from NSCLC patients (*Kim dataset*) (Methods). Following the hyper-parameter selection pipeline (Figure D.2), we selected two different architectures for the scVI models (Supp. Table 1) and approximated the embedding of the VAE latent factors with the Falkon kernel machine using a Laplacian kernel with $\sigma = 15$ (Methods). Each scVI latent factor is approximated very well, with the Spearman correlation between our approximation and the scVI-computed values ranging between 0.97 and 0.99 for the cell lines and between 0.95 and 0.99 for the tumors (Figure D.7A).

The similarities between the SPVs of the cell lines and the tumor data ranged from 0.51 to 0.01 (Figure D.7B) and we retained the top 12 significant SPV pairs (Methods). From the 12 cell line-tumor SPV pairs, we subsequently constructed a 12-dimensional Sobolov Consensus Space. Each dimension of the consensus space consists of a consensus vector obtained by interpolating between the matched cell line and tumor SPV pair (Methods). These consensus vectors represent the best balance between the effects of the cell lines and the tumors. We then projected the cell line and tumor data on the resulting 12 consensus vectors and performed a UMAP projection and observed a reasonably good but not perfect co-clustering of cell lines and tumors (Figure 5.4A). Inspired by Seurat's and LIGER's workflows, we performed a Mutual Nearest Neighbor (MNN) correction to the cell-line and tumor datasets in the Sobolev Consensus Space [185]. Although MNN on the gene expression profiles performs poorly (Figure 5.4B), its combination with SA resulted in a clear improved co-clustering of cell lines and tumors (Figure 5.4C). Such co-clustering is not achieved by either Seurat v3 (Figure D.8A), Harmony (Figure D.8B), or LIGER (Figure D.8C). In subsequent analyses, we will employ the datasets resulting from the projection on the Sobolev Consensus Space with MNN correction, denoted as the SA+MNN space.

To validate the quality of our alignment, we compared the EGFR mutation status of patients from the *Kim dataset* (tumor) with the EGFR signaling activity in the cell lines. We chose this comparison as the cell lines in the Kinker dataset do not harbor any activating EGFR mutations, rendering the direct transfer of mutation status from cell lines to tumors impossible. The cell line EGFR signaling is computed [193], for each cell, using UCell [194] (Methods). As we expect cell line cells with high EGFR signaling activity to be in close proximity to EGFR mutant tumor cells, we predict, for each tumor cell, the EGFR signaling level from the cell lines using a k-Nearest Neighbor (kNN) regression in the SA+MNN space (Figure 5.4D, Figure D.9A). From the cell-line predictions, we observed that mutant EGFR tumors have higher EGFR signaling levels compared to wild-type tumors (Figure 5.4E, Figure D.9B), coherent with a constitutive activation of EGFR signaling pathway. In other words, EGFR-mutated tumor cells indeed preferentially cluster with cell lines harboring a high level of EGFR signaling after aligning the datasets in the SA+MN consensus space. Next, we evaluated whether the cell cycle states of neighboring cell line and tumor cells in the SA+MNN space are comparable. "G2/M" and "G1/S" scores were provided for the *Kinker cell line dataset*. We computed equivalent quantities on the *Kim tumor dataset* using the Seurat v3 cell-cycle regression tool. Comparing the tumor cells' G2/M scores predicted with a kNN regression model from the cell line data

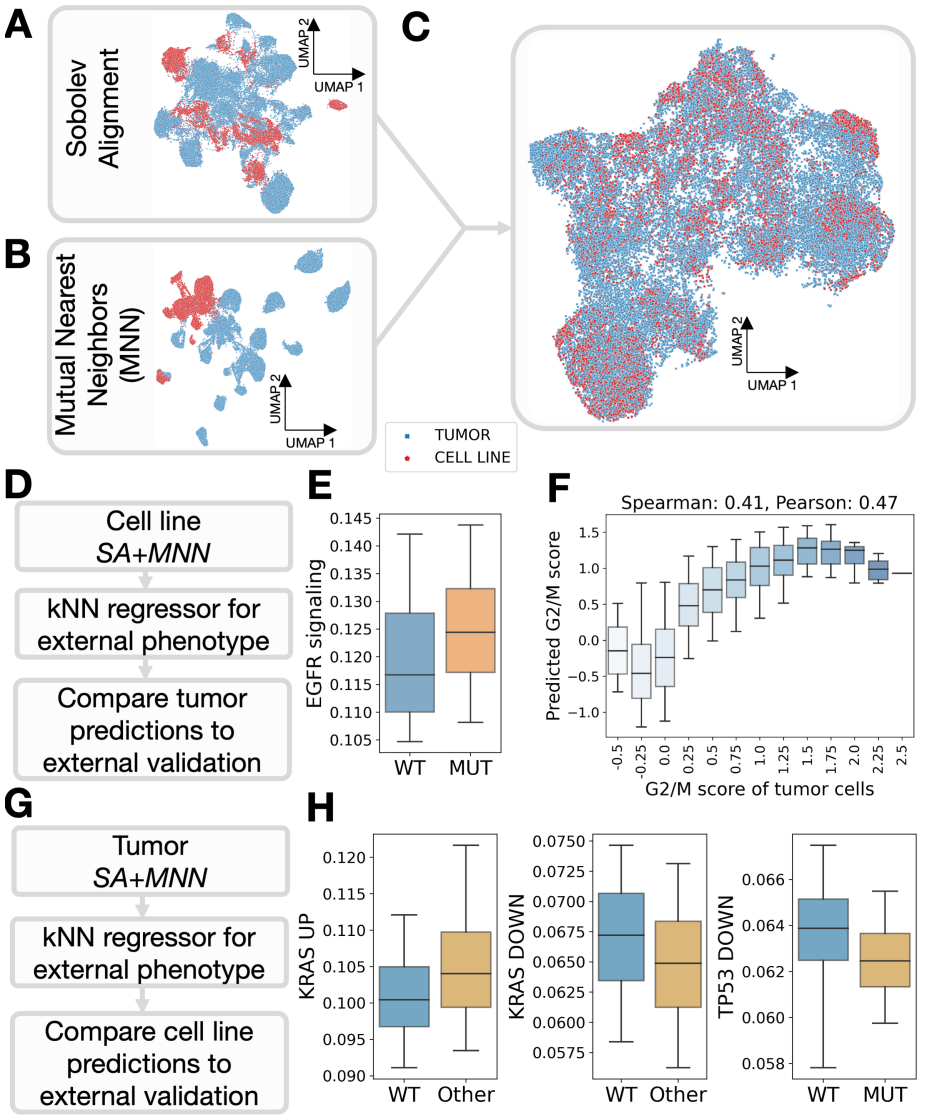


Figure 5.4 – **Sobolev Alignment effectively integrates cell lines and epithelial tumor cells**. We employed Sobolev Alignment to compare untreated epithelial cells from NSCLC cancer patients with untreated NSCLC cell lines. (**A**) UMAP of cell lines and tumors after projection on SPVs and interpolation. (**B**) UMAP of cell lines and tumors after correction with Mutual Nearest Neighbors (MNN). (**C**) UMAP of cell lines and tumors after projection on SPVs, interpolation, and subsequent correction by MNN. (**D**) Workflow of our cell line neighborhood validation, which computes, for each tumor cell, the value of a certain phenotype in neighboring cell line cells. (**E**) Relationship between imputed EGFR signaling levels and EGFR status in tumor cells. (**F**) Relationship in tumor cells between G2/M score imputed from cell lines (y-axis) and G2/M score computed using Seurat v3 (x-axis). (**G**) Workflow of our tumor neighborhood validation, reversing the order from cell line neighborhood validation. (**H**) Relationship between pathway levels and oncogenic form in cell line; left to right: genes up-regulated in KRAS mutated cells ("KRAS UP"), genes down-regulated in KRAS mutated cells ("KRAS DOWN") and genes down-regulated in TP53 mutated cells ("TP53 DOWN") (Figure D.9H-M).

(Figure D.9C) with the cell-cycle scores, we observe a spearman correlation of 0.41 (Figure 5.4F), indicating that our co-clustering captures mitotic entrance. When performing the same experiment with the S scores (Figure D.9D-E), we however only observe a modest spearman correlation of 0.11. Following a similar type of analysis (Figure 5.4G), we find that tumor cells in the neighborhood of KRAS-mutant cell lines have a higher level of expression of genes in KRAS-mutant lung cancer cells according to the "KRAS UP" gene set and a lower level of expression of genes down-regulated in KRAS-mutant lung cancer cells based on the "KRAS DOWN" gene set (Figure 5.4H, Figure D.9F-I). Also, tumor cells in the neighborhood of TP53 mutated cell lines have lower levels of genes down-regulated by TP53 mutations (Figure 5.4H, Figure D.9J); however, we could not find an association for genes up-regulated by TP53 mutant protein (Figure D.9K-L). Taken together, alignment of cells from cell lines and tumors in the SA+MNN space conserves important biomarkers, indicating the biological relevance of our mathematical model.

5.2.5. SOBOLEV ALIGNMENT HIGHLIGHTS THE CONSERVATION OF IMPORTANT INTRINSIC IMMUNE-RELATED PATHWAYS IN CELL LINES

To assess which tumor-related biological processes align well to cell line biology, we analyzed each of the tumor SPVs (Figure 5.4). First, by construction, cell line and tumor cells in the same neighborhood show similar values for all SPVs, indicating that SPVs can be understood as biomarkers, which can help relate clinical samples to cell line models. To interpret the tumor SPVs, we computed the linear (Subsection D.7.3, Supp. Table 2) and interaction loadings (Methods, Supp. Table 3), and subsequently performed a gene set enrichment analysis (Subsection D.7.4, Supp. Table 4). Using GSEA on the linear part (Figure 5.2E), we report, for each SPV, the top 10 most enriched gene sets in the linear portion (FDR < 0.05). To study interaction terms, we created interaction gene sets that correspond to all genes in a pair of gene sets (Methods). For instance, the interaction-gene-set "M Phase x Keratinization" contains all interaction pairs with one gene from the "M Phase" gene set, and one gene from the "Keratinization" gene set. For each SPV, we computed the Normalized Enrichment Scores (NES) of all interaction gene sets using a modified procedure akin to PreRanked GSEA (Methods). The remaining analyses are based on the top 5% interaction terms with the largest absolute NESs values.

We observed that negative coefficients of SPV1 are enriched for gene sets linked to G2/M mitotic gene sets (Figure 5.5A), indicating a conservation of mitosis-related biological processes. Keratin genes, represented by e.g. the "keratinization" set, are also driving the enrichment of several gene sets in the negative coefficients. Four known lung adenocarcinoma markers [195], KRT7-8–18-19, are amongst the 10 most important linear contributors, indicating a conservation of the keratin markers between cell lines and tumors. The two subsequent SPVs (SPV2 and SPV3) are enriched for PI3K-AKT and the MET pathways (Figure 5.5B-C), two important cancer-related pathways expected to be shared between cell lines and tumors. SPV4 is not associated with any enrichment. Positive coefficients of SPV5 (Figure 5.5D) are significantly enriched for immune-related gene sets (interferon signaling, cytokine signaling, interferon gamma, adaptive immune system) and their interactions. The negative coefficients of SPV5 are enriched for keratin-related gene sets. Analysis with the oncogenic signatures from MSigDB (Supp Table 4) indicates an enrichment in the positive portion for genes upregulated in KRAS mutant cells

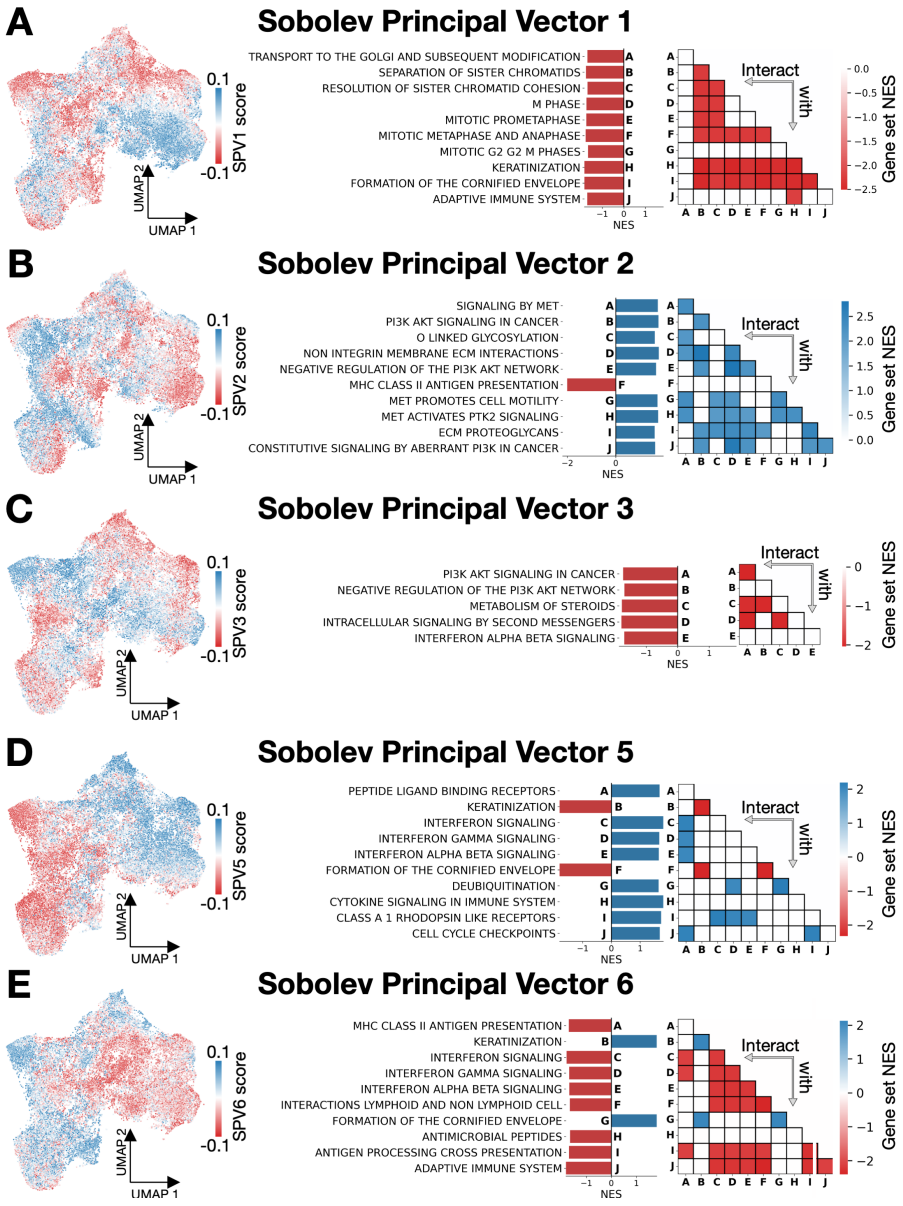


Figure 5.5 – **Analysis of Sobolev Alignment directions show the conservation in cell lines of important mitotic pathways alongside innate immune processes**. For each SPV computed between the Kinker and the Kim datasets (Figure 5.4), we study the tumor processes recapitulated in cell lines by performing a gene set enrichment analysis on the tumor SPVs. (**A**) Results for SPV1. The left UMAP of cell lines and tumors is colored with the scores of SPV1 projected on the cells (SA+MNN). Turning to gene loadings, the subsequent bar plot summarizes the top 10 most enriched gene sets for the Reactome MSigDB collection, represented by their Normalized Enrichment Scores (NES). Finally, interaction terms between these linearly enriched gene sets are reported in a heatmap. A white value indicate that the interaction term is not in the 5% most enriched interaction terms. A negative value (red) indicates that the product of the two pathways drives down the value of the SPV; a positive (blue) value indicates the reverse. The SPV scores (left) and gene-loadings (right) are connected: blue cells in the UMAP harbor a high expression of blue linear and interaction gene sets, and reciprocally. We repeated the experiment for SPV2 (**B**), SPV3 (**C**), SPV5 (**D**) and SPV6 (**E**) with a similar display.

(FDR < 0.035). SPV5 values gradually change across the UMAP, suggesting a higher-level structure shared between cell lines and tumors. More specifically, cells in the lower-left of the UMAP exhibit a high keratinization level, while cells in the upper-right present either KRAS activation, or a high level of activity for immune-related pathways. Analysis of SPV6 (Figure 5.5E) shows a similar pattern as SPV5 with keratinization and immune activity enriched in the positive and negative coefficients, respectively. SPV6 also shows a gradient between high-keratinization and high immune-response, represented here by antigen presentation (MHC class II, microbial peptides, antigen presentation). Altogether, these results highlight the conservation of immune-related pathways in lung cancer cell line models alongside their interplay with keratin-levels and KRAS aberrations.

5.2.6. SOBOLEV ALIGNMENT ALLOWS TO STUDY THE MODE OF ACTION FOR CERTAIN DRUGS

Finally, we employed our methodology to study the different modes of action to anti-cancer drugs. We first computed the SPVs between the McFarland [176] cell line and the Kim tumor datasets. The McFarland dataset consists of a multiplexed perturbation screen on 33 NSCLC cell lines where each cell line was exposed to 19 anti-cancer drugs. Transcriptomic read-outs were measured 6 or 24 hours after drug exposure (Figure D.10). To study this set of perturbations (Figure 5.6A), we first mapped the cell line data to the top cell line SPVs (Figure D.11A-C), thereby restricting the analysis to processes shared with tumors.

For each embedded cell, we computed the gene-loadings corresponding to the combination of projected SPV scores (Methods). For each cell line and exposed drug, we employed the Mann-Whitney test to assess, per gene, whether the gene weights from cells retrieved after drug induction are significantly different from gene weights of cells obtained after vehicle-treatment (DMSO). To include a measure of effect size, we filtered genes based on the mean differences of the associated gene-weights between the two groups (drug induction and DMSO). For this effect size, the threshold is set to the 95% percentile of the effect size distribution resulting from random DMSO-perturbed cells (Figure D.12A). Genes which are significant (FDR < 0.05) and pass the effect size filter are considered to be perturbed by the drug in the cell line under consideration. The set of genes is then further annotated using EnrichR44 to associate them with biological processes (only gene sets with at least 5 perturbed genes are considered).

We first analyzed Idasanutlin (a.k.a. Nutlin-3), an Mdm2 inhibitor which triggers apoptosis selectively in TP53 wild-type cells [196]. When examining the number of genes perturbed after 24 hours for each of the 5 cell lines (Figure 5.6B), we observe a clear difference between the sole TP53 wild-type cell-line (NCIH226) and the four mutated cell lines. While NCIH226 harbors 70 up-regulated genes and 78 down-regulated genes, none of the mutated cell lines show more than 5 perturbed genes. For NCIH226, we observe a clear enrichment for apoptosis and the p53 pathway in the up-regulated genes, while G2M checkpoints and E2F targets are down-regulated, coherent with the known mode of action of Idasanutlin (Figure 5.6C). We then turned to Everolimus, an mTOR inhibitor [197], and performed the analysis presented in Figure 5.6A for each of the 18 perturbed cell lines (Figure D.12B, Supp. Table 5). The enrichment scores are summa-

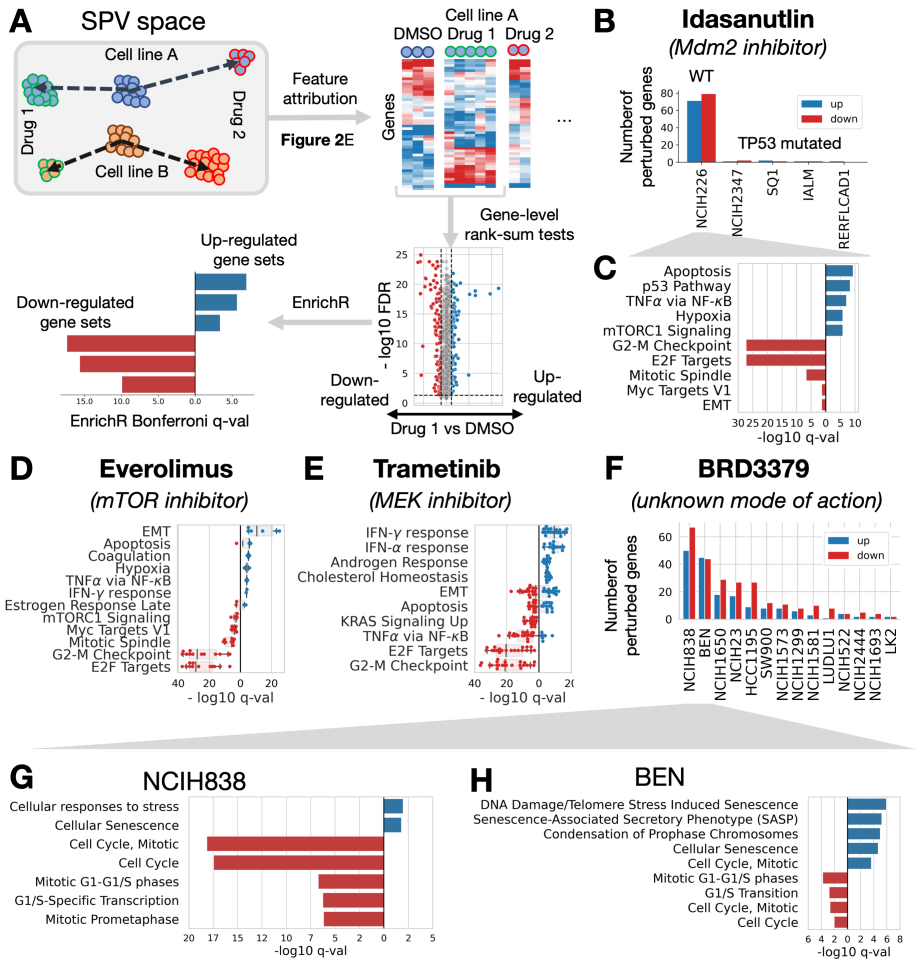


Figure 5.6 – **Sobolev Alignment identifies modes of action from drug perturbations**. Exploiting a large drug-perturbation screen (*McFarland dataset*), we studied the modes of action of several drugs recapitulated by the SPV common with tumors (*Kim dataset*). (**A**) After first embedding all cells to the SPV space corresponding to the top cell lines SPVs, we computed the gene-weights corresponding to each single cell embedding, yielding the pictured heatmaps. Using a rank-sum test (Mann-Whitney, Benjamini-Hochberg FDR multiple-testing correction), we assessed, for each drug, the difference in gene weights between DMSO-treated cells and drug-treated cells; the threshold in effect size was set as the 95% percentile of gene-weights differences observed within the DMSO-treated cells of the studied cell line. Up- and down-regulated genes were then analyzed using EnrichR. (**B**) Number of up- and down-regulated genes for 5 cell lines subjected to Mdm2-inhibitor Idasanutlin alongside their TP53 mutation status. (**C**) Enrichment observed using the MSigDB Hallmarks collection for NCIH226 (blue: enrichment in genes up-regulated by Idasanutlin, red: enrichment in down-regulated genes). (**D**) Boxplot of q-values obtained when analyzing Everolimus (mTOR inhibitor). (**E**) Boxplots of q-values obtained when analyzing Trametinib (MEK-inhibitor). (**F**) Number of up- and down-regulated genes for each cell line perturbed with BRD3379, a drug with an unknown mode of action. Zooming on the two most perturbed cell lines, we performed the EnrichR analysis using the Reactome MSigDB collection for NCIH838 (**G**) and Ben (**H**).

rized in a boxplot (Figure 5.6D), each point corresponding to a significantly enriched cell line. We observe that for 12 cell lines, G2M checkpoints and E2F targets are significantly down-regulated; the 6 other cell lines correspond to the cell lines with the least number of perturbed genes (Figure D.12B). We also observe a down-regulation of mTORC1 signaling for four cell lines (A549, IALM, LU99, NCIH1355), coherent with the known mode of action. We performed a similar experiment for Trametinib, a MEK-inhibitor (Figure 5.6E, Figure D.12C). Upregulation of IFN-γ and IFN-α, down-regulation of genes over-expressed in KRAS mutant cells, and down-regulation of cell cycle pathways are observed, consistent with previous reports [176, 198]. Androgen response and cholesterol homeostasis pathways are enriched, suggesting combination therapies with anti-androgen or cholesterol lowering drugs may yield synergistic effects or reduce resistance to MEK inhibitors [199–201]. Several pathways, such as Apoptosis and EMT, display a duality where up-regulation is observed for some cell lines while others show depletion. While this may represent normal variations in drug response, these results are consistent with the notion that resistance to MEK inhibitors can be achieved through various adaptation mechanisms, including EMT, and highlights the need for combination therapies in NSCLC [202, 203]. Finally, we turned to a drug with an unknown mode of action, BRD3379. When examining the number of perturbed genes per cell line, we observe that two cell lines stand out: NCIH838 and BEN (Figure 5.6F, Figure D.12D). Performing pathway enrichment analysis revealed that both NCIH838 (Figure 5.6G) and BEN (Figure 5.6H) are characterized by cell cycle arrest (down-regulation of cell cycle gene sets). Both cell lines show enrichment for genes involved in senescence, reminiscent of the effects of BRD4 inhibition [204, 205]. Indeed, genes shown to be related to senescence induction by BRD4 inhibition, such as Aurora kinases A/B55 and CDKN1A (p21) [205], show similar patterns of expression upon BRD3379 treatment (Supp. Table 5). Furthermore, JQ1 inhibition, a BRD2-3-4 inhibitor, similarly shows induction of senescence and cell cycle arrest (Figure D.12E-F). Therefore, we posit that BRD3379 mode of action is likely to be similar to that of BRD4 inhibitors.

Taken together, these results demonstrate the ability of our approach to recapitulate clinically validated drug response mechanisms and exemplify the potential of Sobolev Alignment to decipher complex modes of action.

5.3. DISCUSSION

We showed that single cell profiles measured from pre-clinical models and human tumors cannot be aligned using current batch-effect correction tools like Seurat, LIGER and Harmony. To help researchers study the translational potential of pre-clinical models, we derived a novel framework that exploits the power of unsupervised deep generative models (VAE), known to their ability to embed scRNA-seq data, with the benefits of kernel machines, known to be interpretable. In the first step, a VAE is trained to embed both input datasets independently, capturing the different sources of variations into a set of latent factors. In the second step, we approximate the mapping towards the latent factors using Falkon-trained kernel machines, which allows us to calculate the contribution of each gene to each of the latent factors. We then match the latent factors of the two domains by calculating their Sobolev Principal Vectors (SPVs). Finally, we construct a consensus space by interpolation between matched SPVs onto which all data can be

projected.
We applied our approach to a set of synthetic examples and showed that a conserved network of genes contributes to the most important SPVs. We then applied our alignment to NSCLC cell lines and epithelial cancer cells and showed enrichment of known oncogenes in the top SPVs. Further analysis of the top SPVs pointed towards the conservation of mitotic and immune-related pathways in cultured cells. In a last experiment, we aligned a multiplexed drug-perturbation screen with NSCLC cell lines and showed that the common SPVs recapitulate modes of action for several drugs.
Although we restricted our analysis to single cell gene expression data, our computational approach is versatile, and can easily be adapted for other molecular features, such as chromatin accessibility [206], protein levels [207] or ribosomal profiling [208]. Recently developed deep probabilistic models tailored for such data [209, 210] could be for instance employed to adapt our framework. We also solely compare cell line models to human tumors, but more complex models could be studied equally well, such as organoids or patient derived xenografts.
On a more technical note, our approach develops and exploits connections between deep-learning-based algorithms and kernel methods [211]. Recent works have shown theoretical connections [212], demonstrating, for instance, the equivalence between the Laplacian kernel and the so-called Neural Tangent Kernel [213]. We envision that future theoretical works could help improve our approach, and we believe that a promising avenue for improvement lies in replacing the kernel ridge regression step. Indeed, we observe that kernel ridge regression is locally adaptive and is not ideal for approximating functions that exhibit large variations in localized areas, as is often the case with neural networks. Furthermore, our interpretation scheme relies on the decomposition of the Gaussian kernel, which we extended to the Laplacian kernel by exploiting connections between the feature spaces of Gaussian and Laplacian kernels. Derivation of a closed-form solution for the Laplacian kernel RKHS basis could help improve our analysis. Finally, our approach could be used in other areas, for instance to define a similarity measure between neural networks, as spearheaded by recent works [214–216].

5.4. METHODS

5.4.1. DATA DOWNLOAD AND PROCESSING

Cell line and tumor data were downloaded following the protocol indicated in the supported publications [175–177], i.e., using the Broad Institute Single Cell Portal for cell lines, and GEO repository GSE131907 for tumors. Both datasets were processed using Scanpy [217].

5.4.2. BATCH EFFECT CORRECTION

The Seurat R implementation (version 4.0.1) was used and corrected for batch effect using Reciprocal PCA. We used Liger R implementation (version 0.5.0.9). Finally, we use Harmony Python implementation (version 0.0.5). For all methods, we used default parameters.

5.4.3. MODEL SELECTION FOR SCVI

The first step of Sobolev Alignment consists in training two Variational Auto-Encoders (VAEs): one using the source (cell lines) and one using the target (tumor) dataset. For training these two VAEs, we turned to an established implementation, called scVI [182, 218]. As any neural-network-based approach, scVI requires a lot of hyper-parameters to be tuned, e.g., number of hidden layers, number of latent factors, number of neurons per layers, dropout rate, weight decay, likelihood function, early stopping or number of iterations. In order to select a combination of hyper-parameters that best reconstructs the data, we turned to Bayesian Optimization [219] and selected independently for each dataset the best combination of hyper-parameters based on the reconstruction error computed on an held-out test set (Supp. Table 1). The two resulting optimal models were then trained using the complete source and target data, respectively. Furthermore, VAEs are frequently suffering from posterior collapse [220, 221], i.e., the presence of at least one latent factor with zero variance, which causes matrices M_X and M_Y to be singular and renders our alignment strategy unstable. To avoid it, we devised a rejection scheme: after complete training of the model, we computed the variance of each latent vector and restarted the training, should one latent vector happen to be collapsed. If the training fails five times in a row, one latent factor is removed.

Naming d_X and d_Y the number of latent factors for source and target, respectively, this training step generates to set of encoders: $f_1^X, .., f_{d_X}^X$ for source and $f_1^Y, .., f_{d_Y}^Y$ for the target. Each of these functions maps a single cell expression profile to a latent factor score.

5.4.4. APPROXIMATION OF LATENT FACTORS EMBEDDING FUNCTION BY KERNEL RIDGE REGRESSION

In a second step, we approximate each embedding function ($f_1^X, .., f_{d_X}^X$ and $f_1^Y, .., f_{d_Y}^Y$) using Kernel Ridge Regression (KRR), with a view to exploit interesting mathematical properties of kernel methods (Section D.3). To generate enough training data, we exploit the generative nature of the two scVI models. We provide here the procedure for the source model, which can be readily applied to the target model as well (Sections D.4 and D.5).

First, we randomly sample N random noise vectors $z_1^X, .., z_N^X \sim \mathcal{N}\left(0, I_{d_X}\right)$, with I_{d_X} the identity matrix of size d_X. Using the source model decoder, we collect N artificial gene expression profiles $\hat{x}_1^X, .., \hat{x}_N^X \in \mathbb{R}^p$. As VAEs are not bijective, the original vectors $z_1^X, .., z_N^X$ do not perfectly equate the encoder outputs for $\hat{x}_1^X, .., \hat{x}_N^X \in \mathbb{R}^p$. To limit the noise in our model, we input the samples $\hat{x}_1^X, .., \hat{x}_N^X \in \mathbb{R}^p$ in the encoder functions, resulting in new artificial vectors $\hat{z}_1^X, .., \hat{z}_N^X \in \mathbb{R}^{d_X}$; each $\hat{z}_\cdot^X$ corresponds here to the mean of the embedding parametrization of a sample. We subsequently approximate the embedding functions by training, on the artificial data $\{(\hat{x}_1^X, \hat{z}_1^X), .., (\hat{x}_N^X, \hat{z}_N^X)\}$, d_X independent KRR models. To scale to the large amount of data we sampled, we turned to Falkon [180], a stochastic approximation of KRR which relies on the Nyström approximation [222]. Critically, Falkon relies on the choice of a kernel function, which in our case, must provide a universal approximation. We turned to the Laplacian, Matérn and Gaussian kernels, which are defined in Table 1.4.

Let us denote by K the kernel selected. Falkon selects M anchor points ($M << N$) among the training data, denoted $\tilde{x}_1^X, .., \tilde{x}_M^X$ and approximates the embedding functions f_k^X by

computing the matrix $\alpha^X \in \mathbb{R}^{d_X \times M}$ such that:

$$\forall k \in \{1,..,d_X\}, \quad f_k^X\left(x \in \mathbb{R}^p\right) \quad \approx \quad \sum_{i=1}^{M} \alpha_{k,i}^X K\left(\tilde{x}_k^X, x\right) \quad \widehat{=} \quad \theta_k^X(x). \tag{5.1}$$

Equation (5.1) corresponds to the regular basis-expansion form resulting from the Representer Theorem, with the M anchor points instead of the N training samples. It is however important to note that the coefficients α^X are optimized using the $N-M$ non-anchor points, and the optimization therefore exploits the whole dataset.

5.4.5. COMPARISON OF LATENT FACTORS BY SOBOLEV ALIGNMENT

We have approximated the source encoder functions $f_1^X,..,f_{d_X}^X$ by the kernel machines $\theta_1^X,..,\theta_{d_X}^X$, and the target encoder functions $f_1^Y,..,f_{d_Y}^Y$ by the kernel machines $\theta_1^Y,..,\theta_{d_Y}^Y$. These two approximations are then used to compute the cosine similarity matrix $\mathbf{M}^K$ (Definition D.5.10). We here present the computational approach to compute $\mathbf{M}^K$, and refer the reader to Section D.5 for the precise mathematical definition and derivation. Let us define the three following kernel matrices:

$$\begin{aligned} K_X &= \left(K\left(\tilde{x}_i^X, \tilde{x}_j^X\right)\right)_{1\leq i,j\leq M} \\ K_Y &= \left(K\left(\tilde{x}_i^Y, \tilde{x}_j^Y\right)\right)_{1\leq i,j\leq M} \\ K_{X,Y} &= \left(K\left(\tilde{x}_i^X, \tilde{x}_j^Y\right)\right)_{1\leq i,j\leq M} \end{aligned}. \tag{5.2}$$

These three matrices correspond to similarity (or kernel) values between the different anchor points computed by Falkon (Equation 5.1). We then define the three following matrices, referred to as un-normalized cosine similarity matrices (Proposition D.5.9):

$$\begin{aligned} \widetilde{\mathbf{M}}_X &= \alpha^X K_X {\alpha^X}^T \\ \widetilde{\mathbf{M}}_Y &= \alpha^Y K_Y {\alpha^Y}^T \\ \widetilde{\mathbf{M}}_{XY} &= \alpha^X K_{X,Y} {\alpha^Y}^T \end{aligned}. \tag{5.3}$$

The cosine similarity is then computed as follows (Definition D.5.10):

$$\widetilde{\mathbf{M}}^K = \widetilde{\mathbf{M}}_X^{-\frac{1}{2}} \widetilde{\mathbf{M}}_{XY} \widetilde{\mathbf{M}}_Y^{-\frac{1}{2}}. \tag{5.4}$$

5.4.6. COMPUTATION OF PRINCIPAL VECTORS AND PRINCIPAL ANGLES

We now have two sets of vectors which approximate the two VAEs: $\theta_1^X,..,\theta_{d_X}^X$ for source and $\theta_1^Y,..,\theta_{d_Y}^Y$ for target. To compare them, we use the notion of Sobolev Principal Vectors (SPVs) which correspond to pairs of vectors (one from source, one from target) ranked by decreasing similarity (Definition D.6.1). To compute them, we need to decompose the cosine similarity matrix (Equation 5.4) by SVD:

$$\widetilde{\mathbf{M}}^K = U \Sigma V^T. \tag{5.5}$$

The diagonal matrix Σ contains the similarity values between the principal vectors (Corollary D.6.2.1), and can be written as:

$$\Sigma = \begin{bmatrix} \cos\theta_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \cos\theta_{\widehat{d}} \end{bmatrix}, \quad \text{with} \quad \widehat{d} = \min(d_X, d_Y). \tag{5.6}$$

We also define the two matrices $\gamma^X \in \mathbb{R}^{\widehat{d}\times M}$ and $\gamma^Y \in \mathbb{R}^{\widehat{d}\times M}$ as

$$\gamma^X = U^T \widetilde{M}_X^{-\frac{1}{2}} \alpha^X \quad \text{and} \quad \gamma^Y = V^T \widetilde{M}_Y^{-\frac{1}{2}} \alpha^Y. \tag{5.7}$$

The principal vector pairs $\{(s_1, t_1), \cdots, (s_{\widehat{d}}, t_{\widehat{d}})\}$ are finally computed as (Theorem D.6.2):

$$\forall k \in \{1, \cdots \widehat{d}\}, \quad s_k = \sum_{i=1}^{M} \gamma_{k,i}^X K(\tilde{x}_i^X, \cdot) \quad \text{and} \quad t_k = \sum_{i=1}^{M} \gamma_{k,i}^Y K(\tilde{x}_i^Y, \cdot). \tag{5.8}$$

5.4.7. INTERPRETABILITY OF PRINCIPAL VECTORS BY KERNEL TAYLOR EXPANSION

We here present the interpretability scheme which relies on the explicit characterization of the Gaussian kernel RKHS [169] (Section D.7). We define the following univariate basis function e_i^k, for $i \in \{1, \cdots, p\}$ and $k \in \mathbb{N}$, as:

$$\forall x \in \mathbb{R}^p, \quad e_i^k(x) \;\widehat{=}\; \frac{x_i^k}{\sigma^k \sqrt{k!}} \exp\left(-\frac{x_i^2}{2\sigma^2}\right), \tag{5.9}$$

with σ the scale parameter of the Matérn kernel (Table 1.4). An orthonormal basis of the Gaussian RKHS can then be defined by combining these univariate basis functions with an outer product (Proposition D.7.3), yielding *Gaussian basis functions*, defined, for $K = [k_1, \cdots, k_p] \in \mathbb{N}^p$, as:

$$\forall x \in \mathbb{R}^p, \quad G_K(x) \;\widehat{=}\; \prod_{j=1}^{p} e_j^{k_j}(x) = \left[\prod_{j=1}^{p} \frac{x_j^{k_j}}{\sigma^{k_j}\sqrt{k_j!}}\right] \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right). \tag{5.10}$$

Since these Gaussian basis functions form an orthonormal basis, we exploit them to parametrize the latent factors and the SPVs (Subsection D.7.1). Based on Equation (5.10), we define the following offset matrices as the exponential term, i.e.,

$$\mathscr{O}^X = \operatorname{diag}\left[\exp\left(-\frac{\|\tilde{x}_i^X\|^2}{2\sigma^2}\right)\right]_{1\le i\le M} \quad \text{and} \quad \mathscr{O}^Y = \operatorname{diag}\left[\exp\left(-\frac{\|\tilde{x}_i^Y\|^2}{2\sigma^2}\right)\right]_{1\le i\le M}, \tag{5.11}$$

together with the artificial data matrices $\mathscr{A}^X = [\tilde{x}_1^X, \cdot, \tilde{x}_M^X] \in \mathbb{R}^{M\times p}$ and $\mathscr{A}^Y = [\tilde{x}_1^Y, \cdot, \tilde{x}_M^Y] \in \mathbb{R}^{M\times p}$, which correspond to the linear terms. Finally, we define:

- $L_{fact}^X \in \mathbb{R}^{d_X \times p}$ and $L_{fact}^Y \in \mathbb{R}^{d_Y \times p}$, which contain the contributions of each gene (columns) to each source and target latent factor (rows) respectively.

- $L^X_{SPV} \in \mathbb{R}^{\hat{d}\times p}$ and $L^Y_{SPV} \in \mathbb{R}^{\hat{d}\times p}$ containing the contribution of each gene (columns) to each source and target principal vector (rows) respectively.

These matrices of linear weights can be computed as follows (Theorem D.7.11):

$$L^X_{fact} = \frac{1}{\sigma}\alpha^X \mathscr{O}^X \mathscr{A}^X \quad \text{and} \quad L^Y_{fact} = \frac{1}{\sigma}\alpha^Y \mathscr{O}^Y \mathscr{A}^Y. \tag{5.12}$$

$$L^X_{SPV} = \frac{1}{\sigma}\gamma^X \mathscr{O}^X \mathscr{A}^X \quad \text{and} \quad L^Y_{SPV} = \frac{1}{\sigma}\gamma^Y \mathscr{O}^Y \mathscr{A}^Y. \tag{5.13}$$

Following a similar protocol, we can compute the contributions of each interaction term to the latent factors and the SPVs, and we refer the reader to Proposition D.7.3 and Definition D.7.6 for the complete definition.
To extend this interpretability scheme to other Matérn kernels, e.g., the Laplacian kernels, we proved the orthogonality of linear and interaction terms within the RKHS (Lemmas D.7.16 and D.7.17). In order to scale linear and interaction terms, we exploited a heuristic which consists in comparing the amount of non-linearities with results obtained using a Gaussian kernel with same parameter σ (Subsection D.7.5).

5

5.4.8. VISUALIZATION USING SOBOLEV PRINCIPAL VECTOR INTERPOLATION

In order to reduce the SPV pairs to a single vector, we followed the same protocol as in PRECISE and TRANSACT and devised an interpolation scheme (Figure D.6). If we denote by θ_k between vectors of the k$^{\text{th}}$ SPV, i.e., $\theta_k = \arccos\langle s_k, t_k\rangle_{\mathscr{H}}$, the projection on interpolation time $\tau \in [0,1]$ is defined as

$$F_k(\tau) = \frac{\sin(1-\tau)\theta_k}{\sin\theta_k} s_k + \frac{\sin\tau\theta_k}{\sin\theta_k} t_k. \tag{5.14}$$

For each SPV, we discretize $[0,1]$ and project the whole data on all the interpolated vectors. Using a Kolmogorov-Smirnov statistics, we compare source and target projected data and return the interpolation time τ_k which minimizes these statistics. We repeat this step for all SPVs and project the data on all the interpolated vectors.

5.4.9. GENE SET ENRICHMENT ANALYSIS

To connect the linear and interaction weights (Equations (5.12) and (5.13)) to known biological processes, we employ Gene Set Enrichment Analysis (GSEA) [223]. We process the linear and interaction parts in two distinct analyses. First, the linear weights are used as input to a PreRanked analysis with 1000 gene-level permutations. For the interaction terms, we compute enrichment score following a protocol inspired from the original GSEA framework. As interaction terms correspond to pairs of genes, we define for two gene sets G_1 and G_2 the interaction gene set $G_1 \times G_2$ as the set of gene pairs with one gene in G_1 and the other in G_2. To compute the enrichment score of $G_1 \times G_2$, termed $ES(G_1 \times G_2)$, we rank all interaction terms in decreasing order and assign to each interaction 1 if the interaction is in $G_1 \times G_2$ and -1 otherwise. We then divide the positive values by the number of elements in $G_1 \times G_2$ and the negative values by the number of interactions not in $G_1 \times G_2$. We finally compute the cumulative sum over the ranked set

of interactions and define $ES(G_1 \times G_2)$ as the cumulative sum with the largest absolute value.
Taking inspiration from the original GSEA protocol, we generated the null model by sample permutation: once the scVI model is trained and artificial data is generated, we randomly shuffled the sample labels of latent space values. We then ran the Sobolev Alignment protocol and computed the interaction weights. Although the linear analysis has been carried out with 1 000 permutations, we only performed 100 permutations for the interaction terms, out of computational time considerations. We finally computed the enrichment scores for each permutation and computed the normalized enrichment scores (NES) and FDR values as in the original GSEA.

6

DISCUSSION

Exploiting the high versatility and cost-effectiveness of model systems, such as cell lines and patient derived xenografts, the research community has gathered a vast amount of knowledge and data. These pre-clinical discoveries, however, happen to have a limited applicability in a clinical setting and most findings do not readily translate to actionable clinical insights. This is for instance the case for predictive biomarkers of drug response. To enhance the applicability of model-system-based biomarkers, we have developed computational methods to integrate pre-clinical and clinical data. In this final chapter, we summarize the key methodological aspects and findings of each developed method and discuss the limitations of our studies. The methods developed in this thesis contribute to wider scientific fields which we tried to advance. In this chapter, we attempt to position our work in the broader scientific community, allowing us to improve the methodologies we developed but to also scout for potential new application areas. Finally, we propose some general thoughts on were we stand in the development of computational models for drug response prediction.

6.1. USING GENE EXPRESSION PREDICTIVE POWER TO ENHANCE DNA-BASED PREDICTORS

When it comes to drug response prediction, gene expression offers the highest predictive performance, completely over-shadowing other data-types such as mutation, copy-number or methylation [46, 54, 55]. These DNA-based measurements, however, are clinically more actionable and their measurement is generally more robust. To bridge this gap and help transfer gene expression predictive power to DNA-based measurements, we developed **Percolate** (Chapter 2). Percolate takes as input two data-types, e.g. mutations and gene expression, and break these down into a joint and an individual signals. Using the out-of-sample extension we devised, a drug response predictor can be trained on multi-modal data, thereby exploiting the strength of gene expression, but used in practice when only one of the two modalities is present. We here discuss two main limitations of our approach and offer ways to circumvent it. We finally present ideas to scale

up this idea up to other applications.

6.1.1. LIMITATIONS OF THE JOINT SIGNAL

Modelling the joint signal between two data-modalities constitutes the cornerstone of Percolate and we show that, in our specific application, the individual signal is not predictive. However, we observe that the resulting mutation-based signatures do not offer a systematic gain in performance: using the joint signal leads to better predictions for 11 drugs, while it induces a drop in performance for 11 drugs, including 8 MAPK inhibitors. Computing the joint information between gene expression and mutations requires access to both modalities, and the imputation we perform to predict the joint signal solely from mutation seems to be sub-optimal. This sub-optimality could be alleviated by using a larger sample size, for instance using a compendium like TCGA [160], by using another distribution for gene expression or mutation, or by designing a better out-of-sample extension. These signatures were also computed using pan-cancer data and some tissue-specific mutational effects could therefore not be fully retrieved. However, restricting to a single tissue would significantly shrink the sample-size, which is not desirable. A plausible solution could be to train GLM-PCA in a few-shot-learning fashion [124, 224–226], exploiting other cancer types for pre-training and tuning on the tissue of interest. These remarks also apply to copy-number and methylation.

Furthermore, the individual signal is not necessarily irrelevant and can potentially contain information not joined with gene expression. Post-translational modifications form a counter-example where mutations have a limited effect on the gene expression while having a large phenotypic effect. This is the case of mutations in genes coding for histone acetyl and methyl modifiers, like EZH2, which have been shown to exert a direct impact in various pathologies like folicular lymphomas [227] or the Weaver syndrome [228], while having a limited and convoluted impact on gene expression. Such weak signal would require larger samples to be detected efficiently, which can explain the conclusion of our study. Designing predictors which take into account both the joint and the individual signal, while exploiting the stratification, is a potential fruitful avenue.

6.1.2. LIMITATIONS OF THE EXPONENTIAL FAMILY

Another key contribution of our work consists of extending the popular JIVE framework from Gaussian noise models to the broader class of exponential family distributions. Although this constitutes, we believe, an important improvement in multi-omic modelling, our approach has so far been technically limited in two aspects. First, we restricted our study to distributions already used in the literature. Designing an exponential-family distribution based on external data is one avenue to incorporate prior knowledge in Percolate. For instance, one could think of modelling a log-partition function on a wide compendium of external data, allowing for instance to encode prior-knowledge on the interactions between genes. This modelling is however not trivial and would require some theoretical and methodological advances. Provided such methodology is developed, combining the resulting exponential family distribution with Percolate would be relatively straight-forward.

Furthermore, the exponential family distribution itself can potentially be restrictive. For instance, although the choice of a Negative Binomial distribution was natural for gene

expression, technical adaptations were needed as the Negative Binomial is not an exponential family distribution. Employing Percolate in applications where an exponential family distribution is not a relevant modelling choice would thus require to extend our methodology past this hypothesis. We argue that Normalizing Flows (NF) form an attractive solution [229]. Formally, a Normalizing Flow corresponds to a bijective function $f : \mathbb{R}^d \to \mathbb{R}^p$ which reconstructs a data-matrix $X \in \mathbb{R}^{n \times p}$ from d factors assumed to follow a multivariate normal distribution. Taking the notations from Chapter 2, if we consider two trained normalizing flows f_A and f_B, one can compute the scores $Z_A = f_A^{-1}(X_A)$ and $Z_B = f_B^{-1}(X_B)$, subsequently whitened by SVD, i.e., $Z_A = U_A \Sigma_A V_A^T$ and $Z_B = U_B \Sigma_B V_B^T$. The matrix M computed by stacking U_A and U_B can then be decomposed by SVD and the resulting out-of-sample is obtained, using notations from Chapter 2, as:

$$\begin{aligned} \widetilde{U}_{J,A} &= f_A^{-1}(X_A) V_A \Sigma_A^{-1} V_{M,A} \Sigma^{-1} \\ \widetilde{U}_{J,B} &= f_B^{-1}(X_B) V_B \Sigma_B^{-1} V_{M,B} \Sigma^{-1}. \end{aligned} \tag{6.1}$$

6.1.3. IMPLICATIONS AND POTENTIAL NEW APPLICATIONS OF PERCOLATE

The results of our study suggest that the Percolate predictors perform better than a standard model for a subset of drugs. As gene expression is laborious and costly to obtain, and as it suffers from severe batch effects, being able to obtain better predictions from DNA-based measurements can have an important clinical impact. For instance, the Hartwig Medical Foundation (HMF) has spearheaded the systematic whole-exome-sequencing profiling of metastatic patients who have exhausted all lines of treatment. An improved mutation-based predictor of drug response could be a powerful companion for clinicians in their complex decision making process.

Finally, non-gaussian multi-modal data is absolutely ubiquitous: many areas, for instance in climate science or supply chain optimization, study phenomenon which are intrinsically non Gaussian. Should experts in these fields be interested in a similar data aggregation strategy, we believe that our tool could easily be applied. Our software implementation has been specifically thought out to be easily adaptable to new classes of data.

6.2. TRANSFERRING PREDICTORS OF DRUG RESPONSE FROM CELL LINES TO TUMORS

As gene expression offers the highest predictive performance compared to other data-types, we then set out to correct gene expression to translate drug response biomarkers from pre-clinical models to tumors. In chapter 3, we developed **PRECISE**, a linear domain adaptation technique which extracts the biological signal present in both cell lines and tumors. To account for complex biological mechanisms, we subsequently derived **TRANSACT** (chapter 4) exploiting the notion of kernel machines (Subsection 1.3.1). We present two clinical and two technical limitations of these two approaches, and propose a potential road-map to accelerate the improvement, and potential adoption, of such methodologies in a clinical setting.

6.2.1. CLINICAL LIMITATIONS

The results presented in chapters 3 and 4 improve significantly on existing approaches. However, the results are still far from warranting a clinical applicability and several limitations would need to be addressed in future work. First, there is a mismatch between the clinical data used to assess the accuracy of our prediction and the pre-clinical training data (Subsection 4.2.4). The two clinical datasets used in our approach, TCGA and HMF, have been gathered over extensive periods of time and are the product of clinical decisions. To increase the likelihood of response in patients, the standard-of-care usually consists of giving drugs in combination; for instance, BRAF-mutated lung adenocarcinoma are usually treated jointly with Trametinib (MEK inhibitor) and Dabrafrenib (BRAF inhibitor), and metastatic colorectal cancers are treated with a combination of Irinotecan, 5-Fluorouracil and Leucovorin [230]. As a result, an important portion of patients in these datasets have been treated with more than one drug. The comparison we performed between the different method is therefore incomplete, but could not be much improved in absence of large-scale monotherapy-treated patient panels. Recent combination screens, where a comparable number of cell lines are screened for various drug combinations, could help refine our framework [231].

Another limitation stems from cell lines themselves. PRECISE and TRANSACT computationally correct for differences between pre-clinical and clinical data, but these two methods cannot account for complex effects not observed *in-vitro*. Interplay between cancer cells and surrounding stromal and immune cells cannot for instance be modeled, although these are key determinants of response [232]. Recent work attempted to replicate *in-vitro* the effect of the immune system, for instance by co-culturing cancer cells and T-cells [233], paving the way for large-scale datasets which could help improve the performance of our approach. As discussed in Subsection 1.1.4, more advanced model systems like organoids could also be instrumental in improving the clinical relevance of drug response prediction models, but their current cost prohibits the formation of large-scale data.

6.2.2. TECHNICAL LIMITATIONS

TRANSACT is by design a kernel method, and thus heavily relies on the choice of kernel. For our study, we exploited the Gaussian kernel, which forms a very standard choice. The Gaussian kernel has however a major drawback: it is isotropic, meaning that perturbations with equal norms will yield the same similarity to a given sample. For instance, a knock-down of P53 will be as distant to a cancer cell as the same perturbation diluted along all the genes, while one would expect the former to be more distant than the latter. To circumvent this issue, prior knowledge can be incorporated inside the kernel, for example, by considering only interactions between genes and their transcription factors. Another strategy consists of considering kernels involving small subsets of genes known to be involved in similar processes; these kernels could then be combined using Multiple Kernel Learning, as is proposed in PIMKL [154] or PAMOGK [234].

Finally, by design, PRECISE and TRANSACT were restricted to gene expression. Although gene expression recapitulates a significant portion of the signal from DNA-based measurements (Chapter 2), other important molecular features are not taken into account. This is the case for post-translational modifications like ubiquitylation or phosphoryla-

tion which critically measure the activity of important pathways, e.g., the MAPK activity (Subsection 1.2.1). Although complementing large-scale cell line panels with such molecular measurements is hindered by the current cost of existing technologies, we envision that such molecular features could markedly help improve the clinical performance of our method. Sustained efforts in technological development plead in favor of optimism when it comes to the availability of such data in the foreseeable future

6.2.3. A ROAD-MAP TO CLINICAL ADOPTION

Although these limitations are all individually complex, we believe them to be surmountable. To reach clinical applicability, i.e. the massive adoption of such predictive tools in standard clinical practice, we here advocate for the following strategy which entails several iterations between experiments and computational advances. As explained above, a key technical limitation lies in the choice of a kernel and no existing kernel, to the best of our knowledge, would be relevant to our problem. To design a relevant kernel, or similarity function, we must set out to find a vast amount of data. A natural idea consists of increasing the number of cell lines. Unfortunately, this may be out of reach: given the complexity to establish a cell line, the number of available pre-clinical models will most likely not scale to more than a few thousands, at least not in the foreseeable future.
This first issue can be tackled by harnessing the wealth of data provided by recent advances in single cell sequencing. Recently published panels indeed scale to more than 100 000 cells and, as costs rapidly decrease, we believe panels consisting of millions of cells to be soon available. Such datasets are however unlabelled, i.e., the drug response of each cell is unknown, which calls for new methods for designing the kernel. We advocate for a **deep kernel learning** strategy where the kernel, or similarity function, is learned in an unsupervised manner and only then applied in the TRANSACT framework. For this purpose, we propose to exploit the model presented by *Wilson et al* [235] which combines a neural network with a Gaussian kernel (Table [**?**]) in a Gaussian-Process regression model. As our model must be unsupervised, we present in Chapter E an adaptation of the notion of Gaussian Process Latent Variable Model (GP-LVM) [76] which allows to train such deep kernel without labels. This procedure can then first be used using abundant scRNA-seq data to learn the kernel function. The resulting kernel would incorporate as much prior-knowledge as possible and would therefore be much more specific than the vanilla Gaussian kernel we employed. Such kernel, with probable modifications due to differences in library size, would then directly be used in the TRANSACT framework with bulk-RNA-seq data, i.e. computing the consensus features between cell lines and tumors and training a regression models after projection onto these consensus features. As the deep kernel is already low-rank, refinements in the TRANSACT alignment procedure are also possible.
This strategy presents many scientific opportunities as countless questions are still open, for instance:

- How to efficiently translate single-cell-level to bulk-level models ?
- How to integrate different multi-omic measurements in a transferable predictor ?
- How to extend the methodology to combine cell line, mouse and organoids, and subsequently transfer to humans ?

- Is the level of intra-tumor heterogeneity in model systems sufficient to properly model the diversity observed in patients ?
- How to prioritize cell lines and conditions to generate data relevant for improving the model ?

6.3. COMPARING CELL LINES AND TUMORS AT SINGLE-CELL RESOLUTION

6.3.1. SOBOLEV ALIGNMENT RECAPITULATES THE BIOLOGY COMMON TO TUMORS AND CELL LINES

As explained in Subsection 1.1.1, tumors never form monoclonal populations and large deviations are observed within any tumor [236]. To increase the granularity of the previous analyses (Chapters 3 and 4), we compared cell lines and tumors at a single cell resolution, and developed a novel computational tool, **Sobolev Alignment** (Chapter 5). We first discuss some important biological and technical limitations of our approach, alongside some directions to improve our work. As Sobolev Alignment critically relies on the notion of Sobolev space, we then take a step back and discuss the reasons behind this particular technical choice. Finally, we explain how our methodology can be employed in any field where generative models have proven useful.

6.3.2. BIOLOGICAL AND TECHNICAL LIMITATIONS

Similarly to Chapters 3 and 4, our study is limited to gene expression. Recent technologies, such as chromatin accessibility (scATAC-seq), spatial transcriptomics (FISH), (phospho-)proteomics (CITE-seq) or ribosomal activity (Ribo-seq) could offer complementary information. Such combined analysis could for instance be useful when studying drug resistance to BRAF inhibition in skin melanoma. *Shaffer and colleagues* [236] have indeed shown that, upon induction of Vemurafenib (Table 1.1), a small subset of cells undergo an epigenetic reprogramming which renders them insensitive to BRAF inhibition. Drug perturbation screens combined with measurements of gene expression, methylation levels and chromatin accessibility would be instrumental to better understand this mode of drug resistance, and persistence in general [237, 238]. Recent deep generative models have been designed for such combined data, e.g. MultiVI [209] or TotalVI [210], and our framework and software could easily be modified to accommodate such models.

Furthermore, the latent variable model we exploit, scVI, does not encode any prior knowledge about the gene regulatory network. As a result, the model relies on all of the associations present in the data and is therefore prone to spurious correlations ; such a phenomenon is exemplified in computer vision, where neural networks have been shown to exploit *shortcuts* [239]. On top of limiting the generalization of the model, these shortcuts can potentially harm an analysis by selecting wrong associations. Recent work successfully developed ways to inject prior knowledge in the neural network. An example are visible neural networks [240], like DrugCell [241, 242], where gene-ontology is exploited to build the neural network architecture. Associations between non-related genes would therefore be discarded, limiting the impact of spurious correlations. This

line of research has been further developed in other recent work, like BDKANN, expiMap [243] or GLUE [244]. Another example, coming from computer vision, is the study and the design of inductive biases: the Hough transform, for instance, hardcodes the notion of lines in a neural network [245]. Devising inductive biases relevant for genomic data could similarly help the model focus on relevant patterns and limit the impact of shortcuts.

Our study is also limited in scope by the absence of drug response data in patients. Although we could recapitulate known modes of action in the McFarland multiplexed drug screen, such perturbations could not readily be translated to the patients as no drug response was, at the time of study, available. As the costs of scRNA-seq technologies and experiments decrease, we expect large-scale atlases of cells extracted from cancer patients to become available for researchers in a near future, ideally containing pre- and post-treatment data. These datasets could be employed in our workflow, and cell-line perturbations could be subsequently predicted using recent computational approaches like scGen [246] or refinements of recent differential abundance testing methods, e.g. MILO [247] or MELD [248]. One important open problem would, in our opinion, still remain: to have clinical impact, the biological effect of any given perturbation would need to be systematically inferred, e.g., cell-cycle arrest, senescence, immune activation and apoptosis. In Dr.VAE [51], *Rampasek and colleagues* have shown that perturbation data can help improve drug response prediction when using bulk RNA-seq data; extending this model to scRNA-seq data and to more cellular phenotypes could help tackle this first problem.

Finally, we focused our study on NSCLC, a lung cancer sub-type for which a wealth of data is available, rendering it conducive to deep generative modelling. However, for rare cancers where such large datasets are virtually non-existent, we would not recommend using our methodologies, as VAEs require large training datasets to converge to a meaningful representation. Few-shot learning [225, 226], or meta-learning [224], could offer a solution in such small sample regimes. *Ma and colleagues* have for instance proposed a method, TCRP [124], which learns drug response cancer-type by cancer-type and adapts the model to cancer sub-types with a small amount of samples. A similar strategy could replace the current dimensionality reduction step, with a VAE trained on NSCLC, breast and skin cell lines, and subsequently adapted to a rare cancer type.

6.3.3. ON THE TECHNICAL CHOICE OF SOBOLEV SPACES

Our methodology is built on several mathematical concepts, and we here present and discuss the thought-process and explain the decisions which led to our derivation. Starting from the idea that we wish to compare encoder functions of a VAE *à la* PRECISE and TRANSACT, we looked for a way to compute a measure of similarity between such functions. We specifically turned to inner products in Hilbert spaces as these offer attractive mathematical properties (positive-definiteness, linearity, symmetry) and are particularly amenable to the framework developed in Chapter 5. Several functional Hilbert spaces have been proposed in the literature, but these rely on the computation of an integral in a high-dimensional space, thereby suffering from the curse of dimensionality. Interestingly, Reproducing Kernel Hilbert Spaces (RKHS, Subsection 1.3.1), used in the derivation of TRANSACT, have an associated inner product which can be computed in

closed-form for kernel machines. We therefore reasoned that writing VAE encoder functions as kernel machines would be a fruitful avenue.
The choice of an RKHS, and thus its associated kernel, is crucial. Indeed, it conditions the similarity measure used to compare functions: it needs to be computationally tractable, and the associated space must be large enough to provide a good approximation of the encoding functions. Interestingly, Matérn kernels provide a good solution to these three requirements. First, the norm associated to the Matérn kernel is equivalent to a Sobolev norm, which corresponds to the L_2 inner product between high-order derivatives and is frequently used in the literature to compare functions. Second, the Matérn kernels can efficiently be computed for a large number of samples, using the *kernel trick*. Last, but certainly not least, the RKHS of a Matérn kernel, which we denote by $\mathscr{H}$ for the rest of this section, corresponds to a Sobolev space, which is dense in L_2. This last argument means that any function in L_2 can be approximated by a sequence of functions from the Matérn kernel RKHS; formally, if $f \in L_2(\mathbb{R}^p)$, then there exists a sequence $f_1, f_2, .., \in \mathscr{H}$ such that

$$\|f_n - f\|_{L_2(\mathbb{R}^p)} \underset{n\to+\infty}{\longrightarrow} 0. \tag{6.2}$$

If we generate a sequence of kernel machines $f_1, f_2, ..$, then with a large enough n, f_n would provide a good approximation of f in L_2-norm sense, which we could use in a mathematical framework similar to the one developed for TRANSACT. Generating such a sequence faces two other challenges: the poor-scalability of kernel methods and the absence of sufficient training data. The second point is particularly critical: as we exploit VAEs for their high-expressivity, one must not expect a kernel machine trained on the same data to recapitulate the richness of the VAE's representation. To tackle this problem, we exploit the idea that VAEs can generate points from the data distribution it models. We can therefore generate a very large number of cells, possibly as much as one wants, and thereby establish a training data set composed of the artificial gene expression profiles, used as input to construct the encoding function in a kernel ridge regressor (KRR). To tackle the scalability-issue of kernel methods, we turned to randomized kernel methods, and specifically Falkon [179, 180], based on the Nyström approximation. We show that the KRR approximation provides good concordance with the original VAE encoding functions. We finally exploited the attractive properties of RKHS referred to above to compare the encoding functions.
This journey into the construction of Sobolev Alignment highlights key technical decisions we made, which are subject to discussion and improvements. A first very fundamental issue lies in the choice of a Sobolev Space as the approximation RKHS. These spaces constitute, we believe, a natural first choice, especially due to the expressivity of the Laplacian kernel [212, 213] and the explicit characterization of the Gaussian kernel RKHS [169]. A richer RKHS could be used to increase the fidelity with which the VAE encoder functions are approximated, using for instance an exponential power kernel [213, 249], defined, for $\gamma > 0$ and $\sigma > 0$, as:

$$\forall x, y \in \mathbb{R}^p, \quad K_\gamma(x, y) = \exp\left(-\frac{\|x - y\|^\gamma}{\sigma}\right). \tag{6.3}$$

As the RKHS of the exponential power kernel contains, for $\gamma < 1$, the RKHS of the Laplacian kernel [213], we expect the approximation, and therefore the alignment, to be more informative. More theoretically, we could extend the RKHS employed in our alignment by using the notion of Powers of RKHS [250], which is defined based on the Mercer-decomposition of the kernel. Although, to the best of our knowledge, Matérn kernels do not admit analytical closed-form definitions for the kernel associated with the powers of RKHS, random features could offer an interesting technical option [251].
Finally, an important source of improvement lies in the regression model. Although it already provides a good approximation, Kernel ridge regression has a major drawback: its is locally adaptive, meaning that the function will present the same degree of smoothness for all gene expression profiles. Deep neural networks, on the contrary, can be very rough in selected areas of the gene expression space and very smooth in others, as is exemplified by recent works [252, 253]. One way to circumvent the issue while conserving the kernel ridge regression could be to preferentially select points harboring a high gradient's Frobenius norm.

6.3.4. IMPLICATIONS IN MACHINE LEARNING

Although we applied our methodology to a specific application in genomics and computational biology, we believe our approach to have implications beyond bioinformatics or cancer research. Our work indeed provides a general framework for comparing and interpreting sets of deep generative models, which are increasingly being used or researched in fields like computer vision, fluid mechanics [254], inverse problems [255] or physics [256]. Should a scientist want to compare the generative models associated with two different datasets, the modularity of our supporting software package would allow them to apply our methodology with a minimal amount of overhead. Furthermore, although we focused here on Variational Auto-Encoders, a similar strategy to ours could be employed with other generative models, such as Generative Adversarial Networks (GAN). Finally, the interpretability scheme we developed could similarly be readily translated to other applications.
More theoretically perhaps, we believe our framework to be relevant for the empirical study of learning dynamics [215, 257]. As observed in Chapter 4, deep neural networks heavily rely on their initialization, and their optimization is at the heart of current machine learning research. Our framework provides a way to quantitatively assess the similarity between two neural networks, for instance trained on the same data but with different initialization, or at different steps of the learning procedure. Another example of intriguing behavior is the so-called double-descent phenomenon [258–261] which shows that the learning curve for certain classes of learners happen to contradict the classical bias-variance trade-off. Comparing deep generative models with increasing capacity could be an interesting direction to study this phenomenon.

6.4. SPECULATIONS ON THE FUTURE OF DRUG RESPONSE PREDICTION

We believe predicting drug response in patients to be a complex task: state-of-the-art predictive performance in model systems are low, and transferring these predictors lead

to even lower performance. Nonetheless, solving this problem is of utmost importance on the road towards precision medicine, as it would allow physicians to offer cancer patients bespoke lines of treatment with high probability of success. We here discuss two potential future directions in this field.

6.4.1. A SHIFT IN PERSPECTIVE IN THE DESIGN OF MACHINE LEARNING MODELS

Machine learning, and especially deep learning, have recently gathered a significant amount of attention due to impressive breakthrough research in computer vision [262], natural language processing [263] or reinforcement learning [264, 265]. Inspired by these very impressive results, and strongly helped by high-quality software packages like PyTorch [83], we, as researchers in drug response prediction, have naturally attempted to transfer these successes to our field. Although this scientific strategy is sensible, I here argue that there is a fundamental reason why these models do not achieve in drug response prediction, and genomics more broadly, the same success they met in the aforementioned fields.

Computer vision and natural language processing aim at automating a task humans know how to perform. On the contrary, there is no gold-standard on how to generally treat a patient, on how a cell would respond to a given stimuli, or even on how gene expression is regulated. In genomics, machine learning is therefore used to understand and generate knowledge on processes we incompletely understand. The successful models developed in computer vision and natural language processing, on top of relying on significantly larger datasets, are loaded with design biases (often unconscious) which are not necessarily applicable to the problem computational biologists try to solve. As a consequence, complex models from other fields are useful as inspiration, but should not be expected to directly solve any task they are being tried on. Instead, the synergy between expertise is, in our opinion, the best path towards success in our field. The impressive results obtained by the AlphaFold team provide a compelling example [266]: by building a diverse team and using tools from different fields, DeepMind managed to leapfrog advances in protein folding prediction.

6.4.2. GENOMIC EQUIVARIANCES FOR TACKLING THE SMALL-DATA REGIME

Predicting drug response is bound to face a small data-regime, as the number of cancer patients will hopefully never reach the size of datasets like ImageNet. On top of few-shot learning approaches previously discussed, data-augmentation schemes have been recently proposed to tackle this lack of data. For instance, *Yoon and colleagues* [267] proposed VIME, a method tailored to tabular data which elegantly exploits corrupted inputs to increase the size of the training set. This data-augmentation strategy, however, does not exploit much prior knowledge, contrary to recent developments in machine learning which exploit symmetries, and more broadly equivariance [268], with particular success in medical imaging [269].

Translating such success in genomics is however not trivial, as there exists no clear geometrical invariant for gene expression or other genomic profiles. Nonetheless, we argue that such equivariance must exist, starting, for instance, from the cell cycle. Geometric deep learning [270] offers a comprehensive framework for modelling such behaviors,

and we believe that the design of inductive biases for genomic data is one key to unlock the success of drug response prediction models.

6.4.3. THE PATH TOWARDS AN INTEGRATED MODEL FOR CLINICAL PREDICTION

Finally, we would like to highlight a political aspect of drug response prediction which, we believe, will be crucial in the years to come. As discussed at length in this chapter, improvements over existing methods would require the integration of heterogenous datasets coming from different molecular levels, different technologies, but also, and perhaps more importantly, from different labs scattered all across the globe, and oftentimes competing with one another. Such tasks cannot be completed without the establishment of clear guidelines for data generation and processing and a complete cooperation between international research institutes. Only then, in our opinion, will we as a field be ready to bring about an actionable drug response prediction tool which could accurately and cost-effectively find the best cure for each patient.

Most of the research presented in this thesis was performed during the Covid-19 pandemic, during which scientists from all countries have joined forces to accelerate scientific discoveries. Let it provide us with optimism on our collective capacity to tackle the aforementionned political issues in the coming years.

A

SUPPLEMENTARY MATERIAL FOR CHAPTER 2

A.1. NEGATIVE BINOMIAL PARAMETRIZATION

A.1.1. COMMON PARAMETRIZATION

The Negative Binomial distribution is commonly defined using two parameters:

- $r > 0$, called the **inverse-dispersion**.
- $p \in [0,1]$, called the **success probability**.

The negative binomial distribution models the number of successes observed when consecutive Bernoulli experiments with probability p are carried out and stopped after r failures. Its probability distribution is defined on $\mathbb{N}$ as follows,

$$\forall k \in \mathbb{N}, \quad \mathbb{P}(X = k) = \frac{\Gamma(k+r)}{\Gamma(k+1)\Gamma(r)} (1-p)^r p^k, \tag{A.1}$$

where Γ stands for the Gamma-distribution defined as

$$\forall z > 0, \quad \Gamma(z) = \int_0^{+\infty} t^{z-1} e^{-t} dt. \tag{A.2}$$

If we re-write Equation (A.1) in an exponential form, we obtain:

$$\forall k \in \mathbb{N}, \quad \mathbb{P}(X = k) = \frac{\Gamma(k+r)}{\Gamma(k+1)\Gamma(r)} \exp\left[r \ln(1-p) + k \ln p\right]. \tag{A.3}$$

In Equation (A.3), we easily recognise an exponential form when r is fixed with p as parameters. However, this formulation is computationally challenging, as observed values of p tends to accumulate around 1. This leads to instabilities in the optimisation and motivate the utilisation of another parametrization.

A

A.1.2. PARAMETRIZATION EMPLOYED IN OUR APPROACH

We turned to the parametrization proposed by *Risso et al* [181], which construction we present here. This parametrization stems from the observation that, if Z follows a Negative Binomial distribution with parameters p and r, then $\mathbb{E}(Z) = \frac{pr}{1-p}$. This expectation belongs to $\mathbb{R}^+$ which still contains one constraint. This constraint can easily be alleviated by log-transform, leading to the following parametrization:

$$\theta = \ln \frac{pr}{1-p}. \tag{A.4}$$

Equivalently, we have $p = \frac{r}{e^{\theta}+r}$, which combined with Equation (A.3) yields the exponential form presented in the main text.

A.1.3. TECHNICAL IMPLEMENTATION

The Negative Binomial distribution $\mathrm{NB}(\theta, r)$ depends on two parameters $\theta \in \mathbb{R}$ and $r > 0$. When the r parameter, called **inverse-dispersion**, is fixed, the Negative Binomial distribution belongs to the exponential family. We use the parametrization used by *Risso et al* [181] which yields, with fixed-parameter $r > 0$, the exponential-family functions explicited in Supp. Table 1. We compute for each gene (feature) $j \in \{1, .., p\}$ a dispersion parameter r_j using DESeq2 [271]. Once these dispersion parameters set, the parametrization can be exploited in Percolate.

A.2. BETA PARAMETRIZATION

A.2.1. COMMON PARAMETRIZATION

The Beta distribution is commonly parametrized by two parameters, $\alpha > 0$ and $\beta > 0$, called **shape** parameters. If a random-variable Z follows the Beta distribution with parameters α and β, then its probability density function $f(\cdot; \alpha, \beta)$ is defined as

$$\forall z \in [0,1], \quad f(z; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} z^{\alpha-1}(1-z)^{\beta-1}. \tag{A.5}$$

Noting that $\mathbb{E}[Z] = \frac{\alpha}{\alpha+\beta}$, we can use another parametrization which brings the saturated parameters to be the data-expectation, as is the case for other distributions like Gaussian, Bernoulli, Poisson or Negative-Binomial.

A.2.2. PARAMETRIZATION EMPLOYED IN OUR APPROACH

Using the parametrization from Equation (A.5), we define the two parameters θ and ν as:

$$\theta = \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \nu = \alpha + \beta. \tag{A.6}$$

The re-parametrization from Equation (A.6) can be reversed, yielding $\alpha = \theta\nu$ and $\beta = (1-\theta)\nu$. The probability density function from Equation (A.5) then becomes

$$\forall z \in [0,1], \quad f(z; \theta, \nu) = \frac{1}{1-z} \exp\left[\begin{bmatrix} \theta\nu \\ (1-\theta)\nu \end{bmatrix}^T \begin{bmatrix} \ln z \\ \ln(1-z) \end{bmatrix} - \ln \frac{\Gamma(\theta\nu)\Gamma((1-\theta)\nu)}{\Gamma(\nu)} \right]. \tag{A.7}$$

From Equation (A.7), one can easily derive the A, T and η functions listed in Table 1 of the main manuscript.

A.2.3. COMPUTING SATURATED PARAMETERS

Let $x \in [0,1]$ and $\nu > 0$, the saturated parameter $\widetilde{\theta}$ solves the following equation

$$\frac{df}{d\theta}\left(x;\widetilde{\theta},\nu\right) = 0. \tag{A.8}$$

We denote by ψ the **digamma** function defined as the derivative of the log-Gamma function:

$$\forall t > 0, \quad \psi(t) = \frac{d\ln\Gamma}{dt}(t) = \frac{1}{\Gamma(t)}\frac{d\Gamma}{dt}(t). \tag{A.9}$$

Combining Equations (A.7) and (A.9), and noting that $r > 0$, the saturated parameter $\widetilde{\theta}$ is the solution of

$$\ln\frac{x}{1-x} + \psi\left((1-\theta)\nu\right) - \psi(\theta\nu) = 0. \tag{A.10}$$

We define g as $g : \theta \mapsto \ln\frac{x}{1-x} + \psi\left((1-\theta)\nu\right) - \psi(\theta\nu)$. We have:

$$\forall \theta > 0, \quad \frac{dg}{d\theta}(\theta) = -\nu\left[\psi^{(1)}\left((1-\theta)\nu\right) + \psi^{(1)}(\nu\theta)\right], \tag{A.11}$$

where $\psi^{(1)}$ represent the first-order derivative of the digamma function. $\psi^{(1)}$ can be efficiently computed using its integral representation:

$$\forall z > 0, \quad \psi^{(1)}(z) = -\int_0^1 \frac{t^{z-1}}{1-t}\ln t \, dt \tag{A.12}$$

Using Equation (A.12), we obtain

$$\forall \theta > 0, \quad \frac{dg}{d\theta}(\theta) = \nu\left[\int_0^1 \frac{t^{\theta\nu-1} + t^{\nu-\theta\nu-1}}{1-t}\ln t \, dt\right] \geq 0. \tag{A.13}$$

Using the monotonicity of g, we easily solve Equation (A.8) by means of a dichotomy. A Newton method would have also been possible, but would have required to set a learning rate. Our approach allows to avoid such additional hyper-parameters.

A.2.4. TECHNICAL IMPLEMENTATION

The Beta distribution $B(\theta,\nu)$ depends on two parameters $\theta \in [0,1]$ and $\nu > 0$ and belongs to the exponential family, defined, for $x \in [0,1]$ as in Table A.1. Since the Beta distribution has two natural parameters, we took inspiration from the Negative Binomial and fixed ν at the gene-level: since θ can be intuitively understood as the mean of the distribution, we reasoned that it would act as a good parameter for GLM-PCA. We computed the parameters $\nu_1, .., \nu_p$ by maximizing the likelihood for each gene and used the resulting distributions in Percolate.

A

A.3. BERNOULLI PARAMETRIZATION

The natural parameters from the Bernoulli distribution are either $-\infty$ or ∞ and thus not computationally tractable. Following *Landgraf et al* [82], we employ a thresholding at 25 for all entries of the matrix $\widetilde{\theta}$; this choice did not impact results (Supp. Figure 2).

A.4. DERIVATION OF THE OUT-OF-SAMPLE EXTENSION

The matrix M is defined in the main text as:

$$M = \left[\widetilde{U}_A, \widetilde{U}_B\right] = U_M \Sigma_M V_M^T \quad . \tag{A.14}$$

Without loss of generality, we restrict to non-singular directions, i.e. $\Sigma_{i,i} > 0 \forall i$. Using the decomposition $V_M = \left[V_{M,A}^T \; V_{M,B}^T\right]^T$ and the fact that $V_M^T V_M = I$ proposed in the text, we obtain directly:

$$\begin{aligned} U_M &= \left[\widetilde{U}_A, \widetilde{U}_B\right] \begin{bmatrix} V_{M,A} \\ V_{M,B} \end{bmatrix} \Sigma_M^{-1} \\ &= \widetilde{U}_A V_{M,A} \Sigma_M^{-1} + \widetilde{U}_B V_{M,B} \Sigma_M^{-1}. \end{aligned} \tag{A.15}$$

If we further note that $\left(\widetilde{\theta}_A - \widetilde{\mu}_A\right) \widetilde{V}_A^T \widetilde{V}_A = \widetilde{U}_A \Sigma_A W_A^T$ and $\left(\widetilde{\theta}_B - \widetilde{\mu}_B\right) \widetilde{V}_B^T \widetilde{V}_B = \widetilde{U}_B \Sigma_B W_B^T$ (definition of scores), we obtain:

$$\begin{aligned} \widetilde{U}_A &= \left(\widetilde{\theta}_A - \widetilde{\mu}_A\right) \widetilde{V}_A^T \widetilde{V}_A W_A \Sigma_A^{-1} \\ \widetilde{U}_B &= \left(\widetilde{\theta}_B - \widetilde{\mu}_B\right) \widetilde{V}_B^T \widetilde{V}_B W_B \Sigma_B^{-1}. \end{aligned} \tag{A.16}$$

Combining Equations (A.15) and (A.16) yields the out-of-sample projection presented in the main text.

Name	$\mathcal{X}$	$T(x)$	$\eta(\theta)$	$A(\theta)$	g^{-1}	$\mathbb{E}[T(X)]$
Gaussian	$\mathbb{R}$	x	θ	$\theta^2/2$	x	θ
Bernoulli	$\{0,1\}$	x	θ	$\log\left(1+e^{\theta}\right)$	$+\infty$ if $x=1$ else $-\infty$	$\left(1+e^{-\theta}\right)^{-1}$
Poisson	$\mathbb{R}_+$	x	θ	e^{θ}	$\log(x)$	
Negative Binomial	$\mathbb{R}_+$	x	$\log\frac{r}{e^{\theta}+r}$	$r\log\left(1+re^{-\theta}\right)$	$2\log r-\log x$	$r^2e^{-\theta}$
Beta	$[0,1]$	$\begin{bmatrix}\log x\\ \log(1-x)\end{bmatrix}$	$\begin{bmatrix}\theta\nu\\ (1-\theta)\nu\end{bmatrix}$	$\log\frac{\Gamma(\theta\nu)\Gamma((1-\theta)\nu)}{\Gamma(\nu)}$	untractable	$\begin{bmatrix}\psi(\theta\nu)-\psi(\nu)\\ \psi((1-\theta)\nu)-\psi(\nu)\end{bmatrix}$
Log-normal	$\mathbb{R}_+$	$\begin{bmatrix}\log x\\ (\log x)^2\end{bmatrix}$	$\begin{bmatrix}\theta/\nu^2\\ -1/(2\nu^2)\end{bmatrix}$	$\frac{\theta^2}{2\nu^2}+\log\nu$	$\log x$	$\begin{bmatrix}\theta\\ \theta^2-\sigma^2\end{bmatrix}$
Gamma	$\mathbb{R}_+$	$\begin{bmatrix}\log x\\ x\end{bmatrix}$	$\begin{bmatrix}\theta\\ -\nu\end{bmatrix}$	$\log\Gamma(\theta+1)-(\theta+1)\log\nu$	$\psi^{-1}(\log(\nu x))-1$	$\begin{bmatrix}\psi(\theta+1)-\log\nu\\ -(\theta+1)/\nu\end{bmatrix}$

Table A.1 – **Exponential family distributions**. Gaussian distribution is assumed to have unit variance. The inverse-dispersion parameter r is fixed for the Negative Binomial. For the Beta and the log-normal distributions, ν refers to the second parameter computed at the feature-level.

A

A.5. SUPPLEMENTARY FIGURES

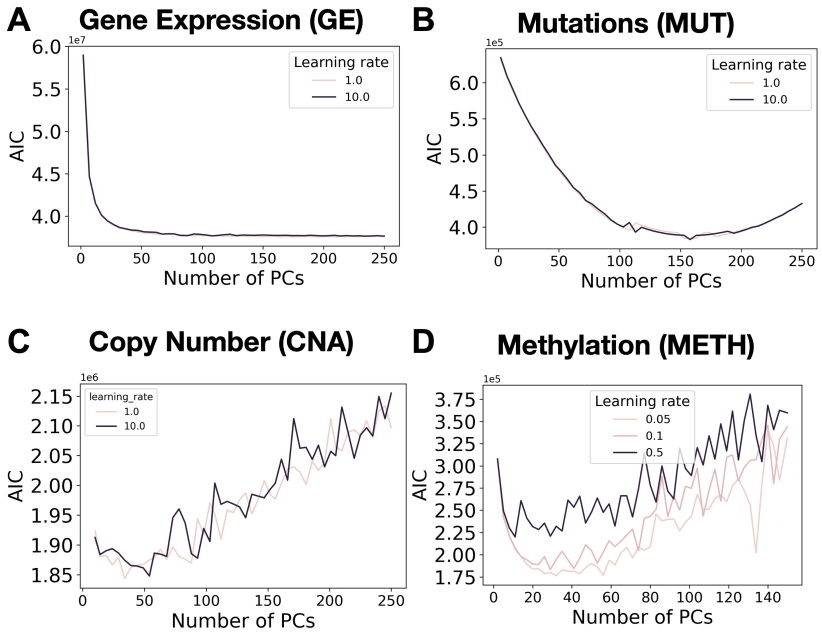


Figure A.1 – **Model selection for each data-type.** For each data-type we compute the AIC on the complete dataset for various hyper-parameters and report the best performance (i.e., lowest AIC) per number of components. This model selection wass performed for (**A**) gene expression (negative binomial), (**B**) mutation (bernoulli), (**C**) copy number (gamma) and (**D**) methylation (beta).

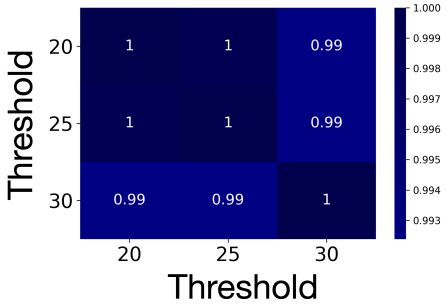


Figure A.2 – **Impact of threhsolding for mutations.** The infinity values in mutations are replaced by a large number, called **threshold**. To measure the impact of this parameter on Bernoulli-GLM-PCA, we used the optimal number of PCs set in Figure A.1 and fit a GLM-PCA for different threshold values. We then compared the loadings (Equation 4) between experiments by reporting the average singular value of the dot-product matrix ; i.e. if $V_{20} \in \mathbb{R}^{d \times p}$ and $V_{30} \in \mathbb{R}^{d \times p}$ contain the loadings for threshold=20 and threshold=30 respectively, we compute the spectrum (singular values) of $V_{20} V_{30}^T$ and returns the mean. Values can range from 0 to 1. For threshold 40 or above, GLM-PCA does not converge due to gradient explosion.

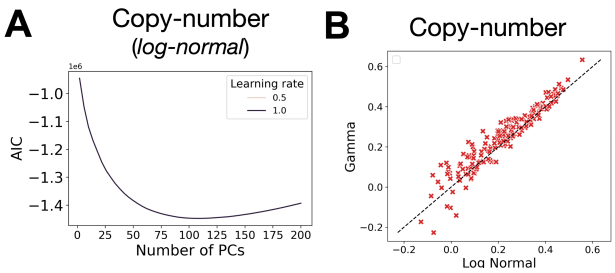


Figure A.3 – **Copy-number results with log normal distribution.** (**A**) AIC model selection for copy-number with log normal noise model (**B**) Comparison of predictive performance for the joint signal between CNA and GE for log-normal (x-axis) and gamma (y-axis) distribution. Each dot represents a single drug.

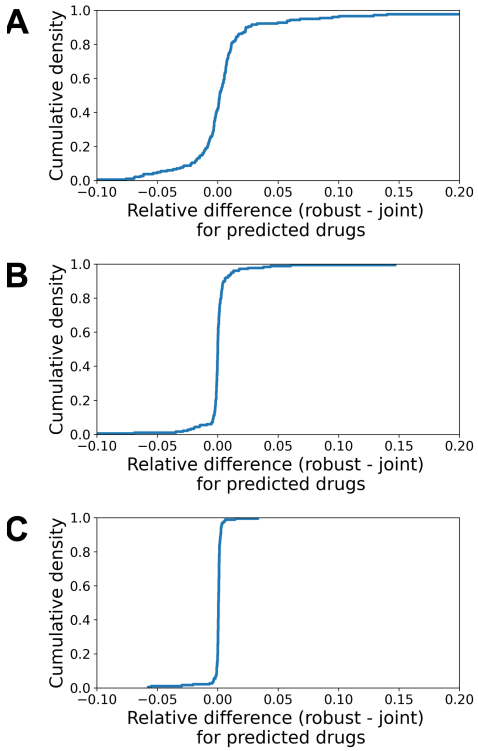


Figure A.4 – **Relative difference in predictive performance between joint signal (full cross-validation) and robust cell-view.** For each data-type, we computed the relative difference in predictive performance between the joint (p_j) and cell-view (p_r) as $(p_j - p_r)/p_j$. We report results for (**A**) mutations, (**B**), copy-number and (**C**) methylation.

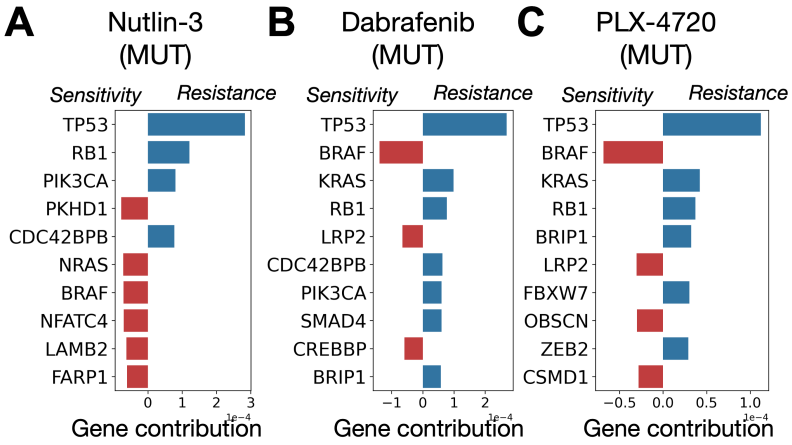


Figure A.5 – **Analysis of mutation joint-signal predictors.** Supporting Figure 5, we here report the predictors build on the joint signal between mutation and gene expression (solely based on mutations). (**A**) Nutlin-3, (**B**) Dabrafenib and (**C**) PLX-4720.

B

SUPPLEMENTARY MATERIAL FOR CHAPTER 3

B.1. ADDITIONAL INFORMATION

B.1.1. LIST OF DRUGS

TO ADD.

B.1.2. NOTES ON TRANSCRIPTOMICS DATA

Transcriptome levels have been measured using RNA-Seq Illumina HTSeq for both cell lines, PDX as well as the tumors. For cell lines and tumors, RNA-Seq data was available as read counts. For PDX and tumors, RNA-Seq data was available as FPKM. Since FPKM values are corrected for gene length at the transcript level and already normalised for library size, they cannot directly be compared to read counts. Consequently, we use two separate pre-processing pipelines, following the recommendation in ([158, 159]). For read counts, data is first normalized using TMM ([89]), then log-transformed and mean-centered. For FPKM, data is log-transformed and mean-centered. Experiments involving cell line to human tumor transfer have been performed using read counts, while PDX to human tumor transfer experiments have been performed using FPKM.

B.2. GEODESIC FLOW DERIVATION

B.2.1. ORIGINAL FORMULATION

We denote by $\mathbb{G}(d,p)$ the Grassmannian of d-dimensional subspaces within a p-dimensional space. This is formally defined as the set with a Riemannian structure of the d-dimensional subspaces within a larger p-dimensional space. The geometry of this space is non-Euclidean and therefore the shortest paths to go from one point to another are referred to as geodesic. The source *domain-specific factors* can be represented by one point in the Grassmannian and so do the target *domain-specific factors*. The idea now is to find this geodesic within $\mathbb{G}(d,p)$ that links the two. An analytical formulation of this

curve is given in ([58]).
A SVD on the cosine similarity matrix yields the matrices $\mathbf{U}_1 \in \mathbb{R}^{d\times d}$ and $\mathbf{U}_2 \in \mathbb{R}^{(p-d)\times d}$ such that

$$\mathbf{P}_s^T\mathbf{P_t} = \mathbf{U}_1\Gamma\mathbf{V}^T \quad \text{where} \quad \Gamma = \text{diag}\left(\cos\left(\theta_1\right),..,\cos\left(\theta_d\right)\right) \tag{B.1}$$

Let $\mathbf{R}_s \in \mathbf{R}^{p\times(p-d)}$ be the orthonormal complement of $\mathbf{P}_s$ (i.e. $\mathbf{P}_s^T\mathbf{R}_s = \mathbf{0}_{p-d,d}$ and $\mathbf{R}_s^T\mathbf{R_s} = \mathbf{I}_{p-d,p-d}$). The cosine similarity matrix between the orthogonal complement of $\mathbf{P}_s$ and the matrix $\mathbf{P}_t$ gives, after a SVD and a column-wise permutation on the right matrix:

$$\mathbf{R}_s^T\mathbf{P_t} = -\mathbf{U}_2\Sigma\mathbf{V}^T \quad \text{where} \quad \Sigma = \text{diag}\left(\sin\left(\theta_1\right),..,\sin\left(\theta_d\right)\right) \tag{B.2}$$

With these quantities, one can now define:

Proposition B.2.1 (Geodesic on the Grassmann manifold)**.** *The geodesic on the Grassmann manifold can be represented by the bases* Φ *defined as:*

$$\begin{aligned}
\Phi: \quad & [0,1] \longrightarrow \mathbb{G}\left(d,p\right) \\
& \tau \longmapsto \mathbf{P}_s\mathbf{U}_1\Gamma\left(\tau\right) - \mathbf{R}_s\mathbf{U}_2\Sigma\left(\tau\right) \\
\textit{where} \quad & \Gamma\left(\tau\right) = \textit{diag}(\cos\left(\tau\theta_1\right),..,\cos\left(\tau\theta_d\right)) \\
\textit{and} \quad & \Sigma\left(\tau\right) = \textit{diag}(\sin\left(\tau\theta_1\right),..,\sin\left(\tau\theta_d\right))
\end{aligned} \tag{B.3}$$

As shown in (Equation B.3), this formulation requires a lot of computation since the orthogonal complement $\mathbf{R}_s$ has to be computed. What is more, it links the *domain-specific factors* together, which is of limited interest for our study. Indeed, we would like to have a formulation that directly links the *principal vectors* instead, in order to filter out irrelevant factors that are too dissimilar to be used in the regression model.

B.2.2. WRITING THE GEODESIC FLOW IN TERMS OF PRINCIPAL VECTORS INSTEAD OF PRINCIPAL COMPONENTS

We here derive a formulation of the geodesic Φ in terms of principal vectors. We only make the assumption that $\theta_d < \frac{\pi}{2}$, which can easily be checked experimentally, which generally holds for all practical purposes. For problems that nevertheless do not satisfy this assumption, orthogonal principal vectors can be removed from the problem. They indeed do not correspond to transferable features and can be discarded.

Proposition B.2.2 (Equivalent definition of the Geodesic)**.** *Let's assume that* $\theta_d < \frac{\pi}{2}$*, then the geodesic can equivalently be defined as*

$$\begin{aligned}
\forall\tau\in[0,1], \quad \Phi\left(\tau\right) \; &= \; \mathbf{Q}_s\Pi\left(\tau\right) + \mathbf{Q}_t\Xi\left(\tau\right) \\
\textit{with} \quad \Pi\left(\tau\right) &= \textit{diag}\left(\frac{\sin\left(\left(1-\tau\right)\theta_i\right)}{\sin\left(\theta_i\right)}\right) \\
\textit{and} \quad \Xi\left(\tau\right) &= \textit{diag}\left(\frac{\sin\left(\tau\theta_i\right)}{\sin\left(\theta_i\right)}\right)
\end{aligned} \quad . \tag{B.4}$$

Proof. Since $[\mathbf{P}_s,\mathbf{R}_s]$ forms a orthogonal basis of $\mathbb{R}^p$, we have $\mathbf{P}_s\mathbf{P}_s^T + \mathbf{R}_s\mathbf{R}_s^T = \mathbf{I}_p$. Summing up then (Equation B.1) and (Equation B.2) yields, after multiplying by $\mathbf{P}_s^T$ and $\mathbf{R}_s^T$ respectively:

$$\mathbf{P}_t\mathbf{V} \quad = \quad \mathbf{P}_s\mathbf{U}_1\Gamma - \mathbf{R}_s\mathbf{U}_2\Sigma \tag{B.5}$$

We find that $\Phi(1) = \mathbf{P}_t\mathbf{V} = \mathbf{Q}_t$, which means that the end point of the geodesic gives us the basis of target principal vectors. Since $\theta_d < \frac{\pi}{2}$, then $\forall i \in \{1,..,d\}, \theta_i < \frac{\pi}{2}$. Σ will thus be invertible and (Equation B.5) yields:

$$-\mathbf{R}_s\mathbf{U}_2 \quad = \quad \mathbf{Q}_t\Sigma^{-1} - \mathbf{Q}_s\Gamma\Sigma^{-1} \tag{B.6}$$

Plugging (Equation B.6) into (B.3) yields the desired formula. □

This way, the geodesic path is computed in $\mathcal{O}(p \times d)$ instead of $\mathcal{O}(p^2)$ and does not require the computation of the orthogonal complement – which can be computationally intensive. This formulation has the interest of taking the principal vectors as inputs, instead of the principal components. It shows that the geodesic interpolates between principal vectors within each pair by taking features forming a rotating arc between the source and the target principal vectors. It therefore proves that our approach using all the principal vectors is strictly similar to the approach proposed in ([58]) and in ([100]).

B.2.3. EQUIVALENCE BETWEEN GEODESIC FLOW SAMPLING AND PRINCIPAL VECTOR REGRESSION

As suggested by ([100]), a domain-invariant drug response predictor can be created by sampling the interval $[0,1]$, i.e. by taking a number $M+1$ of intermediate representations $\{0, \frac{1}{M}..., 1\}$, computing the corresponding intermediate features $\{\Phi(0), \Phi(\frac{1}{M})..., \Phi(1)\}$, and finally projecting source and target data on these intermediate features. We show here that it strictly equivalent to projecting on the principal vectors and learn a linear regression model onto these principal vectors.

Proposition B.2.3 (Equivalence of estimators without penalization)**.** *Let $\hat{y}_S$ be the linear drug response estimator learnt without penalization by minimizing the loss function ℓ on the interpolated coefficients, and let $\hat{y}_{PV}$ be the linear estimator learnt by minimizing the loss function ℓ on the principal vectors. Then, $\hat{y}_S = \hat{y}_{PV}$.*

Proof. Let $x \in \mathbb{R}^p$ be a sample - from either source or target. A linear model learnt on the projected data will give a response of the form:

$$\begin{aligned}
& \hat{y}_S\left(x; (\alpha_{i,j})_{\substack{1\le i\le d \\ 0\le j\le M}}\right) \\
= \quad & \sum_{i=1}^{d}\sum_{j=0}^{M} \alpha_{i,j} x^T \left(\mathbf{Q}_{s,i}\Pi_{i,i}\left(\frac{j}{m}\right) + \mathbf{Q}_{t,i}\Xi_{i,i}\left(\frac{j}{m}\right)\right) \\
= \quad & \sum_{i=1}^{d} x^T \left[\mathbf{Q}_{s,i}\sum_{j=0}^{M}\alpha_{i,j}\Pi_{i,i}\left(\frac{j}{m}\right) \quad + \mathbf{Q}_{t,i}\sum_{j=0}^{M}\alpha_{i,j}\Xi_{i,i}\left(\frac{j}{m}\right)\right] \\
= \quad & \sum_{i=1}^{d} x^T \left[\beta_i^s \mathbf{Q}_{s,i} + \beta_i^t \mathbf{Q}_{t,i}\right] \\
= \quad & \hat{y}_{PV}\left(x; (\beta_i^s, \beta_i^t)_{1\le i\le d}\right)
\end{aligned} \tag{B.7}$$

with:

- $\alpha_{i,j} \in \mathbb{R}$ for all $i \in \{1,..,d\}$ and $j \in \{0,..,M\}$ the coefficients of the linear model fitted on the interpolated features.
- $\forall i \in \{1,..,d\}, \beta_i^s = \sum_{j=0}^{M} \alpha_{i,j} \Pi_{i,i} \left(\frac{j}{m}\right)$
- $\forall i \in \{1,..,d\}, \beta_i^s = \sum_{j=0}^{M} \alpha_{i,j} \Xi_{i,i} \left(\frac{j}{m}\right)$

Therefore, using this reciprocal correspondence, we can state that the non-regularized minimization procedure, using any loss ℓ is equivalent for both set of parameters, namely:

$$\min_{\alpha_{i,j}} \frac{1}{n} \sum_{k=1}^{n} \ell\left(y_k,\ \hat{y}_S\left(x_k; \alpha_{i,j}\right)\right) \ = \ \min_{\beta_i^s, \beta_i^t} \frac{1}{n} \sum_{k=1}^{n} \ell\left(y_k,\ \hat{y}_{PV}\left(x_k; \beta_i^s, \beta_i^t\right)\right) \tag{B.8}$$

□

Penalization may change the matter and the solution of the two minimization procedure might change slightly. However, we advocate for the latter penalized minimization procedure. Indeed, only $2d$ parameters have to be penalized. This in turn makes the minimization procedure easier and numerically more stable. The former formulation would require shrinking on way more features that are expressing the same content (same total rank).

B.2.4. EQUIVALENT FORMULATION OF GEODESIC FLOW KERNEL MATRIX

The original definition of the Geodesic Flow Kernel is ([**?**]):

$$\begin{aligned} &\forall x, y \in \mathbb{R}^p, \quad \int_0^1 \left(\Phi(\tau)^T x\right)^T \left(\Phi(\tau)^T y\right) d\tau \ = \ x^T \mathbf{G} y \\ &\text{with} \quad \mathbf{G} \ = \ \begin{bmatrix} \mathbf{P}_s \mathbf{U}_1 & \mathbf{R}_s \mathbf{U}_2 \end{bmatrix} \begin{bmatrix} \Lambda_1 & \Lambda_2 \\ \Lambda_2 & \Lambda_3 \end{bmatrix} \begin{bmatrix} \mathbf{U}_1^T \mathbf{P}_s^T \\ \mathbf{U}_2^T \mathbf{R}_s^T \end{bmatrix} \end{aligned} \tag{B.9}$$

As shown in (Equation B.9), computing the matrix $\mathbf{G}$ requires quadratic time in the number of covariates, which can be prohibitive in genomics (when $p \sim 20{,}000$). We show here how to improve this computation using the new formulation of (Equation B.4).

Proposition B.2.4 (Equivalent definition of Geodesic Flow Kernel)**.** *If $\theta_d < \frac{\pi}{2}$, then there exists $\sigma_1, .., \sigma_d \in \mathbb{R}$ and $\omega_1, .., \omega_d \in \mathbb{R}$ such that*

$$\mathbf{G} \quad = \quad \begin{bmatrix} \widetilde{\mathbf{Q}_s} \\ \widetilde{\mathbf{Q}_t} \end{bmatrix}^T \begin{bmatrix} \widetilde{\mathbf{Q}_s} \\ \widetilde{\mathbf{Q}_t} \end{bmatrix}$$

with

$$\widetilde{\mathbf{Q}_s} = \begin{bmatrix} \mathbf{Q}_{s,1}\sigma_1 + \mathbf{Q}_{t,1}\omega_1 \\ .. \\ \mathbf{Q}_{s,d}\sigma_d + \mathbf{Q}_{t,d}\omega_d \end{bmatrix} \quad \textit{and} \quad \widetilde{\mathbf{Q}_t} = \begin{bmatrix} \mathbf{Q}_{s,1}\omega_1 + \mathbf{Q}_{t,1}\sigma_1 \\ .. \\ \mathbf{Q}_{s,d}\omega_d + \mathbf{Q}_{t,d}\sigma_d \end{bmatrix}$$

Proof. First, if $x \in \mathbb{R}^p$, we define $x_s = x^T \mathbf{Q_s}$ and $x_t = x^T \mathbf{Q_t}$ as the projection of the point x and the source and target principal vectors. Then, using flow formulation from (Equation B.4), we get:

$$\begin{aligned}
&\int_0^1 x^T \Phi(\tau) \Phi(\tau)^T y \, d\tau \\
&= \int_0^1 x^T \left[\mathbf{Q}_s \Pi(\tau) + \mathbf{Q}_t \Xi(\tau)\right] \left[\Pi(\tau) \mathbf{Q}_s^T + \Xi(\tau) \mathbf{Q}_t^T\right] y \, d\tau \\
&= \quad x_s^T \left[\int_0^1 \Pi^2(\tau) \, d\tau\right] y_s \\
&\quad + x_t^T \left[\int_0^1 \Xi^2(\tau) \, d\tau\right] y_t \\
&\quad + x_s^T \left[\int_0^1 \Pi(\tau) \Xi(\tau) \, d\tau\right] y_t \\
&\quad + x_t^T \left[\int_0^1 \Xi(\tau) \Pi(\tau) \, d\tau\right] y_s \\
&= \begin{bmatrix} x_s^T & x_t^T \end{bmatrix} \begin{bmatrix} \int_0^1 \Pi^2(\tau) \, d\tau & \int_0^1 \Xi(\tau) \Pi(\tau) \, d\tau \\ \int_0^1 \Xi(\tau) \Pi(\tau) \, d\tau & \int_0^1 \Xi^2(\tau) \, d\tau \end{bmatrix} \begin{bmatrix} y_s \\ y_t \end{bmatrix}
\end{aligned} \tag{B.10}$$

With simple trigonometrical identities, we can show that :

$$\int_0^1 \Pi^2(\tau) \, d\tau = \int_0^1 \Xi^2(\tau) \, dt = \left(\frac{\theta_i - \sin(\theta_i)\cos(\theta_i)}{2\theta_i \sin^2(\theta_i)}\right)_i \tag{B.11}$$

$$\int_0^1 \Pi(\tau) \Xi(\tau) \, d\tau = \left(\frac{\sin(\theta_i) - \theta_i \cos(\theta_i)}{2\theta_i \sin^2(\theta_i)}\right)_i \tag{B.12}$$

Since the matrix is diagonal, we now have a formulation that only requires $\mathcal{O}(d + dp)$, faster than the $\mathcal{O}(p^2)$ that we had before since with $d = 50$ and $p = 19000$, we get a 8x speed-up.
We can write the matrix $\mathbf{G}$ as a product of the principal vector instead :

$$\begin{aligned}
\mathbf{G} &= \begin{bmatrix} \mathbf{Q}_s^T & \mathbf{Q}_t^T \end{bmatrix} \begin{bmatrix} \Lambda & \boldsymbol{\mu} \\ \boldsymbol{\mu} & \Lambda \end{bmatrix} \begin{bmatrix} \mathbf{Q}_s \\ \mathbf{Q}_t \end{bmatrix} \\
\Lambda &= \operatorname{diag}\left(\frac{\theta_i - \sin(\theta_i)\cos(\theta_i)}{2\theta_i \sin^2(\theta_i)}\right) \\
\boldsymbol{\mu} &= \operatorname{diag}\left(\frac{\sin(\theta_i) - \theta_i \cos(\theta_i)}{2\theta_i \sin^2(\theta_i)}\right)
\end{aligned} \tag{B.13}$$

Let's denote $(\lambda_1, .. \lambda_d)$ the diagonal coefficients of Λ and $(\mu_1, .. \mu_d)$ the diagonal coefficients of $\boldsymbol{\mu}$. We can now define the coefficients σ_i and ω_i for all $i \in \{1, .., d\}$ as

$$\begin{aligned}
\sigma_i &= \frac{1}{2}\left(\sqrt{\lambda_i + \mu_i} + \sqrt{\lambda_i - \mu_i}\right) \\
\omega_i &= \frac{1}{2}\left(\sqrt{\lambda_i + \mu_i} - \sqrt{\lambda_i - \mu_i}\right)
\end{aligned} \tag{B.14}$$

and the matrix $\mathbf{H}$ as :

$$\mathbf{H} = \begin{bmatrix} \sigma_1 & & & \omega_1 & & \\ & \dots & & & \dots & \\ & & \sigma_d & & & \omega_d \\ \omega_1 & & & \sigma_1 & & \\ & \dots & & & \dots & \\ & & \omega_d & & & \sigma_d \end{bmatrix} \tag{B.15}$$

$\mathbf{H}$ is positive semi-definite (symmetric with eigenvalues $\sigma_i + \omega_i > 0$ and $\sigma_i - \omega_i > 0$) and respect:

$$\begin{bmatrix} \Lambda & \boldsymbol{\mu} \\ \boldsymbol{\mu} & \Lambda \end{bmatrix} = \mathbf{H}^T\mathbf{H} \tag{B.16}$$

Plugging this equality in (Equation B.13), we get:

$$\mathbf{G} = \begin{bmatrix} \mathbf{Q}_s^T & \mathbf{Q}_t^T \end{bmatrix} \mathbf{H}^T\mathbf{H} \begin{bmatrix} \mathbf{Q}_s \\ \mathbf{Q}_t \end{bmatrix} \tag{B.17}$$

Let's now define the two following matrices:

$$\widetilde{\mathbf{Q}_s} = \begin{bmatrix} \mathbf{Q}_{s,1}\sigma_1 + \mathbf{Q}_{t,1}\omega_1 \\ .. \\ \mathbf{Q}_{s,d}\sigma_d + \mathbf{Q}_{t,d}\omega_d \end{bmatrix} \quad \text{and} \quad \widetilde{\mathbf{Q}_t} = \begin{bmatrix} \mathbf{Q}_{s,1}\omega_1 + \mathbf{Q}_{t,1}\sigma_1 \\ .. \\ \mathbf{Q}_{s,d}\omega_d + \mathbf{Q}_{t,d}\sigma_d \end{bmatrix} \tag{B.18}$$

We finally get:

$$\mathbf{G} = \begin{bmatrix} \widetilde{\mathbf{Q}_s} \\ \widetilde{\mathbf{Q}_t} \end{bmatrix}^T \begin{bmatrix} \widetilde{\mathbf{Q}_s} \\ \widetilde{\mathbf{Q}_t} \end{bmatrix} \tag{B.19}$$

□

The geodesic flow kernel is therefore equivalent to projecting on $2d$ vectors that form a basis equivalent to the source and target principal vectors. Using the same idea as in Prop B.2.4, the ordinary least square estimate will be equivalent to the one obtained using principal vectors.

B.3. COMPARISON OF FACTORS BETWEEN SOURCE AND TARGET

B.3.1. COMPARISON RESULTS FOR OTHER TISSUES

Following experiments from Fig. 3.2, we computed the cosine similarity and the variance explained for other tissues. Results can be found in Fig. B.1.

B.3.2. SIGNIFICANCE OF THE COSINE SIMILARITY VALUES

To show that these cosine similarity values are significant, we performed a permutation test at the gene level. These cosine similarity values are supposed to reflect some shared structure in the data. If we permute the source genes while keeping the target data intact, this structure should be destroyed. The source principal components would be different

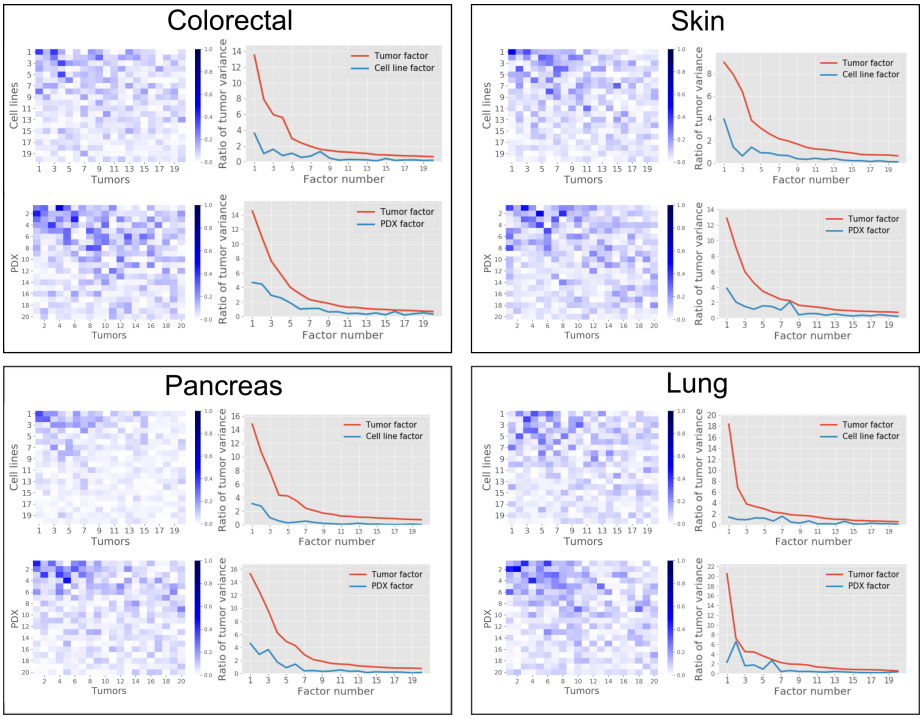


Figure B.1 – **Cosine similarity matrices between pre-clinical models and tumors for different tissues.** Each box represents the analysis carried out on one tissue type. Within each box, the top panel represents the cell line analysis and the bottom one represents the PDX results. Cosine similarity values between source (cell lines or PDXs) and target (tumors) are displayed on the left. Ratio of target variance explained by source principal components is displayed on the right panel.

and the cosine similarity values should be impacted. We permuted the genes order at the source level only and computed the resulting cosine similarity matrix and variance explained 1000 times to create a meaningful comparison on 5 tissues : breast, colorectal, lung, skin and pancreas. The results are displayed in Fig. B.2.

B.3.3. COMPARISON WITH RANDOM SIGNALS

Gene-level permutation, although yielding useful insights as shown in Subsect. B.3.2, restricts the pool of principal components values to the feature-level permutations. To go one step further in the identification, we used a random signal to quantify the commonality. We computed the cosine similarity values and the tumor variance explained for 250 random covariance matrices using the following protocol:

1. A random covariance matrix was sampled uniformly from the positive semi-definite matrices.

2. 1000 data points were drawn from the Gaussian distribution with 0-mean and the covariance matrix drawn in 1.

3. Principal components were computed from the these data points and cosine similarity values were computed alongside the tumor explained variance and compared to real data.

Although the second step could be removed and principal components could be computed directly using the randomly drawn covariance matrix, we decided to use sampled data to be the closest possible to our original setting. 1000 corresponds to the total number of cell lines available and is therefore comparable to our settings. Results are shown in Fig. B.3.

B.4. PRINCIPAL VECTORS ANALYSIS FOR DIFFERENT SET OF TISSUES

B.4.1. BREAST VS BREAST FOR PDX

In Fig. 3.3, we compared breast cancer cell lines to human breast tumors. The same experiment was run using PDXs instead of cell lines and results are shown in Fig. B.5.

B.4.2. BREAST VS ALL

In Fig. 3.4, PRECISE was trained using all cell lines in order to enhance the sample size to around 1000 – only 52 breast cancer cell lines are available. We compared the making of the principal vectors between all cell lines and the breast tumors to make sure that these principal vectors still show some enrichment. Results are shown in Fig. B.5.

B.4.3. SKIN VS SKIN

We repeated the experiment of Fig. 3.3 to another tissue: skin. As shown in Fig. B.6, the same behavior as in breast appears, with immune related pathways mostly enriched in the least similar PVs.

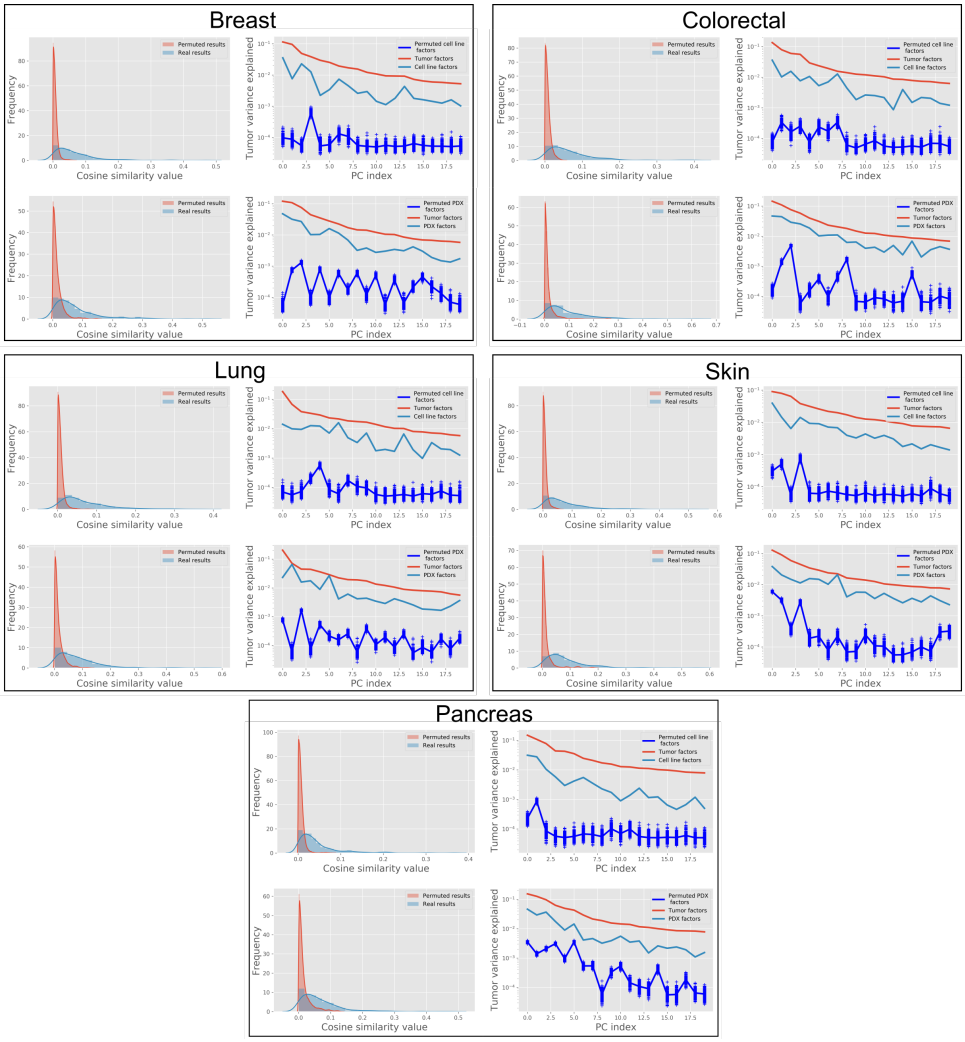


Figure B.2 – **Gene-level permutation results**. For each (source, target) couple, genes have been permuted in the source data. Cosine similarity values and target variance explained have then been computed as in Subsect. 3.3.1. For each tissue, the top row represents results between cell lines and tumors while the bottom one represents results for PDX and tumors. The left column represents the histogram of cosine similarity values while the right column shows the variance explained by target, source and gene-level-permuted source principal components. 1000 permutations have been employed to arrive to these results. For every tissue, the cosine similarity values for the permuted source data range from 0 to 0.05, while certain cosine similarity values are as large as 0.2 for almost every tissue. It suggests that the cosine similarity values encountered in Fig. 3.2 and Fig. B.1 are not the product of non-comparable signals. When it comes to the variance explained, the variance explained by permuted source principal components is consistently two to three orders of magnitude lower than when the tumor data is projected on the non-permuted source data. Two notable exceptions: colorectal PDXs and and Pancreatic PDXs for which some permuted principal components show variance explained only one order of magnitude lower than the non-permuted one.

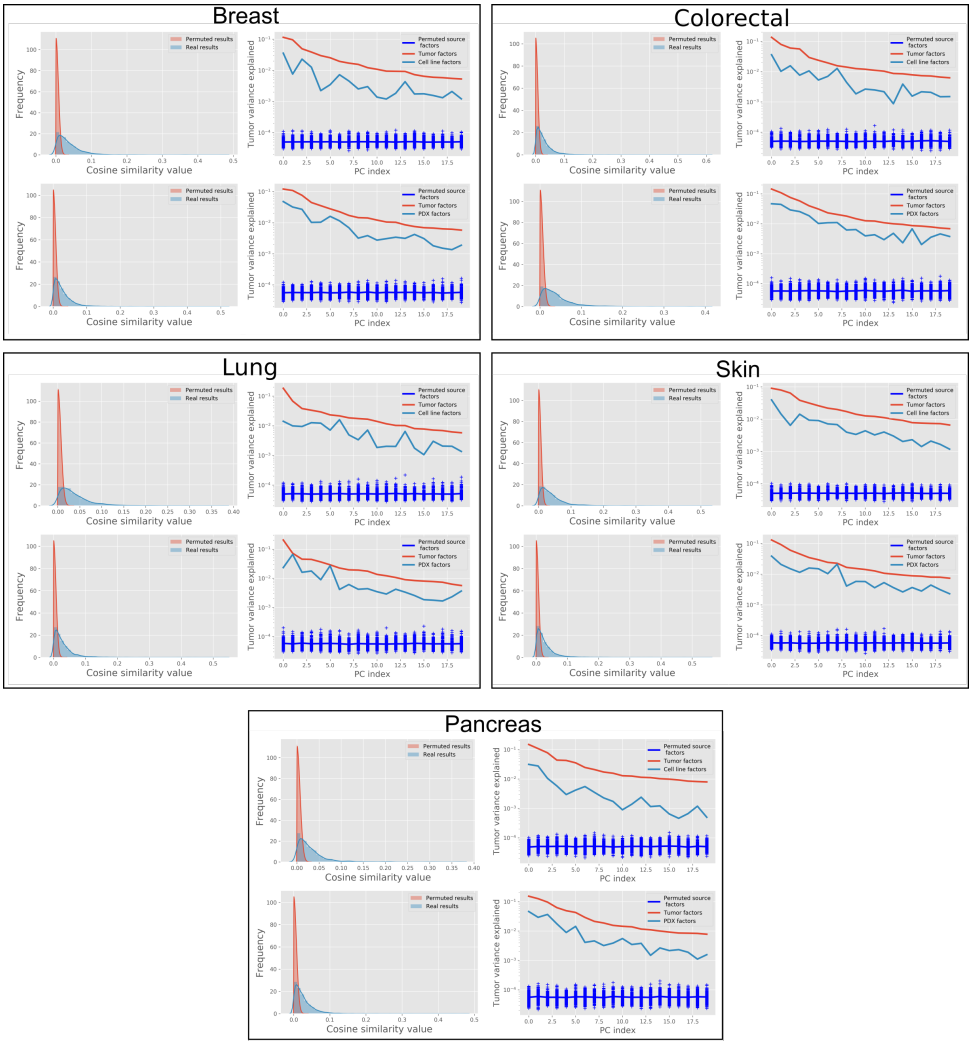


Figure B.3 – **Random signal results**. For each (source, target) couple, 250 positive semi-definite matrices were drawn randomly. For each matrix, 1000 data points were then drawn from a Gaussian distribution with this matrix as covariance. Cosine similarity and tumor variance explained were finally computed. These purely random signals are here compared to the real results. For each tissue, the top panel represents the comparison for cell lines while the bottom represents the results for PDXs. On the left are compared the cosine similarity values and on the right the tumor variance explained ratio. The random cosine similarity values appear to be consistently ranging between 0 and 0.02 while cosine similarity between tumors and real source data are as large as 0.2 for some principal components. It indicates that the similarity values between pre-clinical systems and tumors are not the product of the comparison of two random signals. In terms of variance explained, the variance explained by random principal components is two to five orders of magnitude lower than the tumor variance explained by real source principal components. This result is consistent across all tissue type and once again indicate the existence of some common structure between source and target.

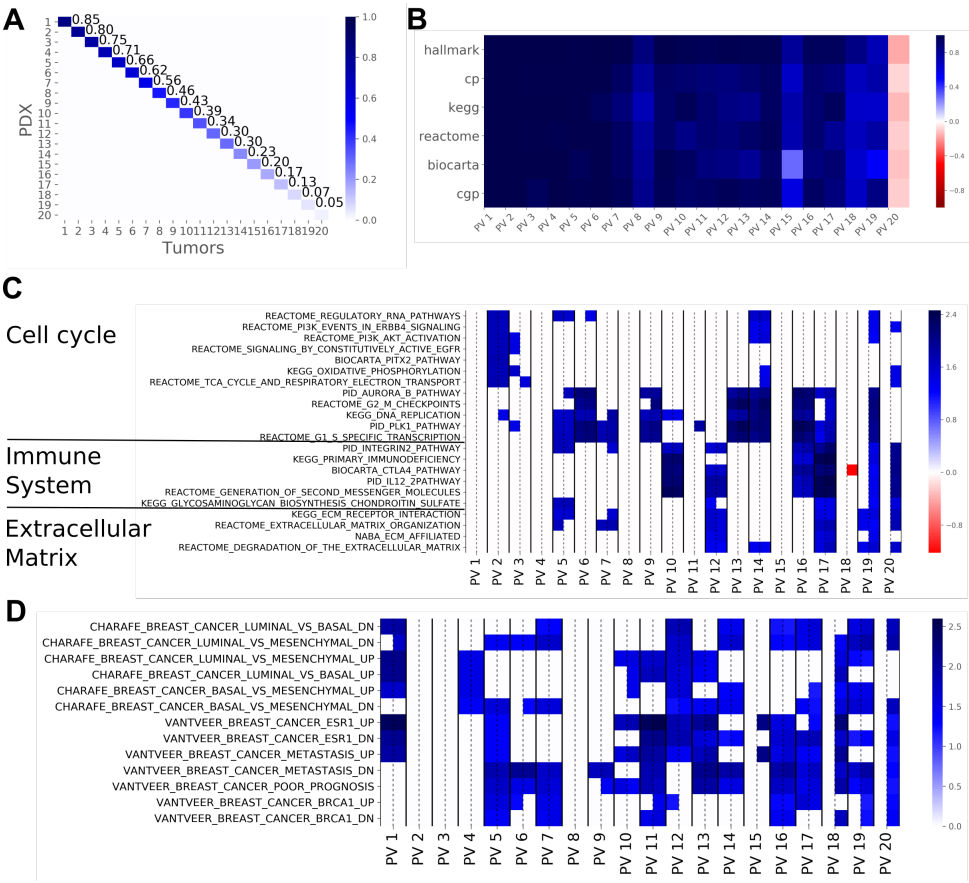


Figure B.4 – **Principal vectors (PVs) computed from breast PDXs and breast tumors from 20 principal components.**. **(A)** Cosine similarity matrix between PDX and tumor principal vectors. As shown on the diagonal, the similarity is higher than in Fig. 3.3A for a similar sample size. This is encouraging since PDXs are expected to mimic human tumors more closely than cell lines. **(B)** Spearman Correlation between PDX and tumor PV Normalized Enrichment Score (NES) for several gene set collections. The spearman correlation is almost 1 up to the 8th PV, suggesting that the same pathways get enriched. The last PV pair shows a negative correlation, in accordance with the almost null similarity. **(C)** The NES based on the Canonical Pathways for each PV pair with the NES for the source PV on the left and the NES for the target PV on the right (separated by a dashed line). Non-significant gene sets are represented as white cells. For this figure panel, we selected the ten gene sets that were most highly enriched in the first five PVs, the ten gene sets that showed the highest enrichment in the bottom PVs as well as all the gene sets related to extra-cellular matrix. The top PVs are exclusively enriched in pathways related to cell cycle. Immune system-related pathways are enriched in the middle and bottom PVs and PVs at the bottom tend to show enrichment for the target PVs only. Compared to results of Fig. 3.2C, the gene sets related to the immune system appear to be again enriched only in the less similar PVs, while extracellular matrix related pathways are this time showing some enrichment for the top PVs. **(D)** The NES for each PV as displayed in **(C)**, for the CHARAFE and VANTVEER gene sets. The top principal vectors are significantly enriched in sets associated with breast cancer subtypes.

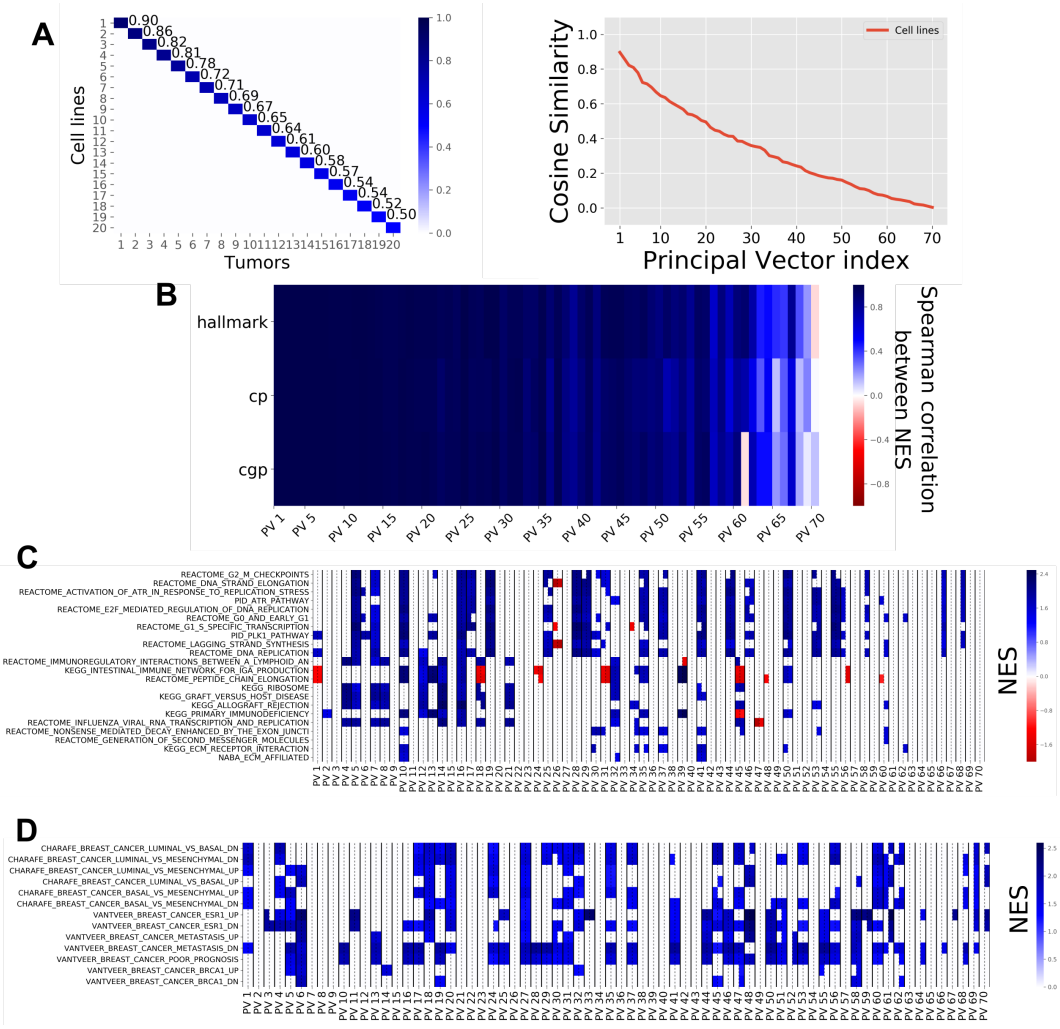


Figure B.5 – **Principal vectors (PVs) computed from breast cell lines and breast tumors from 70 principal components.**.. **(A)** Cosine similarity values between the top 70 principal vectors. A zoom is performed on the top 20, showing that similarity is as high as 90% for the top pair. **(B)** Spearman correlation between Normalised Enrichment Scores (NES) within each pair of PVs. Correlations close to 1 in the top 30 PV show that gene sets get the same enrichment in cell line and tumor PV and indicate an important structural similarity. **(C)** The NES based on the Canonical Pathways for each PV pair with the NES for the source PV on the left and the NES for the target PV on the right (separated by a dashed line). Non-significant gene sets are represented as white cells. For this figure panel, we selected the ten gene sets that were most highly enriched in the first five PVs, the ten gene sets that showed the highest enrichment in the bottom PVs as well as all the gene sets related to extra-cellular matrix. **(D)** The NES for each PV as displayed in **(C)**, for the CHARAFE and VANTVEER gene sets. The top principal vectors are significantly enriched in sets associated with breast cancer subtypes.

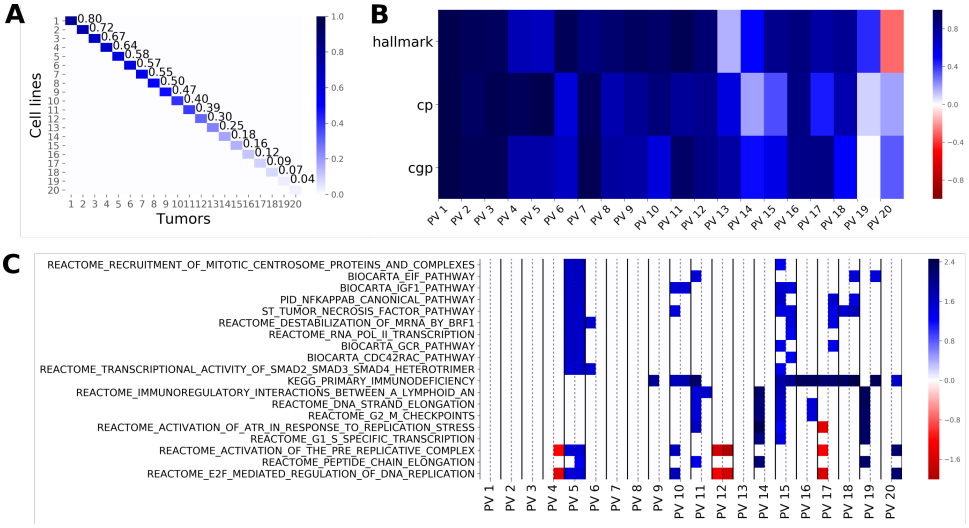


Figure B.6 – **Principal vectors (PVs) computed from skin cell lines and skin tumors from 20 principal components.**. **(A)** Cosine similarity matrix between cell lines and tumor principal vectors. **(B)** Spearman Correlation between PDX and tumor PV Normalized Enrichment Score (NES) for several gene set collections. For the skin, the spearman correlations between NES are lower than for breast, although they remain larger than 0.8 for the top 10 PVs. **(C)** The NES based on the Canonical Pathways for each PV pair with the NES for the source PV on the left and the NES for the target PV on the right (separated by a dashed line). Non-significant gene sets are represented as white cells. For this figure panel, we selected the ten gene sets that were most highly enriched in the first five PVs, the ten gene sets that showed the highest enrichment in the bottom PVs. Although two immune-related pathways are enriched in the top PVs, the same pattern as in Fig. 3.3 appears with several pathways enriched exclusively in the last principal vectors.

B.4.4. OTHER TISSUE

We computed the similarity scores for other tissues: skin, lung, pancreas and colorectal. Results are shown in Fig. B.7 for both cell lines and PDXs.

B.5. CHOICE OF THE HYPER PARAMETERS FOR THE EXPERIMENTS

B.5.1. VARIANCE-BASED APPROACH FOR SELECTING THE NUMBER OF PRINCIPAL COMPONENTS

We selected the number of domain-specific factors (PCs) based on the variance explained by the cell line principal components. Since the sample size is always larger for tumors, this cut-off point is lower for cell lines than for tumors and we only showed the cell line behavior. As shown in Fig. B.8, we took 20 PCs when the same tissue is used for source and for target ; we took 70 PCs when all cell lines are used as source data.

B.5.2. COMPARISON TO THE RANDOMLY-SAMPLED DATA FOR DETERMINING THE SIMILARITY CUT-OFF POINT

Once the number of PCs had been settled, we needed to determine the number of PVs to select. For that purpose, we computed the similarity between tumor data and data drawn from a gaussian distribution with a random covariance matrix. We repeated this experiment 250 times and got 250 similarity profiles. We took as threshold the maximum random similarity and selected PVs with similarity at least as large. As shown in Fig. B.8, it yields 15 PVs when only one tissue is used for source, and 40 when all cell lines are taken into account.

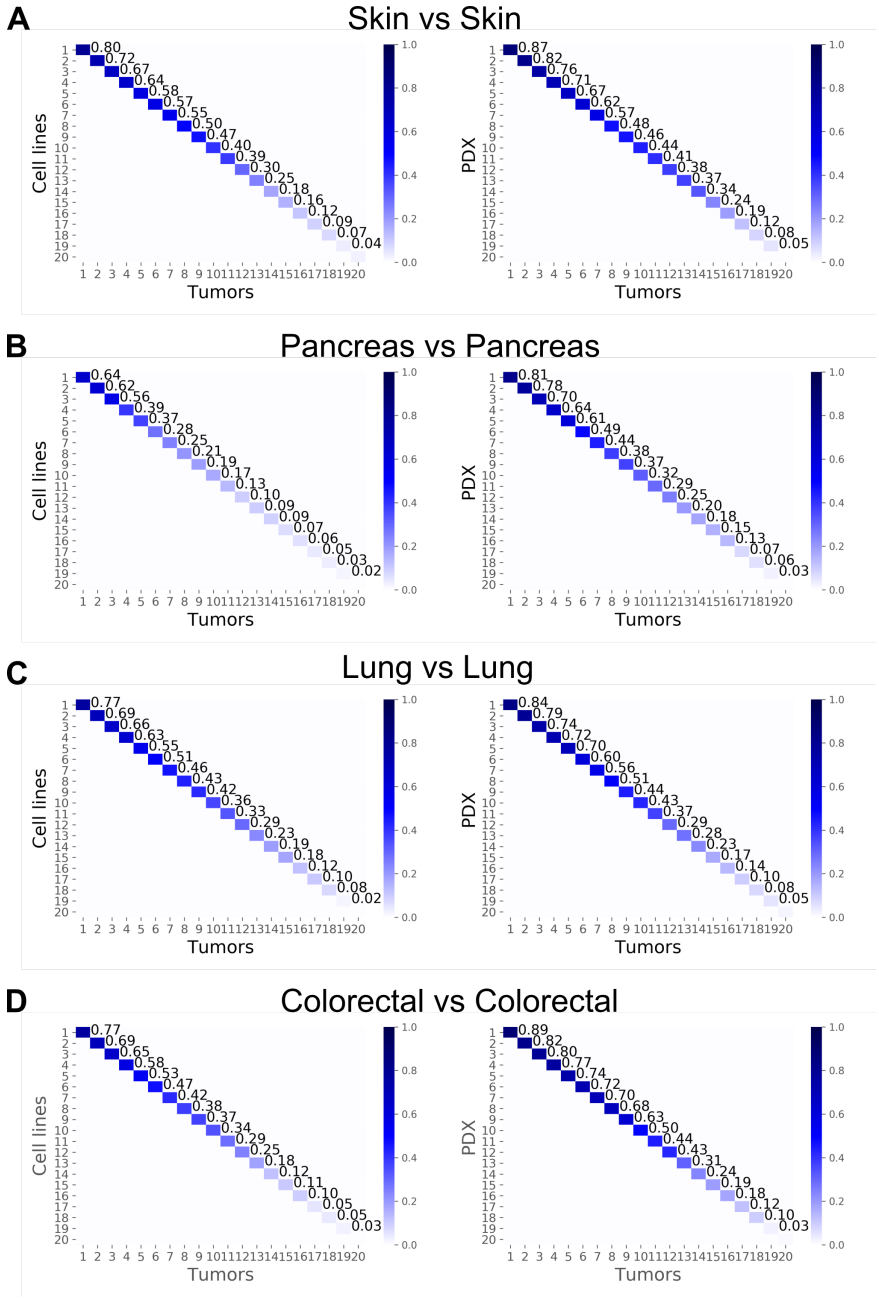


Figure B.7 – **The cosine similarity matrix between principal vectors for other tissues with 20 principal components**. **(A)** Similarity values when target is set as skin tumors and source set as skin cell lines (left) or skin PDXs (right). **(B)** Similarity values when target is set as pancreatic tumors and source set as pancreatic cell lines (left) or pancreatic PDXs (right). **(C)** Similarity values when target is set as lung tumors and source set as lung cell lines (left) or lung PDXs (right). **(D)** Similarity values when target is set as colorectal tumors and source set as colorectal cell lines (left) or colorectal PDXs (right).

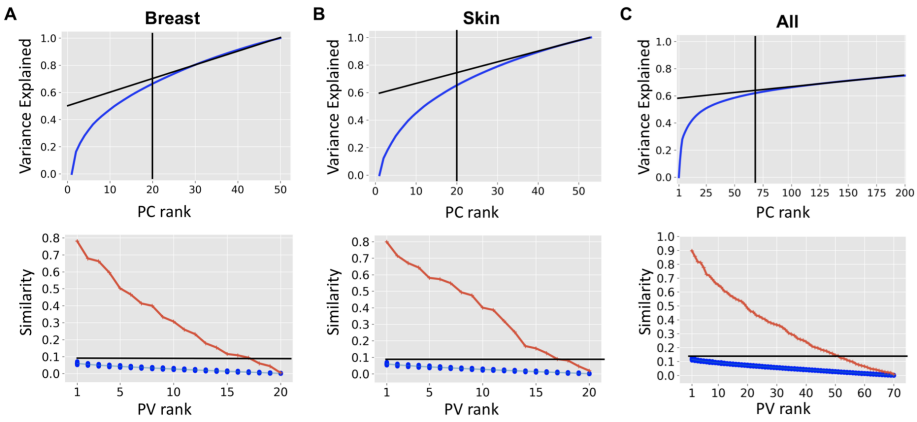


Figure B.8 – **Choice of hyperparameters d_f and d_{pv}.** The top panel shows the cumulative variance explained, while the bottom panel shows the similarity between the resulting PVs computed from the number of PCs found with the top panel. **(A)** shows results for breast cell lines (with breast tumors), **(B)** shows results for skin cell lines (with skin tumors), and **(C)** shows results for all cell lines (with breast tumors). For selecting the number of PCs, we drew a line corresponding to the asymptotic behavior of the cumulative variance and selected the principal components for which the cumulative variance explained does not follow this behavior. This gives a cut-off slightly before 20 for breast, slightly above 20 for skin and around 70 for all. Since we want to use the same number of PCs for all experiment having one tissue for the source, we settled for 20 that makes consensus between skin and breast. We settled to 70 for experiments with all cell lines. Once this number of PCs had been settled, we needed to determine where to put the PV threshold. For that, we sampled data from random covariance matrix 1000 times and compute 1000 similarity profiles following a similar idea than in Fig. B.3. We take the top similarity as cut-off, which yields 15 PVs for breast, slighlty more for skin and around 40 for PVs. Based on this experiment, we decided to settle for 15 PV when one tissue of the cell line is taken and 40 PVs when all cell lines are taken into account.

B.6. COMPARISON WITH KNOWN BIOMARKERS

B.6.1. PRECISE CORRELATION WITH OTHER KNOWN MECHANISMS

We repeated the experiment of Fig. 3.4 with other known biomarker-drug associations. We also repeated the same experiments but took only one tissue for the cell lines. Results shown in Fig. B.9 indicate than PRECISE successfully recapitulates known associations coming from independent data sources.

B.6.2. BIOMARKER CORRELATION FOR RIDGE REGRESSION WITHOUT ANY DOMAIN ADAPTATION OR WITH COMBAT AS PREPROCESSING STEP

We compared PRECISE results to the scenario where no domain adaptation is used and a Ridge regression is trained on the cell lines and directly transferred on the human tumors. We also compared PRECISE to the pipeline used in ([52]), where the difference between cell lines and human tumors is modelled as a batch effect. As shown in Fig. B.10, most of the associations are still recapitulated by the two scenarios, but PRECISE offers a higher discriminative power on most of the biomarkers.

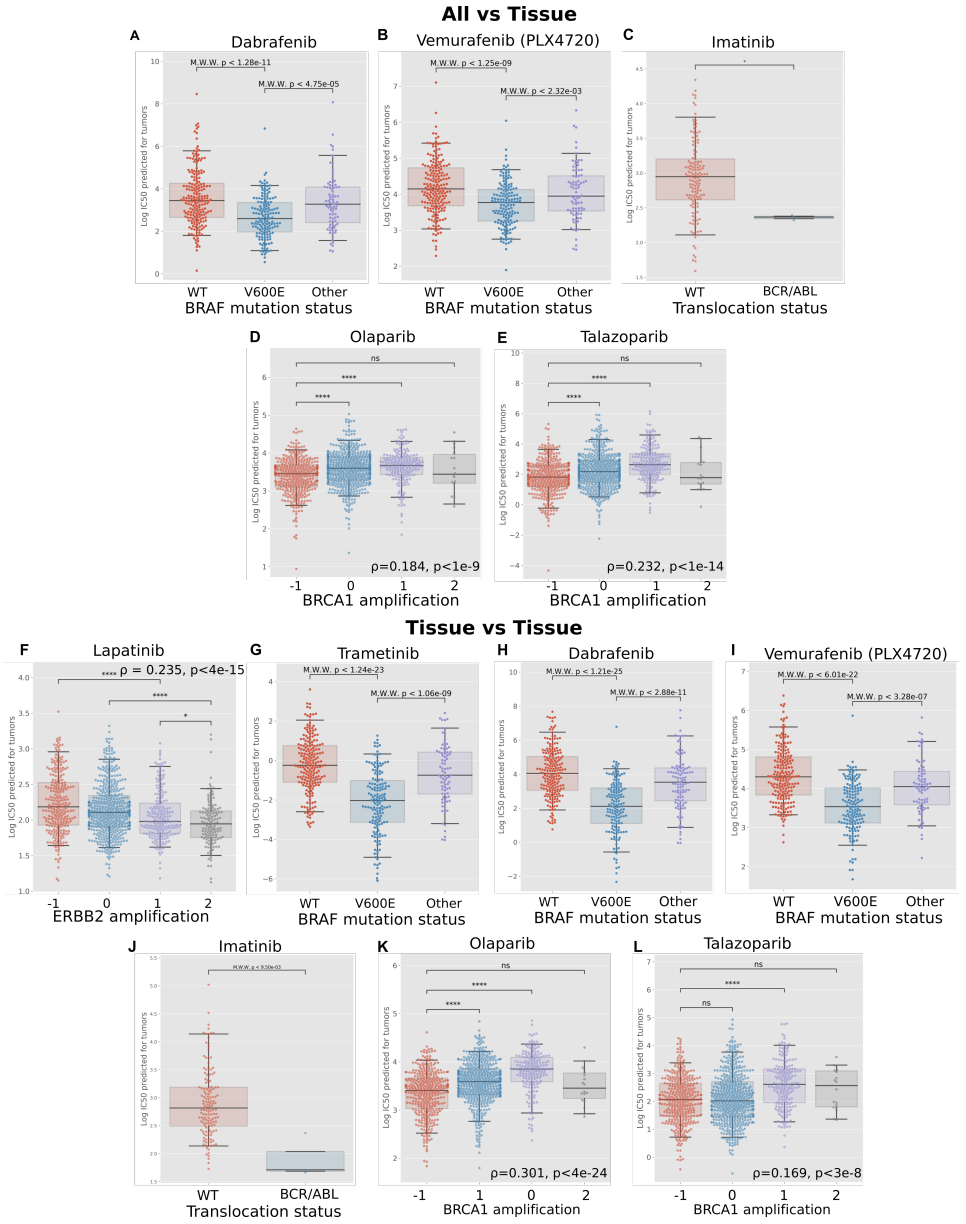


Figure B.9 – **Comparison to other known biomarkers**. Following the same experimental procedure as in Fig. 3.4, we compared IC_{50} predicted by PRECISE with some known biomarkers. Using all cell lines as source data, we show that PRECISE prediction are validated in **(A)** Dabrafenib (sensitive to $BRAF^{V600E}$ mutation), **(B)** Vemurafenib (sensitive to $BRAF^{V600E}$ mutation), **(C)** Imatinib (sensitive to BCR/ABL translocation), **(D)** Olaparib (sensitive to BRCA1 deletion) and **(E)** Talazoparib (sensitive to BRCA1 deletion). We repeated the experiment using only one tissue type in cell lines with all of the investigated drugs. We show that using only breast cell lines reduces the predicted power of ERBB2 in Lapatinib **(F)** and of BRCA1 in Talazoparib **(L)**. However, it increases the power of $BRAF^{V600E}$ mutation in all the MEK inhibitors considered (**G,H,I**), completely discriminates BCR/ABL translocated tumors for Imatinib **(J)** and increases the power of BRCA1 deletion in Olaparib **(K)**.

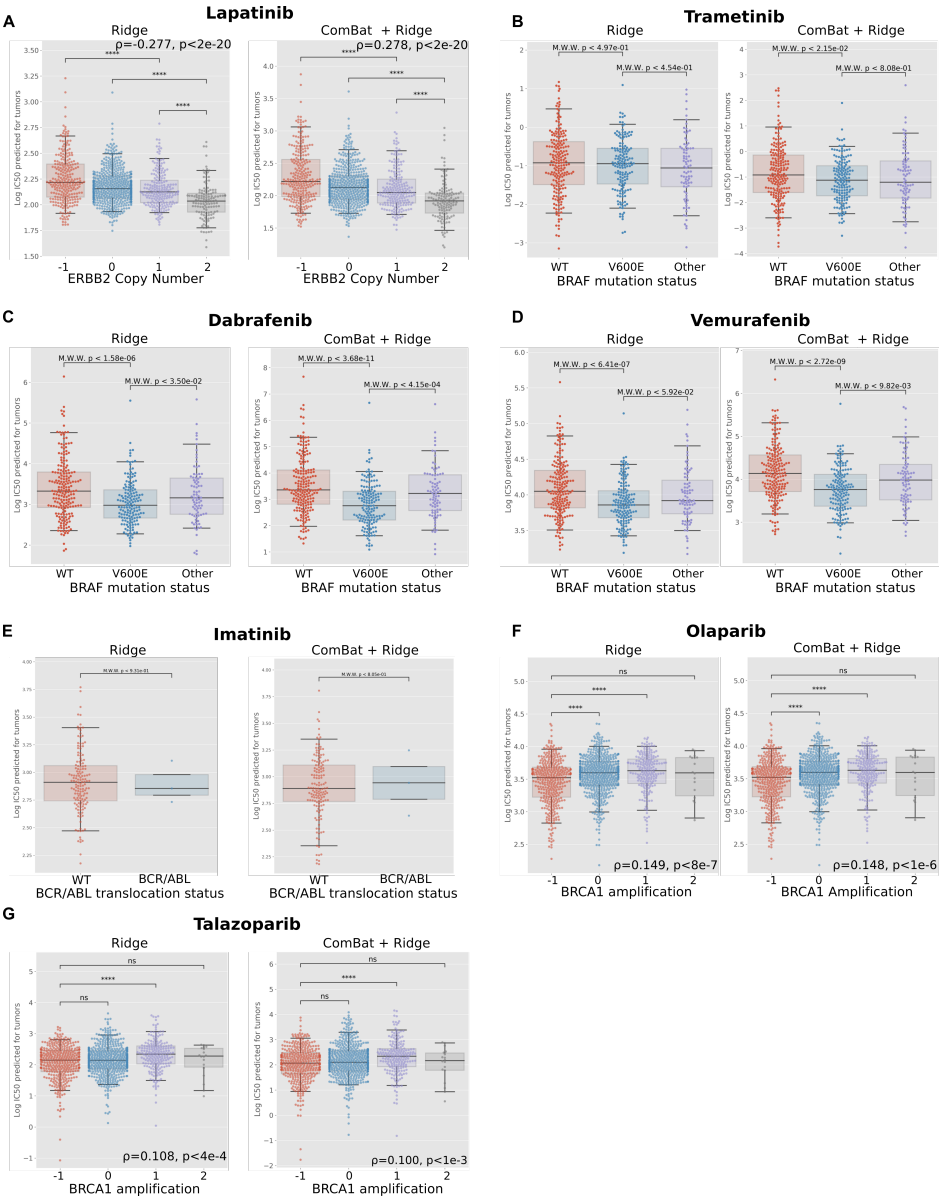


Figure B.10 – **Stratification with a Ridge regression on the bulk set of genes or with ComBat as domain adaptation.** We compared the results of Fig. 3.4 and Fig. B.9 to two scenarios: one without any domain adaptation between cell lines and tumors, and one with ComBat as the domain adaptation step. **(A)** Lapatinib predicted response correlation with ERBB2 amplification is comparable to PRECISE, whether ComBat is used or not. **(B)** Trametinib sensitivity to $BRAF^{V600E}$ mutation, however, is not predicted. When using ComBat, a slight discrimination is observed between wild type and mutated tumors but the regression model fails to discriminates between V600E and other mutations. In Dabrafenib **(C)** and Vemurafenib **(D)**, Ridge regression and ComBat successfully indicate the sensitivity to $BRAF^{V600E}$ mutation, but the power is lower than PRECISE. BCR/ABL is not discriminated by neither Ridge nor ComBat + Ridge **(E)**. Finally, PARP inhibitors Olaparib **(F)** and Talazoparib **(G)** are also recovered, but with correlations two to three times lower than with PRECISE.

C

SUPPLEMENTARY MATERIAL FOR CHAPTER 4

C.1. SUPPLEMENTARY FIGURES

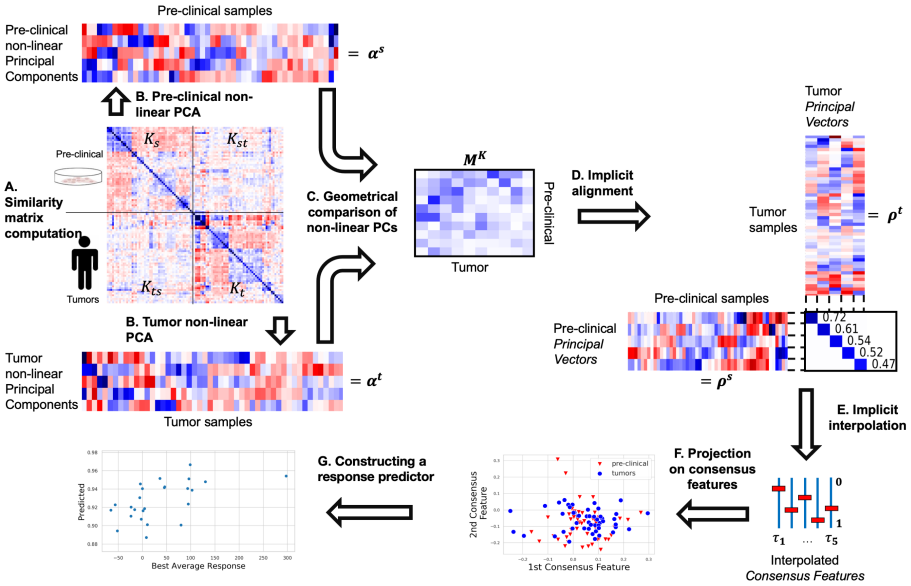


Figure C.1 – **TRANSACT: Generating non-linear manifold representations to transfer predictors of response from pre-clinical models to human tumors**. (**A**) Samples are compared using a similarity function yielding similarity matrices between pre-clinical models (source, K_s), between tumors samples (target, K_t) and between pre-clinical models and tumors (K_{st}). (**A**) Using non-linear PCA, the pre-clinical and tumor similarity matrices are independently decomposed into non-linear principal components (NLPCs) geometrically represented by "sample importance scores" (Supp. Figure C.2A) that represent the importance of each sample in each NLPC (α^s and α^t, for source and target space, respectively). (**C**) Geometrical comparison of pre-clinical and tumor NLPCs results in a non-linear cosine similarity matrix $\mathbf{M}^K$. (**D**) Alignment of NLPCs using the notion of principal vectors (Supp. Figure C.2B). (**E**) Interpolation within each pair of vectors to select one vector per PV-pair that balances the effect of pre-clinical and tumor signals: the consensus features (Supp. Figure C.2C). (**F**) Projection of each tumor and pre-clinical sample on the consensus features to obtain consensus scores: scores that correspond to the activity of processes conserved between tumors and pre-clinical models. (**G**) Finally, these scores can be used as input to any predictive model, for instance to predict drug response based on these consensus scores.

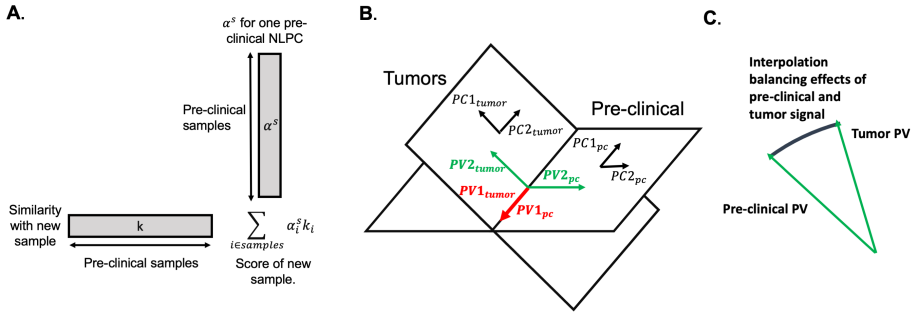


Figure C.2 – **Visual explanation of geometric alignment**. (**A**) Difference between importance scores (α^s,α^t) and projected scores. Since the space induced by the similarity function K is intractable, we use a dual representation of the NLPC in terms of samples: the *importance scores*. To project samples on NLPCs, one needs to compute the similarity between this sample and all of the samples used to gauge the NLPC. The projected score is obtained by taking the vector-product between this similarity vector and the importance scores. The same rational yields principal vectors that are represented by γ^s and γ^t. (**B**) Visual example of principal vectors (PV). We here consider 3 genes (features) and 2 NLPCs. The pre-clinical (source) and tumor (target) NLPCs intersect in one direction, which form the pair of closest vectors: the first PV forms the pair of the two red vectors – although these are identical. The second pair of PVs is defined orthogonally to the red pair. This defines the green vectors (with a swap in direction for visual purposes). These pairs reconstruct the original NLPC spaces and are ordered by similarity. (**C**) Interpolation between PVs. For one pair of PVs – e.g. the green one in (**B**) – source and target vectors are different. In order to generate one robust vector out of these two and avoid redundancy, we draw an arc between these two vectors. We then project source and target datasets onto these interpolated vectors and select one intermediate representation where source and target projected signals are maximally matched. This optimal intermediate vector is called the consensus feature.

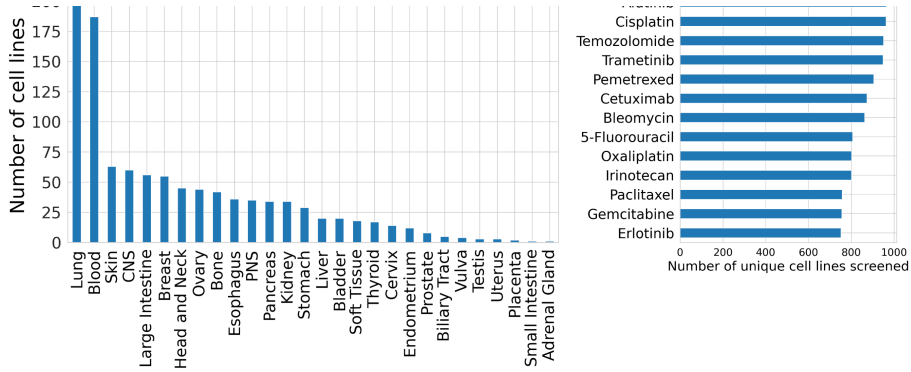


Figure C.3 – **Composition of the GDSC dataset (cell lines)**. We make use of the GDSC1000 cell line panel [9]. (**A**) Number of cell lines per tissue type. (**B**) Number of cell lines screened for each drug that we used in our experiments.

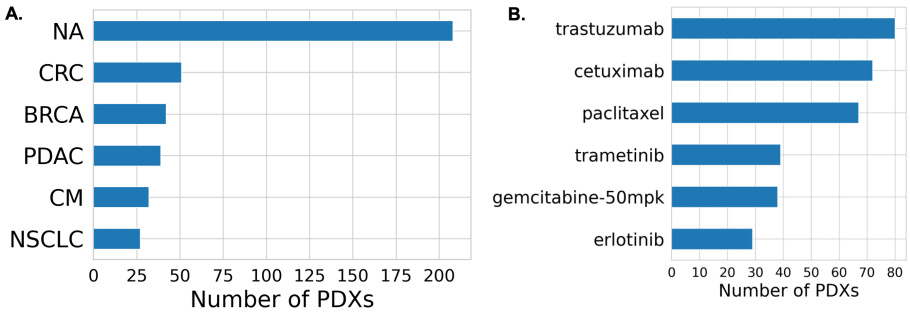


Figure C.4 – **Composition of the NIBR PDXE dataset (patient derived xenografts)**. We make use of the NIBR PDXE patient derived xenograft panel [93]. (**A**) Number of PDXs per tissue type. (**B**) Number of unique PDXs screened for each drug that we used in our experiments.

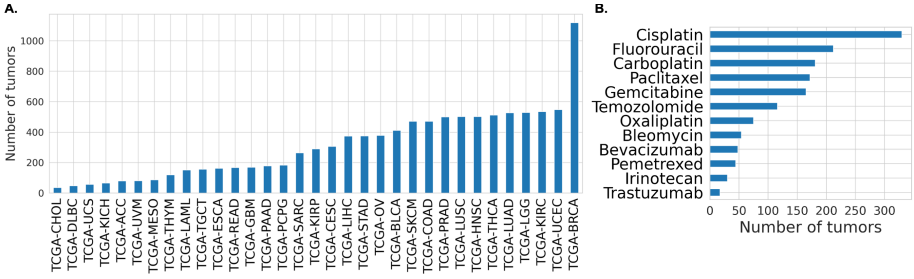


Figure C.5 – **Structure of the TCGA dataset (primary tumors)**. We make use of the TCGA dataset for primary tumors. (**A**) Number of samples per cancer type. (**B**) For each drug, number of samples with known response.

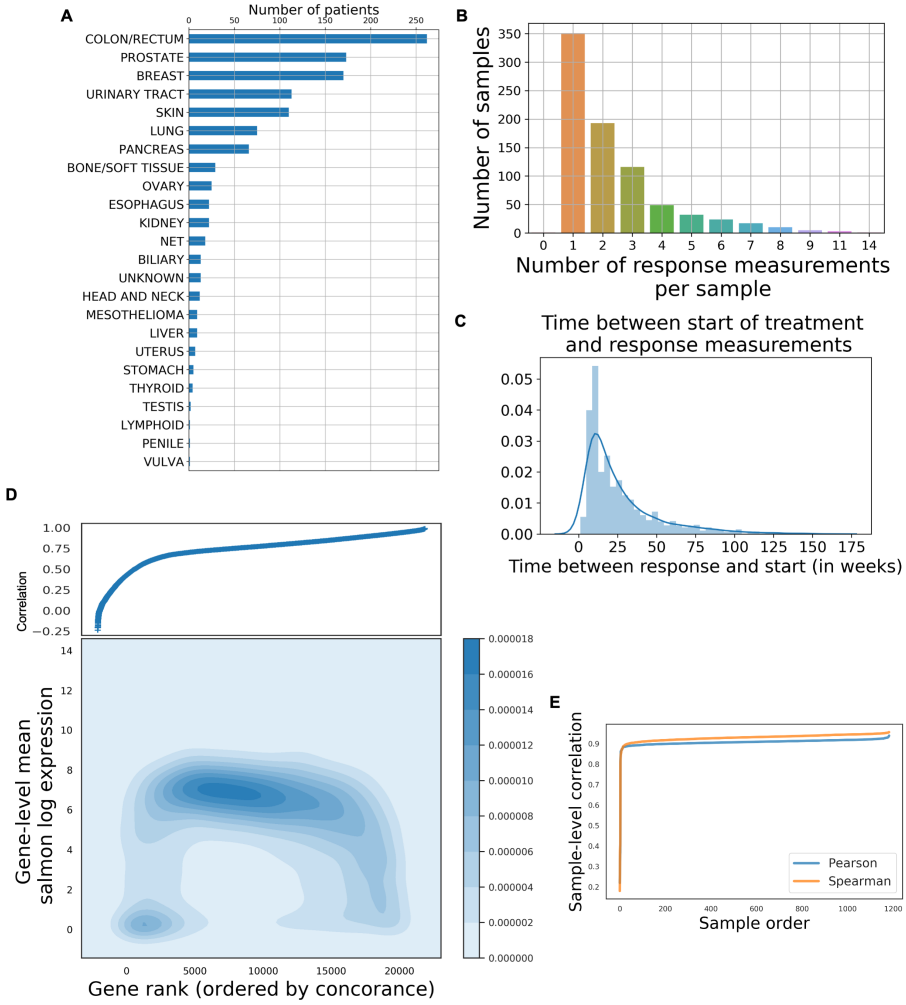


Figure C.6 – **Structure of the HMF dataset (metastatic lesions)**. We make use of the Hartwig Medical Foundation (HMF) dataset for metastatic lesions. (**A**) Number of samples per cancer type (primary tumor location). (**B**)) For each patient, number of response measurements made. For further analysis, we considered the first response measure – i.e. first measure after treatment start. (**C**)) Histogram of number of weeks between treatment start and response measurement. (**D**)) For each protein coding gene, we measure the Spearman correlation between read counts obtained using Salmon and STAR alignment tools using all samples in the HMF dataset. We then ranked genes based on the obtained Spearman correlation (x-axis) and plotted it against the mean-expression of these genes obtained using Salmon (y-axis). Since lowly concordant genes tend to have low expression, we put a threshold at corr=0.5 and discarded genes below this threshold. (**E**)) After the previous selection, we computed the sample-level Pearson and Spearman correlations between read counts obtained with STAR and Salmon. All samples but five show a correlation above 0.8 – these were discarded. We finally further restricted to genes from the mini-cancer genome.

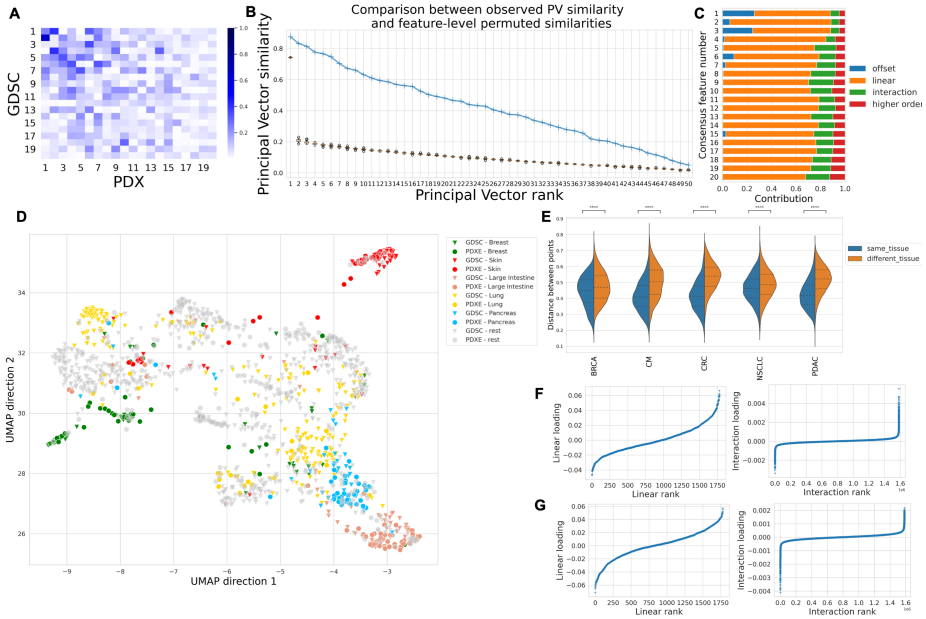


Figure C.7 – **Analysis of consensus features between cell lines (GDSC) and PDXs with** $\gamma = 0.0005$. We use a Gaussian similarity matrix with hyper-parameter $\gamma = 0.0005$ and run TRANSACT. (**A**) Cosine similarity between the 20 top source and target NLPCs. (**B**) Similarity between principal vectors (blue line) alongside the similarity obtained after gene-level permutation on GDSC (boxplots). (**C**) For each consensus feature, proportion of offset, linear and interaction term. (**D**) UMAP of data projected on the consensus features, colored by tissue of origin. (**E**) For each tissue type in PDXs, we compare the distances between corresponding PDXs with cell lines from the same tissue of origin (blue), or from another tissue (orange). (**F**) For the first consensus feature, sorted contribution of each linear features (i.e. gene, left) and interaction terms (right). (**G**) For the second consensus feature, sorted contribution of each linear features (i.e. gene, left) and interaction terms (right).

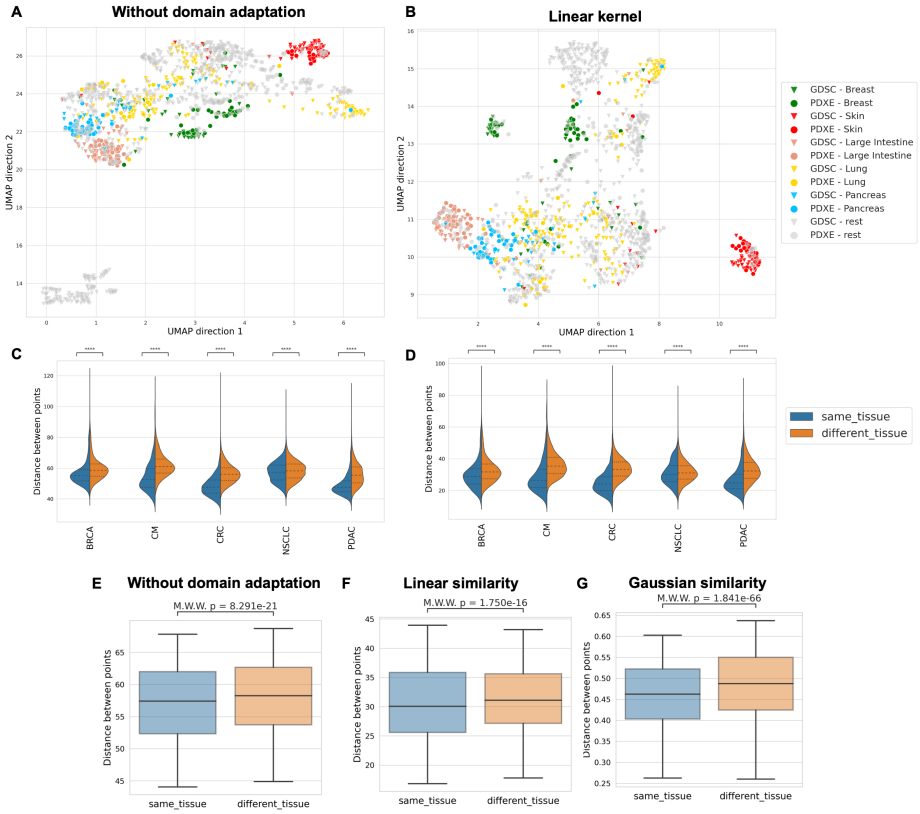


Figure C.8 – **Tissue clustering without domain adaptation and with PRECISE alignment between GDSC and PDXE**. (**A**) UMAP plot of cell lines and PDXs colored by tissue type without any domain-adaptation. Data was normalized prior to performing UMAP: cell lines and PDXs were independently mean-centered and scaled to unit variance. (**B**) UMAP plot of cell lines and PDXs colored by tissue type after projection on consensus features obtained with linear PRECISE. (**C**) Comparison of distances between PDXs and cell lines from the same tissue type (blue) or from a different tissue type (orange) without domain adaptation. (**D**) Comparison of distances when using linear PRECISE. We zoom in on lung (NSCLC) without domain adaptation (**E**), with linear PRECISE (**F**) or with TRANSACT (**G**).

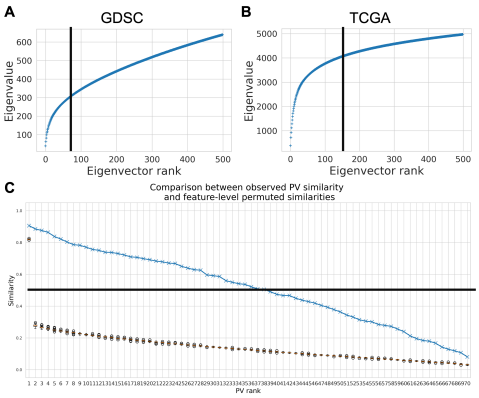


Figure C.9 – **Choice of the number of NLPCs and consensus features between GDSC and TCGA**. (**A**) Cumulative sum of eigenvalues of $(\widetilde{K}_s)$ (GDSC) with $\gamma = 5 \times 10 * *-4$. The cumulative sum increases steeply, reaches an inflection point and then follows an almost-linear behavior. We select all the NLPCs before this almost-linear zone, corresponding to 75 NLPCs. (**B**) Cumulative sum of eigenvalues of $(\widetilde{K}_t)$ (TCGA) with $\gamma = 5 \times 10 * *-4$. Following similar reasoning as in (A), we restrict the study to the first 150 NLPCs. (**C**) Similarity between PVs when 75 NLPCs are considered for GDSC and 150 for TCGA. We observe that the 33 first PVs have a similarity above 0.5 (our cut-off) and round the selection to 30 PVs.

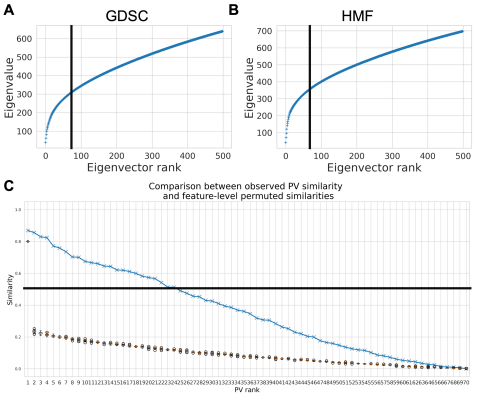


Figure C.10 – **Choice of the number of NLPCs and consensus features between GDSC and HMF**. (**A**) Cumulative sum of eigenvalues of $(\widetilde{K}_s)$ (GDSC) with $\gamma = 5 \times 10 * *-4$. The cumulative sum increases steeply, reaches an inflection point and then follows an almost-linear behavior. We select all the NLPCs before this almost-linear zone, corresponding to 75 NLPCs. (**B**) Cumulative sum of eigenvalues of $(\widetilde{K}_t)$ (HMF) with $\gamma = 5 \times 10 * *-4$. Following similar reasoning as in (A), we restrict the study to the first 150 NLPCs. (**C**) Similarity between PVs when 75 NLPCs are considered for GDSC and 150 for HMF. We observe that the 33 first PVs have a similarity above 0.5 (our cut-off) and round the selection to 30 PVs.

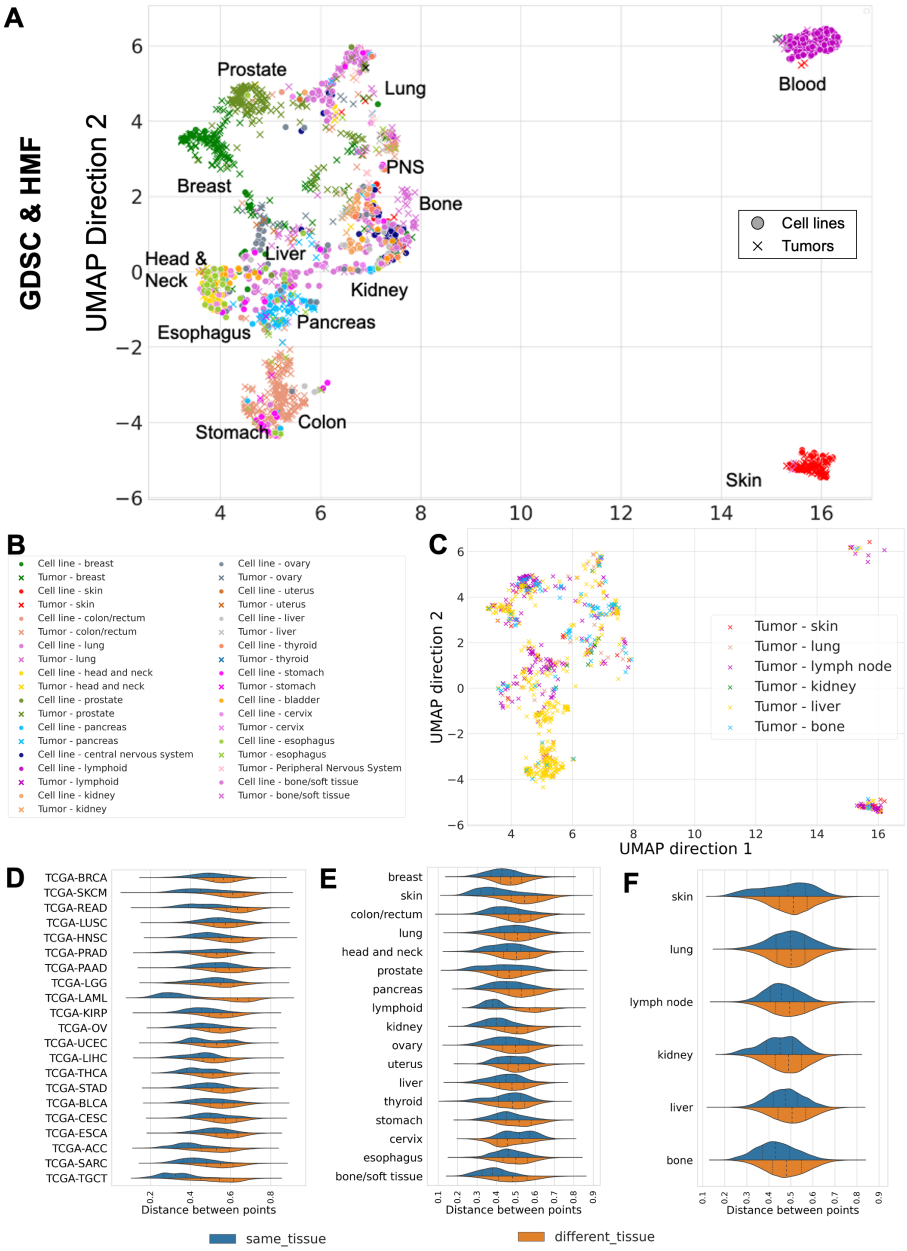


Figure C.11 – **Pan-cancer consensus features between cell lines and tumors conserve tissue type information**. (**A**) UMAP plot of metastatic lesions (HMF) and cell lines, colored by primary tissue for both HMF and GDSC. For both UMAP plots, the full legend can be found in panel B. (**B**) Legend of UMAP plots for Figure 4.3D-E. (**C**) UMAP plot of HMF metastatic lesions (same as Figure 4.3E) colored by metastatic site. (**D**) In TCGA, for each tumor type, distance between tumors and cell lines from similar (blue) and non-similar (orange) tissue. (**E**) In HMF, for each primary tumor type, distance between metastatic sample and cell line from similar and non-similar tissue of origin. (**F**) In HMF, for each metastatic site, distance between metastatic sample and cell line from tissue of origin similar (blue) or dissimilar from the metastatic site.

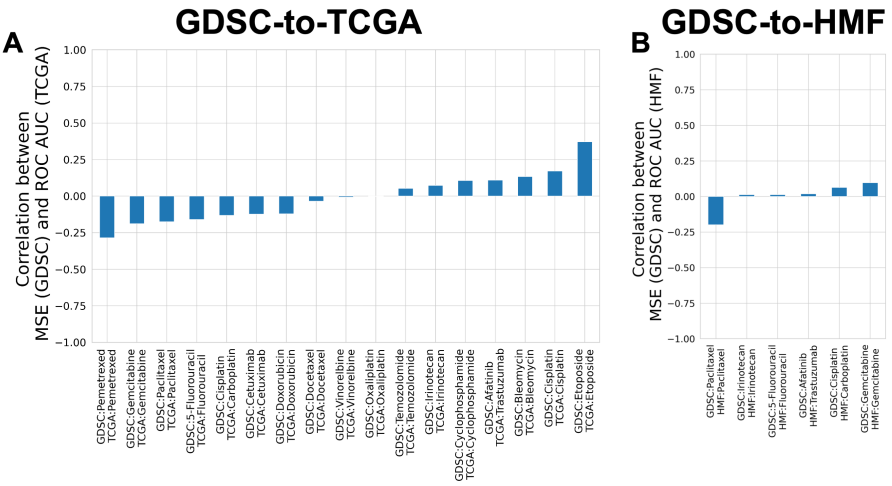


Figure C.12 – **Impact of initialization on results for the Deep Learning (DL) approach**. For each drug on TCGA and HMF, we considered the architecture and the set of hyper-parameters with the lowest Mean Squared Error on GDSC given an initialization. We then randomly generated 50 independent initializations of the resulting networks and trained them using the GDSC data. Each of these trained networks was then employed to predict the TCGA or HMF response. The resulting prediction accuracies (area under the ROC) are plotted for the different drugs on the TCGA and HMF data. (**A**) Pearson correlation of the Mean Square Error of the predictor on GDSC to the Area under the ROC of the same predictor on TCGA. (**B**) Pearson correlation on HMF between MSE (GDSC) and Area under the ROC (HMF).

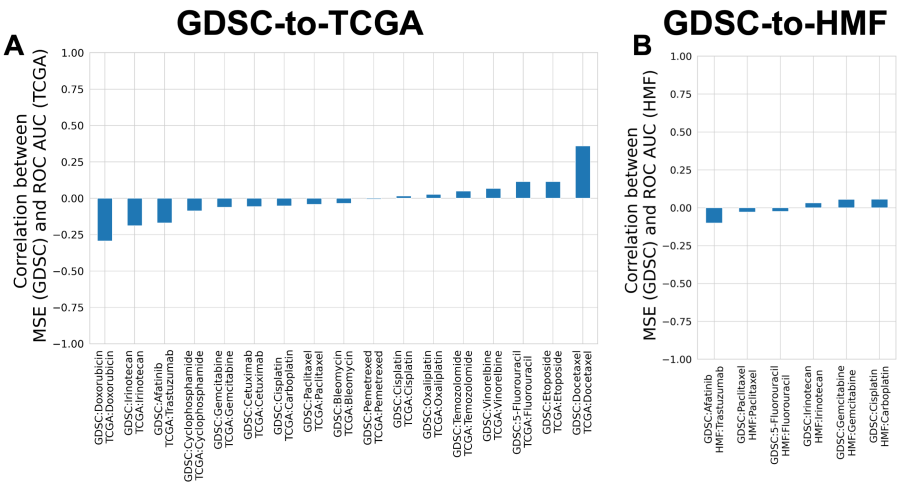


Figure C.13 – **Impact of initialization on results for the ComBat+DL approach**. For each drug on TCGA and HMF, we considered the architecture and the set of hyper-parameters with the lowest Mean Squared Error on GDSC given an initialization. We then randomly generated 50 independent initializations of the resulting networks and trained them using the GDSC data. Each of these trained networks was then employed to predict the TCGA or HMF response. The resulting predictions accuracies (area under the ROC) are plotted for the different drugs on the TCGA and HMF data. (**A**) Pearson correlation of the Mean Square Error of the predictor on GDSC to the Area under the ROC of the same predictor on TCGA. (**B**) Pearson correlation on HMF between MSE (GDSC) and Area under the ROC (HMF).

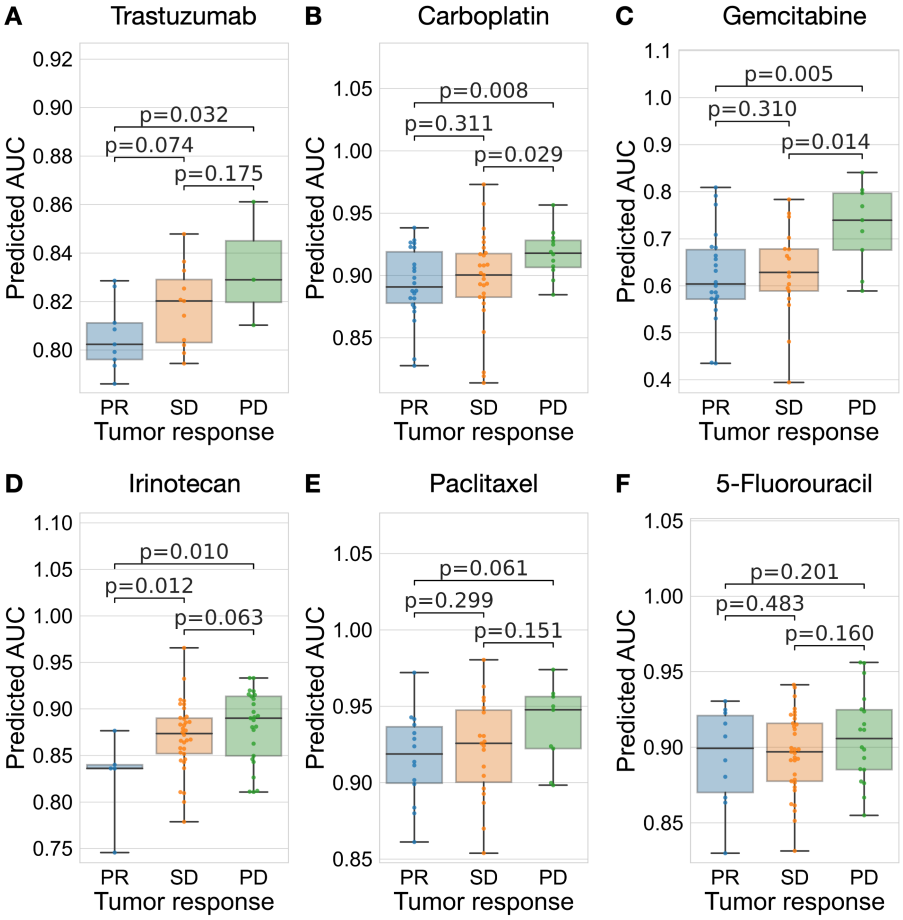


Figure C.14 – **Comparison of clinical status and AUC predicted by TRANSACT for HMF patients**. Using TRANSACT and a predictive model trained solely on GDSC response data, we predicted the response of HMF patients to six different drugs (y-axis). These predicted values are then compared to clinical response which fall into three possible categories: PR (Partial Response), SD (Stable Disease) or PD (Progressive Disease). Patients treated with six drugs were considered: Trastuzumab (**A**), Carboplatin (**B**), Gemcitabine (**C**), Irinotecan (**D**), Paclitaxel (**E**) and 5-Fluorouracil (**D**).

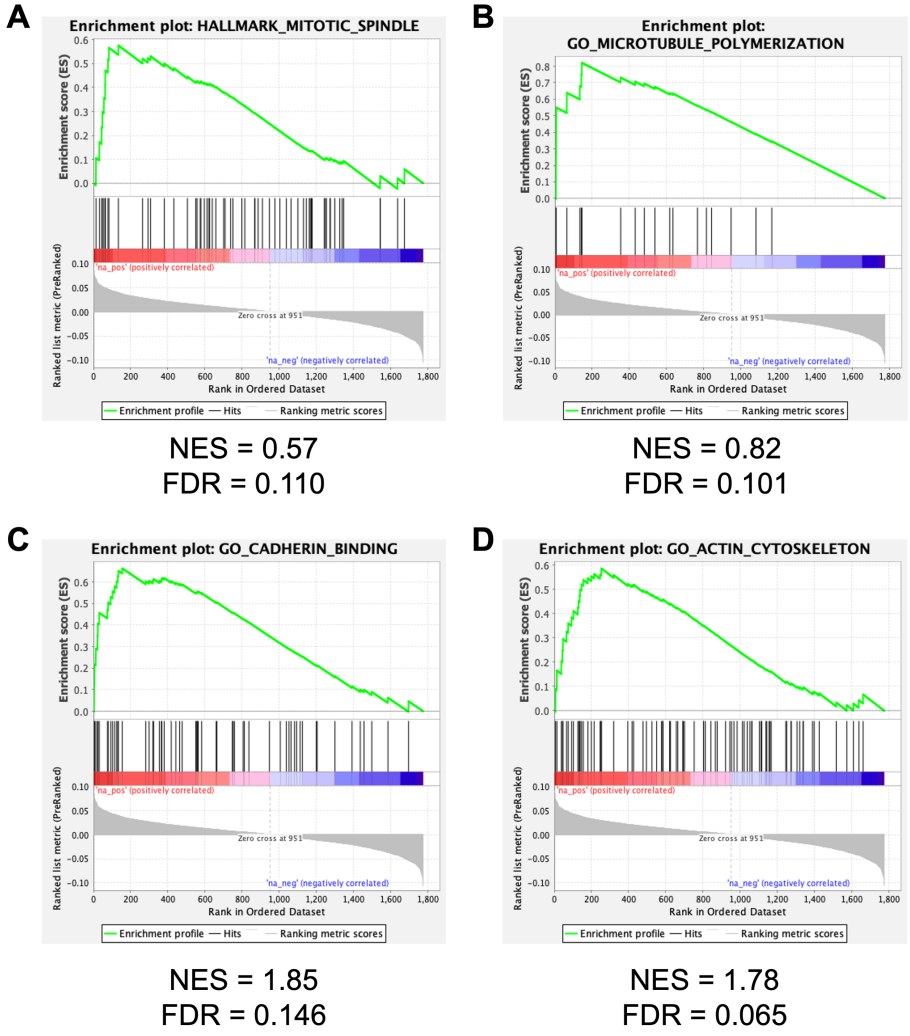


Figure C.15 – **Pathway enriched for resistant linear coefficients in GDSC-to-TCGA Gemcitabine drug response predictor**. Additional pathways significantly enriched in the linear part of the GDSC-to-TCGA predictor.

C.2. NOTATIONS AND SETTINGS

In our scenario, we have two datasets living in the same space – i.e. represented by the same p features (genes, SNPs, methylation probes, ...):

- A source dataset $\mathcal{X}_s = \{x_1^s, x_2^s, .., x_{n_s}^s\} \subset \mathbb{R}^p$, with labels $\mathcal{Y}_s = \{y_1^s, ...y_{n_s}^s\}$.
- A target dataset $\mathcal{X}_t = \{x_1^t, x_2^t, .., x_{n_t}^t\} \subset \mathbb{R}^p$ usually unlabelled.

We represent the source (resp. target) data as a matrix $X_s \in \mathbb{R}^{n_s \times p}$ (resp. $X_t \in \mathbb{R}^{n_t \times p}$) with samples in the rows and features in the columns.
We consider a similarity function, or kernel, $K : \mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R}$ that we will assume for the sequel to be positive semi-definite. Using the theory of Reproducible Kernel Hilbert Space [272], K is represented by the following dual formulation.

Proposition C.2.1 (Reproducing Hilbert Space)**.** *There exists a unique functional Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$, with $\mathcal{H} \subset \mathcal{F}(\mathbb{R}^p, \mathbb{R})$ (functions from $\mathbb{R}^p$ to $\mathbb{R}$), and a mapping function $\varphi : \mathbb{R}^p \mapsto \mathcal{H}$ such that:*

$$\forall x, y \in \mathbb{R}^p, \quad K(x,y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}. \tag{C.1}$$

The mapping φ furthermore satisfies the Reproducing property:

$$\forall f \in \mathcal{H}, \quad f : x \in \mathbb{R}^p \mapsto \langle \varphi(x), f \rangle_{\mathcal{H}}. \tag{C.2}$$

We refer to d_s (resp. d_t) the number of low-rank components we reduced the source data (resp. target data) to. We set d as the maximum number of principal vectors, $d = \min(d_s, d_t)$.
Superscript s is used for source items and superscript t for target items. $K(x,\cdot)$, for $x \in \mathbb{R}^p$, is the function $y \in \mathbb{R}^p \mapsto K(x,y)$. We use the superscript $\cdot^T$ as the transposition operation.
Finally, we define the following kernel matrices:

Definition C.2.2 (Kernel matrices)**.** *We define the following four matrices:*

- ***Source kernel matrix*** $K_s : K_s = \left[K\left(x_i^s, x_j^s\right)\right]_{1 \le i,j \le n_s} \in \mathbb{R}^{n_s \times n_s}$.
- ***Target kernel matrix*** $K_t : K_t = \left[K\left(x_i^t, x_j^t\right)\right]_{1 \le i,j \le n_t} \in \mathbb{R}^{n_t \times n_t}$.
- ***Source-target kernel matrix*** $K_{st} : K_{st} = \left[K\left(x_i^s, x_j^t\right)\right]_{1 \le i \le n_s, 1 \le j \le n_t} \in \mathbb{R}^{n_s \times n_t}$.
- ***Target-source kernel matrix*** *: K_{ts} as $K_{ts} = K_{st}^T \in \mathbb{R}^{n_t \times n_s}$.*

C.3. KERNEL-MEAN CENTERING

We set out to work in the Hilbert space $\mathcal{H}$ after embedding the data with the mapping φ. Prior to any statistical processing, we first need to mean-center the data *in the kernel*

feature space $\mathscr{H}$. For that purpose, we define two means, the *mean source embedding* μ^s and the *mean target embedding* μ^t, as follows:

$$\begin{aligned} \mu^s &= \frac{1}{n_s}\sum_{i=1}^{n_s}\varphi\left(x_i^s\right) = \frac{1}{n_s}\sum_{i=1}^{n_s}K\left(x_i^s,\cdot\right) \\ \mu^t &= \frac{1}{n_t}\sum_{i=1}^{n_t}\varphi\left(x_i^t\right) = \frac{1}{n_s}\sum_{i=1}^{n_t}K\left(x_i^t,\cdot\right) \end{aligned} \tag{C.3}$$

Using the means computed in Equation (C.3), we define two sets of corrected embeddings as follows:

Definition C.3.1 (Mean-centered embedding and kernel function)**.** *The source centered kernel embedding* $\widetilde{\varphi}_s$ *is defined as:*

$$\forall x \in \mathbb{R}^p, \quad \widetilde{\varphi}_s(x) = \varphi(x) - \mu_s = K(x,\cdot) - \mu_s. \tag{C.4}$$

We then defined the source-centered kernel function $\widetilde{K}_s$ *as:*

$$\forall x, y \in \mathbb{R}^p, \quad \widetilde{K}_s\left(x,y\right) = \langle \widetilde{\varphi}_s(x), \widetilde{\varphi}_s\left(y\right).\rangle \tag{C.5}$$

We define equivalently the target centered kernel embedding $\widetilde{\varphi}_t$ *and corresponding target-centered kernel function* $\widetilde{K}_t$.

We use the mean-centered kernel functions defined in Definition C.3.1 to correct the kernel matrices from Definition C.2.2 and define the following four matrices.

Definition C.3.2 (Centered Kernel matrices)**.** *We define the following four matrices:*

- ***Source-centered kernel matrix*** $\widetilde{K}_s$: $\widetilde{K}_s = \left[\widetilde{K}_s\left(x_i^s, x_j^s\right)\right]_{1\leq i,j\leq n_s} \in \mathbb{R}^{n_s\times n_s}$.
- ***Target-centered kernel matrix*** $\widetilde{K}_t$: $\widetilde{K}_t = \left[\widetilde{K}_t\left(x_i^t, x_j^t\right)\right]_{1\leq i,j\leq n_t} \in \mathbb{R}^{n_t\times n_t}$.
- ***Source-target-centered kernel matrix*** $\widetilde{K}_{st}$: $\widetilde{K}_{st} = \left[\langle \widetilde{\varphi}_s\left(x_i^s\right), \widetilde{\varphi}_t\left(x_j^t\right)\rangle\right]_{1\leq i\leq n_s, 1\leq j\leq n_t} \in \mathbb{R}^{n_s\times n_t}$.
- ***Target-source kernel matrix*** : $\widetilde{K}_{ts}$ *as* $\widetilde{K}_{ts} = \widetilde{K}_{st}^T \in \mathbb{R}^{n_t\times n_s}$.

To get a relation between matrices given in Definition C.3.2 and Definition C.2.2, we define the centering matrix of size n, denoted as $\mathbb{C}_n$:

Definition C.3.3 (Centering matrix)**.** *Let* $n \in \mathbb{N}_*$. *We define the centering matrix of size* n, *denoted* C_n *as:*

$$C_n = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T, \tag{C.6}$$

where I_n *is the identity matrix of size* n *and* $\mathbf{1}_n$ *is the* n*-sized vector constituted solely of 1.*

Proposition C.3.4 (Computation of centered kernel matrices)**.** *We have the following equalities:*

$$\begin{aligned} \widetilde{K}_s &= C_{n_s} K_s C_{n_s}, \\ \widetilde{K}_t &= C_{n_t} K_t C_{n_t}, \\ \widetilde{K}_{st} &= C_{n_s} K_{st} C_{n_t}. \end{aligned} \tag{C.7}$$

C.4. KERNEL PCA ON SOURCE AND TARGET

We use Kernel PCA to compute directions of maximum variance in the embedded space [167], yielding kernel Principal Components, also called *non-linear principal components* (NLPCs) in the main text. These NLPCs for source and target are respectively defined as linear combinations of source and target samples' embeddings (after mean-centering) in the kernel feature space.

Definition C.4.1 (Non-linear source and target principal components [167])**.** *Non-linear principal components for source* $\left(f_1^s, .., f_{d_s}^s\right)$ *and target* $\left(f_1^t, .., f_{d_t}^t\right)$ *are defined as linear combinations of source and target embedded samples respectively. Denoting as* α^s *the* d_s *top eigenvectors of* $\widetilde{K_s}$ *and* α^t *the* d_t *top eigenvectors of* $\widetilde{K_t}$*, we have the following equality:*

$$\begin{cases} f_q^s = \sum_{i=1}^{n_s} \alpha_{q,i}^s \widetilde{\varphi}_s\left(x_i^s\right) & \text{for } q \in \{1, .., d_s\}, \\ f_q^t = \sum_{i=1}^{n_t} \alpha_{q,i}^t \widetilde{\varphi}_t\left(x_i^t\right) & \text{for } q \in \{1, .., d_t\}, \end{cases} . \tag{C.8}$$

These non-linear principal directions satisfy some orthogonality constraints on the kernel space $\mathscr{H}$*:*

$$\forall x \in \{s, t\}, \quad \forall k, l \in \{1, .., d\}, \quad \langle f_k^x, f_l^x \rangle_{\mathscr{H}} = \delta_{k,l}, \tag{C.9}$$

where δ *is the equality indicator function. These constraints are equivalent to:*

$$\alpha^s \widetilde{K}_s {\alpha^s}^T = I_{d_s} \quad \text{and} \quad \alpha^t \widetilde{K}_t {\alpha^t}^T = I_{d_t} \tag{C.10}$$

The two matrices $\alpha^s \in \mathbb{R}^{d_s \times n_s}$ and $\alpha^t \in \mathbb{R}^{d_t \times n_t}$ correspond to factors by samples matrices, but do not represent the projected score. Instead, they are equivalent to the feature loadings in linear PCA and correspond to a dual representation of the features in $\mathscr{H}$ that can not be explicitly computed due to the high-dimensions of $\mathscr{H}$. We refer to them as *sample importance loadings* to explicit the difference these have with projected scores.

C.5. VARIATIONAL DEFINITION OF PRINCIPAL VECTORS

We define the first pair of principal vectors between source and target NLPCs as the two unitary vectors s_1 and t_1, with s_1 in source NLPCs span and t_1 in target NLPCs span, such that their similarity is maximized. This extends in $\mathscr{H}$ the principal vectors defined by Golub and Van Loan in [104] and are mathematically formalized using the following variational definition:

$$
\begin{aligned}
s_1, t_1 \quad = \quad & \underset{\substack{s \in \text{span}\left(f_1^s, .., f_{d_s}^s\right), \\ t \in \text{span}\left(f_1^t, .., f_{d_t}^t\right)}}{\text{argmax}} \quad \langle s, t \rangle_{\mathcal{H}} \\
& \text{s.t} \quad \langle s, s \rangle_{\mathcal{H}} = \langle t, t \rangle_{\mathcal{H}} = 1
\end{aligned} \quad . \tag{C.11}
$$

We further define the principal vector by adding an orthogonality constraint, as in [104].

Definition C.5.1 (Kernel Principal Vectors)**.** *We define the d pairs of principal vectors $(s_1, t_1), (s_2, t_2), .., (s_d, t_d)$ as, for all $k \in \{1, .., d\}$:*

$$
\begin{aligned}
s_k, t_k \quad = \quad & \underset{\substack{s \in \text{span}\left(f_1^s, .., f_{d_s}^s\right), \\ t \in \text{span}\left(f_1^t, .., f_{d_t}^t\right)}}{\text{argmax}} \quad \langle s, t \rangle_{\mathcal{H}} \\
& s.t \quad \langle s, s \rangle_{\mathcal{H}} = \langle t, t \rangle_{\mathcal{H}} = 1, \\
& \textit{and } \forall l < k, s_l \perp s, t_l \perp t
\end{aligned} \quad . \tag{C.12}
$$

C.6. COMPUTATION OF PRINCIPAL VECTORS

The first step towards computing principal vectors is to compare the principal components defined in Definition C.4.1. We define for that purpose the cosine similarity matrix between source and target NLPCs and present a closed-form solution for computing it based on centered kernel matrices (Definition C.3.2) and NLPCs' coefficients (Definition C.4.1).

The cosine similarity matrix is a standard way to compare orthonormal basis of vectors and has already been used to compare linear principal components in subspace-based domain adaptation [58–60]. We here extend it to kernel-based non-linear dimensionality reduction.

Definition C.6.1 (Cosine similarity matrix)**.** *We define the cosine similarity matrix $\mathbf{M}^K$ between source and target kernel principal components as:*

$$
\mathbf{M}^K \quad = \quad \left[\langle f_k^s, f_l^t \rangle_{\mathcal{H}}\right]_{1 \leq k \leq d_s, 1 \leq l \leq d_t} \quad \in \mathbb{R}^{d_s \times d_t}. \tag{C.13}
$$

Proposition C.6.2 (Computation of cosine similarity matrix)**.** *$\mathbf{M}^K$ can be computed using the matrices α^s, α^t and K_{ST} as:*

$$
\begin{aligned}
\mathbf{M}^K \quad &= \quad \alpha^s \widetilde{K}_{st} {\alpha^t}^T \\
&= \quad \alpha^s C_{n_s} K_{st} C_{n_t} {\alpha^t}^T.
\end{aligned} \tag{C.14}
$$

Proof. Let $1 \leq k, l \leq d$, then using Equation C.8,

$$
\langle f_k^s, f_l^t \rangle \; = \; \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} \alpha_{k,i}^s \alpha_{l,j}^t \langle \widetilde{\varphi}_s\left(x_i^s\right), \widetilde{\varphi}_t\left(x_j^t\right) \rangle \; = \; {\alpha_{k,:}^s}^T \widetilde{K}_{st} \alpha_{l,:}^t \tag{C.15}
$$

which put together as a matrix gives the wanted result. □

Similarly to the linear setting, we use this cosine similarity matrix to NLPC by means of SVD of $\mathbf{M}^K$.

Theorem C.6.3 (SVD computation of Principal Vectors)**.** *Let $\beta^s \in \mathbb{R}^{d_s \times d}$ (resp. $\beta^t \in \mathbb{R}^{d_t \times d}$) be the first d left (resp. right) singular vectors of $\mathbf{M}^K$, i.e. $\mathbf{M}^K \approx \beta^s \Sigma {\beta^t}^T$. Then, for all $1 \leq q \leq d$:*

$$s_q = \sum_{k=1}^{d_s} \sum_{i=1}^{n_s} \beta_{k,q}^s \alpha_{k,i}^s \widetilde{\varphi}_s\left(x_i^s\right) \quad \textit{and} \quad t_q = \sum_{l=1}^{d_t} \sum_{j=1}^{n_t} \beta_{l,q}^t \alpha_{l,j}^t \widetilde{\varphi}_t\left(x_j^t\right) \tag{C.16}$$

C

Proof. Let $s_1, .., s_d \in \operatorname{span}\left(f_1^s, .., f_{d_s}^s\right)$ and $t_1, .., t_d \in \operatorname{span}\left(f_1^t, .., f_{d_t}^t\right)$ with norm 1, there exists $\beta^s, \in \mathbb{R}^{d_s \times d}$ and $\beta^t \in \mathbb{R}^{d_t \times d}$ such that, for all $q \in \{1, .., d\}$,

$$s_q = \sum_{k=1}^{d_s} \beta_{k,q}^s f_k^s = \sum_{i=1}^{n_s} \sum_{k=1}^{d_s} \alpha_{k,i}^s \beta_{k,q}^s \widetilde{\varphi}_s\left(x_i^s\right) \quad \text{and} \quad t_q = \sum_{l=1}^{d_t} \beta_{l,q}^t f_l^t = \sum_{j=1}^{n_t} \sum_{l=1}^{d_t} \alpha_{l,j}^t \beta_{l,q}^t \widetilde{\varphi}_t\left(x_j^t\right). \tag{C.17}$$

The orthogonality constraint $\langle s_k, s_l \rangle_{\mathcal{H}} = \langle t_k, t_l \rangle_{\mathcal{H}} = \delta_{k,l}$, for $1 \leq k, l \leq d$ coupled with the orthogonality constrains from Equation (C.9) is then equivalent to ${\beta^s}^T \beta^s = {\beta^t}^T \beta^t = I_d$. Computing inner product between source and target PV therefore yields

$$[\langle s_k, t_l \rangle]_{1 \leq k,l \leq d} = {\beta^s}^T {\alpha^s} \widetilde{K}_{st} {\alpha^t}^T {\beta^t}^T = {\beta^s}^T \mathbf{M}^K \beta^t. \tag{C.18}$$

Therefore, the maximization problem from Equation (C.11) is equivalent to the following:

$$\begin{aligned} &\max_{\substack{\beta^s \in \mathbb{R}^{d_s \times d}, \\ \beta^t \in \mathbb{R}^{d_t \times d}}} \quad {\beta^s}^T \mathbf{M}^K \beta^t \\ &\text{s.t.} \quad {\beta^s}^T \beta^s = {\beta^t}^T \beta^t = I_d \end{aligned}, \tag{C.19}$$

which unique solutions are the left and right orthogonal vectors of $\mathbf{M}^K$, obtained by SVD. □

In order to work at the sample-level for each principal vector, we define the PV sample-importance loadings as follows.

Definition C.6.4 (Principal Vector sample importance loadings)**.** *We define the source (resp. target) sample importance loadings $\rho^s \in \mathbb{R}^{d \times n_s}$ (resp. $\rho^t \in \mathbb{R}^{d \times n_t}$) as:*

$$\rho^s = {\beta^s}^T {\alpha^s} \quad \textit{and} \quad \rho^t = {\beta^t}^T {\alpha^t}. \tag{C.20}$$

These PV importance loadings are related to the source and target PVs as follow:

Proposition C.6.5. *Source and target principal vectors have the equivalent following definition:*

$$\forall q \in \{1, ..d\}, \quad \begin{cases} s_q = \sum_{i=1}^{n_s} \rho_{q,i}^s \widetilde{\varphi}_s\left(x_i^s\right), \\ t_q = \sum_{i=1}^{n_t} \rho_{q,i}^t \widetilde{\varphi}_t\left(x_i^t\right). \end{cases} \tag{C.21}$$

We finally defined the similarity between the principal vectors as cosines of angles referred to as *principal angles.*

Definition C.6.6 (Principal Angles)**.** *Let $1 \leq q \leq d$. We define the q-th principal angle as the unique $\theta_q \in \left[0, \frac{\pi}{2}\right]$ that satisfies:*

$$\cos\theta_q = \langle s_q, t_q \rangle_{\mathscr{H}}. \tag{C.22}$$

Proposition C.6.7 (SVD computation of Principal Angles)**.** *Let Σ be the diagonal matrix obtained by SVD of $\mathbf{M}^K$ (as in Proposition C.6.2), then:*

$$\forall q \in \{1,..,d\}, \quad \cos\theta_q = \Sigma_{q,q}. \tag{C.23}$$

Proof.

$$\cos\theta_q = \langle s_q, t_q \rangle_{\mathscr{H}} = {\beta^s_{:,q}}^T \mathbf{M}^K \beta^t_{:,q} = \Sigma_{q,q}, \tag{C.24}$$

by definition of the SVD. □

We showed how to compute the PVs as functions in $\mathscr{H}$ and gave a closed-form solution for the evaluation in $\mathbb{R}^p$. We finally show that the evaluation of PVs correspond to a projection of the embedded vector, keeping the same intuition than in linear setting.

Proposition C.6.8 (Evaluation of principal vectors)**.** *Let $x \in \mathbb{R}^p$. For $q \in \{1,..d\}$, the evaluation of source and target principal vectors s_q and t_q is equivalent to the projection of the embedding of x on these vectors:*

$$s_q(x) = \langle s_q, \varphi(x) \rangle_{\mathscr{H}} \quad \textit{and} \quad t_q(x) = \langle t_q, \varphi(x) \rangle_{\mathscr{H}} \tag{C.25}$$

Proof. Combining Equations (C.3), (C.4) and (C.21), source PV are sum of elements of $\mathscr{H}$:

$$s_q = \sum_{i=1}^{n_s} \rho^s_{q,i} \widetilde{\varphi_s}\left(x^s_i\right) \quad \text{with,} \quad \forall i \in \{1,..,n_s\}, \, \widetilde{\varphi_s}\left(x^s_i\right) \in \mathscr{H}. \tag{C.26}$$

Therefore $s_q \in \mathscr{H}$ since $\mathscr{H}$ is an Hilbert space. Using the reproducing property of the RKHS and the definition of φ (Equation (C.1), we obtain

$$\forall x \in \mathbb{R}^p, \quad s_q(x) = \left\langle s_q, \varphi(x) \right\rangle_{\mathscr{H}}. \tag{C.27}$$

Following the same idea, we obtain the equivalent equality for target PVs. □

C.7. INTERPOLATION SCHEME

The Principal Vectors are pairs of vectors (one form source, one from target) that are geometrically similar. We select only the pairs above a certain threshold of similarity in order to restrict to directions shared by the two signals. Therefore, within each pair, source and target vectors show an important correlation and using the two into a predictive model would not be optimal. We therefore set out to construct a single vector out of each pair by interpolating between the two vectors. This interpolation is the geodesic flow between PVs and is defined as follows.

Definition C.7.1 (Angular interpolation function)**.** *Let $q \in \{1,..,d\}$, we define the angular interpolation functions Γ_q and ξ_q between the q^{th} pair of principal vector as:*

$$\forall \tau \in [0,1], \quad \Gamma_q(\tau) = \frac{\sin\left((1-\tau)\theta_q\right)}{\sin\theta_q} \quad \textit{and} \quad \xi_q(\tau) = \frac{\sin\tau\theta_q}{\sin\theta_q}. \tag{C.28}$$

Definition C.7.2 (Geodesic flow between principal vectors)**.** *Let $q \in \{1,..,d\}$, we define the interpolation ϕ_q between the q^{th} pair of principal vector as:*

$$\forall \tau \in [0,1], \quad \phi_q(\tau) = \Gamma_q(\tau)\, s_q + \xi_q(\tau)\, t_q. \tag{C.29}$$

Since $\mathscr{H}$ is a Hilbert space, $\phi_q \in \mathscr{H}$.

Proposition C.7.3 (Estimation using PV sample importance loadings)**.** *Let $q \in \{1,..,d\}$ and ϕ_q be the geodesic between the q^{th} pair of principal vectors. The geodesic defined in Equation (C.29) has the following equivalent formulation:*

$$\forall \tau \in [0,1], \quad \phi_q(\tau) = \Gamma_q(\tau) \sum_{i=1}^{n_s} \rho^s_{q,i} \widetilde{\varphi}_s\left(x^s_i\right) + \xi_q(\tau) \sum_{j=1}^{n_t} \rho^t_{q,j} \widetilde{\varphi}_t\left(x^t_j\right). \tag{C.30}$$

Proof. Combining the definition of the geodesic from Definition C.7.2 with the equivalent principal vector formulation of Proposition C.6.5 yields the result. □

The formulation of the geodesic from Proposition C.7.3 can easily be written down as a matrix product (for computation purposes) for each sample. We define the matrix angular interpolation function as follow.

Definition C.7.4 (Matrix angular interpolation function)**.** *We define the matrix angular interpolation functions Γ and Ξ*

$$\forall \tau \in [0,1]^d, \quad \Gamma(\tau) = \operatorname{diag}\left[\Gamma_q\left(\tau_q\right)\right]_{1\le q\le d} \quad \textit{and} \quad \Xi(\tau) = \operatorname{diag}\left[\xi\left(\tau_q\right)\right]_{1\le q\le d}. \tag{C.31}$$

Proposition C.7.5 (Matrix estimation of principal vectors)**.** *Let's denote by s (resp. t) the vectors of d source (resp. target) principal vectors ordered by similarity. We define $\mathscr{S}^s$ and $\mathscr{S}^t$ as the matrices that contain the source principal vectors values evaluated on source and target data respectively:*

$$\mathscr{S}^s = \left[s\left(x_1^s\right)^T, .., s\left(x_{n_s}^s\right)^T \right]^T \in \mathbb{R}^{n_s \times d}, \tag{C.32}$$

$$\mathscr{S}^t = \left[s\left(x_1^t\right)^T, .., s\left(x_{n_t}^t\right)^T \right]^T \in \mathbb{R}^{n_t \times d}. \tag{C.33}$$

We define similarly $\mathscr{T}^s \in \mathbb{R}^{n_s \times d}$ as the matrix that contains the target principal vectors evaluated on the source data – and $\mathscr{T}^t \in \mathbb{R}^{n_t \times d}$ as the matrix that contains the target principal vectors evaluated on the target data. These matrices can be computed as follows:

$$\begin{cases} \mathscr{S}^s &= K^s C_{n_s} \rho^{s\,T}, \\ \mathscr{S}^t &= K^{st} C_{n_s} \rho^{s\,T}, \end{cases} \quad \text{and} \quad \begin{cases} \mathscr{T}^t &= K^{ts} C_{n_t} \rho^{t\,T}. \\ \mathscr{T}^t &= K^t C_{n_t} \rho^{t\,T}. \end{cases} \tag{C.34}$$

Proof. Using the definition of principal vectors with ρ coefficients from Equation (C.21), we get, for $l \in \{1, .., n_s\}$ and $q \in \{1, .., d\}$:

$$\begin{aligned} s_q\left(x_l^s\right) &= \sum_{i=1}^{n_s} \rho_{q,i}^s \left[K\left(x_i^s, x_l^s\right) - \frac{1}{n_s} \sum_{j=1}^{n_s} \left(x_j^s, x_l^s\right) \right] \\ &= \sum_{i=1}^{n_s} \left(\rho_{q,:}^s\right)_i \left[K_{i,l}^s - \frac{1}{n_s} \left(\mathbf{1}_{n_s} \mathbf{1}_{n_s}^T K^s\right)_{i,l}. \right] \end{aligned} \tag{C.35}$$

Using the centering matrix defined in Definition C.3.3, we get:

$$s_k\left(x_l^s\right) = \left[\rho^s C_{n_s} K^s\right]_{k,l}, \tag{C.36}$$

and therefore $\mathscr{S}^s = \left(\rho^s C_{n_s} K^s\right)^T$. The other equalities follow from the same proof. □

Let's finally define the geodesic matrix between source and target at interpolation time $\tau \in [0,1]$ as the estimation of both source and target on the geodesic in the kernel feature space.

Theorem C.7.6. *We define as $\mathbf{F}^{st}(\tau)$ for as the matrix of geodesic values evaluated at interpolation time $\tau \in [0,1]^d$, i.e.:*

$$\mathbf{F}^{st}(\tau) = \begin{bmatrix} \phi(\tau)\left(x_1^s\right) \\ \dots \\ \phi(\tau)\left(x_{n_s}^s\right) \\ \phi(\tau)\left(x_1^t\right) \\ \dots \\ \phi(\tau)\left(x_{n_t}^s\right) \end{bmatrix} \in \mathbb{R}^{(n_s+n_t)\times d}. \tag{C.37}$$

. Then $\mathbf{F}^{st}(\tau)$ can be computed as follow:

$$\mathbf{F}^{st}(\tau) = \begin{bmatrix} \mathscr{S}^s & \mathscr{T}^s \\ \mathscr{S}^t & \mathscr{T}^t \end{bmatrix} \begin{bmatrix} \Gamma(\tau) \\ \Xi(\tau) \end{bmatrix}. \tag{C.38}$$

This formulation is equivalent to:

$$\mathbf{F}^{st}(\tau) = \begin{bmatrix} K^s & K^{st} \\ K^{ts} & K^t \end{bmatrix} \begin{bmatrix} C_{n_s} & 0_{n_s \times n_t} \\ 0_{n_t \times n_s} & C_{n_t} \end{bmatrix} \begin{bmatrix} \rho^{s\,T} & 0_{n_s \times d} \\ 0_{n_t \times d} & \rho^{t\,T} \end{bmatrix} \begin{bmatrix} \Gamma(\tau) \\ \Xi(\tau) \end{bmatrix}. \tag{C.39}$$

Proof. Direct by combining Definition C.7.4 and Proposition C.7.5. □

In order to get zero-centered projected source and target samples, we can use two solutions. On one hand, we can perform a consensus-feature-level mean-centering independently on source and target after projection. Equivalently, we can also left-multiply by centering matrix the projected matrix $\mathbf{F}^{st}(\tau)$.
We finally show that the evaluation of the consensus features functions is equivalent to the projection of embedding in the feature space $\mathscr{H}$.

Proposition C.7.7. *Let $x \in \mathbb{R}^p$, $q \in \{1,..d\}$ and $\tau_q \in [0,1]$, then:*

$$\phi_q(\tau_q)(x) = \langle \phi_q(\tau_q), \varphi(x) \rangle_{\mathscr{H}}. \tag{C.40}$$

Proof. Using Proposition C.6.8,

$$\begin{aligned} \phi_q(\tau_q)(x) &= \Gamma_q(\tau_q) s_q(x) + \xi_q(\tau_q) t_q(x) \\ &= \langle \Gamma_q(\tau_q) s_q + \xi_q(\tau_q) t_q, \varphi(x) \rangle_{\mathscr{H}} \\ &= \langle \phi_q(\tau_q), \varphi(x) \rangle_{\mathscr{H}}. \end{aligned} \tag{C.41}$$

□

C.8. GENE SET ENRICHMENT ANALYSIS OF CONSENSUS FEATURES

In order to gain insight into the making of consensus features, we use a Taylor expansion of the Gaussian kernel [169]. The Gaussian kernel can be expressed as outer-product of the following basis functions.

Definition C.8.1. *Let $i \leq 0$ be an integer. We define as $e_i : \mathbb{R} \mapsto \mathbb{R}$ the basis function defined as:*

$$\forall x \in \mathbb{R}, \quad e_i(x) = \sqrt{\frac{2\gamma^i}{i!}} x^i \exp(-\gamma x^2). \tag{C.42}$$

Proposition C.8.2 (Countable orthonormal basis of $\mathscr{H}$ [169])**.** *Let's define for $(i_1,..,i_p) \in \mathbb{N}^p$ the following function*

$$e_{(i_1,..,i_p)} = x \in \mathbb{R}^p \mapsto e_{i_1}(x_1) \times e_{i_2}(x_2) \times \cdots \times e_{i_p}(x_p). \tag{C.43}$$

Then, $\left(e_{(i_1,..,i_p)}\right)_{(i_1,..,i_p) \in \mathbb{N}^p}$ is an orthonormal basis of $\mathscr{H}$, and for $x, y \in \mathbb{R}^p$,

$$\begin{aligned} \exp(-\gamma ||x-y||^2) &= \sum_{i_1,..,i_p \in \mathbb{N}^p} e_{(i_1,..,i_p)}(x)\, e_{(i_1,..,i_p)}(y). \\ &= \widehat{\varphi}(x)^T \widehat{\varphi}(y), \end{aligned} \tag{C.44}$$

with $\widehat{\varphi} : x \mapsto \left(e_{(i_1,..,i_p)}(x)\right)_{(i_1,..,i_p)\in\mathbb{N}^p}$.

Let's consider this approximation map $\widehat{\varphi}$. We extract three different features of interest for our analysis: the offset (sum of indices is 0), the linear terms (sum of indices is 1) and the interaction terms (sum of indices is 2). We define them as follows:

Definition C.8.3 (Offset, linear and interaction terms)**.**
We define the ***offset feature*** e_O *as* $e_{0_{\mathbb{N}^p}}$*, i.e. when all indices are 0.*
For each gene (feature $k \in \{1,..,p\}$*), we define the* k^{th} ***linear feature*** e_k *as* e_{δ_k} *where* δ_k *is the vector of zeros with a single 1 on* k^{th} *position.*
For each combination of genes (feature $k,l \in \{1,..,p\}$*), we define the* $(k,l)^{th}$ ***interaction feature*** $e_{k,l}$ *as* $e_{\delta_{k,l}}$ *where* $\delta_{k,l}$ *is the vector of zero with one 1 on* k^{th} *and* l^{th} *position only if* $k \neq l$*, and 2 on* k^{th} *position if* $k = l$*.*

Definition C.8.4 (Offset, linear and interaction terms for consensus features)**.**
We define the ***offset contribution*** *to consensus feature* q *as* $\mathscr{O}_q = \langle e_O, \phi_q\left(\tau_q^*\right)\rangle$.
For $k \in \{1,..,p\}$*, we define the* k^{th} ***linear contribution*** *to consensus feature* q *as* $\mathscr{L}_{q,k} = \langle e_k, \phi_q\left(\tau_q^*\right)\rangle$.
For $k,l \in \{1,..,p\}$*), we define the* $(k,l)^{th}$ ***interaction contribution*** *to consensus feature* q *as* $\mathscr{I}_{q,k,l} = \langle e_{k,l}, \phi_q\left(\tau_q^*\right)\rangle$.

We now compute the contribution of each of these features to the consensus features. We first rewrite the different contributions to the consensus features for readability.

Definition C.8.5. *For* $q \in \{1,..,d\}$*, we define* $\sigma_q^s = \Gamma_q\left(\tau_q^*\right)\rho_q^s$ *and* $\sigma_q^t = \xi_q\left(\tau_q^*\right)\rho_q^t$.

We finally define the source and target mean centered features.

Definition C.8.6. *We define the source (resp. target) mean-centered offset feature for the* q^{th} *consensus feature* $\widetilde{e}_O^s$ *(resp.* $\widetilde{e}_O^t$*) as:*

$$\widetilde{e}_O^s = e_O - \frac{1}{n_s}\sum_{i=1}^{n_s} e_O\left(x_i^s\right) \quad \text{and} \quad \widetilde{e}_O^t = e_O - \frac{1}{n_t}\sum_{i=1}^{n_t} e_O\left(x_i^t\right). \tag{C.45}$$

For $k \in \{1,..,p\}$*, we define the source (resp. target) mean-centered linear feature for the* q^{th} *consensus feature* $\widetilde{e}_k^s$ *(resp.* $\widetilde{e}_k^t$*) as:*

$$\widetilde{e}_k^s = e_k - \frac{1}{n_s}\sum_{i=1}^{n_s} e_k\left(x_i^s\right) \quad \text{and} \quad \widetilde{e}_k^t = e_k - \frac{1}{n_t}\sum_{i=1}^{n_t} e_k\left(x_i^t\right). \tag{C.46}$$

For $k,l \in \{1,..,p\}$*, we define the source (resp. target) mean-centered linear feature for the* q^{th} *consensus feature* $\widetilde{e}_{k,l}^s$ *(resp.* $\widetilde{e}_{k,l}^t$*) as:*

$$\widetilde{e}_{k,l}^s = e_{k,l} - \frac{1}{n_s}\sum_{i=1}^{n_s} e_{k,l}\left(x_i^s\right) \quad \text{and} \quad \widetilde{e}_{k,l}^t = e_{k,l} - \frac{1}{n_t}\sum_{i=1}^{n_t} e_{k,l}\left(x_i^t\right). \tag{C.47}$$

Proposition C.8.7. *The differents contribution $\mathscr{O}_q$ $\mathscr{L}_{q,i}$ and $\mathscr{I}_{q,i,j}$ for the q^{th} consensus feature can be computed as follow:*

$$\mathscr{O}_q = \sum_{i=1}^{n_s} \sigma^s_{q,i} \widetilde{e}^s_O(x^s_i) + \sum_{i=1}^{n_t} \sigma^t_{q,i} \widetilde{e}^t_O(x^t_i), \tag{C.48}$$

$$\mathscr{L}_{q,k} = \sum_{i=1}^{n_s} \sigma^s_{q,i} \widetilde{e}^s_{q,k}(x^s_i) + \sum_{i=1}^{n_t} \sigma^t_{q,i} \widetilde{e}^t_{q,k}(x^t_i), \tag{C.49}$$

$$\mathscr{I}_{q,k,l} = \sum_{i=1}^{n_s} \sigma^s_{q,i} \widetilde{e}^s_{q,k,l}(x^s_i) + \sum_{i=1}^{n_t} \sigma^t_{q,i} \widetilde{e}^t_{q,k,l}(x^t_i). \tag{C.50}$$

Proof. Combining the expression of consensus features as mean-centered source and target embedding from C.7.3, Definition C.8.5 and Definitions **??** and C.8.6 gives the wanted results. □

Definition C.8.8. *For the q^{th} consensus feature, we define the **offset proportion** as $O_q = \mathscr{O}_q^2$, the **linear contribution** as $L_q = \sum_{k=1}^{p} \mathscr{L}^2_{q,k}$ and the **interaction contribution** as $I_q = \sum_{1\le k\le l\le p} \mathscr{I}^2_{q,k,l}$.*
*Finally, we define the **higher-order contribution** as $R_q = 1 - O_q - L_q - I_q$.*

We now restrict to one gene set to measure the effect of this gene set on interactions and linear effects.
We here restricted to the Gaussian kernel. However, our results would easily be extended to any kernel, provided the feature space $\mathscr{H}$ has a known orthonormal basis.

C.9. EQUIVALENCE WITH GEODESIC FLOW KERNEL

In this section we showed the equivalence with the previously published linear version of the algorithm, the so-called PRECISE model [131]. We recall the main steps of the algorithm.

Definition C.9.1 (Linear Principal Vectors)**.** *Let $P_s \in \mathbb{R}^{d_s \times p}$ and $P_t \in \mathbb{R}^{d_t \times p}$ be two families of orthonormal vectors, i.e. $P_s P_s^T = I_{d_s}$ and $P_t P_t^T = I_{d_t}$. We define the cosine similarity matrix $\mathbf{M}$ as:*

$$\mathbf{M} = P_s P_t^T. \tag{C.51}$$

Let $d \le \min(d_s, d_t)$ and let $U\Sigma^L V^T$ be the $d-$ rank SVD approximation of M. We define the d source (resp. target) principal vectors as the matrix $\mathbf{Q}_s \in \mathbb{R}^{d\times p}$ (resp. $\mathbf{Q}_t \in \mathbb{R}^{d\times p}$) as:

$$\mathbf{Q}_s = U^T P_s \quad \text{and} \quad \mathbf{Q}_t = V^T P_t. \tag{C.52}$$

Samples can be projected on these four matrices ($P_s, P_t, \mathbf{Q}_s$ and $\mathbf{Q}_t$) by inner-product, i.e. canonical projection operator in Euclidean space.

P_s and P_t are here defined generally as two families of orthonormal vectors. In particular, we consider for the rest that they are the results of PCA on respectively the source and the target covariance matrices $X_s^T C_{n_s} C_{n_s}^T X_s$ and $X_t^T C_{n_t} C_{n_t}^T X_t$. Using the linear PVs from Definition C.9.1, we define a linear interpolation scheme as follows.

Definition C.9.2 (Linear Interpolation)**.** *Using notations from Definition C.9.1, we define the linear principal angles as:*

$$\forall q \in \{1,..,d\}, \quad \cos\theta_q^L = \Sigma_{q,q}^L. \tag{C.53}$$

For the PV pair $q \in \{1,..,d\}$, we define the interpolation function ϕ_q^L as follows:

$$\phi_q : \tau_q \in [0,1] \mapsto \frac{\sin(1-\tau_q)\theta_q}{\sin\theta_q}(Q_s)_q + \frac{\sin\tau_q\theta_q}{\sin\theta_q}(Q_t)_q \tag{C.54}$$

Before stating the main result, we need the following well-known lemma.

Lemma C.9.3 (Equivalence of spectrum)**.** *Let $X \in \mathbb{R}^{n\times p}$. We denote by $S^+ = \{(\lambda_1^+, v_1^+), ..,(\lambda_{d^+}^+, v_{d^+}^+)\}$ the non-singular spectrum of XX^T and $S^- = \{(\lambda_1^-, v_1^-),..,(\lambda_{d^-}^-, v_{d^-}^-)\}$ the non-singular spectrum of X^TX, i.e. $\lambda_1 \geq \lambda_2 \geq .. \geq \lambda_d > 0$ in both spectrum. Then $d^+ = d^-$ and*

$$\forall i \in \{1,..,d^+\}, \quad \lambda_i^+ = \lambda_i^-, \quad v_i^+ = Xv_i^- \quad \text{and} \quad v_i^- = X^Tv_i^+ \tag{C.55}$$

We consider two scenario using the same source and target datasets: linear PRECISE, and our kernelized approach with a linear kernel. We consider all other parameters set to the same values.

Proposition C.9.4 (Equality of cosine similarity matrixes)**.** *Let $\mathbf{M}$ and $\mathbf{M}^K$ be the cosine similarity matrices obtained respectively using linear PRECISE (Definition C.9.1) and the kernelized version with a linear kernel (Definition C.6.1), all hyperparameters equal. Then $\mathbf{M} = \mathbf{M}^\mathbf{K}$.*

Proof. We here use notations from Definitions C.6.1 and C.4.1. We define $\widetilde{X_s} = C_{n_s}X_s$ and $\widetilde{X_t} = C_{n_t}X_t$. We also use $\widetilde{K_s} = \widetilde{X_s}\widetilde{X_s}^T$ and $\widetilde{K_t} = \widetilde{X_t}\widetilde{X_t}^T$.
By definition of PCA, P_s contains the top d_s eigenvectors of the matrix $\widetilde{X_s}^T\widetilde{X_s}$, while $\alpha^s\widetilde{K_s}^{\frac{1}{2}}$ contains the top d_s eigenvectors of $\widetilde{X_s}\widetilde{X_s}^T$. Using the result from C.9.3, we have $P_s = \alpha^s\widetilde{K_s}^{\frac{1}{2}}\widetilde{X_s}$. Similarly, we obtain on the target $P_t = \alpha^t\widetilde{K_t}^{\frac{1}{2}}\widetilde{X_t}$. Using Proposition C.6.2,

$$\mathbf{M} = \alpha^s\widetilde{K_s}^{\frac{1}{2}}\widetilde{X_s}\widetilde{X_t}^T\widetilde{K_t}^{\frac{1}{2}}\alpha^{tT} = \alpha^s\widetilde{X_s}\widetilde{X_t}^T\alpha^{tT} = \mathbf{M}^K, \tag{C.56}$$

using the fact that $\alpha^s\widetilde{K_s}^{\frac{1}{2}}$ and $\alpha^t\widetilde{K_t}^{\frac{1}{2}}$ is an eigenvector of $\widetilde{K_s}$ and $\widetilde{K_t}$ respectively. □

From Proposition C.9.4 follows directly this equivalence.

Theorem C.9.5. *With all hyperparameters equal, PRECISE and the kernelized version with a linear kernel are equivalent.*

Theorem C.9.5 shows that all results obtained in linear case [131] hold for TRANSACT with a linear similarity function, and in particular the correspondence with the Geodesic Flow Kernel.

C.10. DIFFERENCE WITH CCA ON THE GENES

Another data-strategy used in single-cell data analysis consists in using the gene-level correspondence to perform a Canonical Correlation Analysis (CCA) on the genes. Using the same notations as in Section C.2, this approach boils down to solve the following maximization procedure:

$$s_1, t_1 = \underset{\substack{s\in\mathbb{R}^{n_s}, s^T s=1,\\ t\in\mathbb{R}^{n_t}, t^T t=1}}{\operatorname{argmax}} s^T X_s X_t^T t \tag{C.57}$$

and the subsequent directions defined orthogonally to these directions. This procedure find directions of maximum covariance at the gene-level between source and target. It will find two combinations of samples (one for source, and one for target) that show the maximum covariance among genes. It differs markedly from our methods on several aspects. First, from a computational standpoint, the SVD-equivalent definition of PVs (Theorem C.6.3) consists in breaking down a relatively small matrix ($d_s \times d_t$) and not a sample-sample similarity matrix. Second, by performing a PCA on source and target independently, we restrict our analysis to a low-rank view of source and target data – which provides a first step filtering. Finally, although there are similarities in the maximization procedures from Equations C.11 and C.57, the product of our maximization procedure gives geometrical weights, and not directly the scores used in the regression. Although we maximize the same objective function, the constraints are different, which would make the final vectors surely different.

We believe our approach to be better suited for our specific problem for several reasons. First because it uses a low-rank representations of source and target. As shown in Figure 1 of main text, the kernel matrices K_s and K_t contain larger values than K_{st} which would increase signal-to-noise ratio. Our sample-size is small – compared to single cell studies at least – and penalization is expected to focus on important signal. Second, our approach gives us a direct access to the geometric components (PV) which we can analyze to understand the making of the common signal. Finally, using PVs allow us to interpolate and get a projection on a single component that would be shared across source and target.

C.11. ALGORITHM WORKFLOW

Algorithm 2 TRANSACT

Require: source data $\mathbf{X}_s$, target data $\mathbf{X}_t$, number of *domain-specific factors* d_s and d_t, p.s.d. kernel K, number of *principal vector* d.

$\mathbf{K}_s \leftarrow$ source kernel matrix.

$\mathbf{K}_t \leftarrow$ target kernel matrix.

$\mathbf{K}_{st} \leftarrow$ source-target kernel matrix.

$\alpha^s \leftarrow$ Kernel Principal Components of source (from K_s).

$\alpha^t \leftarrow$ Kernel Principal Components of source (from K_t).

$\mathbf{M}^K \leftarrow \alpha^s C_{n_s} K_{st} C_{n_t} \alpha^t T$.

$\beta^s \Sigma \beta^t \leftarrow d$-rank SVD of $\mathbf{M}^K$, i.e. $\mathbf{M}^K \approx \beta^s \Sigma {\beta^t}^T$.

$\mathbf{F}^{st} \leftarrow \left[F^{st}(0), F^{st}(0.01), .., F^{st}(1)\right]$ defined as in Theorem C.7.6.

for $q \leftarrow 1$ to d **do**

$\mathbf{S}_q \leftarrow \left[\mathbf{F}^{st}[0]_{1:n_s,q}, \mathbf{F}^{st}[0.01]_{1:n_s,q}, .., \mathbf{F}^{st}[1]_{1:n_s,q}\right]^T$

$\mathbf{T}_q \leftarrow \left[\mathbf{F}^{st}[0]_{n_s:n_s+n_t,q}, \mathbf{F}^{st}[0.01]_{n_s:n_s+n_t,q}, .., \mathbf{F}^{st}[1]_{n_s:n_s+n_t,q}\right]^T$

$D_q \leftarrow \left\{D\left(S_q[0], T_q[0]\right), D\left(S_q[0.01], T_q[0.01]\right), .., D\left(S_q[1], T_q[1]\right)\right\}$.

$\tau_q^* \leftarrow \text{argmin}_\tau D_q$.

end for

$\mathbf{F} \leftarrow [\Phi_1(\tau_1), \Phi_2(\tau_2), .., \Phi_d(\tau_d)]$

$\tau^* \leftarrow \left[\tau_1^*, .., \tau_q^*\right]$.

$X_s^{proj} \leftarrow \mathbf{F}^s t[\tau^*]_{1:n_s}$.

$X_t^{proj} \leftarrow \mathbf{F}^s t[\tau^*]_{n_s:n_s+n_t}$.

Train a regression model on X_s^{proj}

Apply it on the projected target data X_t^{proj}

D

SUPPLEMENTARY MATERIAL FOR CHAPTER 5

D.1. SUPPLEMENTARY FIGURES

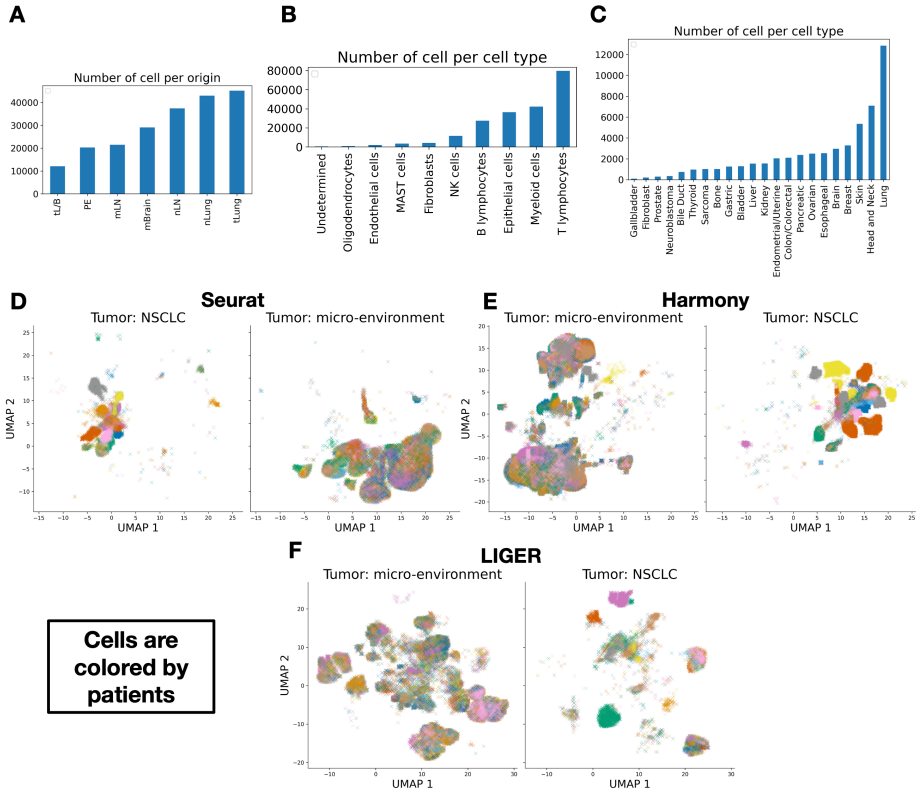


Figure D.1 – **(Supporting Figure 5.1) Existence of systematic differences between cell lines and tumors hinders batch effect correction**. (**E**) Composition of the *Kim dataset* by donors (L/B and tLung: primary lung lesion; PE: pleural fluids; mLN: metastatic lymph node; mBrain: brain metastasis; nLung: healthy lung; nLN: healthy lymph node). (**B**) Composition of the *Kim dataset* by cell types. (**C**) Composition of the *Kinker dataset* by tissue types. (**D**) UMAP of the Seurat corrected gene expression profiles for patient data (*Kim dataset*) divided between epithelial tumor cells (left) and other cells (right), colored by patients. (**E**) UMAP of the Harmony corrected gene expression profiles for patient data colored by patients. (**F**) UMAP of the LIGER corrected gene expression profiles for patient data colored by patients.

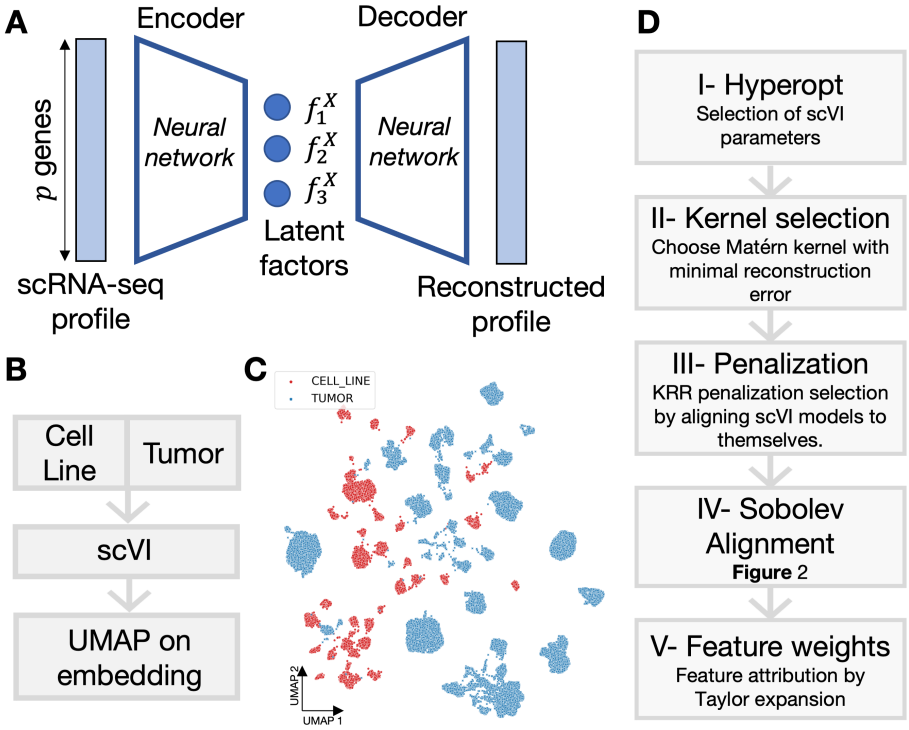


Figure D.2 – **(Supporting Figure 5.2) Technical supplement on the Sobolev Alignment algorithm**. (**A**) General presentation of an Auto-Encoder. A scRNA-seq profile is used as input into a first neural network, called "encoder", which compresses the data to a small number of "latent factors". These latent factors are then fed into a second neural network, called a "decoder", which maps the latent factors back to the original space. The Auto-Encoder is trained by modifying the weights of both neural networks so that the difference between the input and the output is minimal. A Variational Auto-Encoder (VAE) offers a probabilistic extension of this framework with a more complex architecture which allows the incorporation of prior knowledge; it however still relies on this "encoder-decoder" scheme. (**B**) We trained a single scVI model on the concatenated data (Methods) and used scVI native batch correction to account for batch effect, both within and between datasets. (**C**) Using UMAP, we project and visualize the resulting latent factors. (**D**) Complete workflow of Sobolev Alignment. In a first step, the hyperparameters of the scVI models (dropout rate, likelihood, weight decay, network architecture, learning rate, learning rate scheduler, early stopping) are set by minimizing the reconstruction error employing Bayesian Optimisation (Hyperopt). In a second step, we set the parameters of the Matérn kernel to be used in Sobolev Alignment. To do so, we first set the scale parameter (σ) as the median distance observed between source and target single cell profiles. We then train different KRR models on cell line and tumor data with varying values of ν. We use artificial points as training data, and we select ν as the parameter providing the largest Spearman correlation between the embedding values and the KRR values predicted from the scRNA-seq dataset (not used to train the KRR). In a third step, we set the penalty parameter of the cell line KRR model by aligning the trained cell line scVI model to itself. As we align the exact same model to itself, the SPVs should have a similarity close to one. However, small penalization values would lead to overfitting and therefore decrease this observed similarity – overfitting artifacts have a limited chance to be shared between source and target. We train KRR models with different regularization values and select the lowest value past a certain threshold of self-similarity (by default set to 0.9). We proceed similarly for the tumor regularization parameter. Finally, we perform the whole alignment as explained in Figure 5.2.

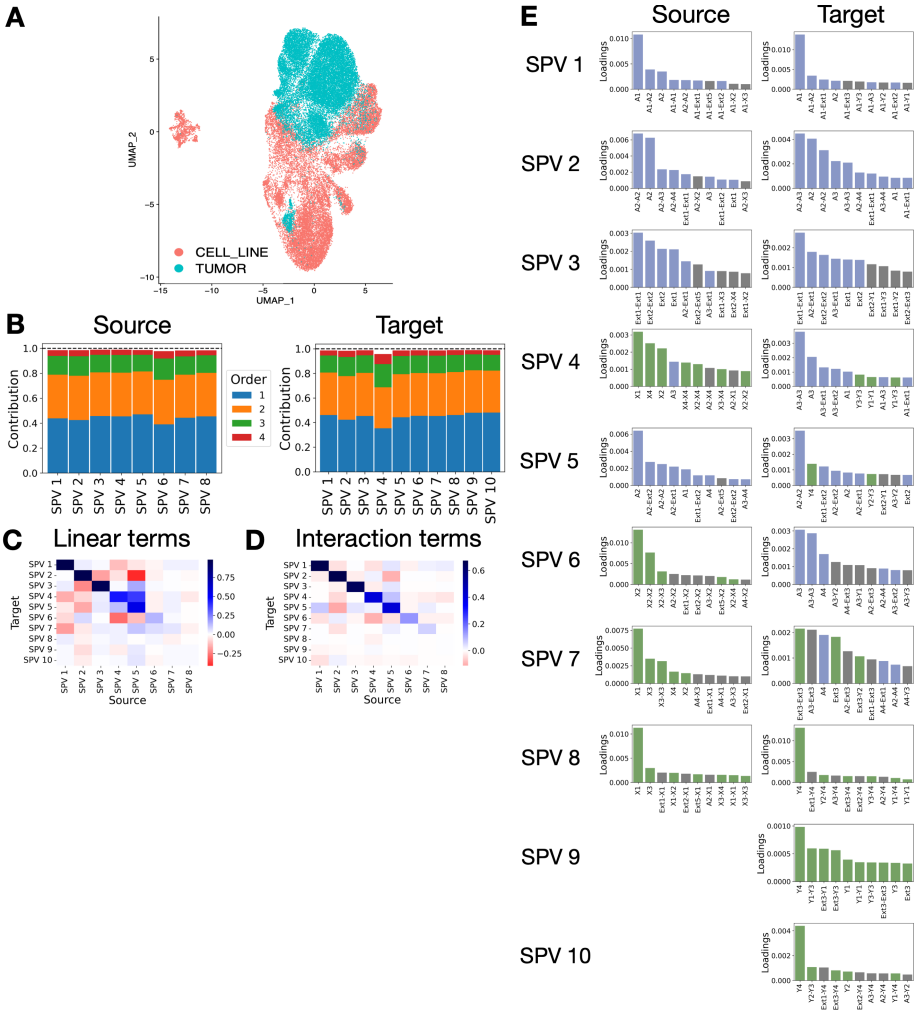


Figure D.3 – **(Supporting Figure 5.3) Analysis of model I**. (**A**) UMAP obtained after Seurat correction between source (cell lines) and target (tumor). (**B**) Contributions of different order features for source (left) and target (right) SPVs. (**C**) Spearman correlations between the linear weights of source SPVs (x-axis) and target SPVs (y-axis). (**D**) Spearman correlations between the interaction weights of source SPVs (x-axis) and target SPVs (y-axis). (**E**) Square contributions of linear and interaction features to each SPV, restricted to the top 10 highest contributions and colored as in Figure 5.3.

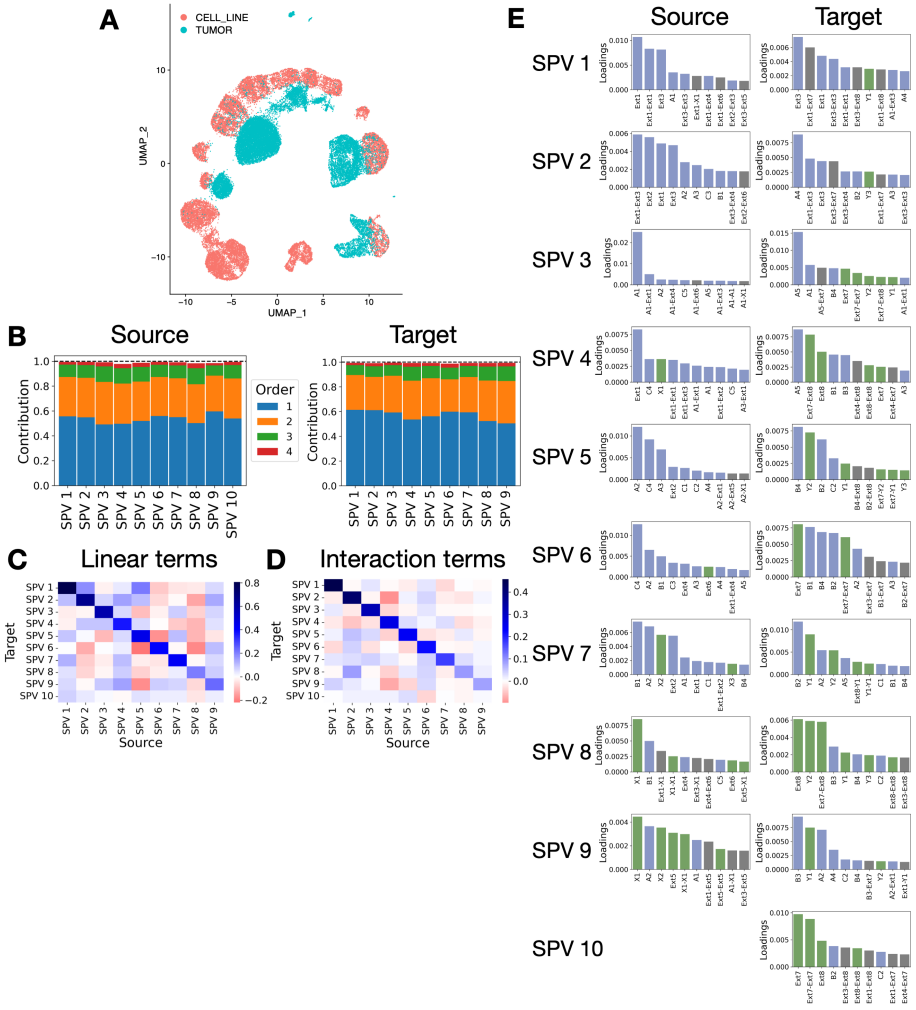


Figure D.4 – **(Supporting Figure 5.3) Analysis of model II**. (**A**) UMAP obtained after Seurat correction between source (cell lines) and target (tumor). (**B**) Contributions of different order features for source (left) and target (right) SPVs. (**C**) Spearman correlations between the linear weights of source SPVs (x-axis) and target SPVs (y-axis). (**D**) Spearman correlations between the interaction weights of source SPVs (x-axis) and target SPVs (y-axis). (**E**) Square contributions of linear and interaction features to each SPV, restricted to the top 10 highest contributions and colored as in Figure 5.3.

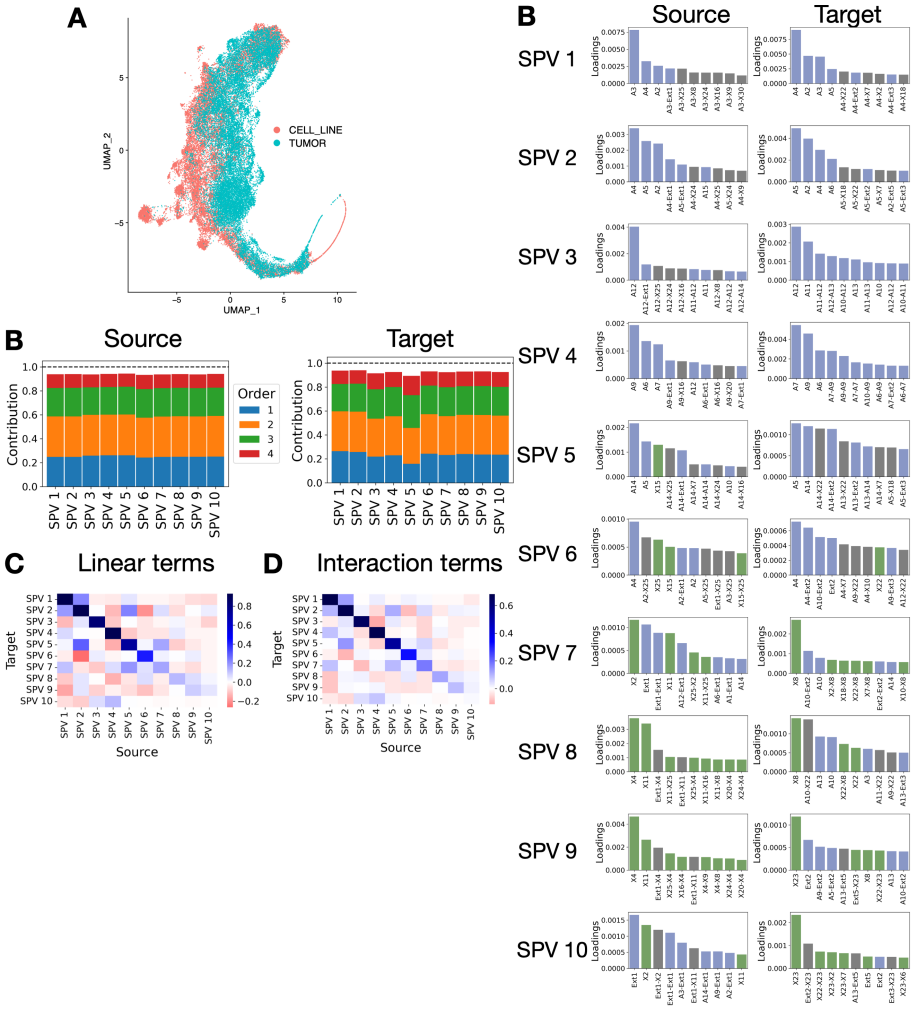


Figure D.5 – **(Supporting Figure 5.3) Analysis of model III**. (**A**) UMAP obtained after Seurat correction between source (cell lines) and target (tumor). (**B**) Contributions of different order features for source (left) and target (right) SPVs. (**C**) Spearman correlations between the linear weights of source SPVs (x-axis) and target SPVs (y-axis). (**D**) Spearman correlations between the interaction weights of source SPVs (x-axis) and target SPVs (y-axis). (**E**) Square contributions of linear and interaction features to each SPV, restricted to the top 10 highest contributions and colored as in Figure 5.3.

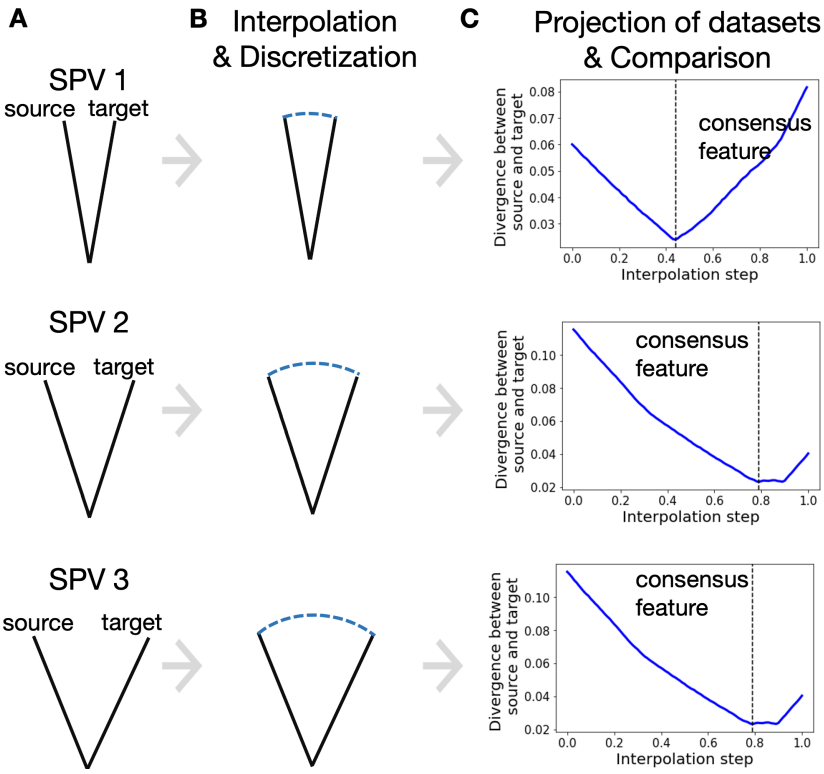


Figure D.6 – **(Supporting Figure 5.4) Schematic of the interpolation scheme between similar Sobolev Principal Vectors (SPV)**. SPVs correspond to pairs of vectors ordered by decreasing similarity. (**A**) When projecting on the top SPVs, it is unclear which of the two vectors to choose: selecting any of the two induces a bias towards either cell lines or tumors. To design a vector which balances the effect of cell lines and tumors, we employ the following interpolation scheme. (**B**) By drawing an arc between the cell line and the tumor vector, we obtain intermediate vectors of same norm. We discretize this arc, e.g., by selecting 100 points equally spaced on this arc, and project cell line and tumor data onto each of these intermediate vectors. (**C**) For each of these intermediate vectors, we compare the cell line and tumor projected data using the Kolmogorov-Smirnoff distance: the lower the distance, the closer the two projected datasets. We select the intermediate vector which minimizes this distance and refer to this vector as the consensus feature. This procedure is performed independently for each SPV pair. Such a procedure is inspired from earlier works and stems from an equivalent definition of the geodesic curves in the Grassmann manifold.

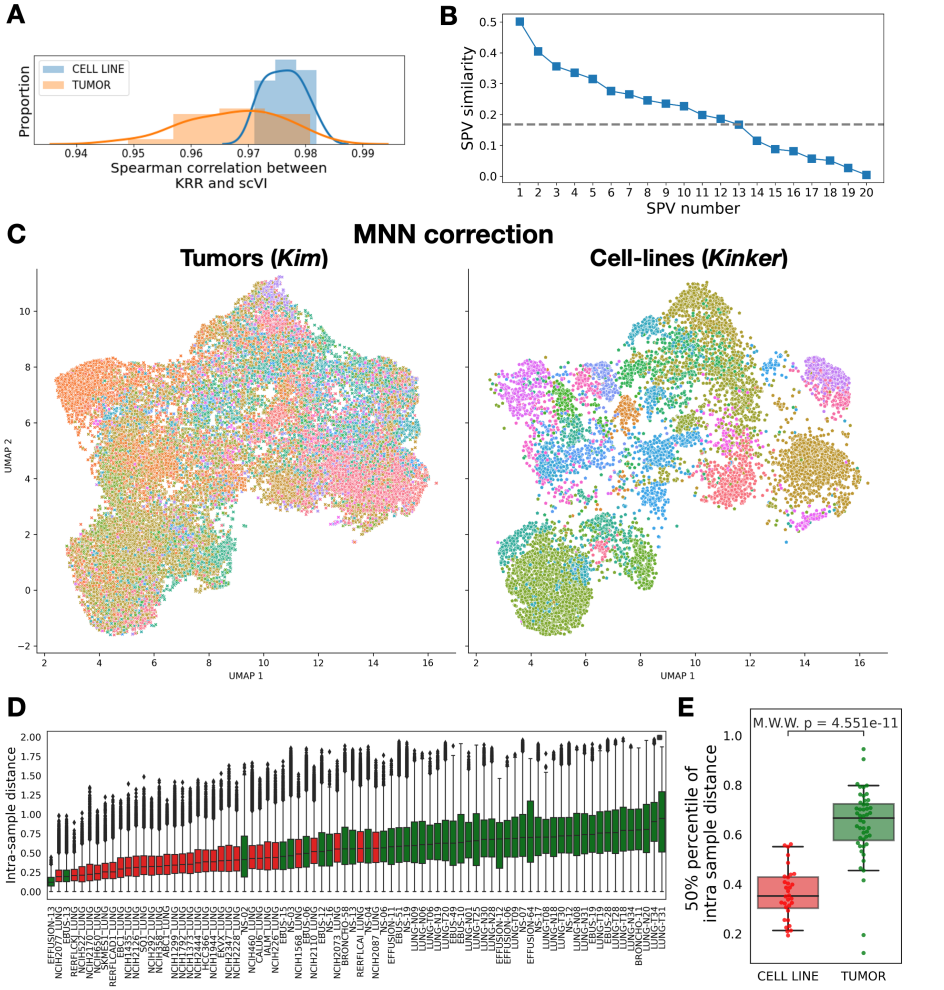


Figure D.7 – **(Supporting Figure 5.4) Sobolev Alignment allows an effective co-clustering of cell lines and tumors**. (**A**) Histograms of Spearman correlations between scVI embeddings and KRR approximations by Falkon. (**B**) Similarity between the Sobolev Principal Vectors (SPV) alongside the maximum similarity value observed after 100 gene permutations (dashed line). (**C**) UMAP visualization of Kinker and Kim datasets after Sobolev Alignment and MNN correction (Figure 5.4C) colored by patients (left) and cell line (right). (**D**) Boxplots of distances between cells from the same patient or cell line. Distances are computed as cosine distances between the cells embedded using Sobolev Alignment and MNN correction. (**E**) Median intra-sample distances observed in panel E.

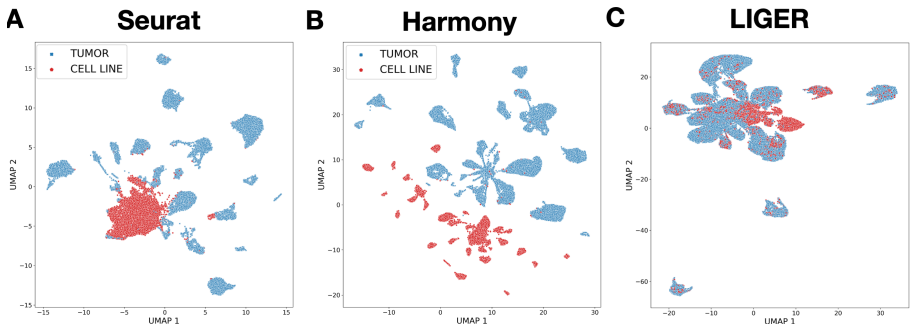


Figure D.8 – **(Supporting Figure 5.4) Comparison of integration obtained using standard batch-effect correction tools**. We performed the same integration task using 3 state-of-the-art batch effect correction tools. After integration with each method, we projected the data in two dimensions using UMAP. (**A**) Results obtained using Seurat v3. (**B**) Results obtained using Harmony. (**C**) Results obtained using LIGER.

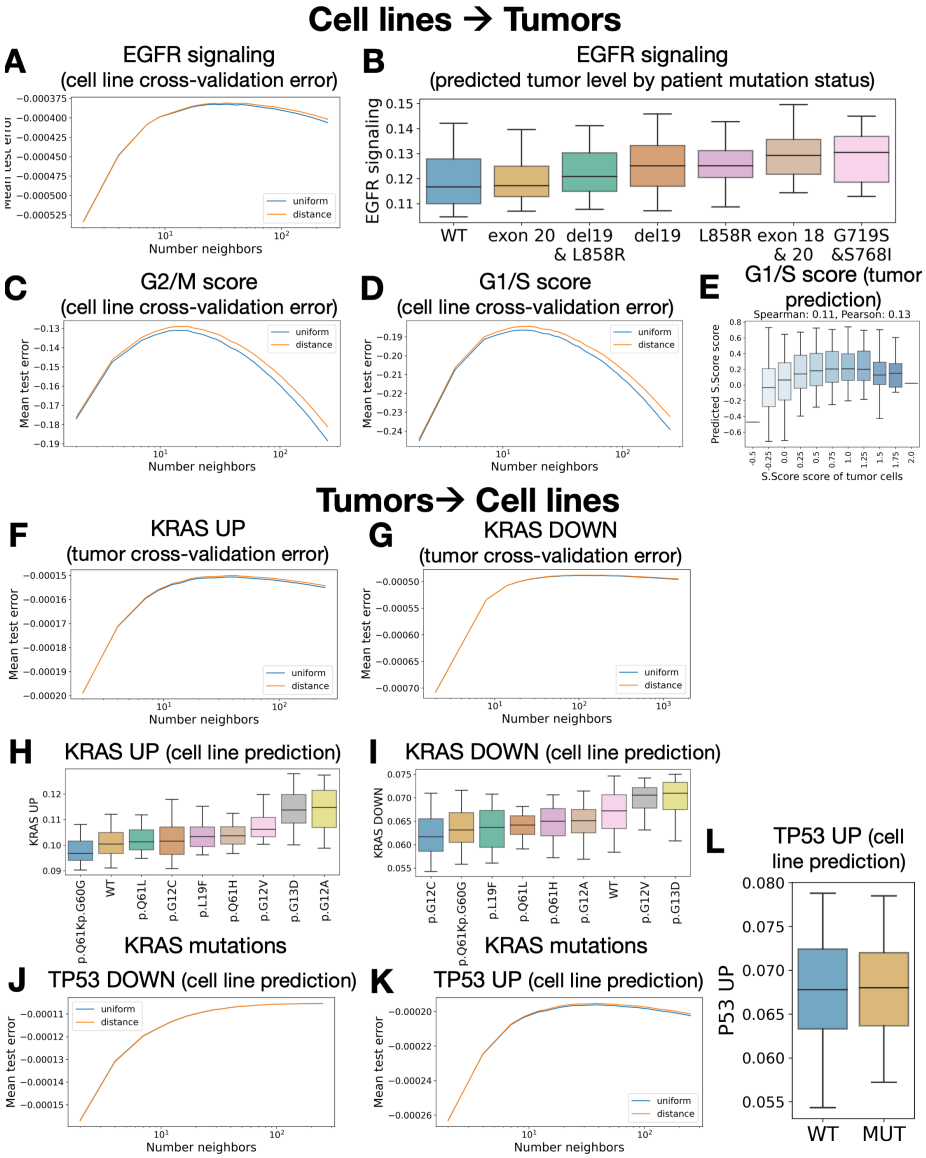


Figure D.9 – **(Supporting Figure 5.4) External biomarkers validation of Sobolev Alignment**. (**A**) 10-fold cross-validation negative mean squared error (NMSE) obtained on cell line data when training k-Nearest-Neighbors (kNN) regression models on "EGFR signaling" levels (Reactome) for various numbers of neighbors. "Uniform" indicates that neighbors were similarly weighted for prediction, while "distance" indicates an inverse-distance weighting. (**B**) Predicted EGFR signaling level for tumor cells broken down by EGFR mutations. (**C**-**D**) 10-fold cross-validation NMSE on cell lines when training kNN regression models on G2/M scores. (**E**) 10-fold cross-validation NMSE on cell lines when training kNN regression models on S-phase scores. (**F**) Boxplots of predicted S-phase score on tumor cells compared to S-phase scored measured using Seurat v3 cell-cycle regression tool. Spearman and Pearson correlation are computed between the continuous (non-binned) values. (**G**) 10-fold cross-validation NMSE on tumors when training kNN regression models on "KRAS UP" levels (Hallmarks). (**H**) Predicted "KRAS UP" level for tumor cells broken down by KRAS mutations.

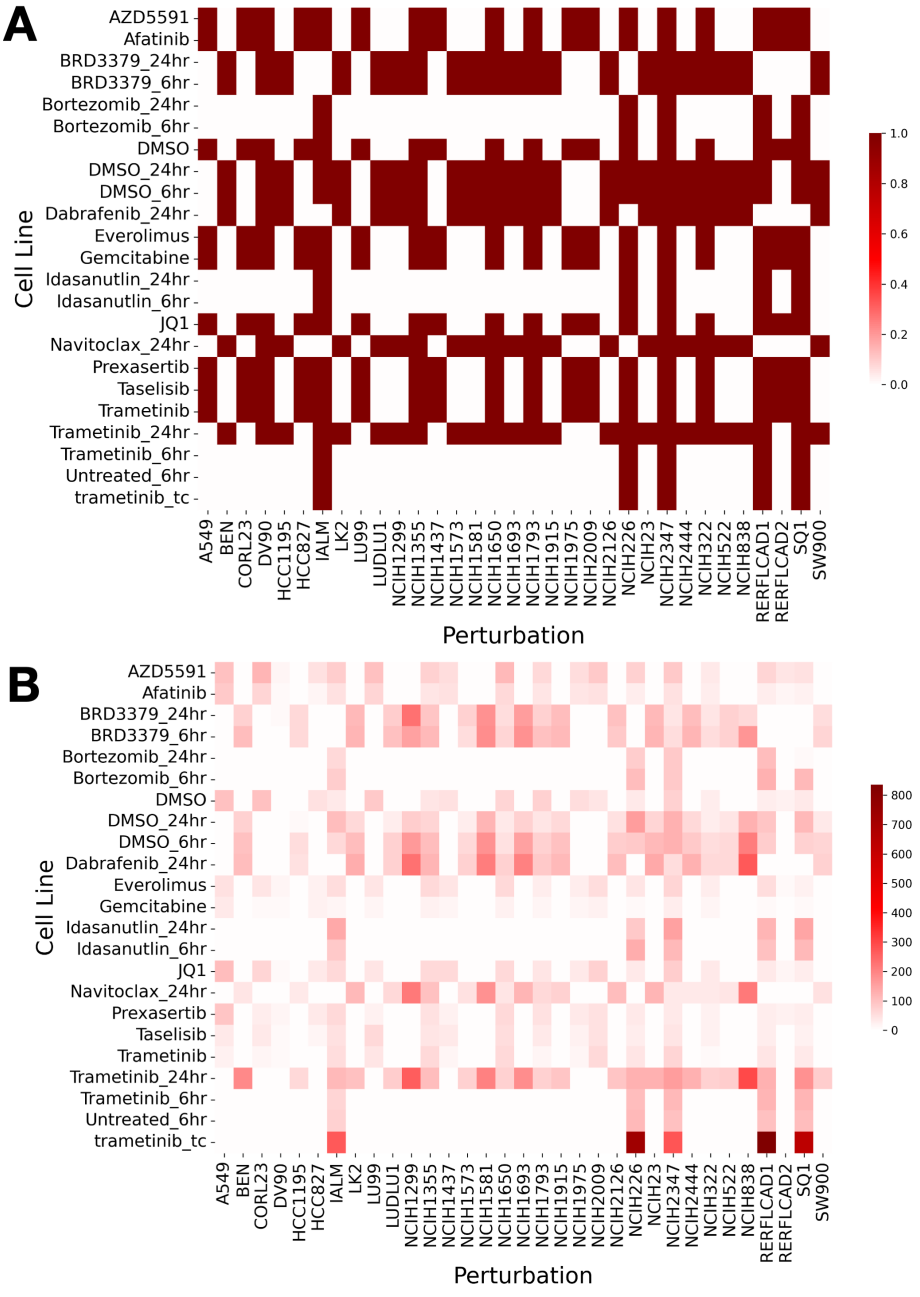


Figure D.10 – **(Supporting Figure 5.6) Structure of the employed perturbation screen (McFarland dataset).** (**A**) Heatmap indicating whether a cell line (column) has been screened for a certain anti-cancer compound (row). (**B**) Heatmap indicating the number of cells retrieved for each condition.

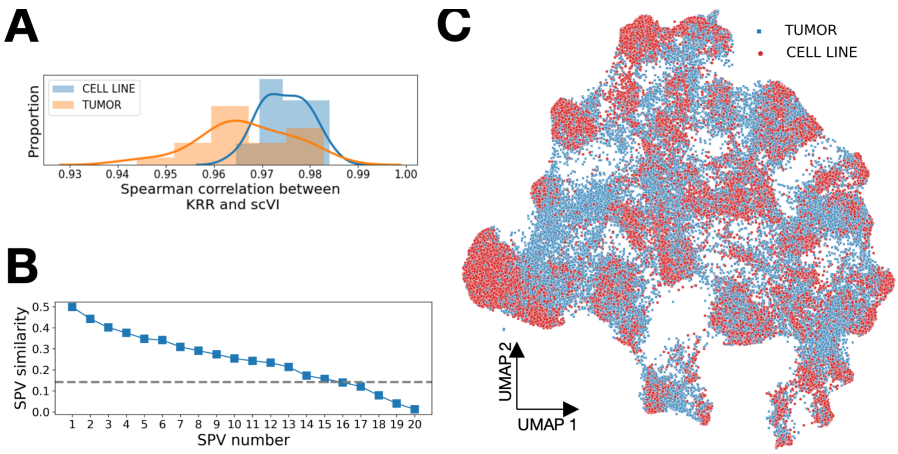


Figure D.11 – **(Supporting Figure 5.6) Sobolev Alignment between the McFarland and Kim datasets**. (**A**) Histograms of spearman correlations between scVI embeddings and KRR approximations by Falkon. (**B**) Similarity between the Sobolev Principal Vectors (SPVs) alongside the maximum similarity value observed after 100 gene permutations (dashed line). (**C**) UMAP of cell lines and tumors after projection on the top SPVs, interpolation and MNN correction.

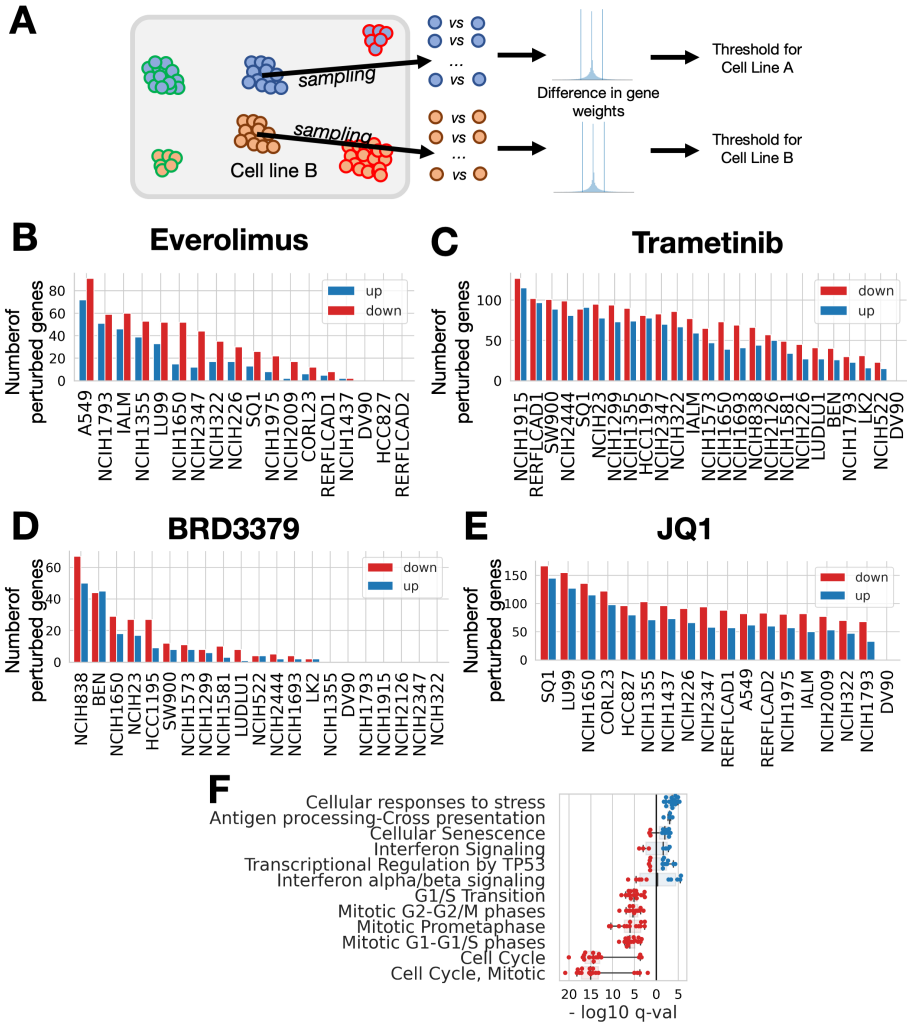


Figure D.12 – **(Supporting Figure 5.6) Analysis of drug perturbations after projection on SPVs common between cell lines and tumors**. (**A**) To determine the significant effect size threshold, we used the DMSO-treated cells. For each cell line, we randomly sampled 1000 pairs of cells from the pool of DMSO-treated cells and computed the absolute difference in gene weights for all genes. We then took the 95% percentile of all these differences as our effect-size threshold. A gene up- or down-regulation is therefore deemed significant if the Mann-Whitney FDR-corrected p-value is below 0.05 and if the effect size lies above random sampling in DMSO-treated cells. (**B**) Number of perturbed genes, for each cell line, after Everolimus induction. (**C**) Number of perturbed genes for each cell line after Trametinib induction. (**D**) Number of perturbed genes for each cell line after BRD3379 induction. (**E**) Number of perturbed genes for each cell line after JQ1 induction (BRD4-inhibitor). (**F**) Boxplot of q-values obtained when analyzing JQ1 using the Reactome gene sets.

D.2. NOTATIONS

In all of this work, we use gene expression profiles characterized by p genes (or features). Elements referring to the source data are characterized by an X subscript/superscript, while elements referring to the target data are characterized by a Y subscript/superscript.
Source data is comprised of n_X samples, yielding a data matrix $\mathscr{X} \in \mathbb{R}^{n_X \times p}$; target data is comprised of n_Y samples, yielding a data matrix $\mathscr{Y} \in \mathbb{R}^{n_Y \times p}$.

D.3. KERNEL METHODS AND ASSOCIATED FEATURE SPACE

We here briefly review mathematical and technical details on kernel methods useful to understand the workflow and derivation of Sobolev Alignment. For a longer and detailed presentation of these approaches, we refer the interested reader to [127, 273, 274]. In particular, [275] contains references and proofs to all results listed below.

D.3.1. KERNEL AND ASSOCIATED FEATURE SPACE (RKHS)

Definition D.3.1 (Positive Definite Kernel)**.** *A kernel K is a function which takes as inputs two samples and return a scalar value, formally: $K : x, y \in \mathbb{R}^p \to \mathbb{R}$.*
A kernel K is positive-definite (p.d. in short) if, and only if:

- *K is symmetric, i.e., for all $x, y \in \mathbb{R}^p, K\left(x, y\right) = K\left(y, x\right)$.*
- *All kernel matrices are positive definite, i.e.,*

$$\forall n \in \mathbb{N}^*, \forall x_1, ..., x_n \in \mathbb{R}^p, \forall \lambda_1, .., \lambda_n \in \mathbb{R}, \quad \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j K\left(x_i, x_j\right) \geq 0. \tag{D.1}$$

A positive-definite kernel can be implicitly understood as an inner product in an Hilbert space. This property is interesting as it allows to perform linear algebra operations in a higher-dimensional space where direct computation would be potentially intractable. Working in an higher-dimensional space allows to incorporate non-linearities which is often needed to model complex processes. This property is usually referred to as the "kernel-trick", and is formalized by the *Aronszajn theorem*.

Theorem D.3.2 (Aronszajn)**.** *K is positive-definite if and only if there exists a Hilbert space $\mathscr{H}$ and a mapping function $\phi : \mathbb{R}^p \mapsto \mathscr{H}$ such that:*

$$\forall x, y \in \mathbb{R}^p, \quad K\left(x, y\right) = \left\langle \phi\left(x\right), \phi\left(y\right) \right\rangle_{\mathscr{H}}. \tag{D.2}$$

This feature space is a **Reproducing Kernel Hilbert Space**, or RKHS in short, and is an intrinsic property of the kernel [1]. By construction, an RKHS corresponds to a functional space and each sample embedding $\phi(x)$ (for $x \in \mathbb{R}^p$) can be understood as a function

[1] The RKHS is actually defined by adding two more hypothesis, which we do not list for sake of simplicity. We refer the reader to [273] for a complete construction of the RKHS.

$\mathbb{R}^p \to \mathbb{R}$. The elements of $\mathscr{H}$ are both vectors in a high-dimensional space and functions. This duality allows us to approximates encoder functions in such an RKHS, and also apply linear algebra and compare encoders as vectors. This is formalized by the **Reproducing property**.

Proposition D.3.3 (Reproducing property)**.** *Let K be a p.d. kernel with RKHS $\mathscr{H}$. Then $\mathscr{H}$ is a set of functions, i.e. $\mathscr{H} \subset \{f : \mathbb{R}^p \to \mathbb{R}\}$ with the two following properties:*

- $\forall x \in \mathscr{H}, \quad K_x \mathrel{\widehat{=}} \quad y \mapsto K(y,x) \quad \in \mathscr{H}$
- $\forall f \in \mathscr{H}, \forall x \in \mathbb{R}^p, \quad f(x) = \langle K_x, f \rangle_{\mathscr{H}}$

We will now define the three (p.d.) kernels we employ in our study, alongside their feature spaces.

D.3.2. GAUSSIAN, MATÉRN AND LAPLACIAN KERNEL AND ASSOCIATED RKHS

We consider three different kernels in our work. We here define them and we will show in a subsequent part how these three are related.

Definition D.3.4 (Laplacian kernel)**.** *Let $\sigma > 0$ we define the Laplacian kernel K^L_σ on $\mathbb{R}^p$ as:*

$$\forall x, y \in \mathbb{R}^p, \quad K^L_\sigma(x,y) = \exp\left(-\frac{\|x-y\|}{\sigma}\right). \tag{D.3}$$

Definition D.3.5 (Matérn kernel [276])**.** *Let $\nu > 0$ and $\sigma > 0$, we define the Matérn kernel $K^M_{\nu,\alpha}$ on $\mathbb{R}^p$ as:*

$$\forall x, y \in \mathbb{R}^p, \quad K^M_{\nu,\sigma}(x,y) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}\|x-y\|}{\sigma}\right)^\nu K_\nu\left(\frac{\sqrt{2\nu}}{\sigma}\|x-y\|\right), \tag{D.4}$$

where Γ is the Gamma function, and K_α the modified Bessel function of second kind of order α.

A few interesting examples of Matérn kernels are the following:

$$\begin{aligned}
&\text{Order } 1/2: \quad k_{\frac{1}{2},\sigma}(x,y) = \exp\left(-\frac{\|x-y\|}{\sigma}\right)\\
&\text{Order } 3/2: \quad k_{\frac{3}{2},\sigma}(x,y) = \left(1+\frac{\sqrt{3}\|x-y\|}{\sigma}\right)\exp\left(-\frac{\sqrt{3}\|x-y\|}{\sigma}\right)\\
&\text{Order } 5/2: \quad k_{\frac{5}{2},h}(x,y) = \left(1+\frac{\sqrt{5}\|x-y\|}{\sigma}+\frac{5\|x-y\|^2}{3\sigma^2}\right)\exp\left(-\frac{\sqrt{5}\|x-y\|}{\sigma}\right)
\end{aligned} \tag{D.5}$$

Definition D.3.6 (Gaussian kernel)**.** *Let $\sigma > 0$, we define the Gaussian kernel K^L_σ on $\mathbb{R}^p$ as:*

$$\forall x, y \in \mathbb{R}^p, \quad K^G_\sigma(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right). \tag{D.6}$$

D.3.3. EQUIVALENCE OF THE THREE KERNELS AND HYPER-PARAMETERS

As already hinted at by the first line of Equation (D.5), the Gaussian, Matérn and Laplacian kernels are related.

Proposition D.3.7 (Equivalence of Gaussian, Matérn and Laplacian kernels)**.** *Let $\sigma > 0$, we have the following equalities:*

$$\textit{Equivalence between Matérn and Laplacian}: \quad K^M_{\frac{1}{2},\sigma} = K^L_\sigma. \tag{D.7}$$

$$\textit{Equivalence between Matérn and Gaussian}: \quad \lim_{\nu \to +\infty} K^M_{\nu,\sigma} = K^G_\sigma. \tag{D.8}$$

D.3.4. RELATIONSHIP BETWEEN MATÉRN FEATURE SPACES AND SOBOLEV SPACES

We first start by a general definition of the RKHS of Matérn kernel, which is related to the Gaussian and Laplacian kernels (Proposition D.3.7). Matérn kernels are related to the so-called *Sobolev spaces* which are functional spaces used in various areas of mathematics and physics.

Definition D.3.8 (Weak differentiation operation)**.** *Let $\beta \in \mathbb{N}^p$ and $|\beta| = \sum_{1 \leq i \leq n} \beta_i$. Let f be a function from $\mathbb{R}^p$ to $\mathbb{R}$, β-weakly differentiable. We denote by $D^\beta f$ the β^{th} weak differential of f.*
In the particular case when f is $|\beta|$-times differentiable, we have the following equality:

$$D^\beta f = \frac{\partial^{|\beta|}}{\partial x_1^{\beta_1} ... \partial x_p^{\beta_p}} f. \tag{D.9}$$

Definition D.3.9 (Space of continuous integrable functions)**.** *We define as $L_2(\mathbb{R}^p)$ the space of continuous functions defined as follows:*

$$L_2\left(\mathbb{R}^p\right) = \left\{ f : \mathbb{R}^p \to \mathbb{R} \mid \int_{\mathbb{R}^p} f(x)^2 dx < +\infty \right\}, \tag{D.10}$$

endowed with the following inner-product.

$$\forall f, g \in L_2\left(\mathbb{R}^p\right), \quad \langle f, g \rangle_{L_2(\mathbb{R}^p)} = \int_{\mathbb{R}^p} f(x) g(x) dx. \tag{D.11}$$

It defines a Hilbert space[2]*, which we denote as L_2 in the sequel. .*

Using these two bricks, we define a Sobolev spaces as follows:

Definition D.3.10 (Sobolev spaces)**.** *Let $s > 0$ be an integer. We define the Sobolev space of order s, denoted noted W_2^s as:*

$$W_2^s = \left\{ f \in L_2 \mid \sum_{\beta \in \mathbb{N}^p || \beta | \leq s} \left\| D^\beta f \right\|^2_{L_2} < +\infty \right\}, \tag{D.12}$$

[2] L_2 is actually a quotient space defined up to an equivalent class related to the Lebesgue-measure used on $\mathbb{R}^p$. We defined it here as a functional space for sake of clarity.

endowed with the following inner product:

$$\forall f, g \in W_2^s, \quad \langle f, g \rangle_{W_2^s} = \sum_{\beta \in \mathbb{N}^p, |\beta| \leq s} \langle D^\beta f, D^\beta g \rangle_{L_2}. \tag{D.13}$$

These Sobolev spaces can approximate any function in L_2, as shown by the following proposition.

Proposition D.3.11 (Density of W_2^s in L_2)**.** *Let $f \in L_2$ and $s > 0$ be an integer. There exists $f_1, f_2, \ldots \in W_2^s$ such that:*

$$\left\| f_n - f \right\|_{L_2} \xrightarrow[n \to +\infty]{} 0 \tag{D.14}$$

Proposition D.3.12 (Matérn feature space)**.** *Let $\nu > 0$ and $\sigma > 0$. If $\nu + \frac{p}{2}$ is an integer, then $K_{\nu,\sigma}^M$ has for RKHS $\mathscr{H}_{\nu,\sigma} = W_2^{\nu + \frac{p}{2}}$, and the associated norms are equivalent, i.e. there exists $c_1, c_2 > 0$ such that:*

$$\forall f \in W_2^s, \quad c_1 \|f\|_{W_2^s} \leq \|f\|_{\mathscr{H}_{\nu,\sigma}} \leq c_2 \|f\|_{W_2^s}. \tag{D.15}$$

The result of Proposition D.3.12 shows that the feature space associated to the Matérn kernel is equivalent to a Sobolev space. Any kernel-based algorithm which employs the Matérn kernel therefore implicitly operates on functions which lie on a Sobolev space.
By definition, the Sobolev inner product of order s between two functions compares all the derivatives of order s or lower using an Euclidean distance. Due to this natural (and standard) way of comparing functions, coupled with the density argument from in L_2 (Proposition D.3.11), we decided to use Sobolev spaces to approximate our encoder functions, and therefore turned to Matérn kernel machines. Although directly using the L_2 space would have been ideal, we were here hindered by the fact that L_2 is **not** an RKHS and therefore not amenable to approximation by kernel machines.

D.4. TRAINING OF VARIATIONAL AUTO-ENCODERS (VAE)

Definition D.4.1 (Space of probability measures)**.** *We define by $P(\mathbb{R}^p)$ the space of all Borel probability measures on $\mathbb{R}^p$.*

Definition D.4.2 (VAE models)**.** *We define the following two sets of models:*

$$\begin{array}{llll} \mu^X : \mathbb{R}^p \mapsto \mathbb{R}^{d_X}, & \Sigma^X : \mathbb{R}^p \mapsto \mathbb{R}^{d_X} & \text{and} & g^X : \mathbb{R}^{d_X} \mapsto P\left(\mathbb{R}^p\right) \\ \mu^Y : \mathbb{R}^p \mapsto \mathbb{R}^{d_Y}, & \Sigma^Y : \mathbb{R}^p \mapsto \mathbb{R}^{d_Y} & \text{and} & g^Y : \mathbb{R}^{d_Y} \mapsto P\left(\mathbb{R}^p\right) \end{array}. \tag{D.16}$$

g^X and g^Y outputs probability distributions on $\mathbb{R}^p$ and are referred to as decoders (or stochastic decoders). The families (μ^X, Σ^X) and (μ^X, Σ^X) are referred to as encoders (or stochastic encoders).

The two functions g^X and g^Y can be tailored to specific problems. In the particular case of scVI, the decoder is constructed based on a zero-inflated negative binomial (ZINB), which models well the scRNA-seq data structure.

Definition D.4.3 (Latent factors sampling)**.** *We assume that the latent factors follow a multivariate normal prior, i.e.,*

$$z_X \sim \mathcal{N}\left(0_{d_X}, I_{d_X}\right) \quad \textit{and} \quad z_Y \sim \mathcal{N}\left(0_{d_Y}, I_{d_Y}\right). \tag{D.17}$$

Given two datapoints $x \in \mathbb{R}^p$ and $y \in \mathbb{R}^p$, its latent factors are computed by sampling from the following multivariate normal posterior:

$$z_X|x \sim \mathcal{N}\left(\mu^X(x), \mathrm{diag}\left(\Sigma^X(x)\right)\right) \quad \textit{and} \quad z_Y|y \sim \mathcal{N}\left(\mu^Y(y), \mathrm{diag}\left(\Sigma^Y(y)\right)\right). \tag{D.18}$$

D

Considering a multivariate normal prior over the latent factors is a standard design choice, particularly attractive due to the so-called re-parametrization trick [178]. The original paper of scVI ([182]) contains additional information on the specific model we used.

D.5. COMPARING ENCODERS BY KERNEL RIDGE REGRESSION APPROXIMATION

The VAE models presented in Section D.4 concentrate a high-dimensional signal into a few latent factors. Understanding how the latent factors from the source model compare to the ones from the target model is however a difficult task. As presented in the main text, we propose to approximate each VAE model by means of Matérn kernel machines and present here our approach in greater mathematical details.

Our approach stems from the rationale that a wide class of function can be approximated by a Matérn kernel machine, provided enough data-points are available – we say that the Matérn kernel is a universal approximator. As generative models can generate their own data points, this allows us to exploit the consistency of kernel machines to our advantage by attempting to get as close as possible to the asymptotic limit.

Once the functions have been approximated, we can work in the Matérn kernel feature space (i.e. Sobolev space, Proposition D.3.12) and align these functions using the Representer Theorem formulation of the approximated functions.

Definition D.5.1 (Mean function of the VAE)**.** *As explicited in Definition D.4.2, the encoder of a VAE is parametrized by two sets of functions: one for the means and one for the standard-deviations. We here consider the mean embedding given to each sample and define, following notations from Equation(D.16):*

$$\forall t \in \{X, Y\},\ \forall i \in \{1, .., d_t\}, \quad f_i^t = \mu_i^t. \tag{D.19}$$

Following the scVI model, we here assume that the latent factors follow a multivariate-normal prior.

We have two sets of encoders we want to align – we will approximate them using two distinct Matérn kernel ridge regression defined as follows.

D.5.1. DEFINITION OF KERNEL RIDGE REGRESSION

In this subsection, we succintly present some results about Kernel Ridge Regression. For the sake of presentation only, we refer to an hypothetical dataset $\mathscr{D} = \{(\widehat{x}_1, \widehat{z}_1), .., (\widehat{x}_N, \widehat{z}_N)\}$ with $\widehat{x}_i \in \mathbb{R}^p$ and $\widehat{z}_i \in \mathbb{R}$. This dataset does not refer to cell lines or tumors and is given purely for illustrative purposes.

Definition D.5.2 (Matérn Kernel Ridge Regression (KRR))**.** *We approximate a function $f : \mathbb{R}^p \to \mathbb{R}$ by performing Kernel Ridge Regression with the Matérn kernel $K^M_{\nu,\sigma}$. This yields, for $\lambda > 0$, the function $\theta^* \in \mathscr{H}_{\alpha,h}$ solution of:*

$$\theta^* \quad = \quad \underset{\theta \in \mathscr{H}_{\alpha,h}}{\operatorname{argmin}} \quad \frac{1}{N} \sum_{i=1}^{N} \left(f(\widehat{x}_i) - \theta(\widehat{x}_i) \right)^2 \; + \; \lambda \|\theta\|^2_{\mathscr{H}_{\nu,\sigma}}. \tag{D.20}$$

The Kernel Ridge Regression problem from Definition D.5.2 can be solved in closed form.

Proposition D.5.3 (Solution of KRR)**.** *The solution of Equation (D.20) is:*

$$\theta^* = \sum_{k=1}^{N} \alpha_k K^M_{\nu,\sigma}(\widehat{x_k}, \cdot), \quad \textit{with} \begin{cases} K &= \left(K^M_{\nu,\sigma}(\widehat{x_i}, \widehat{x_j}) \right)_{1 \le i,j \le N} \\ \widehat{z} &= [\widehat{z_1}, .., \widehat{z_N}] \\ \alpha &= (K + \lambda N I_N)^{-1} \widehat{z} \end{cases} \tag{D.21}$$

Solving the problem from Definition (D.5.2) therefore requires inverting a $N \times N$ matrix, which becomes intractable as soon as N reaches approximately a few tens of thousands. To scale our approach to millions of samples, we exploit recent advances in kernel machines such as Falkon.

Proposition D.5.4 (Falkon approximation of KRR)**.** *The solution of Equation (**??**) can be approximated by:*

$$\hat{\theta} = \sum_{k=1}^{M} \alpha_k K^M_{\nu,\sigma}(\widetilde{x}_k, \cdot), \tag{D.22}$$

*with $M < N$ and α coefficients computed by the Nyström approximation. The M points $\widetilde{x}_1, .., \widetilde{x}_M$ are referred to as **Falkon anchor points**, and correspond to a subset of the training points $\widehat{x}_1, .., \widehat{x}_N$.*

Proposition D.5.4 shows that the weight vector α from Proposition D.5.3 can be approximated using the Nyström approximation, which is the strategy employed in the Falkon package. We refer the reader to the original Falkon paper for details on how to perform this approximation, as it is not necessary for our derivation. The sum expansion of Equation (D.21) is the cornerstone of all these methods.
Importantly, the Nyström approximation does **not** correspond to performing KRR by restricting to M subsampled points. The remaining $N - M$ points are influencing the solution and are present in the weights α: their influence has been factored in during the training process.

D.5.2. APPROXIMATION OF ENCODERS BY LARGE-SCALE KRR

We consider one Matérn kernel $K^M_{\nu,\sigma}$ which we refer to as K in the rest of the text. The RKHS $\mathscr{H}_{\nu,\sigma}$ is also referred to as $\mathscr{H}$ for ease of notation.

Definition D.5.5 (Model Samples)**.** *Let $N \in \mathbb{N}$. Using cell line (resp. tumor) VAE (Section D.4), we sample N points, called **Model Samples**, as follows:*

- $z_1^X, .., z_N^X \sim \mathcal{N}\left(0, I_{d_X}\right)$ *(resp.* $z_1^Y, .., z_N^Y \sim \mathcal{N}\left(0, I_{d_Y}\right)$*).*
- *Passing the random points through the decoder, we obtain* $\widehat{x}_1^X, .., \widehat{x}_N^X \in \mathbb{R}^p$ *(resp.* $\widehat{x}_1^Y, .., \widehat{x}_N^Y$*).*
- *We pass these sample points into the encoders to get the embeddings* $\widehat{z}_1^X, .., \widehat{z}_N^X \in \mathbb{R}^{d_X}$ *(resp.* $\widehat{z}_1^Y, .., \widehat{z}_N^Y \in \mathbb{R}^{d_Y}$*).*

D

A VAE is not bijective: we should therefore expect that the $\widehat{z}_i^X$ will differ from the z_i^X. Since we are here interested to approximate the encoder functions, we computed the output of the Model Points by each encoder function. Furthermore, using the latent values sampled from the multivariate normal prior distribution would have led to errors in the approximation due to the stochasticity of the VAE: the random points are sampled from a distribution specific to each latent value and are not deterministic.

Definition D.5.6 (KRR Approximation)**.** *We use the Model Samples (Definition D.5.5) as training data to train d_X (resp. d_Y) Falkon KRR models for cell lines (resp. tumors), which we termed* $\theta_1^X, .., \theta_{d_X}^X$ *(resp.* $\theta_1^Y, .., \theta_{d_Y}^Y$*) (Definition D.5.2, Propositions D.5.3 and D.5.4), with $M < N$. We define the two matrices $\alpha^X \in \mathbb{R}^{d_X \times M}$ and $\alpha^Y \in \mathbb{R}^{d_Y \times M}$ as the KRR sample weights:*

$$\forall t \in \{X, Y\}, \ \forall k \in \{1, .., d_t\}, \quad \theta_k^t = \sum_{i=1}^{M} \alpha_{k,i}^t K\left(\widetilde{x}_i^t, \cdot\right), \tag{D.23}$$

where $\widetilde{x}_1^X, \cdots, \widetilde{x}_M^X$ *and* $\widetilde{x}_1^Y, \cdots, \widetilde{x}_M^Y$ *are selected from the training data.*

In Definition D.5.6, the M Falkon anchor points correspond to M model points selected at random from the training data ($\widehat{x}_1^X, \cdots, \widehat{x}_M^X$ and $\widehat{x}_1^Y, \cdots, \widehat{x}_M^Y$). We refer the reader to [180] for a presentation and discussion on the methods used to sample these points.

D.5.3. COSINE SIMILARITY MATRIX

Definition D.5.7 (Un-normalized cosine similarity)**.** *We define the un-normalized cosine similarity matrices between X and Y as the matrix $\widetilde{\mathbf{M}}_{X,Y}$:*

$$\widetilde{\mathbf{M}}_{X,Y} = \left[\langle \theta_i^X, \theta_j^Y \rangle_{\mathscr{H}}\right]_{\substack{1 \le i \le d_X \\ 1 \le j \le d_Y}}, \tag{D.24}$$

and the cosine similarity matrices of X (resp. Y), denoted $\widetilde{\mathbf{M}}_X$ (resp. $\widetilde{\mathbf{M}}_Y$), as:

$$\widetilde{\mathbf{M}}_X = \left[\langle \theta_i^X, \theta_j^X \rangle_{\mathscr{H}}\right]_{\substack{1 \le i \le d_X \\ 1 \le j \le d_X}} \quad \text{and} \quad \widetilde{\mathbf{M}}_Y = \left[\langle \theta_i^Y, \theta_j^Y \rangle_{\mathscr{H}}\right]_{\substack{1 \le i \le d_Y \\ 1 \le j \le d_Y}}. \tag{D.25}$$

Definition D.5.8 (Similarity matrices)**.** *We define* $K_X \in \mathbb{R}^{N\times N}$, $K_Y \in \mathbb{R}^{N\times N}$ *and* $K_{X,Y} \in \mathbb{R}^{M\times M}$ *as:*

$$\begin{aligned} K_X &= \left(K\left(\tilde{x}_i^X, \tilde{x}_j^X\right)\right)_{1\leq i,j\leq M} \\ K_Y &= \left(K\left(\tilde{x}_i^Y, \tilde{x}_j^Y\right)\right)_{1\leq i,j\leq M} \\ K_{X,Y} &= \left(K\left(\tilde{x}_i^X, \tilde{x}_j^Y\right)\right)_{1\leq i,j\leq M} \end{aligned}. \tag{D.26}$$

Proposition D.5.9 (Computation of un-normalized cosine similarity matrices)**.** *We have the following equalities:*

$$\begin{aligned} \widetilde{\mathbf{M}}_X &= \alpha^X K_X {\alpha^X}^T \\ \widetilde{\mathbf{M}}_Y &= \alpha^Y K_Y {\alpha^Y}^T \\ \widetilde{\mathbf{M}}_{X,Y} &= \alpha^X K_{X,Y} {\alpha^Y}^T \end{aligned}. \tag{D.27}$$

Proof. We here show the proof for $\widetilde{\mathbf{M}_X}$; the two other equalities follow from the same idea.
We first recall the first reproducing property of the kernel K:

$$\forall x, y \in \mathbb{R}^p, \quad \left\langle K(x,\cdot), K(y,\cdot)\right\rangle_{\mathscr{H}} = K(x,y). \tag{D.28}$$

Let $i, j \in \{1,..,d_X\}$, we have:

$$\langle \theta_i^X, \theta_j^X \rangle_{\mathscr{H}} = \sum_{k=1}^{M}\sum_{l=1}^{M} \alpha_{i,k}^X \alpha_{j,l}^X K\left(\tilde{x}_k^X, \tilde{x}_l^X\right), \tag{D.29}$$

using the bi-linearity of the Hilbertian norm, the fact that α_X is real-valued, and the first reproducing property.
Combining Equation (D.29) with the definition of the similarity matrix K_X (Definition D.5.8), we obtain the desired property. □

Using these notations, we define the **cosine similarity matrix** as follows.

Definition D.5.10 (Cosine similarity matrix)**.** *We define the cosine similarity matrix* $\mathbf{M}^K$ *as:*

$$\mathbf{M}^K = \left(\widetilde{\mathbf{M}}_X\right)^{-1/2} \widetilde{\mathbf{M}}_{X,Y} \left(\widetilde{\mathbf{M}}_Y\right)^{-1/2}. \tag{D.30}$$

D.6. ENCODER ALIGNMENT BY KERNEL PRINCIPAL VECTORS

D.6.1. GENERAL DEFINITION OF PRINCIPAL VECTORS

We define the **Principal Vectors** (PVs) between source and target VAEs as the pairs of vectors (one from source, one from target) with a maximal inner-product in $\mathscr{H}$.

Definition D.6.1 (Principal Vectors). *Let $\hat{d} = \min(d_X, d_Y)$. We define the $\hat{d}$ Principal Vectors (PVs) $(s_1, t_1), .., (s_{\hat{d}}, t_{\hat{d}}) \in \mathcal{H} \times \mathcal{H}$ as the functions that maximise the similarity between source and target subspaces, i.e.,*

$$\forall k \in \{1, .., \hat{d}\}, \quad s_k, t_k = \underset{\substack{s \in \text{span}\left(\theta_1^X, .., \theta_{d_X}^X\right), \\ t \in \text{span}\left(\theta_1^Y, .., \theta_{d_Y}^Y\right)}}{\text{argmax}} \langle s, t \rangle_{\mathcal{H}} \quad s.t \quad \begin{cases} \langle s, s \rangle_{\mathcal{H}} = \langle t, t \rangle_{\mathcal{H}} = 1 \\ \forall i < k, \ s_i \perp s \\ \forall i < k, \ t_i \perp t \end{cases} \tag{D.31}$$

D.6.2. COMPUTATION OF PVS

Theorem D.6.2 (Matérn PVs). *Let $\mathbf{M}^K = U \Sigma V^T$ be the Singular Value Decomposition (SVD) of the cosine similarity matrix (Definition D.5.10). Let's define γ^X and γ^Y as:*

$$\gamma^X = U^T \left(\widetilde{\mathbf{M}}_{\mathbf{X}}\right)^{-1/2} \alpha^X \quad \text{and} \quad \gamma^Y = V^T \left(\widetilde{\mathbf{M}}_{\mathbf{Y}}\right)^{-1/2} \alpha^Y \tag{D.32}$$

Then the PVs (Definition D.6.1) can be computed as follows:

$$\forall k < \hat{d}, \quad s_k = \sum_{i=1}^{M} \gamma_{k,i}^X K\left(\tilde{x}_i^X, \cdot\right) \quad \text{and} \quad t_k = \sum_{i=1}^{M} \gamma_{k,i}^Y K\left(\tilde{x}_i^Y, \cdot\right). \tag{D.33}$$

Proof. By definition of $s_1, .., s_{\hat{d}}$ and $t_1, .., t_{\hat{d}}$ (Equation (D.31)), there exist $\xi^X \in \mathbb{R}^{\hat{d} \times d_X}$ and $\xi^Y \in \mathbb{R}^{\hat{d} \times d_Y}$ such that:

$$\forall k \in \{1, .., d\}, \quad s_k = \sum_{l=1}^{d_X} \xi_{k,l}^Y \theta_k^X \quad \text{and} \quad t_k = \sum_{l=1}^{d_Y} \xi_{k,l}^Y \theta_k^Y. \tag{D.34}$$

Using the bi-linearity of the inner product, we have:

$$\forall k, l \in \{1, .., d\}, \quad \begin{cases} \langle s_k, t_k \rangle &= {\xi_k^X}^T \widetilde{\mathbf{M}}_{X,Y} \xi_k^Y \\ \langle s_k, s_l \rangle &= {\xi_k^X}^T \widetilde{\mathbf{M}}_X \xi_l^X \\ \langle t_k, t_l \rangle &= {\xi_k^Y}^T \widetilde{\mathbf{M}}_Y \xi_l^Y \end{cases}. \tag{D.35}$$

Combining Equations (D.34) and (D.35) in Equation (D.31) yields the following:

$$\forall k \in \{1, .., \hat{d}\}, \quad \xi_k^X, \xi_k^Y = \underset{\substack{u^X \in \mathbb{R}^{d_X}, \\ u^Y \in \mathbb{R}^{d_Y},}}{\text{argmax}} \ {u^X}^T \widetilde{\mathbf{M}}_{X,Y} u^Y \quad \text{s.t} \quad \begin{cases} {u^X}^T \widetilde{\mathbf{M}}_X u^X = 1 \\ {u^Y}^T \widetilde{\mathbf{M}}_Y u^Y = 1 \\ \forall i < k, \ {u^X}^T \widetilde{\mathbf{M}}_X \xi_i^X = 0 \\ \forall i < k, \ {u^Y}^T \widetilde{\mathbf{M}}_Y \xi_i^Y = 0 \end{cases}. \tag{D.36}$$

$\mathcal{H}$ is an Hilbert space, therefore, assuming that basis functions are linearly independent, $\widetilde{\mathbf{M}}_X$ and $\widetilde{\mathbf{M}}_Y$ are symmetric and positive-definite and thus admit a square-root and

an inverse-square-root.
Let's define $\kappa^X = \gamma^X \widetilde{\mathbf{M}}_X^{1/2}$ and $\kappa^Y = \gamma^Y \widetilde{\mathbf{M}}_Y^{1/2}$. The PV definition by linearisation (Equation (D.36)) is then equivalent to the SVD of $\widetilde{\mathbf{M}}_X^{-1/2} \widetilde{\mathbf{M}}_{X,Y} \widetilde{\mathbf{M}}_Y^{-1/2} = \mathbf{M}^K$ defined as $\mathbf{M}^K = U\Sigma V^T$. We therefore have $\kappa^X = U^T$ and $\kappa^Y = V^T$, which turns to $\xi^X = U^T \widetilde{\mathbf{M}}_X^{-1/2}$ and $\xi^Y = V^T \widetilde{\mathbf{M}}_Y^{-1/2}$.
Using the Representer Theorem (Definition D.5.6), we obtained desired formula. □

Corollary D.6.2.1. *Using same notation as Theorem D.6.2 with* $\mathbf{M}^K = U\Sigma V^T$*:*

$$\forall k \in \{1,..,\hat{d}\}, \quad \langle s_k, t_k \rangle_{\mathscr{H}} = \Sigma_{k,k} \in [0,1]. \tag{D.37}$$

These values can be understood as the cosine values of angles, called **principal angles**.

D

D.7. INTERPRETING THE LATENT FACTORS BY TAYLOR EXPANSION OF KERNEL APPROXIMATION

We here present our interpretability scheme, which aims at understanding which genes, or combinations thereof, contribute the most to the latent factors. Our scheme relies on the Taylor expansion of the kernel used for the approximation. Unfortunately, to the best of our knowledge, no analytical form of an orthonormal basis for the Matérn RKHS exists in the literature. We are therefore limited to the Gaussian kernel of length-scale σ, which we refer to as K_σ in the sequel ; we refer to its RKHS as $\mathscr{H}_\sigma$.

D.7.1. ORTHONORMAL BASIS OF GAUSSIAN FEATURE SPACE

We here summarise the construction of the orthonormal basis (ONB) which we exploit. Its complete derivation can be found in [169].

Definition D.7.1 (Univariate basis function)**.** *Let* $j \in \{1,..,p\}$ *and* $k \in \mathbb{N}$*. We define the univariate basis function* $e_j^k : \mathbb{R}^p \mapsto \mathbb{R}$ *as:*

$$\forall x \in \mathbb{R}^p, \quad e_j^k(x) = \frac{x_j^k}{\sigma^k \sqrt{k!}} \exp\left(-\frac{x_j^2}{2\sigma^2}\right) \tag{D.38}$$

Definition D.7.2 (Gaussian basis function)**.** *Let* $I = (I_1,\ldots,I_p) \in \mathbb{N}^p$*, we define the Gaussian basis function* G_I *as:*

$$\forall x \in \mathbb{R}^p, \quad G_I(x) = \prod_{j=1}^p e_j^{I_j}(x) = \left[\prod_{j=1}^p \frac{x_j^{I_j}}{\sigma^{I_j}\sqrt{I_j!}}\right] \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right). \tag{D.39}$$

Proposition D.7.3 (Orthonormal basis)**.** *Let* $I, J \in \mathbb{N}^p$ *and* $\mathbb{1}_{I,J}$ *be the dirac function (equals to one if all values of* I *and* J *are equal, zero otherwise). We have:*

$$\langle G_I, G_J \rangle_{\mathscr{H}_\sigma} = \mathbb{1}_{I,J}, \tag{D.40}$$

which shows that the $(G_I)_{I\in\mathbb{N}^p}$ *defines an orthonormal family of* $\mathcal{H}_\sigma$.
This family furthermore defines a basis of $\mathcal{H}$*, i.e.,*

$$\forall x, y \in \mathbb{R}^p, \quad K_\sigma(x,y) = \sum_{I\in\mathbb{N}^p} G_I(x)\, G_I(y) = \langle \mathcal{G}_\sigma(x)\, \mathcal{G}_\sigma(y)\rangle_{\mathcal{H}_\sigma}, \tag{D.41}$$

with $\mathcal{G}_\sigma(x) = \sum_{I\in\mathbb{N}^p} G_I(x)\, G_I$.

Proof. See Theorem 3 and Theorem 4 of *Steinwart et al* [169] for the proof of this proposition. □

D.7.2. FEATURE ATTRIBUTION SCORES

Using results from Proposition D.7.3, we highlight three kinds of Gaussian basis functions (Definition D.7.2) of interest: linear, interaction and higher-order interaction terms.

Definition D.7.4 (Dirac vector)**.** *Let* $k \in \{1,..,p\}$*. We define as* $\delta_k \in \mathbb{N}^p$ *the vectors of zeros, with a single one at the* k*-th position.*

Definition D.7.5 (Linear weights)**.** *Let* $\theta \in \mathcal{H}_\sigma$ *and* $k \in \{1,..,p\}$*. We define the* k*-th feature weight of* θ*, termed* $\mathcal{L}_k$*, as the projection on* G_{δ_k}*:*

$$\mathcal{L}_k(\theta) = \langle \theta, G_{\delta_k}\rangle_{\mathcal{H}_\sigma}. \tag{D.42}$$

This quantity intuitively corresponds to the weight of a single-feature (the k-th) modulated by the squared exponential term.

Definition D.7.6 (Interaction weight)**.** *Let* $\theta \in \mathcal{H}_\sigma$ *and let* $k, l \in \{1,..,p\}$*. We define the* k,l*-interaction weight of* θ*, termed* $\mathcal{I}_{k,l}$*, as the projection on* $G_{\delta_k+\delta_l}$*:*

$$\mathcal{I}_{k,l}(\theta) = \langle \theta, G_{\delta_k+\delta_l}\rangle_{\mathcal{H}_\sigma}. \tag{D.43}$$

This quantity intuitively corresponds to the weight of an interaction term, i.e. the product of two genes, modulated by the squared exponential term. We can observe that, when $k = l$, the interaction terms corresponds to a quadratic term.

Definition D.7.7 (I-higher order weight)**.** *Let* $\theta \in \mathcal{H}_\sigma$ *be an element of the Gaussian kernel RKHS. Let* $I \in \mathbb{N}^p$ *with* $|I| = \sum_{i=1}^p I_i \geq 3$*. The* I*-higher-order-interaction weight of* θ*, termed* $\mathcal{I}_I^+$*, is defined as the projection on* G_I*:*

$$\mathcal{I}_I^+(\theta) = \langle \theta, G_I\rangle_{\mathcal{H}_\sigma}. \tag{D.44}$$

Using these three kinds of contributions, we can define the global contribution of linear, interaction and higher order interactions as follows.

Definition D.7.8 (Linear, interactions and higher order interactions contributions)**.** *Let $\theta \in \mathcal{H}_\sigma$. We define the linear, interactions and higher-order interactions contributions, called $\mathscr{L}$, $\mathscr{I}$ and $\mathscr{I}^+$ respectively, as the squared norm of the associated vector weights, i.e.:*

$$\mathscr{L}(\theta) = \sum_{k=1}^{p} \mathscr{L}_k(\theta)^2, \quad \mathscr{I}(\theta) = \sum_{1\leq k\leq l\leq p} \mathscr{I}_{k,l}(\theta)^2 \quad \text{and} \quad \mathscr{I}^+(\theta) = \sum_{\substack{I\in\mathbb{N}^p \\ |I|\geq 3}} \mathscr{I}_I^+(\theta)^2. \tag{D.45}$$

D.7.3. COMPUTATION OF GENE-LEVEL CONTRIBUTION (LINEAR)

We now present how to compute the linear weights. These rely on the artificial samples (Definition D.5.5) used for approximating the latent factor (Definition D.5.6).

Definition D.7.9 (Offset matrices)**.** *We define the two offset matrices $\mathscr{O}^X \in \mathbb{R}^M$ and $\mathscr{O}^Y \in \mathbb{R}^M$ as*

$$\mathscr{O}^X = \operatorname{diag}\left[\exp\left(-\frac{\|\widetilde{x}_i^X\|^2}{2\sigma^2}\right)\right]_{1\leq i\leq M} \quad \text{and} \quad \mathscr{O}^Y = \operatorname{diag}\left[\exp\left(-\frac{\|\widetilde{x}_i^Y\|^2}{2\sigma^2}\right)\right]_{1\leq i\leq M} \tag{D.46}$$

Definition D.7.10 (Model points matrices)**.** *We define the two artificial sample matrices $\mathscr{A}^X \in \mathbb{R}^{M\times p}$ and $\mathscr{A}^Y \in \mathbb{R}^{M\times p}$ as:*

$$\mathscr{A}^X = \left[\widetilde{x}_1^X, .., \widetilde{x}_M^X\right]^T \quad \text{and} \quad \mathscr{A}^Y = \left[\widetilde{x}_1^Y, .., \widetilde{x}_M^Y\right]^T \tag{D.47}$$

Theorem D.7.11 (Computing individual contributions)**.** *We have the following equalities, for all $t \in \{X, Y\}$:*

$$\begin{aligned}
\mathscr{L}_{fact}^t &= \left(\mathscr{L}_j\left(\theta_i^t\right)\right)_{\substack{1\leq i\leq d_t \\ 1\leq j\leq p}} = \frac{1}{\sigma}\alpha^t\mathscr{O}^t\mathscr{A}^t. \\
\mathscr{L}_{SPV}^X &= \left(\mathscr{L}_j(s_i)\right)_{\substack{1\leq i\leq d_X \\ 1\leq j\leq p}} = \frac{1}{\sigma}\gamma^X\mathscr{O}^X\mathscr{A}^X. \\
\mathscr{L}_{SPV}^Y &= \left(\mathscr{L}_j(t_i)\right)_{\substack{1\leq i\leq d_Y \\ 1\leq j\leq p}} = \frac{1}{\sigma}\gamma^Y\mathscr{O}^Y\mathscr{A}^Y.
\end{aligned} \tag{D.48}$$

Proof. We first recall the second reproducing property of the kernel K_σ:

$$\forall f \in \mathscr{H}, \forall x \in \mathbb{R}^p, \quad \left\langle K_\sigma^G(x,\cdot), f\right\rangle_{\mathscr{H}_\sigma} = f(x). \tag{D.49}$$

Let $1 \leq i \leq d_t$ and $1 \leq j \leq p$. Using Definitions D.5.6 Definition D.7.5 and the aforementionned second reproducing property, we have:

$$\mathscr{L}_j\left(\theta_i^t\right) = \sum_{k=1}^{M} \alpha_{i,k}^t \left\langle K_\sigma^G\left(\widetilde{x}_k^t,\cdot\right), G_{\delta_j}\right\rangle_{\mathscr{H}_\sigma} = \sum_{k=1}^{M} \alpha_{i,k}^t G_{\delta_j}\left(\widetilde{x}_k^t\right). \tag{D.50}$$

Using the definition of the Gaussian basis functions (Definition D.7.2), we obtain:

$$\mathscr{L}_j\left(\theta_i^t\right) = \sum_{k=1}^{M} \alpha_{i,k}^t \frac{\widetilde{x}_{k,j}^t}{\sigma} \exp\left(-\frac{\left\|\widetilde{x}_k^t\right\|^2}{2\sigma^2}\right), \tag{D.51}$$

which, put in matrix format, gives the desired result.
Same idea gives the other two results, using the expansion of Theorem D.6.2 instead of Definition D.5.6. □

It follows from Theorem D.7.11 that the global linear contribution can be computed as follows.

Proposition D.7.12 (Computing global contributions)**.** *We have the following equalities:*

$$\forall t \in \{X, Y\}, \forall i \in \{1, .., d_t\}, \quad \mathscr{L}\left(\theta_i^t\right) = \frac{1}{\sigma^2}\left(\alpha^t \mathscr{O}^t \mathscr{A}^t \mathscr{A}^{t\,T} \mathscr{O}^t \alpha^{t\,T}\right)_{i,i}. \tag{D.52}$$

$$\forall i \in \{1, .., d_X\}, \quad \mathscr{L}\left(s_i\right) = \frac{1}{\sigma^2}\left(\gamma^X \mathscr{O}^t \mathscr{A}^t \mathscr{A}^{t\,T} \mathscr{O}^t \gamma^{X\,T}\right)_{i,i}. \tag{D.53}$$

$$\forall i \in \{1, .., d_Y\}, \quad \mathscr{L}\left(t_i\right) = \frac{1}{\sigma^2}\left(\gamma^Y \mathscr{O}^t \mathscr{A}^t \mathscr{A}^{t\,T} \mathscr{O}^t \gamma^{Y\,T}\right)_{i,i}. \tag{D.54}$$

Proof. Immediate by combining Theorem D.7.11 with Definition D.7.8. □

D.7.4. COMPUTATION OF INTERACTION WEIGHTS

Definition D.7.13 (Gene expression product matrix)**.** *We define the matrices* $\overline{\mathscr{A}^X}$ *and* $\overline{\mathscr{A}^Y}$ *as*

$$\begin{aligned} \overline{\mathscr{A}}^X &= \left[\mathscr{A}_{:,i}^X \circ \mathscr{A}_{:,j}^X\right]_{1\le i\le j\le p} \in \mathbb{R}^{M\times\frac{p(p+1)}{2}} \\ \overline{\mathscr{A}}^Y &= \left[\mathscr{A}_{:,i}^Y \circ \mathscr{A}_{:,j}^Y\right]_{1\le i\le j\le p} \in \mathbb{R}^{M\times\frac{p(p+1)}{2}}, \end{aligned} \tag{D.55}$$

where $\circ$ *is the Hadamard (piece-wise) product between two vectors.*

The product matrices presented in Definition D.7.13 correspond to the products of columns of $\mathscr{A}^X$ and $\mathscr{A}^Y$. The indices i and j are ordered by first setting i and then varying j in increasing order, i.e. $(1,1),(1,2),..,(1,p),(2,2),..,(2,p),..,(p-1,p-1),(p-1,p),(p,p)$.
In order to keep track of the different interaction terms and avoid computing the same interaction terms twice, we introduce the following interaction indexing.

Definition D.7.14 (Interaction indexing)**.** *We define the interaction indexing function* Υ *as:*

$$\forall i, j \in \{1, .., p\}, \quad \Upsilon\left(i, j\right) = \begin{cases} (i-1)\left(p+1-\frac{i}{2}\right)+j & \text{if } i \le j \\ \Upsilon\left(j, i\right) & \text{otherwise} \end{cases}. \tag{D.56}$$

Proposition D.7.15. *The interaction weights can be computed as follows, for $t \in \{X, Y\}$ and $k \in \{1, .., d_t\}$:*

$$\forall 1 \leq i < j \leq p, \quad \begin{aligned} \mathscr{I}_{i,j}\left(\theta_k^t\right) &= \frac{1}{\sigma^2}\left(\alpha^t \mathscr{O}^t \overline{\mathscr{A}}^t\right)_{k,\Upsilon(i,j)} \\ \mathscr{I}_{i,j}\left(s_k\right) &= \frac{1}{\sigma^2}\left(\gamma^X \mathscr{O}^X \overline{\mathscr{A}}^X\right)_{k,\Upsilon(i,j)} \\ \mathscr{I}_{i,j}\left(t_k\right) &= \frac{1}{\sigma^2}\left(\gamma^Y \mathscr{O}^Y \overline{\mathscr{A}}^Y\right)_{k,\Upsilon(i,j)} \end{aligned}. \tag{D.57}$$

D.7.5. INTERPRETRATION IN THE LAPLACIAN KERNEL

The orthonormal basis of the the Gaussian kernel of length-scale σ relies on the Hilbertian structure of the Gaussian RKHS σ. However, the Gaussian kernel K_σ and the Laplacian kernel K_σ^L correspond to the two extreme of the Matérn family: with ν going to infinity for the Gaussian and to $\frac{1}{2}$ for the Laplacian kernel. We use this identity to transfer the feature weights from the Gaussian setting to the Laplacian setting, which we present here. This identification relies on two lemmas.

Lemma D.7.16. *Let $k, l \in \{1, .., p\}$ with $k \neq l$, then*

$$\left\langle G_{\delta_k}, G_{\delta_l} \right\rangle_{\mathscr{H}_\sigma^L} = 0 \tag{D.58}$$

Proof. Let's denote by $\mathscr{F}[f]$ the Fourier transform of a function f. Following [Kimeldorf et al] [277] (Lemma 3.1), the inner-product of the Laplacian RKHS is computed as follows, for f and g two real functions:

$$\langle f, g \rangle_{\sigma^L} = \frac{1}{(2\pi)^{p/2} C_{\sigma,p}} \int_{\mathbb{R}^p} \mathscr{F}[f] \overline{\mathscr{F}[g]} \left(\frac{2}{\sigma^2} + 4\pi^2 \|\omega\|^2\right)^{\frac{1+p}{2}} d\omega, \tag{D.59}$$

with $\bar{\cdot}$ indicating complex conjugate. Noting that $\mathscr{F}\left[G_{\delta_k}\right](\omega) = -i\frac{\omega_k}{2\sigma^2} \exp\left(-\frac{\|\omega\|^2}{2\sigma^2}\right)$ by partial derivation, we can write:

$$\left\langle G_{\delta_k}, G_{\delta_l} \right\rangle_{\mathscr{H}_\sigma^L} = \frac{-1}{(2\pi)^{p/2} C_{\sigma,p}} \int_{\mathbb{R}^p} \omega_k \omega_l \left(\frac{2}{\sigma^2} + 4\pi^2 \|\omega\|^2\right)^{\frac{1+p}{2}} \exp\left(-\frac{\|\omega\|^2}{\sigma^2}\right). \tag{D.60}$$

The integrated function is odd with regard to the plane $\omega_k = 0$, so $\left\langle G_{\delta_k}, G_{\delta_l} \right\rangle_{\mathscr{H}_\sigma^L} = 0$. □

Lemma D.7.17. *Let $k_1, k_2, l_1, l_2 \in \{1, .., p\}$ with $(k_1, l_1) \neq (k_2, l_2)$, $k_1 < l_1$, $k_2 < k_2$, then*

$$\left\langle G_{\delta_{k_1}+\delta_{l_1}}, G_{\delta_{k_1}+\delta_{l_1}} \right\rangle_{\mathscr{H}_\sigma^L} = 0 \tag{D.61}$$

Proof. If $k_1 \neq l_1$, we have:

$$\mathscr{F}\left[G_{\delta_{k_1}+\delta_{l_1}}\right] = -\frac{\omega_{k_1}\omega_{l_1}}{4\sigma^2} \exp\left(-\frac{\|\omega\|^2}{2\sigma^2}\right). \tag{D.62}$$

□

These two lemmas shows that the linear and interaction features – except quadratic terms – form an orthogonal family of the Laplacian kernel RKHS. Their norm, however, is not unit and computation thereof is analytically challenging and computationally untractable due to the high dimensionality of the integration problem. In order to correct and obtain a unit orthogonal family, we correct by using the linear and interaction contributions from the Gaussian kernel (Definition D.7.8). Specifically, for a function f, we multiply all the linear terms by $\frac{\mathscr{L}(f)}{\mathscr{L}^L(f)}$ where L superscripts refers to the contribution in the Laplacian kernel (no super-script for the gaussian kernel). We similarly multiply all interaction terms by $\frac{\mathscr{I}(f)}{\mathscr{I}^L(f)}$.

D.8. DIFFERENCES AND COMPARISON WITH CANONICAL CORRELATION ANALYSIS (CCA)

Our methodology crucially relies on the notion of Principal Vectors (PV) in Hilbert spaces. Since this notion also forms the backbone of Canonical Correlation Analysis (CCA), and its kernel extension (k-CCA), it is natural to question the differences between our approach and k-CCA. Apart from sharing the conceptual notion of Principal Vectors, we believe our approach to markedly differ from CCA for two main reasons: a difference in the input-pairing, and important differences in the maximization problem.

D.8.1. K-CCA IS UNSUPERVISED AND REQUIRES SAMPLE PAIRING

SHORT PRESENTATION OF KERNEL-CCA

Kernel CCA (k-CCA) takes as input:

- Two paired sets of samples $x_1, .., x_n \in \mathscr{X}$ and $y_1, .., y_n \in \mathscr{Y}$, with $\mathscr{X}$ and $\mathscr{Y}$ two sets.
- Two kernels $K_A : \mathscr{X} \times \mathscr{X} \longrightarrow \mathbb{R}$ and $K_B : \mathscr{Y} \times \mathscr{Y} \longrightarrow \mathbb{R}$ with respective RKHS $\mathscr{H}_A$ (Theorem D.3.2).

K-CCA looks for the functions $f_A \in \mathscr{H}_A$ and $f_B \in \mathscr{H}_B$ with maximal correlation across the samples, i.e., maximising

$$\frac{\frac{1}{n}\sum_{i=1}^n f_A(x_i) f_B(y_i)}{\sqrt{\frac{1}{n}\sum_{i=1}^n f_A(x_i)^2}\sqrt{\frac{1}{n}\sum_{i=1}^n f_B(y_i)^2}}. \tag{D.63}$$

The solutions of this problem are of the form $f_A = \sum_{1\leq i\leq n} \alpha_i K_A(x_i, \cdot)$ and $f_B = \sum_{1\leq i\leq n} \beta_i K_B(y_i, \cdot)$, with $\alpha, \beta \in \mathbb{R}^n$. If we denote by $K_X = \left(K_A(x_i, x_j)\right)_{1\leq i,j\leq n}$ and $K_Y = \left(K_B(y_i, y_j)\right)_{1\leq i,j\leq n}$, α and β are the results of the following maximisation.

$$\max_{\substack{\alpha,\beta\in\mathbb{R}^n,\\ \alpha^T K_X^2 \alpha = 1\\ \beta^T K_Y^2 \beta = 1}} \alpha^T K_X K_Y \beta. \tag{D.64}$$

This formulation shows that k-CCA is completely unsupervised, and does not readily take into account the trained VAE. It can however be used between the model samples which provide a larger training data for k-CCA.

DIFFERENCE IN PAIRING

As shown in the previous section, k-CCA requires a sample-pairing. In our scenario, cell line and tumor model samples (Definition D.5.5) constitute two independent sets; they cannot therefore directly be used in k-CCA. Original cells from cell lines and tumors are also not related, showing a first difference between k-CCA and Sobolev Alignment.
Cell line and tumor cells (and model samples), however, show a gene-level pairing. If we consider the input data (Section D.2), we could exploit this pairing in CCA, yielding functions f_A and f_B which take as input the values of a single gene across all cell line and tumor samples, respectively. Such functions are therefore not living in the same functional space as the one used in Sobolev Alignment (Definition D.5.6), and cannot be easily used for assessing the similarity between cells from cell lines and tumors. Finally, such procedure would require two sets of kernels (one for cell lines, one for tumors), while Sobolev Alignment only requires one kernel.

D.8.2. DIFFERENCE IN OPTIMISATION PROBLEM

Even if we assume cells from cell lines and tumors to be paired, we argue that the optimisation problems in kernel-CCA and Sobolev Alignment markedly differ. Let us consider the definitions of kernel-CCA (Equation D.64) and Sobolev Alignment (particularly, Equation D.36). Both optimisations rely on finding a left- and a right-vector which maximise the spectrum of a certain non-symmetric matrix given some constraints. However, the dimensions of the vectors differ between k-CCA and Sobolev Alignment, and, more profoundly, Sobolev Alignment requires the computation of K_{XY}, which k-CCA does not: it only relies on $K_X K_Y$. Without further assumption on the hypothetical pairing between cell line and tumor cells, it is not clear to us that these two matrices can be somewhat related.
Finally, the constraints of the two problems strongly differ: K_X^2 and K_Y^2 are used in k-CCA while only K_X and K_Y are used for Sobolev Alignment.

D.9. ALGORITHM

D.10. GLOSSARY

Algorithm 3 Sobolev Alignment

Require: Datasets $\mathscr{X}$ and $\mathscr{Y}$, scVI parameters, number anchors M, number of artificial points N, penalizations λ_X and λ_Y, Matérn kernel $K_{\nu,\sigma}$.
Train scVI (VAE) model $\mathscr{M}_X$ on $\mathscr{X}$ (d_X hidden neurons).
Train scVI (VAE) model $\mathscr{M}_X$ on $\mathscr{Y}$ (d_Y hidden neurons).
$Z_X \leftarrow N$ vectors sampled from $\mathscr{N}\left(0, I_{d_X}\right)$.
$\widehat{X}_X \leftarrow$ decoding of Z_X using decoder of $\mathscr{M}_X$
$\widehat{Z}_X \leftarrow$ encoding of $\widehat{X}_X$ using encoder of $\mathscr{M}_X$
$Z_Y \leftarrow N$ vectors sampled from $\mathscr{N}\left(0, I_{d_Y}\right)$.
$\widehat{X}_Y \leftarrow$ decoding of Z_Y using decoder of $\mathscr{M}_Y$.
$\widehat{Z}_Y \leftarrow$ encoding of $\widehat{X}_Y$ using encoder of $\mathscr{M}_Y$.
$\theta_1^X, ...\theta_{d_X}^X \leftarrow$ KRR models between $\widehat{X}_X$ (input) and $\widehat{Z}_X$ (label).
$\theta_1^Y, ...\theta_{d_Y}^Y \leftarrow$ KRR models between $\widehat{X}_Y$ (input) and $\widehat{Z}_Y$ (label).
$\alpha^X \leftarrow$ Sample coefficients of $\theta_1^X, ...\theta_{d_X}^X$.
$\alpha^Y \leftarrow$ Sample coefficients of $\theta_1^Y, ...\theta_{d_Y}^Y$.
$K_X, K_Y, K_{XY} \leftarrow$ Kernel matrices using $K_{\nu,\sigma}$ on $\widehat{X}_X$ and $\widehat{X}_Y$. {Definition D.5.8}.
$\widetilde{\mathbf{M}}_X \leftarrow \alpha^X K_X {\alpha^X}^T$.
$\widetilde{\mathbf{M}}_Y \leftarrow \alpha^Y K_Y {\alpha^X}^Y$.
$\widetilde{\mathbf{M}}_{X,Y} \leftarrow \alpha^X K_{X,Y} {\alpha^Y}^T$.
$\mathbf{M} \leftarrow \widetilde{\mathbf{M}}_X^{-1/2} \mathbf{M}_{X,Y} \widetilde{\mathbf{M}}_Y^{-1/2}$. {Definition D.5.10}.
$U, \Sigma, V \leftarrow$ SVD decomposition of $\mathbf{M}$ $\{\mathbf{M} = U\Sigma V^T\}$.
$\gamma^X \leftarrow U^T \widetilde{\mathbf{M}}_X^{-1/2} \alpha^X$.
$\gamma^Y \leftarrow V^T \widetilde{\mathbf{M}}_Y^{-1/2} \alpha^Y$. {Theorem D.6.2.}

Symbol	Meaning	Reference
p	Number of genes (features).	
n_X and n_Y	Number of source (cell-lines) and target (tumors) samples.	
$\mathbb{R}$	Real numbers.	
I_d	Identity matrix of size d.	
diag	Diagonal matrix.	
$\langle \cdot, \cdot \rangle$	Inner-product.	
$\mathcal{N}(\mu, \Sigma)$	Multivariate normal distribution, μ: mean, Σ: covariance.	
$\mathcal{X}_X$ and $\mathcal{X}_Y$	Source and target datasets, of sizes $n_X \times p$ and $n_Y \times p$.	
K_σ^L	Laplacian kernel, σ:lengthscale.	Definition D.3.4
$K_{\nu,\sigma}^M$	Matérn kernel, ν:smoothness, σ:lengthscale.	Definition D.3.5
K_σ^G	Gaussian kernel, σ:lengthscale.	Definition D.3.6
Γ	Gamma function	
K_α	Modified Bessel function of second kind of order α.	
$L_2(\mathbb{R}^p)$	Space of continuous integrable functions.	Definition D.3.9
$W_2^s(\mathbb{R}^p)$	Sobolev space of order s.	Definition D.3.10
$\mathcal{H}_{\nu,\sigma}$	RKHS associated to the Matérn kernel $K_{\nu,\sigma}^M$.	Proposition D.3.12
d_X and d_Y	Number of source and target latent variables.	Definition D.4.2
$\mu_1^X, .., \mu_{d_X}^X$	Mean embedding function for source VAE.	Definition D.4.2
$\Sigma_1^X, .., \Sigma_{d_X}^X$	Standard-deviation embedding function for source VAE.	Definition D.4.2
$f_1^X, .., f_{d_X}^X$	Encoding (mean) functions for cell lines scVI model.	Definition D.5.1
$z_1^X, .., z_N^X$	Points randomly sampled from noise (cell).	Definition D.5.5
$\widehat{x}_1^X, .., \widehat{x}_N^X$	Decoded values of $z_1^X, .., z_N^X$ using cell line scVI model	Definition D.5.5
$\widehat{z}_1^X, .., \widehat{z}_N^X$	Encoded values of $\widehat{x}_1, .., \widehat{x}_N^X$ using cell line scVI model	Definition D.5.5
$\widehat{\theta}_k^t$	KRR approximations for encoding functions $\widehat{f}_k^t$	Definition D.5.6
M	Number of anchor points (Falkon approximation)	Proposition D.5.4
α^X and α^Y	Sample weights for KRR approximations	Definition D.5.6
$\widetilde{\mathbf{M}}_{XY}$	Un-normalized cosine similarity matrix	Definition D.5.10
$\widetilde{\mathbf{M}}_X$ and $\widetilde{\mathbf{M}}_Y$	Inner-product matrices between latent factors	Definition D.5.10
K_X, K_Y, and K_{XY}	Similarity matrices (between samples)	Definition D.5.8
$\mathbf{M}^K$	Cosine similarity matrix	Definition D.5.10
$\widehat{d}$	Number of principal vectors (PVs)	Definition D.6.1
$s_1, ..., s_{\widehat{d}}$	Source principal vectors	Definition D.6.1
$t_1, ..., t_{\widehat{d}}$	Target principal vectors	Definition D.6.1
γ^X, γ^Y	Sample-weights for source and target PVs	Theorem D.6.2
$G_I, I \in \mathbb{N}^p$	Gaussian basis function	Definition D.7.2
$\mathscr{L}_k, k \in \{1, .., p\}$	Linear weights	Definition D.7.5
$\mathscr{I}_{k,l}, k, l \in \{1, .., p\}$	Interaction weights	Definition D.7.6
$\mathscr{O}^X, \mathscr{O}^Y$	Offset matrices	Definition D.7.9
$\mathscr{A}^X, \mathscr{A}^Y$	Model points matrices	Definition D.7.10
$\overline{\mathscr{A}}^X, \overline{\mathscr{A}}^Y$	Gene expression products matrix	Definition D.7.14
Υ	Interaction indexing	Definition D.7.13
$\mathscr{F}$	Fourier transform.	

E

PERSPECTIVE: LEARNING A SUITABLE DISTANCE BETWEEN GENE EXPRESSION PROFILES USING DEEP KERNEL LEARNING

E.1. DEEP KERNEL LEARNING ON SINGLE CELL RNA-SEQ DATA

In this thesis, we have made an extensive use of kernel methods and have shown their relevance when integrating cell line and tumor gene expression data. These non-parametric methods offer several attractive properties which we exploited in TRANSACT (Chapter 4) and Sobolev Alignment (Chapter 5). In both chapters, we employed kernels from the Matérn family, a standard choice used in other studies [47, 73]. These Matérn kernels however do not encode any prior knowledge on the data, which, we reasoned, could help increase performance in drug response prediction by constraining the regression model to focus on pre-selected patterns. Beyong drug response prediction, the design of a biologically-informed kernel, understood as a similarity function, would have implications for other analysis, e.g., clustering or lineage tracing.

As presented in Chapter 5, scRNA-seq datasets form a wealth of unlabelled data which recapitulate important patterns of gene expression. Exploiting such expansive data to design a similarity measure, or kernel, between gene expression profiles therefore offers an enticing prospect. In order to model such similarity measure, we propose to use Deep Kernel Learning, a learning framework which combines a standard kernel function with a parametric neural network [235]. In the rest of this section, we present our approach: we first define what a deep kernel consists of, and then present the cost-function to be used.

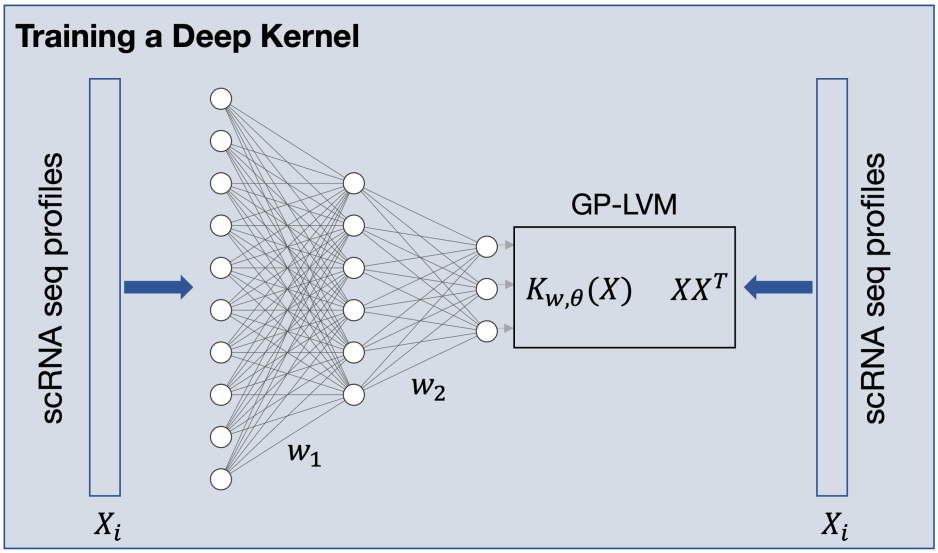


Figure E.1 – **Learning a Deep Kernel on scRNA-seq data.**[1] We propose an unsupervised extension of Deep Kernel Learning which relies on a refinement of GP-LVMs. scRNA-seq profiles are given as input to a neural network, yielding low-dimensional embeddings. These embeddings are then compared using a GP-LVM with a Gaussian kernel on both input and embedding. As this complete procedure is differentiable, the cost function can be minimized by back-propagation. The end-product of this computational routine is a similarity function which can be used in other tasks, e.g., TRANSACT.

E.2. DEEP KERNEL

Definition E.2.1 (Deep Kernel)**.** *Let p the number of genes studied and $d \in \mathbb{N}$. Let f_ω be a neural network parametrized by the coefficients ω, of input-size p and output-size d. Let K_θ be a kernel with parameter θ of input-size d. We define the deep kernel $K_{\omega,\theta}$ as:*

$$\forall x, y \in \mathbb{R}^p, \quad K_{\omega,\theta} = K_\theta\left(f_\omega(x), f_\omega(y)\right). \tag{E.1}$$

We show that, provided the kernel K_θ is positive definite, then a deep kernel based on $K_{\omega,\theta}$ would also be positive definite. Such property is important, as it shows that K_θ corresponds to the inner product of two samples after embedding in a feature space.

Proposition E.2.2 ($K_{\omega,\theta}$ is p.d.)**.** *Let assume that K_θ is positive definite (p.d.). Then for $\omega \in \Omega$, $K_{\omega,\theta}$ is p.d.*

Proof. K_θ is p.d. Following Aronszajn thereom, there exists a Hilbert space $\mathscr{H}$ and a function $\phi : \mathbb{R}^d \mapsto \mathscr{H}$ such that:

$$\forall x, y \in \mathbb{R}^d, \quad K_\theta(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathscr{H}}. \tag{E.2}$$

Let $\omega \in \Omega$. By combining Equations E.1 and E.2, we obtain:

$$\begin{aligned} \forall x, y \in \mathbb{R}^p, \quad K_{\omega,\theta}(x, y) &= K_\theta\left(f_\omega(x), f_\omega(y)\right) \\ &= \langle \phi\left(f_\omega(x)\right), \phi\left(f_\omega(y)\right) \rangle_{\mathscr{H}}. \end{aligned} \tag{E.3}$$

Therefore, $K_{\omega,\theta}$ can be represented in space $\mathscr{H}$ with mapping $\phi \circ f_\omega$. It is thus p.d. . □

In their seminal paper, *Wilson and colleagues* propose a supervised learning approach to learn the parameters corresponding to a deep kernel $K_{\omega,\theta}$, and specifically the neural network weights ω. Unfortunately, the datasets we exploit here do not contain any continuous phenotype, preventing us from applying this strategy. In order to nonetheless utilize the notion of deep kernels, we derived an unsupervised methodology based on the notion of *Gaussian-Process Latent-Variable Models* (GP-LVM) [76].

E.3. REFINING THE GP-LVM LOSS FUNCTION FOR UNSUPERVISED TRAINING OF DEEP KERNELS

GP-LVM is a dimensionality reduction method which extends Kernel PCA, the backbone of TRANSACT (Chapter 4). A GP-LVM is defined as follows.

Definition E.3.1 (Gaussian Process Latent Variable Model)**.** *Let $d \in \{1, .., p\}$ be an integer. Let K_p be a kernel on $\mathbb{R}^p$ and K_d a kernel on $\mathbb{R}^d$, both p.d. We define by $S = K_p(X)$ the kernel matrix of the data using kernel K_p, and for any $Y \in \mathbb{R}^{n \times d}$, $K = K_d(Y)$ the kernel matrix of Y using kernel K_d. The embedding Y^* are computed by matching the Gaussian Processes associated to K_p and K_d:*

$$\begin{aligned} Y^* &= \underset{Y \in \mathbb{R}^{n \times d}}{\operatorname{argmin}} \, \mathrm{KL}\left[\mathcal{N}(0, S) \,||\, \mathcal{N}(0, K)\right] \\ &= \frac{1}{2}\log|K| - \frac{1}{2}\log|S| + \frac{1}{2}\mathrm{Tr}\left(SK^{-1}\right) - \frac{n}{2} \end{aligned} \tag{E.4}$$

with KL the Kullback-Leibler divergence between distributions, and $\mathcal{N}$ the normal distribution.

In the case where K_p and K_d are linear kernels, this formulation is strictly equivalent to PCA. The KL-divergence from Equation E.4 can therefore be understood as a reconstruction error. In the special case when K_d is linear, the solution is exactly equivalent to Kernel PCA with kernel K_p.
In Equation E.4, the embedding computed by the GP-LVM is the solution of the optimisation problem. To learn our deep kernel $K_{\omega,\theta}$, we reasoned that this embedding Y could be replaced with the embedding produced by the neural network f_ω. The kernel K would then be substituted by the matrix $\left(K_{\omega,\theta}\left(X_i, X_j\right)\right)_{1\leq i,j\leq n}$. As a direct consequence, Equation (E.4) is replaced by a minimization with regards to the neural network weights ω and the kernel hyper-parameter γ.

Definition E.3.2 (GP-LVM training for deep kernels)**.** *We define the objective function (or loss function) for the GP-LVM estimation of the Deep Kernel, $\mathscr{L}_{GP-LVM}$ defined as:*

$$\begin{aligned}\forall \gamma \in \Omega \times \Theta, \quad \mathscr{L}_{GP-LVM}(\gamma) &= \mathrm{KL}\left[\mathcal{N}\left(0, XX^T\right) || \mathcal{N}\left(0, K_\gamma\right)\right] \\ &= \frac{1}{2}\log\left|K_\gamma + \sigma^2 I_n\right| - \frac{1}{2}\log\left|XX^T\right| \\ &\quad + \frac{1}{2}\mathrm{Tr}\left(XX^T\left(K_\gamma + \sigma^2 I_n\right)^{-1}\right) - \frac{n}{2}.\end{aligned} \tag{E.5}$$

Removing the terms that are independent from γ, we get the following estimator:

$$\gamma^{GP-LVM} = \underset{\omega,\theta\in\Omega\times\Theta}{\operatorname{argmin}} \ \mathrm{Tr}\left(XX^T\left(K_\gamma + \sigma^2 I_n\right)^{-1}\right) + \log\left|K_\gamma + \sigma^2 I_n\right| \tag{E.6}$$

We summarise the complete mathematical workflow in Figure E.1.

REFERENCES

[1] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021.

[2] Peter C. Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 1976.

[3] Mel Greaves and Carlo C. Maley. Clonal evolution in cancer. *Nature*, 481(7381):306–313, 2012.

[4] Keren Yizhak, François Aguet, Jaegil Kim, Julian M. Hess, Kirsten Kübler, Jonna Grimsby, Ruslana Frazer, Hailei Zhang, Nicholas J. Haradhvala, Daniel Rosebrock, Dimitri Livitz, Xiao Li, Eila Arich-Landkof, Noam Shoresh, Chip Stewart, Ayellet V. Segrè, Philip A. Branton, Paz Polak, Kristin G. Ardlie, and Gad Getz. RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science*, 364(6444), 2019.

[5] Douglas Hanahan and Robert A. Weinberg. The Hallmarks of Cancer. *Cell*, 100:57–70, 2000.

[6] Douglas Hanahan and Robert A. Weinberg. Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674, 2011.

[7] Douglas Hanahan. Hallmarks of Cancer: New Dimensions. *Cancer discovery*, 12(1):31–46, 2022.

[8] Mathew J. Garnett, Elena J. Edelman, Sonja J. Heidorn, Chris D. Greenman, Anahita Dastur, King Wai Lau, Patricia Greninger, I. Richard Thompson, Xi Luo, Jorge Soares, Qingsong Liu, Francesco Iorio, Didier Surdez, Li Chen, Randy J. Milano, Graham R. Bignell, Ah T. Tam, Helen Davies, Jesse A. Stevenson, Syd Barthorpe, Stephen R. Lutz, Fiona Kogera, Karl Lawrence, Anne McLaren-Douglas, Xeni Mitropoulos, Tatiana Mironenko, Helen Thi, Laura Richardson, Wenjun Zhou, Frances Jewitt, Tinghu Zhang, Patrick O'Brien, Jessica L. Boisvert, Stacey Price, Wooyoung Hur, Wanjuan Yang, Xianming Deng, Adam Butler, Hwan Geun Choi, Jae Won Chang, Jose Baselga, Ivan Stamenkovic, Jeffrey A. Engelman, Sreenath V. Sharma, Olivier Delattre, Julio Saez-Rodriguez, Nathanael S. Gray, Jeffrey Settleman, P. Andrew Futreal, Daniel A. Haber, Michael R. Stratton, Sridhar Ramaswamy, Ultan McDermott, and Cyril H. Benes. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570–575, 2012.

[9] Francesco Iorio, Theo A. Knijnenburg, Daniel J. Vis, Graham R. Bignell, Michael P. Menden, Michael Schubert, Nanne Aben, Emanuel Gonçalves, Syd Barthorpe, Howard Lightfoot, Thomas Cokelaer, Patricia Greninger, Ewald van Dyk, Han Chang, Heshani de Silva, Holger Heyn, Xianming Deng, Regina K. Egan, Qingsong Liu, Tatiana Mironenko, Xeni Mitropoulos, Laura Richardson, Jinhua Wang, Tinghu Zhang, Sebastian Moran, Sergi Sayols, Maryam Soleimani, David Tamborero, Nuria Lopez-Bigas, Petra Ross-Macdonald, Manel Esteller, Nathanael S. Gray, Daniel A. Haber, Michael R. Stratton, Cyril H. Benes, Lodewyk F.A. Wessels, Julio Saez-Rodriguez, Ultan McDermott, and Mathew J. Garnett. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*, 2016.

[10] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A. Margolin, Sungjoon Kim, Christopher J. Wilson, Joseph Lehár, Gregory V. Kryukov, Dmitriy Sonkin, Anupama Reddy, Manway Liu, Lauren Murray, Michael F. Berger, John E. Monahan, Paula Morais, Jodi Meltzer, Adam Korejwa, Judit Jané-Valbuena, Felipa A. Mapa, Joseph Thibault, Eva Bric-Furlong, Pichai Raman, Aaron Shipway, Ingo H. Engels, Jill Cheng, Guoying K. Yu, Jianjun Yu, Peter Aspesi, Melanie De Silva, Kalpana Jagtap, Michael D. Jones, Li Wang, Charles Hatton, Emanuele Palescandolo, Supriya Gupta, Scott Mahan, Carrie Sougnez, Robert C. Onofrio, Ted Liefeld, Laura MacConaill, Wendy Winckler, Michael Reich, Nanxin Li, Jill P. Mesirov, Stacey B. Gabriel, Gad Getz, Kristin Ardlie, Vivien Chan, Vic E. Myer, Barbara L. Weber, Jeff Porter, Markus Warmuth, Peter Finan, Jennifer L. Harris, Matthew Meyerson, Todd R. Golub, Michael P. Morrissey,

William R. Sellers, Robert Schlegel, and Levi A. Garraway. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, 2012.

[11] Mahmoud Ghandi, Franklin W. Huang, Judit Jané-Valbuena, Gregory V. Kryukov, Christopher C. Lo, E. Robert McDonald, Jordi Barretina, Ellen T. Gelfand, Craig M. Bielski, Haoxin Li, Kevin Hu, Alexander Y. Andreev-Drakhlin, Jaegil Kim, Julian M. Hess, Brian J. Haas, François Aguet, Barbara A. Weir, Michael V. Rothberg, Brenton R. Paolella, Michael S. Lawrence, Rehan Akbani, Yiling Lu, Hong L. Tiv, Prafulla C. Gokhale, Antoine de Weck, Ali Amin Mansour, Coyin Oh, Juliann Shih, Kevin Hadi, Yanay Rosen, Jonathan Bistline, Kavitha Venkatesan, Anupama Reddy, Dmitriy Sonkin, Manway Liu, Joseph Lehar, Joshua M. Korn, Dale A. Porter, Michael D. Jones, Javad Golji, Giordano Caponigro, Jordan E. Taylor, Caitlin M. Dunning, Amanda L. Creech, Allison C. Warren, James M. McFarland, Mahdi Zamanighomi, Audrey Kauffmann, Nicolas Stransky, Marcin Imielinski, Yosef E. Maruvka, Andrew D. Cherniack, Aviad Tsherniak, Francisca Vazquez, Jacob D. Jaffe, Andrew A. Lane, David M. Weinstock, Cory M. Johannessen, Michael P. Morrissey, Frank Stegmeier, Robert Schlegel, William C. Hahn, Gad Getz, Gordon B. Mills, Jesse S. Boehm, Todd R. Golub, Levi A. Garraway, and William R. Sellers. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature*, 569(7757):503–508, 2019.

[12] Amrita Basu, Nicole E. Bodycombe, Jaime H. Cheah, Edmund V. Price, Ke Liu, Giannina I. Schaefer, Richard Y. Ebright, Michelle L. Stewart, Daisuke Ito, Stephanie Wang, Abigail L. Bracha, Ted Liefeld, Mathias Wawer, Joshua C. Gilbert, Andrew J. Wilson, Nicolas Stransky, Gregory V. Kryukov, Vlado Dancik, Jordi Barretina, Levi A. Garraway, C. Suk Yee Hon, Benito Munoz, Joshua A. Bittker, Brent R. Stockwell, Dineo Khabele, Andrew M. Stern, Paul A. Clemons, Alykhan F. Shamji, and Stuart L. Schreiber. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell*, 154(5):1151–1161, 2013.

[13] Brinton Seashore-Ludlow, Matthew G. Rees, Jaime H. Cheah, Murat Coko, Edmund V. Price, Matthew E. Coletti, Victor Jones, Nicole E. Bodycombe, Christian K. Soule, Joshua Gould, Benjamin Alexander, Ava Li, Philip Montgomery, Mathias J. Wawer, Nurdan Kuru, Joanne D. Kotz, C. Suk-Yee Hon, Benito Munoz, Ted Liefeld, Vlado Dančik, Joshua A. Bittker, Michelle Palmer, James E. Bradner, Alykhan F. Shamji, Paul A. Clemons, and Stuart L. Schreiber. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer Discovery*, 5(11):1210–1223, 2015.

[14] Yutaka Shimada, Masato Maeda, Go Watanabe, Seiji Yamasaki, Izumi Komoto, Junichi Kaganoi, Takatsugu Kan, Yosuke Hashimoto, Issei Imoto, Johji Inazawa, and Masayuki Imamura. Cell culture in esophageal squamous cell carcinoma and the association with molecular markers. *Clinical Cancer Research*, 9(1 I):243–249, 2003.

[15] Uri Ben-David, Benjamin Siranosian, Gavin Ha, Helen Tang, Yaara Oren, Kunihiko Hinohara, Craig A. Strathdee, Joshua Dempster, Nicholas J. Lyons, Robert Burns,

Anwesha Nag, Guillaume Kugener, Beth Cimini, Peter Tsvetkov, Yosef E. Maruvka, Ryan O'Rourke, Anthony Garrity, Andrew A. Tubelli, Pratiti Bandopadhayay, Aviad Tsherniak, Francisca Vazquez, Bang Wong, Chet Birger, Mahmoud Ghandi, Aaron R. Thorner, Joshua A. Bittker, Matthew Meyerson, Gad Getz, Rameen Beroukhim, and Todd R. Golub. Genetic and transcriptional evolution alters cancer cell line drug response. *Nature*, 560(7718):325–330, 2018.

[16] Krijn K. Dijkstra, Chiara M. Cattaneo, Fleur Weeber, Myriam Chalabi, Joris van de Haar, Lorenzo F. Fanchi, Maarten Slagter, Daphne L. van der Velden, Sovann Kaing, Sander Kelderman, Nienke van Rooij, Monique E. van Leerdam, Annekatrien Depla, Egbert F. Smit, Koen J. Hartemink, Rosa de Groot, Monika C. Wolkers, Norman Sachs, Petur Snaebjornsson, Kim Monkhorst, John Haanen, Hans Clevers, Ton N. Schumacher, and Emile E. Voest. Generation of Tumor-Reactive T Cells by Co-culture of Peripheral Blood Lymphocytes and Tumor Organoids. *Cell*, 174(6):1586–1598.e12, 2018.

[17] Uri Ben-David, Gavin Ha, Yuen Yi Tseng, Noah F. Greenwald, Coyin Oh, Juliann Shih, James M. McFarland, Bang Wong, Jesse S. Boehm, Rameen Beroukhim, and Todd R. Golub. Patient-derived xenografts undergo mouse-specific tumor evolution. *Nature Genetics*, 49(11):1567–1575, 2017.

[18] Vivien Veninga and Emile E. Voest. Tumor organoids: Opportunities and challenges to guide precision medicine. *Cancer Cell*, 39(9):1190–1201, 2021.

[19] Prior Ian and Hancock John. Ras trafficking, localization and compartmentalized signalling Ian. *Semin. Cell. Dev. Biol.*, 23(2):145–143, 2012.

[20] Nicolas Nassar, Gudrun Horn, Christian Herrmann, Anna Scherer, Frank McCormick, and Alfred Wittinghofer. Nassar - the 2 2 A crystal structure of the RAS binding domain of the serine threonine kinase cRaf1 in complex with Rap1A and a GTP analogue.pdf. *Nature*, 375:554–560, 1995.

[21] Lufen Chang and Michael Karin. Mammalian MAP kinase signaling cascades. *Nature*, 410(6824):37–40, 2001.

[22] John Brognard and Tony Hunter. Protein kinase signaling networks in cancer. *Current Opinion in Genetics and Development*, 21(1):4–11, 2011.

[23] Michael E. Pacold, Sabine Suire, Olga Perisic, Samuel Lara-Gonzalez, Colin T. Davis, Edward H. Walker, Phillip T. Hawkins, Len Stephens, John F. Eccleston, and Roger L. Williams. Crystal structure and functional analysis of Ras binding to its effector phosphoinositide 3-kinase γ. *Cell*, 103(6):931–944, 2000.

[24] Brendan D. Manning and Lewis C. Cantley. AKT/PKB Signaling: Navigating Downstream. *Cell*, 129(7):1261–1274, 2007.

[25] Reuben Shaw and Lewis C. Cantley. Ras, PI(3)K and mTOR signalling controls tumour cell growth. *Nature*, 441:424–430, 2006.

[26] Arne Östman and Frank Böhmer. Regulation of receptor tyrosine kinase signaling by protein tyrosine phosphatase. *Trends in Cell Biology*, 11(6):258–266, 2001.

[27] Hong Joo Kim and Dafna Bar-Sagi. Modulation of signalling by sprouty: A developing story. *Nature Reviews Molecular Cell Biology*, 5(6):441–450, 2004.

[28] Yosef Yarden and Mark X Sliwkowski. Untangling the ErbB network. *Nature reviews Molecular Cell Biology*, 2(February):127–137, 2001.

[29] Peter J. O'Donovan and David M. Livingston. BRCA1 and BRCA2: Breast/ovarian cancer susceptibility gene products and participants in DNA double-strand break repair. *Carcinogenesis*, 31(6):961–967, 2010.

[30] Sreenath V. Sharma, Daphne W. Bell, Jeffrey Settleman, and Daniel A. Haber. Epidermal growth factor receptor mutations in lung cancer. *Nature Reviews Cancer*, 7(3):169–181, 2007.

[31] G. C. Burmer and L. A. Loeb. Mutations in the KRAS2 oncogene during progressive stages of human colon carcinoma. *Proceedings of the National Academy of Sciences of the United States of America*, 86(7):2403–2407, 1989.

[32] Sérgia Velho, Carla Oliveira, Ana Ferreira, António Carlos Ferreira, Gianpaolo Suriano, Simó Schwartz, Alex Duval, Fátima Carneiro, José Carlos Machado, Richard Hamelin, and Raquel Seruca. The prevalence of PIK3CA mutations in gastric and colon cancer. *European Journal of Cancer*, 41(11):1649–1654, 2005.

[33] Hiromasa Yamamoto, Hisayuki Shigematsu, Masaharu Nomura, William W. Lockwood, Mitsuo Sato, Naoki Okumura, Junichi Soh, Makoto Suzuki, Ignacio I. Wistuba, Kwun M. Fong, Huei Lee, Shinichi Toyooka, Hiroshi Date, Wan L. Lam, John D. Minna, and Adi F. Gazdar. PIK3CA mutations and copy number gains in human lung cancers. *Cancer Research*, 68(17):6913–6921, 2008.

[34] Kevin Kalinsky, Lindsay M. Jacks, Adriana Heguy, Sujata Patil, Marija Drobnjak, Umeshkumar K. Bhanot, Cyrus V. Hedvat, Tiffany A. Traina, David Solit, William Gerald, and Mary Ellen Moynahan. PIK3CA mutation associates with improved outcome in breast cancer. *Clinical Cancer Research*, 15(16):5049–5059, 2009.

[35] Marcia S. Brose, Patricia Volpe, Michael Feldman, Madhu Kumar, Irum Rishi, Renee Gerrero, Eugene Einhorn, Meenhard Herlyn, John Minna, Andrew Nicholson, Jack A. Roth, Steven M. Albelda, Helen Davies, Charles Cox, Graham Brignell, Philip Stephens, P. Andrew Futreal, Richard Wooster, Michael R. Stratton, and Barbara L. Weber. BRAF and RAS mutations in human lung cancer and melanoma. *Cancer Research*, 62(23):6997–7000, 2002.

[36] Karen H. Vousden and Xin Lu. Live or let die: The cell's response to p53. *Nature Reviews Cancer*, 2(8):594–604, 2002.

[37] Alfonso Quintás-Cardama and Jorge Cortes. Molecular biology of bcr-abl1-positive chronic myeloid leukemia. *Blood*, 113(8):1619–1630, 2009.

[38] Oriol Pich, Albert Cortes-Bullich, Ferran Muiños, Marta Pratcorona, Abel Gonzalez-Perez, and Nuria Lopez-Bigas. The evolution of hematopoietic cells under cancer therapy. *Nature Communications*, 12(1):1–11, 2021.

[39] David M. Hyman, Barry S. Taylor, and José Baselga. Implementing Genome-Driven Oncology. *Cell*, 168(4):584–599, 2017.

[40] Daniel J Vis, Lorenzo Bombardelli, Howard Lightfoot, Francesco Iorio, Mathew J Garnett, and Lodewyk FA Wessels. Pharmacogenomics Multilevel models improve precision and. *Pharmacogenomics*, 17:691–700, 2016.

[41] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning - Second Edition*. 2008.

[42] Christopher Bishop. *Pattern Recognition and Machine Learning*. 2006.

[43] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[44] Wouter Kouw and Marco Loog. A review of domain adaptation without target labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(c):1–1, 2019.

[45] In Sock Jang, Elias Chaibub Neto, Justin Guinney, Stephen H Friend, and Adam A Margolin. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. In *Pacific Symposium for Biocomputing*, volume 23, pages 1–7, 2013.

[46] Nanne Aben, Daniel J. Vis, Magali Michaut, and Lodewyk F.A. Wessels. TANDEM: A two-stage approach to maximize interpretability of drug response models based on multiple molecular data types. *Bioinformatics*, 32(17):i413–i420, 2016.

[47] Muhammad Ammad-Ud-Din, Suleiman A. Khan, Disha Malani, Astrid Murumägi, Olli Kallioniemi, Tero Aittokallio, and Samuel Kaski. Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. *Bioinformatics*, 32(17):i455–i463, 2016.

[48] Xiao He, Lukas Folkman, and Karsten Borgwardt. Kernelized rank learning for personalized drug recommendation. *Bioinformatics*, 34(16):2808–2816, 2018.

[49] Chayaporn Suphavilai, Denis Bertrand, and Niranjan Nagarajan. Predicting Cancer Drug Response using a Recommender System. *Bioinformatics*, 34(22):3907–3914, 2018.

[50] Betül Güvenç Paltun, Hiroshi Mamitsuka, and Samuel Kaski. Improving drug response prediction by integrating multiple data sources : matrix factorization , kernel and network-based approaches. 00(July):1–14, 2019.

[51] Ladislav Rampášek, Daniel Hidru, Petr Smirnov, Benjamin Haibe-Kains, and Anna Goldenberg. Dr.VAE: Improving drug response prediction via modeling of drug perturbation effects. *Bioinformatics*, 35(19):3743–3751, 2019.

[52] Paul Geeleher, Nancy J. Cox, and R. Stephanie Huang. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biology*, 15(3):1–12, 2014.

[53] Paul Geeleher, Zhenyu Zhang, Fan Wang, Robert F. Gruener, Aritro Nath, Gladys Morrison, Steven Bhutra, Robert L. Grossman, and R. Stephanie Huang. Discovering novel pharmacogenomic biomarkers by imputing drug response in cancer patients from large genomics studies. *Genome Research*, 27(10):1743–1751, 2017.

[54] Nanne Aben, Johan A. Westerhuis, Yipeng Song, Henk A.L. Kiers, Magali Michaut, Age K. Smilde, and Lodewyk F.A. Wessels. ITOP: Inferring the topology of omics data. *Bioinformatics*, 34(17):i988–i996, 2018.

[55] Joshua M Dempster, John M Krill-burger, James M Mcfarland, Allison Warren, S Jesse, Francisca Vazquez, William C Hahn, Todd R Golub, and Aviad Tsherniak. Gene expression has more power for predicting in vitro cancer cell vulnerabilities than genomics. *bioRxiv*, 1:1–42, 2020.

[56] Eric F. Lock, Katherine A. Hoadley, J. S. Marron, and Andrew B. Nobel. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Annals of Applied Statistics*, 7(1):523–542, 2013.

[57] Qing Feng, Meilei Jiang, Jan Hannig, and J. S. Marron. Angle-based joint and individual variation explained. *Journal of Multivariate Analysis*, 166:241–265, 2018.

[58] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2066–2073, 2012.

[59] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. *Proceedings of the IEEE International Conference on Computer Vision*, pages 2960–2967, 2013.

[60] Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Unsupervised Adaptation Across Domain Shifts by Generating Intermediate Data Representations. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2288–2302, 2014.

[61] Iñigo Martincorena and Peter J. Campbell. Somatic mutation in cancer and normal cells. *Science*, 349(6255):1483–1489, 2015.

[62] Howard L. McLeod. Cancer pharmacogenomics: Early promise, but concerted effort needed. *Science*, 340(6127):1563–1566, 2013.

[63] Mary V. Relling and William E. Evans. Pharmacogenomics in the clinic. *Nature*, 526(7573):343–350, 2015.

[64] Matteo Bersanelli, Ettore Mosca, Daniel Remondini, Enrico Giampieri, Claudia Sala, Gastone Castellani, and Luciano Milanesi. Methods for the integration of multi-omics data: Mathematical aspects. *BMC Bioinformatics*, 17(2), 2016.

[65] Vessela N Kristensen, Ole Christian Lingjærde, Hege G Russnes, Hans Kristian M. Vollan, Arnoldo Frigessi, and Anne-Lise Børresen-Dale. Kristensen - Principles and methods of integrative genomic analyses in cancer.pdf. *Nature Reviews Cancer*, 14:299–313, 2014.

[66] James C. Costello, Laura M. Heiser, Elisabeth Georgii, Mehmet Gönen, Michael P. Menden, Nicholas J. Wang, Mukesh Bansal, Muhammad Ammad-Ud-Din, Petteri Hintsanen, Suleiman A. Khan, John Patrick Mpindi, Olli Kallioniemi, Antti Honkela, Tero Aittokallio, Krister Wennerberg, James J. Collins, Dan Gallahan, Dinah Singer, Julio Saez-Rodriguez, Samuel Kaski, Joe W. Gray, Gustavo Stolovitzky, Jean Paul Abbuehl, Jeffrey Allen, Russ B. Altman, Shawn Balcome, Alexis Battle, Andreas Bender, Bonnie Berger, Jonathan Bernard, Madhuchhanda Bhattacharjee, Krithika Bhuvaneshwar, Andrew A. Bieberich, Fred Boehm, Andrea Califano, Christina Chan, Beibei Chen, Ting Huei Chen, Jaejoon Choi, Luis Pedro Coelho, Thomas Cokelaer, James C. Collins, Chad J. Creighton, Jike Cui, Will Dampier, V. Jo Davisson, Bernard De Baets, Raamesh Deshpande, Barbara DiCamillo, Murat Dundar, Zhana Duren, Adam Ertel, Haoyang Fan, Hongbin Fang, Robinder Gauba, Assaf Gottlieb, Michael Grau, Yuriy Gusev, Min Jin Ha, Leng Han, Michael Harris, Nicholas Henderson, Hussein A. Hejase, Krisztian Homicsko, Jack P. Hou, Woochang Hwang, Adriaan P. IJzerman, Bilge Karacali, Sunduz Keles, Christina Kendziorski, Junho Kim, Min Kim, Youngchul Kim, David A. Knowles, Daphne Koller, Junehawk Lee, Jae K. Lee, Eelke B. Lenselink, Biao Li, Bin Li, Jun Li, Han Liang, Jian Ma, Subha Madhavan, Sean Mooney, Chad L. Myers, Michael A. Newton, John P. Overington, Ranadip Pal, Jian Peng, Richard Pestell, Robert J. Prill, Peng Qiu, Bartek Rajwa, Anguraj Sadanandam, Francesco Sambo, Hyunjin Shin, Jiuzhou Song, Lei Song, Arvind Sridhar, Michiel Stock, Wei Sun, Tram Ta, Mahlet Tadesse, Ming Tan, Hao Tang, Dan Theodorescu, Gianna Maria Toffolo, Aydin Tozeren, William Trepicchio, Nelle Varoquaux, Jean Philippe Vert, Willem Waegeman, Thomas Walter, Qian Wan, Difei Wang, Wen Wang, Yong Wang, Zhishi Wang, Joerg K. Wegner, Tongtong Wu, Tian Xia, Guanghua Xiao, Yang Xie, Yanxun Xu, Jichen Yang, Yuan Yuan, Shihua Zhang, Xiang Sun Zhang, Junfei Zhao, Chandler Zuo, Herman W.T. Van Vlijmen, and Gerard J.P. Van Westen. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology*, 32(12):1202–1212, 2014.

[67] Matthew G. Rees, Brinton Seashore-Ludlow, Jaime H. Cheah, Drew J. Adams, Edmund V. Price, Shubhroz Gill, Sarah Javaid, Matthew E. Coletti, Victor L. Jones, Nicole E. Bodycombe, Christian K. Soule, Benjamin Alexander, Ava Li, Philip Montgomery, Joanne D. Kotz, C. Suk Yee Hon, Benito Munoz, Ted Liefeld, Vlado Dancík, Daniel A. Haber, Clary B. Clish, Joshua A. Bittker, Michelle Palmer, Bridget K. Wagner, Paul A. Clemons, Alykhan F. Shamji, and Stuart L. Schreiber. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nature Chemical Biology*, 12(2):109–116, 2016.

[68] Yifeng Li, Fang Xiang Wu, and Alioune Ngom. A review on machine learning principles for multi-view biological data integration. *Briefings in bioinformatics*, 19(2):325–340, 2018.

[69] Laura Cantini, Pooya Zakeri, Celine Hernandez, Aurelien Naldi, Denis Thieffry, Elisabeth Remy, and Anaïs Baudot. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nature Communications*, 12(1):1–12, 2021.

[70] Ricard Argelaguet, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6):1–13, 2018.

[71] Yongsoo Kim, Tycho Bismeijer, Wilbert Zwart, Lodewyk F A Wessels, and J Daniel. WON-PARAFAC : a genomic data integration method to identify interpretable factors for predicting drug-sensitivity in-vivo. pages 1–30.

[72] Hossein Sharifi-Noghabi, Olga Zolotareva, Colin C. Collins, and Martin Ester. MOLI: Multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics*, 35(14):i501–i509, 2019.

[73] Bo Wang, Aziz M. Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11(3):333–337, 2014.

[74] Christos Sagonas, Yannis Panagakis, Alina Leidinger, and Stefanos Zafeiriou. Robust joint and individual variance explained. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:5739–5748, 2017.

[75] Hai Shu, Xiao Wang, and Hongtu Zhu. D-CCA: A Decomposition-Based Canonical Correlation Analysis for High-Dimensional Datasets. *Journal of the American Statistical Association*, 115(529):292–306, 2020.

[76] Neil Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, 6:1783–1816, 2005.

[77] Michael E Tipping and Christopher M. Bishop. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society, Series B*, 61(3):611–622, 1999.

[78] E. J.G. Pitman. Sufficient Statistics and Intrinsic Accuracy. *Mathematical Proceedings of the Cambridge Philosophical Society*, 32(4):567–579, 1936.

[79] Michael Collins, Sanjoy Dasgupta, and Robert E. Schapire. A generalization of principal component analysis to the exponential family. *Advances in Neural Information Processing Systems*, (1), 2002.

[80] Jun Li and Dacheng Tao. Simple exponential family PCA. *IEEE Transactions on Neural Networks and Learning Systems*, 24(3):485–497, 2013.

[81] Lydia T. Liu, Edgar Dobriban, and Amit Singer. ePCA: High dimensional exponential family PCA. *Annals of Applied Statistics*, 12(4):2121–2150, 2018.

[82] Andrew J. Landgraf and Yoonkyung Lee. Dimensionality reduction for binary data through the projection of natural parameters. *Journal of Multivariate Analysis*, 180(1999), 2020.

[83] Adam Paske, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in prose. In *NeurIPS*, 2017.

[84] Mayank Meghwanshi, Pratik Jawanpuria, Anoop Kunchukuttan, Hiroyuki Kasai, and Bamdev Mishra. McTorch, a manifold optimization library for deep learning. pages 1–5, 2018.

[85] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[86] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net Hui. *Journal of the Statistical Society, Series B*, 67(2):301–320, 2005.

[87] Aaron M. Smith, Jonathan R. Walsh, John Long, Craig B. Davis, Peter Henstock, Martin R. Hodge, Mateusz Maciejewski, Xinmeng Jasmine Mu, Stephen Ra, Shanrong Zhao, Daniel Ziemek, and Charles K. Fisher. Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. *BMC Bioinformatics*, 21(1):1–18, 2020.

[88] Marlous Hoogstraat, Mirjam S. De Pagter, Geert A. Cirkel, Markus J. Van Roosmalen, Timothy T. Harkins, Karen Duran, Jennifer Kreeftmeijer, Ivo Renkens, Petronella O. Witteveen, Clarence C. Lee, Isaac J. Nijman, Tanisha Guy, Ruben Van't Slot, Trudy N. Jonges, Martijn P. Lolkema, Marco J. Koudijs, Ronald P. Zweemer, Emile E. Voest, Edwin Cuppen, and Wigard P. Kloosterman. Genomic and transcriptomic plasticity in treatment-naïve ovarian cancer. *Genome Research*, 24(2):200–211, 2014.

[89] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2009.

[90] Karin Hientz, André Mohr, Dipita Bhakta-Guha, and Thomas Efferth. The role of p53 in cancer drug resistance and targeted chemotherapy. *Oncotarget*, 8(5):8921–8946, 2017.

[91] Seung Tae Kim, Do Hyoung Lim, Kee Taek Jang, Taekyu Lim, Jeeyun Lee, Yoon La Choi, Hye Lim Jang, Jun Ho Yi, Kyung Kee Baek, Se Hoon Park, Young Suk Park, Ho Yeong Lim, Won Ki Kang, and Joon Oh Park. Impact of KRAS mutations on clinical outcomes in pancreatic cancer patients treated with first-line gemcitabine-based chemotherapy. *Molecular Cancer Therapeutics*, 10(10):1993–1999, 2011.

[92] Ella A Eklund, Clotilde Wiel, Henrik Fagman, Levent M Akyürek, Jan Nyman, Andreas Hallqvist, and Volkan I Sayin. KRAS mutations impact clinical outcome in metastatic non-small cell lung cancer Department of Surgery , Institute of Clinical Sciences , Sahlgrenska Center for Cancer Research , University of Gothenburg , Gothenburg , Sweden ; 2 Wallenberg Centre for Mole. 2022.

[93] Hui Gao, Joshua M. Korn, Stéphane Ferretti, John E. Monahan, Youzhen Wang, Mallika Singh, Chao Zhang, Christian Schnell, Guizhi Yang, Yun Zhang, O. Alejandro Balbin, Stéphanie Barbe, Hongbo Cai, Fergal Casey, Susmita Chatterjee, Derek Y. Chiang, Shannon Chuai, Shawn M. Cogan, Scott D. Collins, Ernesta Dammassa, Nicolas Ebel, Millicent Embry, John Green, Audrey Kauffmann, Colleen Kowal, Rebecca J. Leary, Joseph Lehar, Ying Liang, Alice Loo, Edward Lorenzana, E. Robert McDonald, Margaret E. McLaughlin, Jason Merkin, Ronald Meyer, Tara L. Naylor, Montesa Patawaran, Anupama Reddy, Claudia Röelli, David A. Ruddy, Fernando Salangsang, Francesca Santacroce, Angad P. Singh, Yan Tang, Walter Tinetto, Sonja Tobler, Roberto Velazquez, Kavitha Venkatesan, Fabian Von Arx, Hui Qin Wang, Zongyao Wang, Marion Wiesmann, Daniel Wyss, Fiona Xu, Hans Bitter, Peter Atadja, Emma Lees, Francesco Hofmann, En Li, Nicholas Keen, Robert Cozens, Michael Rugaard Jensen, Nancy K. Pryer, Juliet A. Williams, and William R. Sellers. High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nature Medicine*, 21(11):1318–1325, 2015.

[94] Jean Pierre Gillet, Sudhir Varma, and Michael M. Gottesman. The clinical relevance of cancer cell lines. *Journal of the National Cancer Institute*, 105(7):452–458, 2013.

[95] Gabriela Csurka. A comprehensive survey on domain adaptation for visual applications. In Gabriela Csurka, editor, *Advances in Computer Vision and Pattern Recognition*, number 9783319583464, pages 1–35. Springer Series, 2017.

[96] Le Song, Alex Smola, Karsten Borgwardt, and Arthur Gretton. Colored Maximum Variance Unfolding. *Nips*, pages 1–8, 2007.

[97] Sinno Jialin Pan, James T Kwok, and Qiang Yang. Transfer Learning via Dimensionality Reduction. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2008.

[98] Lixin Duan, Ivor W. Tsang, and Dong Xu. Domain transfer multiple kernel learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):465–479, 2012.

[99] Sinno Jialin Pan, Ivor W. Tsang, James T. Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.

[100] Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Domain adaptation for object recognition: An unsupervised approach. *Proceedings of the IEEE International Conference on Computer Vision*, pages 999–1006, 2011.

[101] Jianjiong Gao, Bülent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S. Onur Sumer, Yichao Sun, Anders Jacobsen, Rileen Sinha, Erik Larsson, Ethan Cerami, Chris Sander, and Nikolaus Schultz. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling*, 6(269):1–20, 2013.

[102] Xin Hu, Qianghu Wang, Ming Tang, Floris Barthel, Samirkumar Amin, Kosuke Yoshihara, Frederick M. Lang, Emmanuel Martinez-Ledesma, Soo Hyun Lee, Siyuan Zheng, and Roel G.W. Verhaak. TumorFusions: An integrative resource for cancer-associated transcript fusions. *Nucleic Acids Research*, 46(D1):D1144–D1149, 2018.

[103] L J P Van Der Maaten, E O Postma, and H J Van Den Herik. Dimensionality Reduction: A Comparative Review. *Journal of Machine Learning Research*, 10:1–41, 2009.

[104] Gene H Golub and Charles F Van Loan. *Matrix Computations*. 2013.

[105] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.

[106] G. Varoquaux, L. Buitinck, G. Louppe, O. Grisel, F. Pedregosa, and A. Mueller. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 19(1):29–33, 2015.

[107] Tycho Bismeijer, Sander Canisius, and Lodewyk F.A. Wessels. Molecular characterization of breast and lung tumors by integration of multiple data types with functional sparse-factor analysis. *PLoS Computational Biology*, 14(10):1–27, 2018.

[108] Rui Caseiro, João F. Henriques, Pedro Martins, and Jorge Batista. Beyond the shortest path: Unsupervised domain adaptation by Sampling Subspaces along the Spline Flow. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June:3846–3854, 2015.

[109] James T. Webber, Swati Kaushik, and Sourav Bandyopadhyay. Integration of Tumor Genomic Data with Cell Lines Using Multi-dimensional Network Modules Improves Cancer Pharmacogenomics. *Cell Systems*, 7(5):526–536.e6, 2018.

[110] Rachelly Normand, Wenfei Du, Mayan Briller, Renaud Gaujoux, Elina Starosvetsky, Amit Ziv-Kenet, Gali Shalev-Malul, Robert J. Tibshirani, and Shai S. Shen-Orr. Found In Translation: a machine learning model for mouse-to-human inference. *Nature Methods*, 15(12):1067–1073, 2018.

[111] Dennis J. Slamon, Brian Leyland-Jones, Steven Shak, Hank Fuchs, Virginia Paton, Alex Bajamonde, Thomas Fleming, Wolfgang Eiermann, Janet Wolter, Mark Pegram, and Jose Baselga. Use of Chemotherapy Plus a Monoclonal Antibody Against

Her2 for metastatic breast cancer that overexpressed Her2. *The New England Journal of Medicine*, 344(11):783–792, 2001.

[112] Helen Davies, Graham R. Bignell, Charles Cox, Philip Stephens, Sarah Edkins, Sheila Clegg, Jon Teague, Hayley Woffendin, Mathew J. Garnett, William Bottomley, Neil Davis, Ed Dicks, Rebecca Ewing, Yvonne Floyd, Kristian Gray, Sarah Hall, Rachel Hawes, Jaime Hughes, Vivian Kosmidou, Andrew Menzies, Catherine Mould, Adrian Parker, Claire Stevens, Stephen Watt, Steven Hooper, Hiran Jayatilake, Barry A. Gusterson, Colin Cooper, Janet Shipley, Darren Hargrave, Katherine Pritchard-Jones, Norman Maitland, Georgia Chenevix-Trench, Gregory J. Riggins, Darell D. Bigner, Giuseppe Palmieri, Antonio Cossu, Adrienne Flanagan, Andrew Nicholson, Judy W.C. Ho, Suet Y. Leung, Siu T. Yuen, Barbara L. Weber, Hilliard F. Seigler, Timothy L. Darrow, Hugh Paterson, Richard Wooster, Michael R. Stratton, and P. Andrew Futreal. Mutations of the BRAF gene in human cancer. *Nature*, 417(6892):949–954, 2002.

[113] Tim P. Hughes, Jaspal Kaeda, Susan Branford, Zbigniew Rudzki, Andreas Hochhaus, Martee L. Hensley, Insa Gathmann, Ann E. Bolton, Iris C. Van Hoomissen, John M. Goldman, and Jerald P. Radich. Frequency of major molecular responses to imatinib or interferon alfa plus cytarabine in newly diagnosed chronic myeloid leukemia. *New England Journal of Medicine*, 349(15):1423–1432, 2003.

[114] Angeliki Kalamara, Luis Tobalina, and Julio Saez-Rodriguez. How to find the right drug for each patient? Advances and challenges in pharmacogenomics. *Current Opinion in Systems Biology*, 10:53–62, 2018.

[115] Anirudh Prahallad, Chong Sun, Sidong Huang, Federica Di Nicolantonio, Ramon Salazar, Davide Zecchin, Roderick L. Beijersbergen, Alberto Bardelli, and René Bernards. Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. *Nature*, 483(7387):100–104, 2012.

[116] Paul Valery. *Mauvaises pensées et autres*. 1941.

[117] Jean Pierre Gillet, Anna Maria Calcagno, Sudhir Varma, Miguel Marino, Lisa J. Green, Meena I. Vora, Chirayu Patel, Josiah N. Orina, Tatiana A. Eliseeva, Vineet Singal, Raji Padmanabhan, Ben Davidson, Ram Ganapathi, Anil K. Sood, Bo R. Rueda, Suresh V. Ambudkar, and Michael M. Gottesman. Redefining the relevance of established cancer cell lines to the study of mechanisms of clinical anti-cancer drug resistance. *Proceedings of the National Academy of Sciences of the United States of America*, 108(46):18708–18713, 2011.

[118] Isabella W.Y. Mak, Nathan Evaniew, and Michelle Ghert. Lost in translation: Animal models and clinical trials in cancer treatment. *American Journal of Translational Research*, 6(2):114–118, 2014.

[119] Douglas K. Brubaker and Douglas A. Lauffenburger. Translating preclinical models to humans. *Science*, 367(6479):742–743, 2020.

[120] Mehreen Ali and Tero Aittokallio. Machine learning and feature selection for drug response prediction in precision oncology applications. *Biophysical Reviews*, 11(1):31–39, 2019.

[121] Roman Kurilov, Benjamin Haibe-Kains, and Benedikt Brors. Assessment of modelling strategies for drug response prediction in cell lines and xenografts. *Scientific reports*, 10(1):2849, 2020.

[122] Theodore Sakellaropoulos, Konstantinos Vougas, Sonali Narang, Filippos Koinis, Athanassios Kotsinas, Alexander Polyzos, Tyler J. Moss, Sarina Piha-Paul, Hua Zhou, Eleni Kardala, Eleni Damianidou, Leonidas G. Alexopoulos, Iannis Aifantis, Paul A. Townsend, Mihalis I. Panayiotidis, Petros Sfikakis, Jiri Bartek, Rebecca C. Fitzgerald, Dimitris Thanos, Kenna R. Mills Shaw, Russell Petty, Aristotelis Tsirigos, and Vassilis G. Gorgoulis. A Deep Learning Framework for Predicting Response to Therapy in Cancer. *Cell Reports*, 29(11):3367–3373.e4, 2019.

[123] Hossein Sharifi Noghabi, Shuman Peng, Olga Zolotareva, Colin C Collins, and Martin Ester. AITL: Adversarial Inductive Transfer Learning with input and output space adaptation for pharmacogenomics. *bioRxiv*, page 2020.01.24.918953, 2020.

[124] Jianzhu Ma, Samson H. Fong, Yunan Luo, Christopher J. Bakkenist, John Paul Shen, Soufiane Mourragui, Lodewyk F.A. Wessels, Marc Hafner, Roded Sharan, Jian Peng, and Trey Ideker. Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nature Cancer*, 2021.

[125] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Subspace Alignment For Domain Adaptation. 2014.

[126] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *Annals of Statistics*, 36(3):1171–1220, 2008.

[127] Jean-Philippe Vert, Koji Tsuda, and Bernhard Schölkopf. *Kernel methods in computational biology*. MIT Press, 2004.

[128] K. Yu, B. Chen, D. Aran, J. Charalel, C. Yau, D. M. Wolf, L. J. van 't Veer, A. J. Butte, T. Goldstein, and M. Sirota. Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types. *Nature Communications*, 10(1), 2019.

[129] Allison Warren, Yejia Chen, Andrew Jones, Tsukasa Shibue, William C. Hahn, Jesse S. Boehm, Francisca Vazquez, Aviad Tsherniak, and James M. McFarland. Global computational alignment of tumor and cell line transcriptional profiles. *Nature Communications*, 12(1):1–12, 2021.

[130] Da Peng, Rachel Gleyzer, Wen Hsin Tai, Pavithra Kumar, Qin Bian, Bradley Issacs, Edroaldo Lummertz da Rocha, Stephanie Cai, Kathleen DiNapoli, Franklin W. Huang, and Patrick Cahan. Evaluating the transcriptional fidelity of cancer models. *bioRxiv*, 2020.

[131] Soufiane Mourragui, Marco Loog, Mark A van de Wiel, Marcel J T Reinders, and Lodewyk F A Wessels. PRECISE: a domain adaptation approach to transfer predictors of drug response from pre-clinical models to tumors. *Bioinformatics*, 35(14):i510–i519, jul 2019.

[132] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive Integration of Single-Cell Data. *Cell*, 177(7):1888–1902.e21, 2019.

[133] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018.

[134] E. Charafe-Jauffret, C. Ginestier, F. Monville, P. Finetti, J. Adélaïde, N. Cervera, S. Fekairi, L. Xerri, J. Jacquemier, D. Birnbaum, and F. Bertucci. Gene expression profiling of breast cell lines identifies potential new basal markers. *Oncogene*, (25):2273–2284, 2006.

[135] Chien Hui Weng, Li Yu Chen, Yu Chin Lin, Jin Yuan Shih, Yun Chieh Lin, Ruo Yu Tseng, An Chieh Chiu, Yu Hsuan Yeh, Chi Liu, Yi Ting Lin, Jim Min Fang, and Ching Chow Chen. Epithelial-mesenchymal transition (EMT) beyond EGFR mutations per se is a common mechanism for acquired resistance to EGFR TKI. *Oncogene*, 38(4):455–468, 2019.

[136] Bruce Cuevas, Yiling Lu, Steven Watt, Rakesh Kumar, Jinyi Zhang, Katherine A. Siminovitch, and Gordon B. Mills. SHP-1 regulates Lck-induced phosphatidylinositol 3-kinase phosphorylation and activity. *Journal of Biological Chemistry*, 274(39):27583–27589, 1999.

[137] F. J. Rodríguez-Ubreva, A. E. Cariaga-Martinez, M. A. Cortés, M. Romero-De Pablos, S. Ropero, P. López-Ruiz, and B. Colás. Knockdown of protein tyrosine phosphatase SHP-1 inhibits G1/S progression in prostate cancer cells through the regulation of components of the cell-cycle machinery. *Oncogene*, 29(3):345–355, 2010.

[138] Qian Liu, Shengnan Yu, Weiheng Zhao, Shuang Qin, Qian Chu, and Kongming Wu. EGFR-TKIs resistance via EGFR-independent signaling pathways. *Molecular Cancer*, 17(1):1–9, 2018.

[139] Julien Cau and Alan Hall. Cdc42 controls the polarity of the actin and microtubule cytoskeletons through two distinct signal transduction pathways. *Journal of Cell Science*, 118(12):2579–2587, 2005.

[140] Yuna Guo, S. Ray Kenney, Carolyn Y. Muller, Sarah Adams, Teresa Rutledge, Elsa Romero, Cristina Murray-Krezan, Rytis Prekeris, Larry A. Sklar, Laurie G. Hudson, and Angela Wandinger-Ness. R-ketorolac targets Cdc42 and Rac1 and alters ovarian cancer cell behaviors critical for invasion and metastasis. *Molecular Cancer Therapeutics*, 14(10):2215–2227, 2015.

[141] María Del Mar Maldonado and Suranganie Dharmawardhane. Targeting rac and Cdc42 GT pases in cancer. *Cancer Research*, 78(12):3101–3111, 2018.

[142] S. R. Murugesan, C. R. King, R. Osborn, W. R. Fairweather, E. M. O'Reilly, M. O. Thornton, and L. L. Wei. Combination of human tumor necrosis factor-alpha (hTNF-α) gene delivery with gemcitabine is effective in models of pancreatic cancer. *Cancer Gene Therapy*, 16(11):841–847, 2009.

[143] Y. Basaki, F. Hosoi, Y. Oda, A. Fotovati, Y. Maruyama, S. Oie, M. Ono, H. Izumi, K. Kohno, K. Sakai, T. Shimoyama, K. Nishio, and M. Kuwano. Akt-dependent nuclear localization of Y-box-binding protein 1 in acquisition of malignant characteristics by human ovarian cancer cells. *Oncogene*, 26(19):2736–2746, 2007.

[144] Björn C. Frye, Sarah Halfter, Sonja Djudjaj, Philipp Muehlenberg, Susanne Weber, Ute Raffetseder, Abdelaziz En-Nia, Hanna Knott, Jens M. Baron, Steven Dooley, Jürgen Bernhagen, and Peter R. Mertens. Y-box protein-1 is actively secreted through a non-classical pathway and acts as an extracellular mitogen. *EMBO Reports*, 10(7):783–789, 2009.

[145] Genevieve Housman, Shannon Byler, Sarah Heerboth, Karolina Lapinska, McKenna Longacre, Nicole Snyder, and Sibaji Sarkar. Drug resistance in cancer: An overview. *Cancers*, 6(3):1769–1792, 2014.

[146] L. J. Goldstein. MDR1 gene expression in solid tumours. *European Journal of Cancer*, 32(6):1039–1050, 1996.

[147] Aparajitha Vaidyanathan, Lynne Sawers, Anne Louise Gannon, Probir Chakravarty, Alison L. Scott, Susan E. Bray, Michelle J. Ferguson, and Gillian Smith. ABCB1 (MDR1) induction defines a common resistance mechanism in paclitaxel- and olaparib-resistant ovarian cancer cells. *British Journal of Cancer*, 115(4):431–441, 2016.

[148] Elizabeth L. Christie, Swetansu Pattnaik, Jessica Beach, Anthony Copeland, Nineveh Rashoo, Sian Fereday, Joy Hendley, Kathryn Alsop, Samuel L. Brady, Greg Lamb, Ahwan Pandey, Anna DeFazio, Heather Thorne, Andrea Bild, and David D.L. Bowtell. Multiple ABCB1 transcriptional fusions in drug resistant high-grade serous ovarian and breast cancer. *Nature Communications*, 10(1):5–14, 2019.

[149] Dongshao Chen, Xiaoting Lin, Cheng Zhang, Zhentao Liu, Zuhua Chen, Zhongwu Li, Jingyuan Wang, Beifang Li, Yanting Hu, Bin Dong, Lin Shen, Jiafu Ji, Jing Gao, and Xiaotian Zhang. Dual PI3K/mTOR inhibitor BEZ235 as a promising therapeutic strategy against paclitaxel-resistant gastric cancer via targeting PI3K/Akt/mTOR pathway article. *Cell Death and Disease*, 9(2), 2018.

[150] Limin Hu, Judith Hofmann, Yiling Lu, Gordon B. Mills, and Robert B. Jaffe. Inhibition of phosphatidylinositol 3-kinase increases efficacy of paclitaxel in in vitro and in vivo ovarian cancer models. *Cancer Research*, 62(4):1087–1092, 2002.

[151] Lei Zhang, Yidong Li, Qianchao Wang, Zhuo Chen, Xiaoyun Li, Zhuoxun Wu, Chaohua Hu, Dan Liao, Wei Zhang, and Zhe Sheng Chen. The PI3K subunits, P110α and P110β are potential targets for overcoming P-gp and BCRP-mediated MDR in cancer. *Molecular Cancer*, 19(1):1–18, 2020.

[152] Yuebo Gan, M. Guillaume Wientjes, and Jessie L.S. Au. Expression of basic fibroblast growth factor correlates with resistance to paclitaxel in human patient tumors. *Pharmaceutical Research*, 23(6):1324–1331, 2006.

[153] Se Hyun Kim, Haram Ryu, Chan Young Ock, Koung Jin Suh, Ji Yun Lee, Ji Won Kim, Jeong Ok Lee, Jin Won Kim, Yu Jung Kim, Keun Wook Lee, Soo Mee Bang, Jee Hyun Kim, Jong Seok Lee, Joong Bae Ahn, Kui Jin Kim, and Sun Young Rha. BGJ398, a pan-FGFR inhibitor, overcomes paclitaxel resistance in urothelial carcinoma with FGFR1 overexpression. *International Journal of Molecular Sciences*, 19(10), 2018.

[154] Matteo Manica, Joris Cadow, Roland Mathis, and María Rodríguez Martínez. PIMKL: Pathway-Induced Multiple Kernel Learning. *npj Systems Biology and Applications*, 5(1), 2019.

[155] Xudong Lin, Yueqi Duan, Qiyuan Dong, Jiwen Lu, and Jie Zhou. Deep variational metric learning. In *ECCV*, pages 714–729, 2018.

[156] Ladislav Rampášek. *Latent-variable models for drug response prediction and genetic testing*. PhD thesis, 2020.

[157] Mark D. Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3), 2010.

[158] Marie Agnès Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, Nicolas Servant Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Laloë, Caroline Le Gall, Brigitte Schaëffer, Stéphane Le Crom, Mickaël Guedj, and Florence Jaffrézic. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, 14(6):671–683, 2013.

[159] Isabella Zwiener, Barbara Frisch, and Harald Binder. Transforming RNA-Seq data to improve the performance of prognostic gene signatures. *PLoS ONE*, 9(1):1–13, 2014.

[160] John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R.Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Chris Sander, Joshua M. Stuart, Kyle Chang, Chad J. Creighton, Caleb Davis, Lawrence Donehower, Jennifer Drummond, David Wheeler, Adrian Ally, Miruna Balasundaram, Inanc Birol, Yaron S.N. Butterfield, Andy Chu, Eric Chuah, Hye Jung E. Chun, Noreen Dhalla, Ranabir Guin, Martin Hirst, Carrie Hirst, Robert A. Holt, Steven J.M. Jones, Darlene Lee, Haiyan I. Li, Marco A. Marra, Michael Mayo, Richard A. Moore, Andrew J. Mungall, A. Gordon Robertson, Jacqueline E. Schein, Payal Sipahimalani, Angela Tam, Nina Thiessen, Richard J. Varhol, Rameen Beroukhim, Ami S. Bhatt, Angela N. Brooks, Andrew D.

Cherniack, Samuel S. Freeman, Stacey B. Gabriel, Elena Helman, Joonil Jung, Matthew Meyerson, Akinyemi I. Ojesina, Chandra Sekhar Pedamallu, Gordon Saksena, Steven E. Schumacher, Barbara Tabak, Travis Zack, Eric S. Lander, Christopher A. Bristow, Angela Hadjipanayis, Psalm Haseley, Raju Kucherlapati, Semin Lee, Eunjung Lee, Lovelace J. Luquette, Harshad S. Mahadeshwar, Angeliki Pantazi, Michael Parfenov, Peter J. Park, Alexei Protopopov, Xiaojia Ren, Netty Santoso, Jonathan Seidman, Sahil Seth, Xingzhi Song, Jiabin Tang, Ruibin Xi, Andrew W. Xu, Lixing Yang, Dong Zeng, J. Todd Auman, Saianand Balu, Elizabeth Buda, Cheng Fan, Katherine A. Hoadley, Corbin D. Jones, Shaowu Meng, Piotr A. Mieczkowski, Joel S. Parker, Charles M. Perou, Jeffrey Roach, Yan Shi, Grace O. Silva, Donghui Tan, Umadevi Veluvolu, Scot Waring, Matthew D. Wilkerson, Junyuan Wu, Wei Zhao, Tom Bodenheimer, D. Neil Hayes, Alan P. Hoyle, Stuart R. Jeffreys, Lisle E. Mose, Janae V. Simons, Mathew G. Soloway, Stephen B. Baylin, Benjamin P. Berman, Moiz S. Bootwalla, Ludmila Danilova, James G. Herman, Toshinori Hinoue, Peter W. Laird, Suhn K. Rhie, Hui Shen, Timothy Triche, Daniel J. Weisenberger, Scott L. Carter, Kristian Cibulskis, Lynda Chin, Jianhua Zhang, Carrie Sougnez, Min Wang, Gad Getz, Huyen Dinh, Harsha Vardhan Doddapaneni, Richard Gibbs, Preethi Gunaratne, Yi Han, Divya Kalra, Christie Kovar, Lora Lewis, Margaret Morgan, Donna Morton, Donna Muzny, Jeffrey Reid, Liu Xi, Juok Cho, Daniel Dicara, Scott Frazer, Nils Gehlenborg, David I. Heiman, Jaegil Kim, Michael S. Lawrence, Pei Lin, Yingchun Liu, Michael S. Noble, Petar Stojanov, Doug Voet, Hailei Zhang, Lihua Zou, Chip Stewart, Brady Bernard, Ryan Bressler, Andrea Eakin, Lisa Iype, Theo Knijnenburg, Roger Kramer, Richard Kreisberg, Kalle Leinonen, Jake Lin, Yuexin Liu, Michael Miller, Sheila M. Reynolds, Hector Rovira, Ilya Shmulevich, Vesteinn Thorsson, Da Yang, Wei Zhang, Samirkumar Amin, Chang Jiun Wu, Chia Chin Wu, Rehan Akbani, Kenneth Aldape, Keith A. Baggerly, Bradley Broom, Tod D. Casasent, James Cleland, Deepti Dodda, Mary Edgerton, Leng Han, Shelley M. Herbrich, Zhenlin Ju, Hoon Kim, Seth Lerner, Jun Li, Han Liang, Wenbin Liu, Philip L. Lorenzi, Yiling Lu, James Melott, Lam Nguyen, Xiaoping Su, Roeland Verhaak, Wenyi Wang, Andrew Wong, Yang Yang, Jun Yao, Rong Yao, Kosuke Yoshihara, Yuan Yuan, Alfred K. Yung, Nianxiang Zhang, Siyuan Zheng, Michael Ryan, David W. Kane, B. Arman Aksoy, Giovanni Ciriello, Gideon Dresdner, Jianjiong Gao, Benjamin Gross, Anders Jacobsen, Andre Kahles, Marc Ladanyi, William Lee, Kjong Van Lehmann, Martin L. Miller, Ricardo Ramirez, Gunnar Rätsch, Boris Reva, Nikolaus Schultz, Yasin Senbabaoglu, Ronglai Shen, Rileen Sinha, S. Onur Sumer, Yichao Sun, Barry S. Taylor, Nils Weinhold, Suzanne Fei, Paul Spellman, Christopher Benz, Daniel Carlin, Melisssa Cline, Brian Craft, Mary Goldman, David Haussler, Singer Ma, Sam Ng, Evan Paull, Amie Radenbaugh, Sofie Salama, Artem Sokolov, Teresa Swatloski, Vladislav Uzunangelov, Peter Waltman, Christina Yau, Jing Zhu, Stanley R. Hamilton, Scott Abbott, Rachel Abbott, Nathan D. Dees, Kim Delehaunty, Li Ding, David J. Dooling, Jim M. Eldred, Catrina C. Fronick, Robert Fulton, Lucinda L. Fulton, Joelle Kalicki-Veizer, Krishna Latha Kanchi, Cyriac Kandoth, Daniel C. Koboldt, David E. Larson, Timothy J. Ley, Ling Lin, Charles Lu, Vincent J. Magrini, Elaine R. Mardis, Michael D. McLellan, Joshua F. McMichael, Christopher A.

Miller, Michelle O'Laughlin, Craig Pohl, Heather Schmidt, Scott M. Smith, Jason Walker, John W. Wallis, Michael C. Wendl, Richard K. Wilson, Todd Wylie, Qunyuan Zhang, Robert Burton, Mark A. Jensen, Ari Kahn, Todd Pihl, David Pot, Yunhu Wan, Douglas A. Levine, Aaron D. Black, Jay Bowen, Jessica Frick, Julie M. Gastier-Foster, Hollie A. Harper, Carmen Helsel, Kristen M. Leraas, Tara M. Lichtenberg, Cynthia McAllister, Nilsa C. Ramirez, Samantha Sharpe, Lisa Wise, Erik Zmuda, Stephen J. Chanock, Tanja Davidsen, John A. Demchok, Greg Eley, Ina Felau, Margi Sheth, Heidi Sofia, Louis Staudt, Roy Tarnuzzer, Zhining Wang, Liming Yang, Jiashan Zhang, Larsson Omberg, Adam Margolin, Benjamin J. Raphael, Fabio Vandin, Hsin Ta Wu, Mark D.M. Leiserson, Stephen C. Benz, Charles J. Vaske, Houtan Noushmehr, Denise Wolf, Laura Van T. Veer, Dimitris Anastassiou, Tai Hsien Ou Yang, Nuria Lopez-Bigas, Abel Gonzalez-Perez, David Tamborero, Zheng Xia, Wei Li, Dong Yeon Cho, Teresa Przytycka, Mark Hamilton, Sean McGuire, Sven Nelander, Patrik Johansson, Rebecka Jörnsten, and Teresia Kling. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013.

[161] Zijian Ding, Songpeng Zu, and Jin Gu. Evaluating the molecule-based prediction of clinical drug responses in cancer. *Bioinformatics*, 32(19):2891–2895, 2016.

[162] Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048, 2016.

[163] Rob Patro, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4):417–419, 2017.

[164] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[165] Yang Liao, Gordon K. Smyth, and Wei Shi. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 2014.

[166] Patrick Therasse. tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. *J natl Cancer inst*, 92(3):205–16, 2000.

[167] Bernhard Schölkopf, Alexander J. Smola, and Klaus-Robert Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10:1299–1319, 1998.

[168] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010.

[169] Ingo Steinwart, Don Hush, and Clint Scovel. An Explicit Description of the Reproducing Kernel Hilbert Spaces of Gaussian RBF Kernels. *IEEE Transactions on Information Theory*, 52(10):4635–4643, 2006.

[170] Natacha Turck, Laszlo Vutskits, Paola Sanchez-Pena, Xavier Robin, Alexandre Hainard, Marianne Gex-Fabry, Catherine Fouda, Hadiji Bassem, Markus Mueller, Frédérique Lisacek, Louis Puybasset, and Jean-Charles Sanchez. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 8:12–77, 2011.

[171] Fiona M. Behan, Francesco Iorio, Gabriele Picco, Emanuel Gonçalves, Charlotte M. Beaver, Giorgia Migliardi, Rita Santos, Yanhua Rao, Francesco Sassi, Marika Pinnelli, Rizwan Ansari, Sarah Harper, David Adam Jackson, Rebecca McRae, Rachel Pooley, Piers Wilkinson, Dieudonne van der Meer, David Dow, Carolyn Buser-Doepner, Andrea Bertotti, Livio Trusolino, Euan A. Stronach, Julio Saez-Rodriguez, Kosuke Yusa, and Mathew J. Garnett. Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens. *Nature*, 568(7753):511–516, 2019.

[172] Piet Borst and Lodewyk F.A. Wessels. Borst - Do predictive signature really predict response to cancer chemotherapy.pdf. *Cell cycle*, 9(24):4836–4840, 2010.

[173] Soufiane M.C. Mourragui, Marco Loog, Daniel J. Vis, Kat Moore, Anna G. Manjon, Mark A. van de Wiel, Marcel J.T. Reinders, and Lodewyk F.A. Wessels. Predicting patient response with models trained on cell lines and patient-derived xenografts by nonlinear transfer learning. *Proceedings of the National Academy of Sciences of the United States of America*, 118(49):1–12, 2021.

[174] Da Peng, Rachel Gleyzer, Wen Hsin Tai, Pavithra Kumar, Qin Bian, Bradley Isaacs, Edroaldo Lummertz da Rocha, Stephanie Cai, Kathleen DiNapoli, Franklin W. Huang, and Patrick Cahan. Evaluating the transcriptional fidelity of cancer models. *Genome Medicine*, 13(1):1–27, 2021.

[175] Gabriela S. Kinker, Alissa C. Greenwald, Rotem Tal, Zhanna Orlova, Michael S. Cuoco, James M. McFarland, Allison Warren, Christopher Rodman, Jennifer A. Roth, Samantha A. Bender, Bhavna Kumar, James W. Rocco, Pedro A.C.M. Fernandes, Christopher C. Mader, Hadas Keren-Shaul, Alexander Plotnikov, Haim Barr, Aviad Tsherniak, Orit Rozenblatt-Rosen, Valery Krizhanovsky, Sidharth V. Puram, Aviv Regev, and Itay Tirosh. Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. *Nature Genetics*, 52(11):1208–1218, 2020.

[176] James M. McFarland, Brenton R. Paolella, Allison Warren, Kathryn Geiger-Schuller, Tsukasa Shibue, Michael Rothberg, Olena Kuksenko, William N. Colgan, Andrew Jones, Emily Chambers, Danielle Dionne, Samantha Bender, Brian M. Wolpin, Mahmoud Ghandi, Itay Tirosh, Orit Rozenblatt-Rosen, Jennifer A. Roth, Todd R. Golub, Aviv Regev, Andrew J. Aguirre, Francisca Vazquez, and Aviad Tsherniak. Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. *Nature Communications*, 11(1), 2020.

[177] Nayoung Kim, Hong Kwan Kim, Kyungjong Lee, Yourae Hong, Jong Ho Cho, Jung Won Choi, Jung Il Lee, Yeon Lim Suh, Bo Mi Ku, Hye Hyeon Eum, Soyean Choi, Yoon La Choi, Je Gun Joung, Woong Yang Park, Hyun Ae Jung, Jong Mu Sun, Se Hoon Lee, Jin Seok Ahn, Keunchil Park, Myung Ju Ahn, and Hae Ock Lee. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nature Communications*, 11(1), 2020.

[178] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, (Ml):1–14, 2014.

[179] Alessandro Rudi, Luigi Carratino, and Lorenzo Rosasco. FALKON: An optimal large scale kernel method. *Advances in Neural Information Processing Systems*, 2017-Decem:3889–3899, 2017.

[180] Giacomo Meanti, Luigi Carratino, Lorenzo Rosasco, and Alessandro Rudi. Kernel methods through the roof: handling billions of points efficiently. In *NeurIPS*, 2020.

[181] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean Philippe Vert. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications*, 9(1):1–17, 2018.

[182] Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018.

[183] Gökcen Eraslan, Lukas M. Simon, Maria Mircea, Nikola S. Mueller, and Fabian J. Theis. Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications*, 10(1):1–14, 2019.

[184] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–420, 2018.

[185] Laleh Haghverdi, Aaron T.L. Lun, Michael D. Morgan, and John C. Marioni. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427, 2018.

[186] Uri Shaham, Kelly P. Stanton, Jun Zhao, Huamin Li, Khadir Raddassi, Ruth Montgomery, and Yuval Kluger. Removal of batch effects using distribution-matching residual networks. *Bioinformatics*, 33(16):2539–2546, 2017.

[187] Joshua D. Welch, Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, and Evan Z. Macosko. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell*, 177(7):1873–1887.e17, 2019.

[188] Hoa Thi, Nhu Tran, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee, Shin Lee, Michelle Goh, and Jinmiao Chen. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biology*, pages 1–32, 2020.

[189] Felix Raimundo, Celine Vallot, and Jean Philippe Vert. Tuning parameters of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biology*, pages 1–17, 2020.

[190] Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12):1289–1296, 2019.

[191] Etienne Becht, Leland McInnes, John Healy, Charles Antoine Dutertre, Immanuel W.H. Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W. Newell. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, 37(1):38–47, 2019.

[192] Robrecht Cannoodt, Wouter Saelens, Louise Deconinck, and Yvan Saeys. Spearheading future omics analyses using dyngen, a multi-modal simulator of single cells. *Nature Communications*, 12(1):1–9, 2021.

[193] Bijay Jassal, Lisa Matthews, Guilherme Viteri, Chuqiao Gong, Pascual Lorente, Antonio Fabregat, Konstantinos Sidiropoulos, Justin Cook, Marc Gillespie, Robin Haw, Fred Loney, Bruce May, Marija Milacic, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Solomon Shorser, Thawfeek Varusai, Joel Weiser, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome pathway knowledgebase. *Nucleic Acids Research*, 48(D1):D498–D503, 2020.

[194] Massimo Andreatta and Santiago J. Carmona. UCell: Robust and scalable single-cell gene signature scoring. *Computational and Structural Biotechnology Journal*, 19:3796–3798, 2021.

[195] V. Karantza. Keratins in health and cancer: More than mere epithelial cell markers. *Oncogene*, 30(2):127–138, 2011.

[196] Lyubomir T. Vassilev, Binh T. Vu, Bradford Graves, Daisy Carvajal, Frank Podlaski, Zoran Filipovic, Norman Kong, Ursula Kammlott, Christine Lukacs, Christian Klein, Nader Fotouhi, and Emily A. Liu. In Vivo Activation of the p53 Pathway by Small-Molecule Antagonists of MDM2. *Science*, 303(5659):844–848, 2004.

[197] Mary Ann Bjornsti and Peter J. Houghton. The TOR pathway: A target for cancer therapy. *Nature Reviews Cancer*, 4(5):335–348, 2004.

[198] Steve Wagner, Georgios Vlachogiannis, Alexis De Haven Brandon, Melanie Valenti, Gary Box, Liam Jenkins, Caterina Mancusi, Annette Self, Floriana Manodoro, Ioannis Assiotis, Penny Robinson, Ritika Chauhan, Alistair G. Rust, Nik Matthews, Kate Eason, Khurum Khan, Naureen Starling, David Cunningham, Anguraj Sadanandam, Clare M. Isacke, Vladimir Kirkin, Nicola Valeri, and Steven R. Whittaker. Suppression of interferon gene expression overcomes resistance to MEK inhibition in KRAS-mutant colorectal cancer. *Oncogene*, 38(10):1717–1733, 2019.

[199] Ali Naderi, Kee Ming Chia, and Ji Liu. Synergy between inhibitors of androgen receptor and MEK has therapeutic implications in estrogen receptor-negative breast cancer. *Breast Cancer Research*, 13(2), 2011.

[200] Mahiro Iizuka-Ohashi, Motoki Watanabe, Mamiko Sukeno, Mie Morita, Ngoc Thi Hong Hoang, Takahiro Kuchimaru, Shinae Kizaka-Kondoh, Yoshihiro Sowa, Koichi Sakaguchi, Tetsuya Taguchi, and Toshiyuki Sakai. Blockage of the mevalonate pathway overcomes the apoptotic resistance to MEK inhibitors with suppressing the activation of Akt in cancer cells. *Oncotarget*, 9(28):19597–19612, 2018.

[201] Grace H. McGregor, Andrew D. Campbell, Sigrid K. Fey, Sergey Tumanov, David Sumpton, Giovanny Rodriguez Blanco, Gillian Mackay, Colin Nixon, Alexei Vazquez, Owen J. Sansom, and Jurre J. Kamphorst. Targeting the metabolic response to statin-mediated oxidative stress produces a synergistic antitumor response. *Cancer Research*, 80(2):175–188, 2020.

[202] Hidenori Kitai, Hiromichi Ebi, Shuta Tomida, Konstantinos V. Floros, Hiroshi Kotani, Yuta Adachi, Satoshi Oizumi, Masaharu Nishimura, Anthony C. Faber, and Seiji Yano. Epithelial-to-mesenchymal transition defines feedback activation of receptor tyrosine kinase signaling induced by MEK inhibition in KRAS-mutant lung cancer. *Cancer Discovery*, 6(7):754–769, 2016.

[203] Jieru Meng, Bingliang Fang, Yong Liao, Christine M. Chresta, Paul D. Smith, and Jack A. Roth. Apoptosis induction by MEK inhibition in human lung cancer cells is mediated by Bim. *PLoS ONE*, 5(9), 2010.

[204] Nilgun Tasdemir, Ana Banito, Jae Seok Roe, Direna Alonso-Curbelo, Matthew Camiolo, Darjus F. Tschaharganeh, Chun Hao Huang, Ozlem Aksoy, Jessica E. Bolden, Chi Chao Chen, Myles Fennell, Vishal Thapar, Agustin Chicas, Christopher R. Vakoc, and Scott W. Lowe. BRD4 connects enhancer remodeling to senescence immune surveillance. *Cancer Discovery*, 6(6):613–629, 2016.

[205] Xingchen Dong, Xiangming Hu, Jinjing Chen, Dan Hu, and Lin Feng Chen. BRD4 regulates cellular senescence in gastric cancer cells via E2F/miR-106b/p21 axis. *Cell Death and Disease*, 9(2), 2018.

[206] Jason D. Buenrostro, Beijing Wu, Ulrike M. Litzenburger, Dave Ruff, Michael L. Gonzales, Michael P. Snyder, Howard Y. Chang, and William J. Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490, 2015.

[207] Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K. Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868, 2017.

[208] Michael VanInsberghe, Jeroen van den Berg, Amanda Andersson-Rolf, Hans Clevers, and Alexander van Oudenaarden. Single-cell Ribo-seq reveals cell cycle-dependent translational pausing. *Nature*, 597(7877):561–565, 2021.

[209] Tal Ashuach, Mariano I. Gabitto, Michael I. Jordan, and Nir Yosef. MultiVI: deep generative model for the integration of multi-modal data. *bioRxiv*, page 2021.08.20.457057, 2021.

[210] Adam Gayoso, Zoë Steier, Romain Lopez, Jeffrey Regier, Kristopher L. Nazor, Aaron Streets, and Nir Yosef. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature Methods*, 18(3):272–282, 2021.

[211] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18:1–53, 2017.

[212] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. *35th International Conference on Machine Learning, ICML 2018*, 2:874–882, 2018.

[213] Chen Lin and Xu Sheng. Deep neural tangent kernel and laplace kernel have the same RKHS. In *ICLR 2021*, 2021.

[214] Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*, volume 2018-Decem, 2018.

[215] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips):6077–6086, 2017.

[216] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. *36th International Conference on Machine Learning, ICML 2019*, 2019-June:6156–6175, 2019.

[217] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. Scanpy: large scale single cell gene expression data analysis. *Genome Biology*, 19(15), 2018.

[218] Adam Gayoso, Romain Lopez, Galen Xing, Pierre Boyeau, and Katherine Wu. scvi-tools : a library for deep probabilistic analysis of single-cell omics data. 2021.

[219] J. Bergstra, D. Yamins, and D.D. Cox. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In *International Conference on Machine Learning*, volume 28, 2013.

[220] Bin Dai and David Wipf. Diagnosing and enhancing VAE models. In *Internation Conference on Learning Representations*, 2019.

[221] Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. Lagging inference networks and posterior collapse in variational autoencoders. In *Internation Conference on Learning Representations*, 2019.

[222] Christopher K.I. Williams and Matthias Seeger. Using the Nyström Method to Speed Up Kernel Machines. In *NeurIPS*, 2001.

[223] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: Supporting Text. *Pnas*, 102(43):15545–15550, 2005.

[224] Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tnenbaum. Program Induction. *Science*, 350(6266):1332–1338, 2015.

[225] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-Learning with Memory-Augmented Neural Networks. *33rd International Conference on Machine Learning, ICML 2016*, 4:2740–2751, 2016.

[226] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *34th International Conference on Machine Learning, ICML 2017*, 3:1856–1868, 2017.

[227] S. Huet, L. Xerri, B. Tesson, S. Mareschal, S. Taix, L. Mescam-Mancini, E. Sohier, M. Carrère, J. Lazarovici, O. Casasnovas, L. Tonon, S. Boyault, S. Hayette, C. Haioun, B. Fabiani, A. Viari, F. Jardin, and G. Salles. EZH2 alterations in follicular lymphoma: Biological and clinical correlations. *Blood Cancer Journal*, 7(4):e555–8, 2017.

[228] William T. Gibson, Rebecca L. Hood, Shing Hei Zhan, Dennis E. Bulman, Anthony P. Fejes, Richard Moore, Andrew J. Mungall, Patrice Eydoux, Riyana Babul-Hirji, Jianghong An, Marco A. Marra, David Chitayat, Kym M. Boycott, David D. Weaver, and Steven J.M. Jones. Mutations in EZH2 cause weaver syndrome. *American Journal of Human Genetics*, 90(1):110–118, 2012.

[229] Johann Brehmer and Kyle Cranmer. Flows for simultaneous manifold learning and density estimation. *arXiv*, (NeurIPS), 2020.

[230] Leonard B. Saltz, John V. Cox, Charles Blanke, and Lee S. Rosen. Irinotecan plus Fluorouracil and Leucovorin for metastatic colorectal cancer. *New England Journal of Medicine*, 343(13):905–914, 2000.

[231] Patricia Jaaks, Elizabeth A. Coker, Daniel J. Vis, Olivia Edwards, Emma F. Carpenter, Simonetta M. Leto, Lisa Dwane, Francesco Sassi, Howard Lightfoot, Syd Barthorpe, Dieudonne van der Meer, Wanjuan Yang, Alexandra Beck, Tatiana Mironenko, Caitlin Hall, James Hall, Iman Mali, Laura Richardson, Charlotte Tolley, James Morris, Frances Thomas, Ermira Lleshi, Nanne Aben, Cyril H. Benes, Andrea Bertotti, Livio Trusolino, Lodewyk Wessels, and Mathew J. Garnett. Effective drug combinations in breast, colon and pancreatic cancer cells. *Nature*, 603(July 2021), 2022.

[232] Arne Östman. The tumor microenvironment controls drug sensitivity. *Nature Medicine*, 18(9):1332–1334, 2012.

[233] Chiara M. Cattaneo, Krijn K. Dijkstra, Lorenzo F. Fanchi, Sander Kelderman, Sovann Kaing, Nienke van Rooij, Stieneke van den Brink, Ton N. Schumacher, and

E

Emile E. Voest. Tumor organoid–T-cell coculture systems. *Nature Protocols*, 15(1):15–39, 2020.

[234] Yasin Ilkagan Tepeli, Ali Burak Ünal, Furkan Mustafa Akdemir, and Oznur Tastan. PAMOGK: A pathway graph kernel-based multiomics approach for patient clustering. *Bioinformatics*, 36(21):5237–5246, 2020.

[235] Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P. Xing. Deep kernel learning. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016*, 51:370–378, 2016.

[236] Sydney M. Shaffer, Margaret C. Dunagin, Stefan R. Torborg, Eduardo A. Torre, Benjamin Emert, Clemens Krepler, Marilda Beqiri, Katrin Sproesser, Patricia A. Brafford, Min Xiao, Elliott Eggan, Anastopoulos Ioannis N., Cesar A. Vargas-Garcia, Abhyudai Singh, Katherine L. Nathanson, Meenhard Herlyn, and Arjun Raj. Shaffer - Rare cell variability and drug induced reprogramming as a mode of cancer drug resistance.pdf. *Nature*, 546(7658):431, 2017.

[237] Yaara Oren, Michael Tsabar, Michael S. Cuoco, Liat Amir-Zilberstein, Heidie F. Cabanos, Jan Christian Hütter, Bomiao Hu, Pratiksha I. Thakore, Marcin Tabaka, Charles P. Fulco, William Colgan, Brandon M. Cuevas, Sara A. Hurvitz, Dennis J. Slamon, Amy Deik, Kerry A. Pierce, Clary Clish, Aaron N. Hata, Elma Zaganjor, Galit Lahav, Katerina Politi, Joan S. Brugge, and Aviv Regev. Cycling cancer persister cells arise from lineages with distinct programs. *Nature*, 596(7873):576–582, 2021.

[238] Sumaiyah K. Rehman, Jennifer Haynes, Evelyne Collignon, Kevin R. Brown, Yadong Wang, Allison M.L. Nixon, Jeffrey P. Bruce, Jeffrey A. Wintersinger, Arvind Singh Mer, Edwyn B.L. Lo, Cherry Leung, Evelyne Lima-Fernandes, Nicholas M. Pedley, Fraser Soares, Sophie McGibbon, Housheng Hansen He, Aaron Pollet, Trevor J. Pugh, Benjamin Haibe-Kains, Quaid Morris, Miguel Ramalho-Santos, Sidhartha Goyal, Jason Moffat, and Catherine A. O'Brien. Colorectal Cancer Cells Enter a Diapause-like DTP State to Survive Chemotherapy. *Cell*, 184(1):226–242.e21, 2021.

[239] Robert Geirhos, Jörn Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[240] Michael K. Yu, Jianzhu Ma, Jasmin Fisher, Jason F. Kreisberg, Benjamin J. Raphael, and Trey Ideker. Visible Machine Learning for Biomedicine. *Cell*, 173(7):1562–1565, 2018.

[241] Jianzhu Ma, Michael Ku Yu, Samson Fong, Keiichiro Ono, Eric Sage, Barry Demchak, Roded Sharan, and Trey Ideker. Using deep learning to model the hierarchical structure and function of a cell. *Nature Methods*, 15(4):290–298, 2018.

[242] Brent M. Kuenzi, Jisoo Park, Samson H. Fong, Kyle S. Sanchez, John Lee, Jason F. Kreisberg, Jianzhu Ma, and Trey Ideker. Predicting Drug Response and Synergy

Using a Deep Learning Model of Human Cancer Cells. *Cancer Cell*, 38(5):672–684.e6, 2020.

[243] Mohammad Lotfollahi, Sergei Rybakov, Karin Hrovatin, and Soroor Hediyeh-zadeh. Biologically informed deep learning to infer gene program activity in single cells. *bioRxiv*, 2022.

[244] Zhi-jie Cao and Ge Gao. Multi- omics single cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, pages 23–25, 2022.

[245] Yancong Lin, Silvia L. Pintea, and Jan C. van Gemert. Deep Hough-Transform Line Priors. In *European Conference on Computer Vision*, 2020.

[246] Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. scGen predicts single-cell perturbation responses. *Nature Methods*, 16(8):715–721, 2019.

[247] Emma Dann, Neil C. Henderson, Sarah A. Teichmann, Michael D. Morgan, and John C. Marioni. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nature Biotechnology 2021*, pages 1–9, 2021.

[248] Daniel B. Burkhardt, Jay S. Stanley, Alexander Tong, Ana Luisa Perdigoto, Scott A. Gigante, Kevan C. Herold, Guy Wolf, Antonio J. Giraldez, David van Dijk, and Smita Krishnaswamy. Quantifying the effect of experimental perturbations at single-cell resolution. *Nature Biotechnology*, 39(5):619–629, 2021.

[249] B. G. Giraud and R. Peschanski. On positive functions with positive Fourier transforms. *Acta Physica B*, 37(2):331–346, 2006.

[250] Ingo Steinwart and Clint Scovel. Mercer's Theorem on General Domains: On the Interaction between Measures, Kernels, and RKHSs. *Constructive Approximation*, 35(3):363–417, 2012.

[251] Ali Rahimi and Ben Recht. Random Features for Large-Scale Kernel Machines. In *Advances in Neural Information Processing Systems*, 2009.

[252] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in Neural Information Processing Systems*, 2018-Decem(NeurIPS 2018):6389–6399, 2018.

[253] Carlo Baldassi, Fabrizio Pittorino, and Riccardo Zecchina. Shaping the learning landscape in neural networks around wide flat minima. *Proceedings of the National Academy of Sciences of the United States of America*, 117(1):161–170, 2020.

[254] Byungsoo Kim, Vinicius C. Azevedo, Nils Thuerey, Theodore Kim, Markus Gross, and Barbara Solenthaler. Deep Fluids: A Generative Network for Parameterized Fluid Simulations. *Computer Graphics Forum*, 38(2):59–70, 2019.

[255] Lukas Mosser, Olivier Dubrule, and Martin J. Blunt. Stochastic Seismic Waveform Inversion Using Generative Adversarial Networks as a Geological Prior. *Mathematical Geosciences*, 52(1):53–79, 2020.

[256] Ehsan Kharazmi, Zhongqiang Zhang, and George Em. hp -VPINNs : Variational Physics-Informed Neural Networks With Domain Decomposition. *Computer Methods in Applied Mechanics and Engineering*, 374:1–24, 2021.

[257] Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for mean squared error regression with wide neural networks. In *Conference on Learning Theory*, volume 125, pages 1–34, 2020.

[258] Marco Loog, Tom Viering, and Alexander Mey. Minimizers of the empirical risk and risk monotonicity. *Advances in Neural Information Processing Systems*, 32(NeurIPS), 2019.

[259] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences of the United States of America*, 116(32):15849–15854, 2019.

[260] Marco Loog, Tom Viering, Alexander Mey, Jesse H. Krijthe, and David M.J. Tax. A brief prehistory of double descent. *Proceedings of the National Academy of Sciences of the United States of America*, 117(20):10625–10626, 2020.

[261] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent where bigger models and more data hurts. In *Internation Conference on Learning Representations*, 2020.

[262] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *NeurIPS*, 2015.

[263] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1(Mlm):4171–4186, 2019.

[264] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

[265] Peter R. Wurman, Samuel Barrett, Kenta Kawamoto, James MacGlashan, Kaushik Subramanian, Thomas J. Walsh, Roberto Capobianco, Alisa Devlic, Franziska Eckert, Florian Fuchs, Leilani Gilpin, Piyush Khandelwal, Varun Kompella, Hao Chih Lin, Patrick MacAlpine, Declan Oller, Takuma Seno, Craig Sherstan, Michael D. Thomure, Houmehr Aghabozorgi, Leon Barrett, Rory Douglas, Dion Whitehead, Peter Dürr, Peter Stone, Michael Spranger, and Hiroaki Kitano. Outracing champion Gran Turismo drivers with deep reinforcement learning. *Nature*, 602(7896):223–228, 2022.

[266] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A.A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[267] Jinsung Yoon, Yao Zhang, James Jordon, and Mihaela Van Der Schaar. VIME: Extending the Success of Self-and Semi-supervised Learning to Tabular Domain. *Advances in Neural Information Processing Systems*, 33(NeurIPS):1–11, 2020.

[268] Taco S Cohen and Max Welling. Group Equivariant Convolutional Networks. In *International Conference on Machine Learning*, 2016.

[269] Jim Winkens and Max Welling. Improved Semantic Segmentation for Histopathology using Rotation Equivariant Convolutional Networks. *Miccai '18*, (Midl):2–4, 2018.

[270] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. 2021.

[271] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):1–21, 2014.

[272] N. Aronszajn. Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, 68(3):337, 1950.

[273] Bernhard Schölkopf and Alexander J. Smola. *Learning with kernels*. 2002.

[274] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. 2004.

[275] Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences. pages 1–64, 2018.

[276] Bertil Matern. *Spatial Variation.*, volume 36. 1970.

[277] George S. Kimeldorf and Grace Wahba. A Correspondence Between Bayesian Estimation on Stochastic Processes and Smoothing by Splines. *The Annals of Mathematical Statistics*, 41(2):495–502, 1970.

CURRICULUM VITÆ

Soufiane Marc Charles MOURRAGUI

06-12-1993	Born in Rouen, France.	

EDUCATION

2008–2011	Scientific High-School	Lycée Pierre Corneille, Rouen, France.
2011–2013	BSc in Mathematics and Physics (*Classe preparatoire aux Grandes Ecoles*).	Lycée Pierre Corneille, Rouen, France.
2013–2017	MSc in Science and Executive Engineering.	Mines ParisTech, Paris Sciences et Lettres, Paris, France.
2017–2022	PhD in Computer Science.	Delft University in Technology, Delft, Netherlands, and Netherlands Cancer Institute, Amsterdam, Netherlands.

PROFESSIONAL EXPERIENCE

Jun. 2015–Jan. 2016	Business Intelligence Intern	Criteo, Munich, Germany.
Feb. 2016–August 2016	R&D Intern.	LEAP Energy, Kuala Lumpur, Malaysia.
2022–now	Postdoctoral researcher.	Hubrecht Institute, Utrecht, Netherlands.

List of Publications

1. **S.M.C. Mourragui**, M. Loog, M.A. van de Wiel, M.J.T. Reinders, L.F.A. Wessels, *PRECISE: a domain adaptation approach to transfer predictors of drug response from pre-clinical models to tumors*, Bioinformatics (ISMB/ECCB 2019) **35** (14), i510-i519 (2019).

2. T. Abdelaal, **S.M.C. Mourragui**, A. Mahfouz, M.J.T. Reinders, *SpaGE: Spatial Gene Enhancement using scRNA-seq*, Nucleic Acids Research **48** (18), e107 (2020).

3. J. Ma, S.H. Fong, Y. Luo, C.J. Bakkenist, J.P. Shen, **S.M.C. Mourragui**, L.F.A. Wessels, M. Hafner, R. Sharan, J. Peng, T. Ideker, *Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients*, Nature Cancer **2** (2), 233-244 (2021).

4. **S.M.C. Mourragui**, M. Loog, D.J. Vis, K. Moore, A.G. Manjon, M.A. van de Wiel, M.J.T. Reinders, L.F.A. Wessels, *Predicting patient response with models trained on cell lines and patient-derived xenografts by nonlinear transfer learning*, Proceedings of the National Academy of Sciences (PNAS) **118**, (49) (2021).

5. **S.M.C. Mourragui**, J.C. Siefert, M.J.T. Reinders, M. Loog, L.F.A. Wessels, *Identifying commonalities between cell lines and tumors at the single cell level using Sobolev Alignment of deep generative models*, bioRxiv, 2022 (under revision at Genome Biology).

6. **S.M.C. Mourragui**, M. Loog, M. van Nee, M. A van de Wiel, M.J.T. Reinders, L.F.A. Wessels, *Percolate: an exponential family JIVE model to design DNA-based predictors of drug response*, bioRxiv, 2022 (accepted for presentation at RECOMB 2023, under review at Genome Research).

ACKNOWLEDGEMENTS

Writing a PhD thesis is a paradoxical exercise: although it is by essence very solitary, it is also the direct product of many discussions and countless great moments enjoyed over the four and a half years spanned by this research. In this last section, I would like to acknowledge all people who made the research presented in this thesis possible, and the journey to get there so enjoyable.

First of all, this research would not have been possible without the guidance of my three supervisors. **Lodewyk**, thank you very much for adding me to the group and help me pivot to cancer research. I have learnt a great deal from you and deeply enjoyed the *guided freedom* this position gave me. Inspite of a very busy schedule, you always managed to regularly find time to help me take critical decisions, which I am very grateful for. More importantly, you created a great environment to work in at B7, where many young scientists like me had the chance to thrive. **Marco**, thanks a lot for the many hours of technical discussions we had. These shaped all of the algorithms presented in this thesis and, more importantly, helped me grow as a computer scientist. Your critical approach to science has been of important value to this whole thesis. Finally, thanks a lot **Marcel** for your guidance and for providing a fertile ground to the PRB department. Your sharp feedback has been very instrumental to the completion of this thesis.

During the whole completion of this thesis, I had the chance to work alongside great collaborators at the VUmc. **Mirrelin** and **Mark**, thanks a lot for all the bi-monthly meetings we held. I profoundly enjoyed the wide-ranging discussions we had and learnt a lot from you. Specifically, your introduction to Bayesian statistics has been extremely useful for me, and I use this framework on a very regular basis.

I would like to present my highest gratitude to the committee members, **Emmanuel Barillot**, **Emile Voest**, **Jeroen de Ridder**, **Geurt Jongbloed** and **Boudewijn Lelieveldt**. Many thanks to have devoted your time to assessing the work I have been doing over these years.

Over these four and a half years, I benefited greatly from the amazing support provided by the two greatest office managers there is, **Patty Lagerweij** and **Saskia Peters**. You made the experience for all members in B7 and PRB so much easier and guided dozens of people through complex bureaucratic processes. I am pretty sure that most people on this Acknowledgments section are grateful for the amazing and relentless support you have been providing us over the years.

The research presented in the previous pages was embedded within two institutions: the Netherlands Cancer Institute (Amsterdam) and Delft University of Technology. I

would like here to thank the countless talented and dedicated colleagues I met and worked with during these years. Starting from Delft, a huge thanks to **Stavros**, who kindly accepted to be one of my paranymphs. It was a real joy to share an office in the *new building* (not so new anymore). I had a lot of great moments with you, whether discussing science, supervising students, playing football, going to conferences, or just discussing about the wide range of topics you are interested in. Attending your wedding with **Valentina** in Greece is definitely a highlight of my PhD years, and it is a great pleasure to see you grow as a successful dad and scientist. **Ahmed**, you are one of the most accomplished people I know, and you cleverly manage the difficult task to be both a fun and ambitious scientist. The many hours you spent explaining me the ins-and-outs of single-cell technologies definitely triggered my current scientific interest. **Tamim**, thanks a lot for all of our discussions, especially the time you dedicated to apply PRECISE for your problem. I am really happy to see SPaGE top the leaderboard of most spatial transcriptomics benchmarks. **Alex**, although our paths overlapped for only two years, I have countless great memories with you. As I would like my thesis to be still referenced by Google, I hope you will understand that I cannot list them here. **Osman**, it was great to work with you on the protein localization Kaggle challenge. I have had a lot of fun playing football with you. **Chirag**, many thanks to have converted me to biking and introduced me to Variational Inference. You are a fantastic scientist and I learnt tons from you. **Christian**, I have been sharing an office with you during all my PhD years (well, except Covid of course). It was great to gossip so much about Germany with you. **Tom (Mokveld)**, we started our PhD together but have been working on drastically different topics. I have learnt a lot from you, especially on graph theory. Your dedication, energy and curiosity are very contagious and I am happy to have spent time within the contagion radius. **Tom (Viering)**, you are the first person I got in contact with in the PRB group, when I was still contemplating joining. I really enjoyed all of our technical discussions on various machine learning topics. **Ramin**, you joined the group about a year later than me, but it feels like you have been in Delft forever. You are a great mate to chat with and I had great fun discussing about Persian and French cultures with you. I wish you all the best for your new scientific endeavors in Leiden. **Christine**, many thanks for your energy and for making the *Party City* a great office to work in. **Lieke**, you are arguably one of the most gifted scientists I know, and I am humbled to have had the chance to be in your MSc thesis committee. Your scientific contributions command the highest respect and I am very grateful to have learnt so much from you. **Aysun**, I remember when I met you at the DBL retreat in Lille, feels like yesterday. Funnily, we ended up discussing more *after* I left the group. Thanks for all the great vibes and all the culinary recommendations. I hope you can excuse me for the conversion to cheesecake. **Wouter**, many thanks for introducing me to transfer learning and pointing me to the right references earlier on during my PhD. I have unfortunately never seen the legendary *party Wouter* I heard so many tales of. **Ziqi**, it was great to brainstorm with you about domain generalization, and discuss about nice places to work in France. **Nicco**, my Tuesday morning train buddy, thanks for making the Amsterdam-to-Delft commute so enjoyable. **Meng**, thanks a lot for all the interesting words you taught me in chinese and for introducing me to the challenges of brain research. I now know what a Braak stage is, and how to tell strangers that I am their father. There are many other scientists in

the PRB department I had less contact with, but whose passion and energy made the experience memorable. In this regard, many thanks to **Mostafa**, **Mo**, **Yasin**, **Arlin**, **Erik**, **Marc**, **Attila**, **Joana**, **Thomas**, **Jan**, **Erik**, **Yunqiang**, **Jin**, **Laura**, **Taygun**, **Hayley**, **David**, **Jesse**, **Alexander**, **Thies Lucas**, and the whole department.

While working at the Netherlands Cancer Institute, I met a lot of fantastic colleagues, many of which became great friends. First of all, I would like to thank **Tom** who made me the honor with Stavros to accept to be my paranymphs. We met when you joined the group as a MSc student and you later came back to work in Lodewyk's and Emile's group. During all of these years, I had the great pleasure to grow scientifically and professionally alongside you, and learnt a lot about cancer biology and bioinformatics from you. Outside of the NKI, I will forever remember the great parties, after the borrel, at King's day or casually at random bars in Amsterdam West. You are an amazing friend and a very successful scientist, I am deeply grateful to have met you in Amsterdam. I wish you all the best together with **Elselien**. When I started my PhD, I benefitted greatly from the guidance of two senior group members, **Gergana** and **Nanne**. Many thanks to both of you for all the time you spent explaning me basics of bioinformatics and cancer research. **Maarten** (vegetarisch) **Slagter**, thanks a lot for all the great philosophical-political-gossipy discussions we had over the years. Although I still have no clue what the heck a CD$\mathscr{U}_{\{1,..,483\}}+\alpha-\epsilon$ EGFR$^{-/-}$ cyto-exotic T-cell is, I really enjoyed our collaboration on DISTINCT and learnt tons from you on immuno-therapy. **Marie**, je te remercie pour tous les bons moments passés au NKI. C'était trés chouette de travailler avec toi, et je te souhaite tous mes voeux de réussite pour ta thèse. Hélas, il semblerait que la lecture de cette phrase t'ai mise *I must survive* dans la tête. **Tesa**, please replace my French accent with a strong Scottisch-Twitter vibe when you read these few sentences. It was great to share an office with you and I have learnt many important things with you, from American politics to the mechanisms of androgen-resistance in prostate cancer. **Alberto**, many thanks for all the great moments in the office with Tesa and Marie and all the memorable Vrijdag-borrels at Tiffany's. **Joe**, many thanks for all the support on the Sobolev Alignment story. I had a lot of great evenings discussing with you about books and I am still looking for one you have not read yet. You are an impressively curious scientist and an absolutely atypical character. I am glad to have met you in Amsterdam. **Tim**, vielleicht nicht so ernst dich aus Deutch zu bedanken, aber werde ich probieren. Obwohl wir nicht viel Zeit zusammen in der Gruppe gearbeitet haben, bin ich super froh, dass ich dich kennengelernt habe. Danke für deine Hilfe um mein Deutsch zu verbessern und für all die sehr unterschiedlichen Discussionen. Ich bin sicher, dass du viel Erfolg mit deiner Doktorarbeit haben wirst. **Kat**, thank you very much for your help on the TRANSACT manuscript, and for all the discussions at the Friday Zoom coffee. **Duco**, thanks a lot for all the bioinformatics discussions that made me rethink key aspects of the field. I will long remember your introduction to blockchain and the provocative results of the beer-tasting experiment. **Lindsay**, it feels a bit sad to have overlapped so little time with you. You are a great scientist and a fun friend to hang out with. Cheers to great Christmas dinners and amazing Instagram stories. **Daniel**, many thanks for all your technical support and for your help on the TRANSACT manuscript. My technical skills owe a great debt to your mentorship. I still do not give up and will

eventually get you to say something nice about Python. **Sander**, many thanks for all the discussions, be them technical or cancer-research-related. I learnt a great deal from you. **Torben**, many thanks for all the RHPC support you have been providing over these years. All of the scripts and software developed in this thesis would not have been possible without your help. **Jinhyuk**, you are the first person I shared an office with. I am very much inspired by your scientific dedication. **Tycho** and **Bram**, you were the senior PhD students when I was just starting my own PhD. You have been crucially helping me navigate the technical landscape of computational biology and have set a high-level example to follow. **Joris**, our local superstar, many thanks for the time you spent explaining me the basics of clinical research. **Alex**, I unfortunately did not have much the occasion to work with you along the years. I really enjoyed going with you to BioSB a few years ago. **Kathy** and **Silvana**, thanks for all the great vibes you brought to the group. Although I did not know anything about experimental work, you took the time to explain me (very) basic things. **Evert**, thanks for all the technical discussions we had at the early stage of my PhD and for discussing our work on deep kernel learning. **Yongsoo** (aka Prof. Kim), I am very grateful for your introduction to PARAFAC, which opened me many doors and ultimately led to TRANSACT and Percolate. **Ewald**, many thanks for all the mathematical discussions and the Friday evening borrels. **Michael**, I sadly have not been climbing much with you these last years. I greatly enjoyed all of our discussions, you are a smart colleague to work with. **Marlous**, thanks for your help with the CBio portal (which you introduced me a few days after I started my PhD) and your support in navigating all the online databases. **Maksim**, we only overlapped a few months, but I am very happy to have got to know you. You joined the institute in difficult times and I am very impressed by the resilience you are showing. **Olga**, I remember perfectly you giving an interview talk at B7 a few years back, when I thought that you would be a great fit for a post-doc...not knowing you were applying for a PhD. You are a greatly gifted colleague and I am pleased to have learnt so much from you. Finally, **Roebi**, you are a very impressive person. I wish you all the best for the years ahead and I hope the best is yet to come for you.

I had the chance to (co-)supervise several students over the years. First, **Eline**, I highly enjoyed working with you on deep kernel learning. You were a fantastic student, always motivated, curious and hard-working. I wish you all the best for your very bright future. **Eduards**, we unfortunately did not have the occasion to meet physically, as Covid constrained your internship to be virtual. Thanks a lot for introducing me to graph neural networks and SMILES embeddings. Finally, **Vanessa**, I learnt enormously from you. It is always challenging to supervise somebody better than yourself, but working with you was very enjoyable. I am really happy to see you now continuing as a PhD student and have no doubt that you will have a lot of success. Some other projects were rather short and I can not here list of all the great people involved in them. Nonetheless, I am grateful to have learnt from each one of you.

During these years in Amsterdam, I had the chance to spend great times with amazing friends. First of all, many thanks to my amazing flatemates who I had the chance to live with at the beginning of my PhD. **Julia**, many thanks for all the great moments

and discussions at Sluisstraat. **Claudia**, you somehow managed the impossible task to make the Covid times enjoyable. I will long cherish these evenings spent discussing (or gossiping) about absolutely everything. I am extremely impressed and inspired by the strength you showed at the end of your Amsterdam experience. **Maria** et **Kris**, un grand merci pour toutes ces soirées mémorables passées à Amsterdam et pour toute l'aide que vous nous avez fournie durant notre déménagement. Même si nous n'avons finalement pas fait de week-end *Tesla-Autobahn* en Allemagne, c'était très sympa de vous rendre visite en Suisse. Cheers to many more great moments to come. A big thank to the Mafia, **Anna**, **Jeremy**, **Antonio**, **Mathias**, **Ziva**, **Angela**, **Ronak**, **Joao**, **Simon**, **Xabier**, **Yulia**, **Myriam**, **Alberto**, **Clara**, **Isabelle**, **Mariana** and **Eric** for all the amazing parties. I had so much fun with all of you folks!! I will forever remember the amazing parties in the *NKI ghetto* and the many passionate discussions we had. **Dorine**, **Xabier** and **Anoek**, many thanks for helping us move to our new place. The Graduate School would unfortunately not accept me to list all the lovely Dutch words you taught me Dorine, so *geen lekkere schaap*. **Felipe** and **Fernando**, I actually met you after my PhD but had great moments with you both. I am acknowledging you together by fear of mixing up memories.

Pendant ces années loin de ma chère France natale, j'ai eu la chance de revenir régulièrement et de partager des soirées et vacances littéralement incroyables. Mon expérience aux Pays-Bas n'aurait pas été aussi agréable sans vous. **Sandrine** et **Marc**, **Aurelien**, **Marco**, **Marie**, **Xavier**, **Léa**, **Laurent**, **Grandeau**, **Gaël**, **Antoine** et **Léa**, merci pour tous les diners sur Paris, toutes les discussions passionnées sur des sujets divers, et tous les week-ends et vacances passés à vos côtés. **Alexandre**, **Faustine**, **Justine**, **Lucie**, **Alex** et **Marie**, je suis super fier que l'on ait réussi à se voir aussi frequemment, même aussi longtemps après nos années lycéennes. Mes retours sur Rouen ont toujours aidé à me requinquer.

Tous les résultats presentés dans cette thèse n'auraient jamais été atteignables sans le soutien de ma famille. Un infini merci à ma **mère**, mon **père** et mes deux soeurs, **Camilia** et **Nour**. Un grand merci également à **Souad** et **Thierry** pour leur soutien ces dernières années, ainsi qu'à mes **grands-parents**, mes **oncles** et **tantes** et mes **cousines** et **cousins**, que je remercie ici chaudement. La fin de mon expérience de thèse a hélas été marquée par la disparition de ma **grand-mère paternelle**, qui laisse notre grande famille dans une profonde tristesse, et à qui je souhaite ici rendre hommage. J'espère de tout mon coeur que cette thèse l'aurait rendue fière.

Last, but *auf jeden fall* not least, my warmest acknowledgment goes to **Lisa**. I remember as if it was yesterday when we met at Anna's and Antonio's. You are the best partner I could have dreamt of and your relentless support over these last years has been instrumental in the success of this thesis. Although the coming years look challenging, I am so much looking forward to spending them with you. Thanks for making my life so amazing.