Dissertation submitted to the
Combined Faculty of Natural Sciences and Mathematics
of the
Ruperto Carola University Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

Presented by
M.Sc. Nicolàs Palacio-Escat
born in: Tarragona, Spain
Date of examination: October 25th 2021

Modeling intra- and intercellular communication in the context of human cancer from high-throughput data

Referees: Prof. Dr. Julio Saez-Rodriguez
Prof. Dr. Ursula Klingmüller

To my parents for their constant support
and to Taro for giving me happiness

Declaration

I hereby declare that I am the true and only author of this thesis and that I used no other sources or materials than those expressely indicated.

Acknowledgments

I would like to first thank my supervisors Prof. Dr. Julio Saez-Rodriguez and Prof. Dr. Ursula Klingmüller for their support and invaluable input, without whom this thesis would not have been possible. Thanks also to our group secretaries Gabriele Lützeler and Erika Schulz, for their constant help with bureaucracy and life in Germany in general.

Thanks to all my colleagues during these years, especially Dr. Attila Gabor, Dr. Rosa Hernansaiz-Ballesteros, Dr. Jovan Tanevski, Dr. Martín Garrido, Aurelien Dugourd, Dr. MD Bence Szalai, Dr. Dénes Türei, Dr. Federica Eduati, Ricardo Ramírez, Jakob Wirbel, Dr. Luís Tobalina, Dr. Enio Gjerga, Francesco Ceccarelli, Hyojin Kim, Dr. Javier Perales-Patón, Christian Holland, Olga Ivanova, Anika Liu and Dr. Ana Victoria Ponce-Bobadilla, for their input, help and just being great teammates to work with throughout all these years.

Finally, I would like to thank all the collaborators that I had the pleasure to work with and provided the data presented in this thesis and other studies I participated in, especially Dr. Magdalena Szczygieł, Dr. Vito Zanotelli, Prof. Dr. Jesper Velgaard Olsen, Dr. Kristina Bennet Emdal, Dr. Caroline Wigerup, Dr. Kristina Masson, Dr. MD Peter Blume-Jensen, and Dr. MD Markus Rinschen.

Publications

Späth MR, Bartram MP, Palacio-Escat N, Hoyer KJR, Debes C, Demir F, Schroeter CB, Mandel AM, Grundmann F, Ciarimboli G, Beyer A, Kizhakkedathu JN, Brodesser S, Göbel H, Becker JU, Benzing T, Schermer B, Höhne M, Burst V, Saez-Rodriguez J, Huesgen PF, Müller RU, Rinschen MM. The proteome microenvironment determines the protective effect of preconditioning in cisplatin-induced acute kidney injury. Kidney International 2019; 95(2):333-349. DOI: 10.1016/j.kint.2018.08.037.

Rinschen MM, Palygin O, Guijas C, Palermo A, Palacio-Escat N, Domingo-Almenara X, Montenegro-Burke R, Saez-Rodriguez J, Staruschenko A, Siuzdak G. Metabolic rewiring of the hypertensive kidney. Science Signaling 2019; 12(611):eaax9760. DOI: 10.1126/scisignal.aax9760.

Hoyt CT, Domingo-Fernández D, Mubeen S, Llaó JM, Konotopez A, Ebeling C, Birkenbihl C, Muslu Ö, English B, Müller S, de Lacerda MP, Ali M, Colby S, Türei D, Palacio-Escat N, Hofmann-Apitius M. Integration of Structured Biological Data Sources using Biological Expression Language. bioRxiv 2019; 631812. DOI: 10.1101/631812.

Türei D, Valdeolivas A, Gul L, Palacio-Escat N, Klein M, Ivanova O, Ölbei M, Gábor A, Theis F, Módos D, Korcsmáros T, Saez-Rodriguez J. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. Molecular Systems Biology 2021; 17(3):e9923. DOI: 10.15252/msb.20209923.

Song C, Roller J, Ponce-Bobadilla AV, Palacio-Escat N, Saez-Rodriguez J, Heuveline V. Spatial effect on separatrix of two-cell system and parameter sensitivity analysis. *Manuscript in preparation*

Emdal KB*, Palacio-Escat N*, Wigerup C, Bekker-Jensen DB, Rönnstrand

L, Uddin K, Puissant A, Itzykson R, Masson K, Saez-Rodriguez J, Olsen JV, Blume-Jensen P. Phosphoproteomics of primary AML patient samples reveals rationale for AKT combination therapy and p53 context to overcome selinexor resistance. *Manuscript submitted*

Szczygieł M, Schneider M, Palacio-Escat N, Meister M, Buck-Wiese M, Postawa L, Schwarzmüller L, Heming S, Vlasov A, Böhm M, Kister B, Kazdal D, Richtmann S, Helm B, Kriegsmann M, Kriegsmann K, Lehmann W-D, Stenzinger A, Schlesner M, Muley T, Saez-Rodriguez J, Schilling M, Klingmüller U. Extracellular Matrix Proteome recapitulated in vitro reveals lung cancer progression mechanisms. *Manuscript in preparation*

* Equal contribution

Abstract

Understanding cell signaling is probably one of the biggest challenges in modern biology. Thanks to the advance in new technologies, like next generation sequencing and other high throughput techniques, commonly referred as omics technologies, researches can generate great amounts of comprehensive biological data. More recently, these technologies have advanced to the point where one can analyze genes or proteins in single cells, even with spatial resolution. The ability of these approaches to generate great amounts of data requires of complementary techniques to analyze it. Analyzing and contextualizing this amounts of data has provided great insight and development in our current understanding and treatment of many diseases. Current research on disease mechanisms focuses mainly on molecular processes in order to understand the underlying systems driving them. Many approaches have so far focused on intracellular signaling and it has not been until recent years that the role of cell-to-cell communication in health disorders has gained importance. With this goal in mind, I will be presenting approaches to model and analyze biological data accounting for intra- and intercellular communication. The models and analyses presented combine omics data with prior biological knowledge. For this, I rely and our in-house resource OmniPath to extract the relevant intra- and intercellular interactions.The analytical approaches presented in this thesis range from the more classical differential expression and gene set enrichment analysis to more advanced and recent machine learning methods. Thanks to the different strategies applied across different settings, I was able to extract relevant biological insights with applications in clinical and biological research. Despite the approaches presented in this thesis being mainly focused on cancer, these surely can be further extended and applicable in many other contexts.

Zusammenfassung

Das Verständnis der Signalübertragung in Zellen ist wahrscheinlich eine der größten Herausforderungen in der modernen Biologie. Aufgrund des Fortschritts neuer Technologien, wie beispielsweise Next Generation Sequencing und anderen Hochdurchsatztechniken, ist es Forschenden möglich große Mengen an umfassenden biologischen Daten zu generieren. Diese Technologien werden als Omics-Technologien bezeichnet. In letzter Zeit sind diese Methoden so fortschreitend entwickelt worden, dass es nun möglich ist Gene oder Proteine in einzelnen Zellen zu analysieren, so wie ihre räumliche Auflösung zu erfahren. Diese Fähigkeit, große Datenmengen zu generieren, erfordert hingegen komplementären Verfahren, um diese zu analysieren. Die Analyse und Kontextualisierung dieser Datenmengen hat zu großen Erkenntnissen und Entwicklungen in unserem aktuellen Verständnis und Behandlungen vieler Krankheiten geführt. Die aktuelle Forschung zu Krankheitsmechanismen konzentriert sich hauptsächlich auf molekulare Prozesse, um die zugrundeliegenden Systeme zu verstehen, die sie antreiben. Viele Ansätze haben sich bisher auf die intrazelluläre Signalübertragung konzentriert und erst in den letzten Jahren hat die Rolle der Zell-zu-Zell-Kommunikation bei Erkrankungen an Bedeutung gewonnen. Mit diesem Ziel vor Augen war es mir möglich Methoden zur Modellierung und Analyse biologischer Daten zu erstellen, die die intra- sowie interzelluläre Kommunikation von Zellen berücksichtigen. Die in dieser Arbeit vorgestellten Modelle und Analysen kombinieren Omics-Daten mit biologischem Hintergrundwissen. Hierzu wurde die hauseigene Datenbank OmniPath verwendet, um die relevanten intra- und interzellulären Interaktionen zu ermitteln. Diese analytischen Ansätze reichen von der klassischen differentiellen Expressions- und Gensatzanreicherungsanalyse bis hin zu fortgeschrittenen und neueren Methoden des machine learnings. Durch die verschiedenen Strategien, die in unterschiedlichen Zusammenhängen angewandt wurden, war es möglich, relevante biologische Erkenntnisse mit Anwendungen in der klinischen und biologis-

chen Forschung zu erhalten. Obwohl sich die in dieser Arbeit vorgestellten Ansätze hauptsächlich auf Krebserkrankungen fokussiert, lassen sie sich sicherlich weiter erweitern und auf vielen andere Bereiche anwenden.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Biology of human disease

The central dogma of molecular biology defines the basic flow of genetic information in living organisms. Briefly, all the genetic information for sustaining life of a given organism is encoded in the genome of every one of its cells. This information is encoded in the form of DeoxyriboNucleic Acid (DNA), which is stored in the cell nucleus in Eukaryotes. Yet, this material is not functional per se, instead, cells make copies of fragments of this information in the form of RiboNucleic Acid (RNA). This fragments or regions of the genome are known as genes, and the process of copying them in the form of RNA is known as transcription. Finally, the genetic information copied in the form of RNA is then converted into proteins in a process known as translation. It is these proteins that will exert the functional activity in the cell. It is important to note that different cells in an organism express different genes depending on their function or type as well as their stage on their life cycle. While some proteins are common to all the cells, like the very same proteins responsible of executing the processes of transcription and translation described above or the ones responsible for generating energy for the cell. Other proteins are expressed in very specific situations or cells. Some examples include antibodies, secreted by a selected population of immune cells in the presence of an infection, or insulin which is produced by specific cells in the pancreas.

Yet, this is just the basis of the actual information processing cells are capable of. Even though cell capabilities are ultimately encoded in the genome, which part of it is expressed, when and and how is mainly controlled by

cell signaling. Regulatory networks control cellular state and behavior via integration of both internal and external stimuli to adapt their response accordingly. In summary, the flow of information in cells is not as linear as it may seem but actually is more like a dynamic cycle where signaling regulates which and how genes are expressed (and translated) and these in turn affect how the cell processes and reacts to the different signals and stimuli (Figure 1).

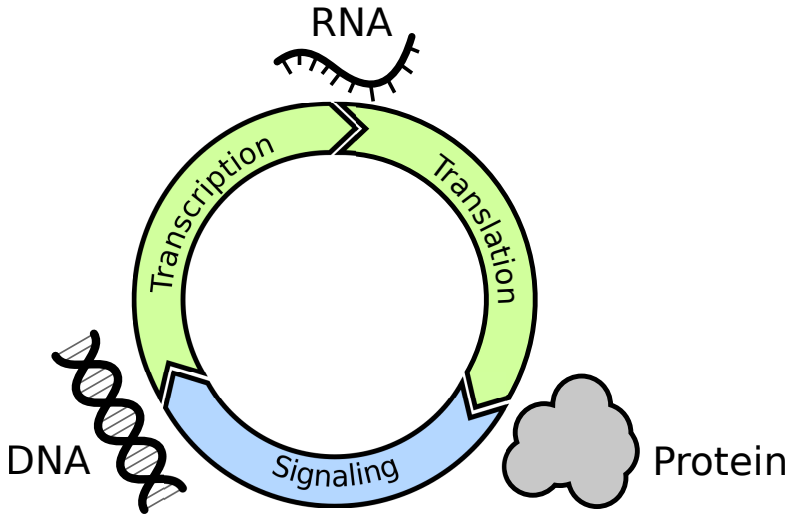

Figure 1: Basic scheme depicting the flow of information in cells. In light green are represented those processes involving synthesis of new molecules while in light blue involves modification or modulation of molecules (i.e. regulation).

Cell signaling is a tightly regulated process and is of utmost importance, specially in multicellular organisms like humans, since cells need to synchronize and orchestrate collective responses in order to maintain homeostasis. Despite this tight regulation, evolution has granted cells with certain robustness against perturbations up to a certain point thanks to several mechanisms. Some well known examples of these processes include the immune system, which counteracts and neutralizes external organisms like viral or bacterial infections, or the DNA damage detection and repair mechanisms that are triggered when genetic material is altered. Nevertheless, these systems are not perfect and can fall short or fail completely, and therefore diseases can occur and develop.

At this point, let us differentiate between infectious from non-infectious diseases. Infectious diseases comprise any disease caused by the presence of

an external entity and consequent disruption of correct function of the infected organism. Generally speaking, this type of diseases have a relatively simple solution, this is, neutralization and elimination of the infectious agent. Of course, this is not always as simple as it seems. Bacterial infections are nowadays more or less easy to tackle thanks to the development of antibiotics. This is, as long as the infection is not extensively spread within the organism or the specific bacteria is not extremely aggressive or resistant to current antibiotics (few rare cases). On the other hand, viral infections tend to be a bit harder to counteract. While the the strategy follows the same principle of eliminating the pathogen from the body, virus present further challenges. First, unlike bacteria, they are not independent living organisms and therefore, cannot be "killed" per se. Their method of infection, which in general involves incorporating their genetic material in the host cells and use their machinery to replicate and extend to other cells, makes them harder to eliminate. Nevertheless, in many cases there are possible strategies to inactivate them, although the most common strategy is to prevent their infection via vaccination. This allows the training of the immune system to quickly recognize the virus and eliminate it before it spreads extensively. Despite all this efforts, the main problem with viral diseases is their high capacity of spread and mutation. A clear example of this is the seasonal flu. Other more severe examples in recent years are the Human Immunodeficiency Virus (HIV), Ebola or the current pandemic of the Severe Acute Respiratory Syndrome CoronaVirus 2 (SARS-CoV-2).

When looking into non-infectious diseases, classification prove more challenging, since many factors can be involved. In a general sense, one can say that non-infectious diseases stem from individuals themselves rather than from other organisms. Broadly, we can distinguish between environmental and genetic diseases, although sometimes the line between these can be blurry like is the case of cancer or some chronic diseases. Environmental diseases include diseases caused by the environment itself (e.g. exposure to toxic chemicals or radiation) as well as the lifestyle (e.g. substance abuse or diet). On the other hand, genetic diseases include any diseases caused by genetic alterations either inherited or acquired later in life. As mentioned previously, the latter can be tightly linked to environmental factors and therefore, some diseases cannot be explicitly classified as one or the other. Furthermore, most of the current biomedical research is focused in the treatment of these type of diseases. This fact highlights the need of a better understanding of how cells behave under normal and diseased conditions, from the molec-

ular to the systemic level. Nevertheless, thanks to recent developments in medicine and molecular biology this knowledge gap is being steadily reduced. Of important note is the advent of Next-Generation Sequencing (NGS) and other omics technologies, which allow us to study hundreds to thousands of molecules in single experiments or conditions. The high throughput data generated from these technologies go hand in hand with the requirement of the adequate tools and methods for their analysis, which allow us to extract the biologically and clinically relevant information in the subject(s) of study.

1.2 Cancer biology

Among the biggest challenges in modern medicine is to tackle cancer, arguably one of the most complex diseases known to date. Part of this complexity stems from the fact that it can develop in any part of the body be it organ, tissue or cell type. Furthermore, even within same cancer types, these can be very heterogeneous, ranging from non-malignant tumors to very aggressive and rapidly-progressing cancers. Broadly speaking, cancer is defined as the disease where a subpopulation of cells starts growing uncontrollably. This transformation generally originates when cells acquire malignant mutations that alter their regulation, allowing them to continuously proliferate. The subsequent deregulation of their information processing networks allow them to undergo through further malignant transformations that alter their phenotype. These changes ultimately involve the aforementioned uncontrolled growth, higher mutation rates and invasion of other tissues mainly. The main characteristics cancer cells develop during their malignant transformation are very well described in the well-known review "Hallmarks of Cancer" by Hanahan and Weinberg [17]. The underlying mechanism common to most of these traits involves the cell information processing machinery, be it by direct mutation of the genes involved or by other regulatory activity. This is, cellular signaling and communication processes. This holds true not only for the malignant cells themselves, but also with other cell types like fibroblasts, immune cells or the extracellular matrix (ECM) [7].

Cancer is a very complex disease and many efforts have been made to discover the different mechanisms involved in this malignancy. Furthermore, current approaches in clinical applications include the discovery of early detection (or even prediction) methods and patient profiling and personalized treatment (also known as personaluzed medicine). New technologies in recent

years like NGS or high-throughput transcriptomics and proteomics enabled huge improvements in our understanding of cancer biology. With the increasing amount and size of these data sets, comes the necessity of powerful yet understandable analytical tools. Furthermore, given the inherent heterogeneity of cancer, flexibility in the application of this tools is also necessary.

1.3 Cell signaling in cancer

A canonical signaling cascade starts by the interaction of a ligand with their corresponding receptor, generally in the cell membrane. This triggers the transmission of the signal to the intracellular domain. Once inside the cell, the information is integrated and transmitted by a cascade of protein-protein interactions which typically involve addition or removal of phosphate groups by kinases or phosphatases respectively. This ultimately leads to the activation or inhibition of transcription factor(s) that induce changes in the cellular expression and therefore, their phenotype. These changes allow cancer cells to acquire different characteristics that drive their malignant transformation. The main characteristic of cancer, continuous growth and proliferation of cells can be achieved by different means. As described by Hanahan and Weinberg [17], these include: sustaining proliferative signaling (e.g. mutation of the KRAS gene that allow its constitutive activation), evasion of growth suppression and/or cell death (e.g. loss of function of the TP53 gene, which is responsible of integrating stress signals and DNA damage which can suppress growth or trigger apoptosis depending on the degree of stress/damage) and acquisition of replicative immortality (e.g. expression of telomerase, which repairs the extremes of the chromosomes enabling infinite replication). Other characteristics that cancer cells acquire during malignant transformation are linked to the immune system. Immune cells are infiltrated within the tumor microenvironment in order to neutralize malignant cells, but some cancers have shown the capability to secrete immune-suppressing factors, for example through TGFβ [45]. Furthermore, they are also capable of secreting pro-inflammatory signals which recruit fibroblasts and immune inflammatory cells that in turn secrete growth factors and other tumor-promoting signals. The pro-inflammatory signals along with others like vascular endothelial growth factor (VEGF) also trigger *de novo* vascularization of the tumor, a process known as angiogenesis. Finally, one of the most dangerous traits of malignancy development is metastasis, the process in which cancer

cells acquire the motility and the ability to travel through blood vessels (facilitated thanks to angiogenesis) or the lymphatic system. This is generally achieved by changes in cell adhesion as well as modification of the surrounding ECM. When this occurs, cancer cells are able to invade other tissues in distant parts of the body and the prognosis worsens significantly. The traits described above highlight the complexity of cancer development. Not only the changes in signaling within the cancer cells but also how these affect their environment and these in turn to the malignant cells (Figure 2). Therefore, in order to understand and treat cancer effectively, we need to study not only malignant cells alone but along with their microenvironmental context all together.

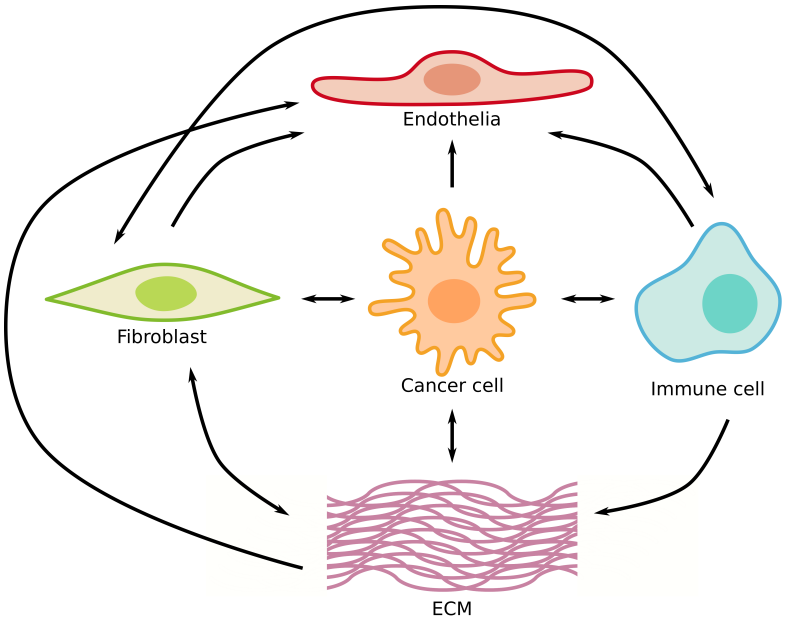

Figure 2: Simplistic representation of intercellular signaling and interactions between cancer cells and their microenvironment. Major players in the tumor microenvironment include several types of immune cells (grouped here for the sake of simplicity), endothelial cells, fibroblasts (also known as cancer-associated fibroblasts or CAFs) and the extracellular matrix (ECM).

1.4 Background

In this thesis I will present my contribution on three of the main projects I have been working on during my time as a doctoral candidate. In the course

of these studies, I applied different data analysis and modeling approaches to study signaling in cancer. These encompass from basic descriptive statistics and linear models to more sophisticated machine learning methods. I applied these to different data types like time- or spatially-resolved and contexts like clonal cells, patient samples and multiple cellular type cultures. Learning and tailoring the different methods to the varying settings and questions allowed us to obtain relevant biological insights and generate novel clinically-relevant hypotheses.

One of the most known aphorisms in statistics and scientific modeling states that "All models are wrong, but some are useful". The usefulness of a model not only depends on how well the model explains the reality (e.g. data), but also in the simplicity of the model. This balance between accuracy/precision and interpretability/explainability is what defines a good model. Whether this balance shifts more towards one side or the other will always depend on two main factors, namely the question one wants to answer and the available data or observations. With this idea in mind, I applied different analytical and modeling approaches on the different projects accordingly. While in some cases this only required well-known and simpler approaches like linear models and least squares method, other instances required more sophisticated and novel approaches.

Thanks to the projects presented herein and others not shown, I was able to learn many different data analysis and modeling methods. Furthermore, I also learned how to conduct my research independently and critically. This allowed me to contribute to different research projects about cancer signaling and obtain insights from high-throughput data.

The work presented here mainly focuses on the computational analysis of different sources of biological data in different cancer types. The different data sets were generated and obtained from our collaborators as described in the Methods section. In the following sections, I will introduce the different projects covered in this dissertation.

1.5 Unveiling drug-resistance mechanisms in acute myeloid leukemia

In the first project about acute myeloid leukemia (AML) drug resistance, I combined the analysis of drug response from survival data with the analysis of phosphoproteomic data from the same samples. Acute Myeloid Leukemia

(AML) is one of the highest incidence cases of leukemia and with high relapse rates [28]. AML is characterized by high genetic and phenotypic diversity [33]. Many efforts have been focused in finding proper drug-based treatments, since this malignancy does not form solid tumors and cannot be treated with surgery. One approach is targeting nuclear export mechanisms, since key proteins regulating cell cycle are required to be located in the nucleus for their activity. Considering that cancer cells are proliferating uncontrollably, modulation or inhibition of nuclear export processes is a valuable therapeutic strategy. Among the nuclear export protein family, Exportin-1 (XPO1, also known as chromosome region maintenance 1 or CRM1) is one of the better characterized. XPO1 is responsible for the nuclear transport of many proteins including several tumor suppressor proteins and oncogenes as well as proteins involved in cell growth and proliferation from the nucleus to the cytoplasm [25]. Yet, traditional XPO1-targeted drugs like leptomycin B (LMB) show great cytotoxicity due to permanent XPO1 inhibition [43]. To circumvent this, a new generation of synthetically-developed drugs termed Selective Inhibitors of Nuclear Export (SINE) have been designed based on molecular modeling. SINEs have shown promising results in their capacity to trigger cell-cycle arrest and apoptosis in highly proliferating (malignant) cells sparing the non-malignant cells [13, 14]. Despite these efforts, it is well known that cancer patients harbor or can develop *de novo* resistance to drug treatments.

In our study we focus on Selinexor, an inhibitor of XPO1 which has just recently been approved by the U.S. Food and Drug Administration (FDA). We analyzed 20 *ex vivo* AML patient samples as well as four cell line models. Cells were cultured with or without drug (control) and analyzed using tandem mass tag (TMT) labeling and Mass Spectrometry (MS). For both sample types, cell viability assay and phosphoproteomic data were obtained from our collaborators. First, the survival of these samples to different concentrations of Selinexor was analyzed. Then I fitted dose-response curves to determine whether a given sample is responding to the drug or it shows resistance to the treatment by using the least squares method. This allowed us to classify the samples into responders and non-responders.

On the other hand, the phosphoproteomics data was firstly normalized with the variance stabilization method (VSN) which transforms the data such that the variance becomes approximately independent of the mean [19]. This transformation is well-suited for multivariate analysis and stabilizes the signal-to-noise ratio in contrast to other more classical methods. After ap-

plying VSN to each experimental set (run of TMT-MS experiment), batch effect was removed by fitting a linear model to the data including the batch information and then having this component is removed. Once the data has been processed and normalized, a differential phosphorylation analysis was performed in order to point out which phosphosites are significantly regulated by the drug. The analysis was performed by fitting a linear model on each gene and condition (treated vs. untreated samples in our case) and comparing the coefficients. These were contrasted for the responder and non-responder samples independently. The models used for the batch effect removal and differential phosphorylation were fitted using the R package limma [34], which uses least squares method to fit the parameters. This allowed us to look for common and unique effects the drug had on the different response groups. The results from the differential phosphorylation ($\log_2$ ratios) were then used to perform both a gene set and a kinase enrichment analyses (GSEA and KSEA respectively). The gene sets for the GSEA were retrieved form the molecular signatures database (MSigDB) [37] which contains a comprehensive collection of annotated gene sets from widely used databases like KEGG [23], BioCarta [31] or Reactome [21] among others. The analysis was performed using the R package piano [41], which uses uses a set of different statistical methods to compute the enrichment scores, which are then ranked in order to render them comparable between the methods. The final enrichment is then computed as a consensus score across the different methods, which greatly reduces potential method biases and is therefore more robust. For the KSEA, the kinase-substrate interactions were retrieved from the OmniPath database, a comprehensive collection of biological prior knowledge which integrates data from more than 100 resources [40]. The enrichment was computed with the the Python package kinact [44], developed in our group.

Finally the results of this analysis were used to perform a functional assessment of the drug-induced changes in the different groups of samples. This allowed us to infer key processes that could lead some insights in the resistance to the treatment and then propose potential combination treatments. Thanks to the results obtained from computational analysis, we proposed a rational drug combination which was then validated experimentally. In this case, cell viability was measured in the cell lines at increasing concentrations of the drug of interest in combination with an AKT inhibitor. The survival data from these drug combination experiments was then used to compute the synergy scores of each drug combination based on the dose-response

model and Loewe additivity principle [38]. Thanks to these analyses we could conclude that the proposed drug combination showed promising results and potential to reverse drug resistance and calls for further validation.

The main aim of the study was to unveil the mechanism of resistance to the drug from the low- or non-responding samples. Within this project, I was able to contribute in the discovery of potential drug resistance mechanism of AML towards Selinexor. Thanks to the combination of dose-response and phosphoproteomics data along with mathematical modeling and data analysis applied, I was able to identify target candidates for drug combination therapy. Follow-up experiments based on our discoveries showed promising results. Based on the obtained insights, we proposed a drug combination to tackle such resistance which showed positive and interesting results. Further experimental validation will demonstrate whether our discoveries can be applied to patients in the clinic and improve their outcomes.

1.6 Discerning cancer cell communication with associated fibroblasts and the extracellular matrix

Cancer is a complex disease in which not only the malignant cells play an important role but also the tumor microenvironment, also called tumor stroma. This environment is mainly composed of immune cells, fibroblasts and extracellular matrix (ECM) which is mainly regulated and modeled by the fibroblasts. Furthermore, fibroblasts are also responsible for the release of growth factors including TGFβ, which has been shown to induce Epithelial to Mesenchymal transition (EMT) in cancer cells [7]. Yet, the exact mechanisms and communication between cancer cells, fibroblasts and ECM is still not fully understood. In this project we therefore sought to study the intercommunication between lung cancer cells, fibroblasts and the ECM. In order to do so, we analyzed peptidomics data from lung cancer cell line H1975 in monoculture as well as in coculture with the lung fibroblast cell line HFL1. Cells were cultivated in ECM developed from the lung fibroblast cell line which was cultivated with and without (control) TGFβ. The cultures were then decellularized and lung cancer cell lines were seeded alone or in combination with the fibroblasts. From both of these cultures, samples were separated and analyzed as cell lysates (i.e. solid phase, cellular and

ECM components) and the supernatants (SN, i.e. liquid phase, secretome). Therefore, cell mass and secretome could be analyzed separately. Each sample and condition were sampled in duplicates and for three consecutive days (see Figure 3). The different samples were then measured using TMT-based proteomics. This setup of conditions allowed us to study the effect of TGFβ-stimulated ECM on the cancer cell lines in both culture types.

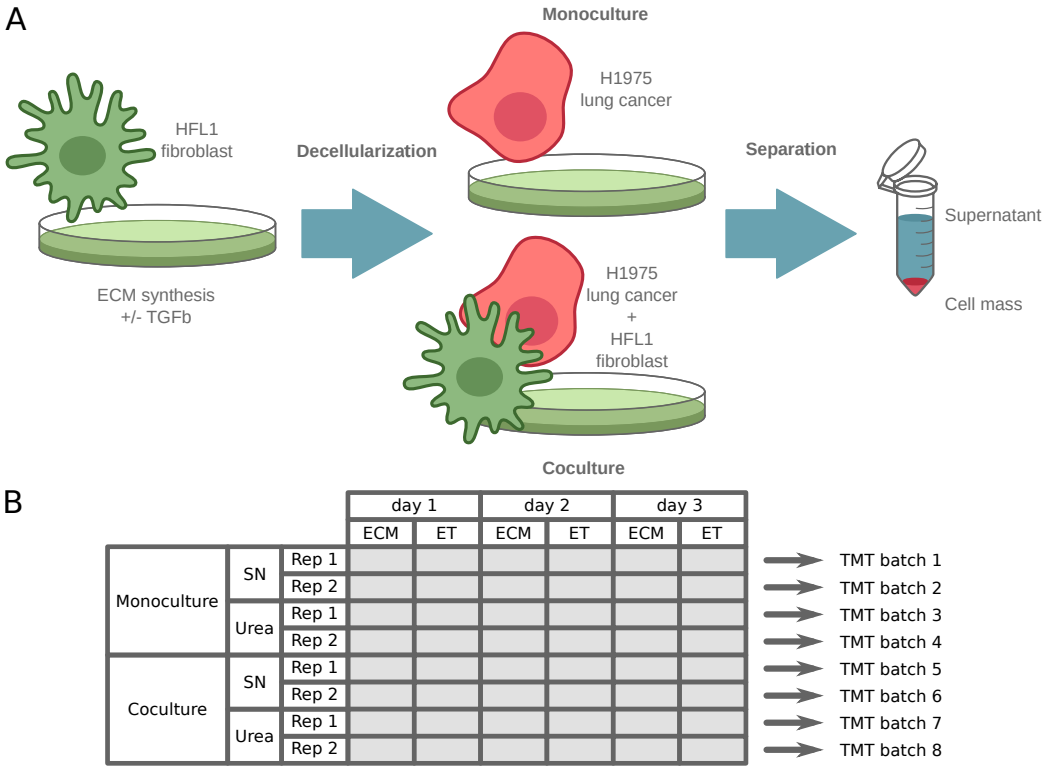

Figure 3: (A) Graphical representation of the experimental design used to analyze the intercellular communication between cancer cells, fibroblasts and ECM. (B) Disposition of the samples across the different TMT batches analyzed using mass spectrometry (MS).

Given the nature of the experimental design, batch effect removal was not performed. Since each of the TMT-MS batches contained different cell culture samples each, application of batch effect removal would also remove biologically-relevant information. In this study I modeled the protein abundance across time, based on linear models for each of the different conditions. As a first step, a linear model was fit for each protein and condition based on the measured $\log_2$ intensities and time points. The model parameters were

optimized using the least squares method (self-implemented). The production rates (i.e. slopes of the linear models) were then compared between the TGFβ-treated ECM and the normal ECM samples by using Welch's t-test for unequal variance. This allowed us to identify the significantly changing proteins in the TGFβ-derived ECM cultures when compared to a normally-generated ECM. From the list of significantly changing proteins, I sought to reconstruct the ligand receptor network between the cancer cell lines and the ECM or fibroblasts in both cell lysates and SN samples. Therefore, I retrieved the ligand-receptor network from the OmniPath database [40], which has been recently updated with several resources focused on intercellular interactions, and retrieved any interaction that contained any of our significantly changing proteins. From the resulting network, any interaction involving non-measured proteins was removed as well as intermediate nodes with degree one (i.e. only one incoming or outgoing edge). Thanks to this I was able to reconstruct the ligand-receptor network based on the potential interactions derived from the significantly changing proteins. Ultimately, this allowed us to obtain interesting and meaningful insights in the intercellular communication of lung cancer cells with the ECM and the associated fibroblasts. Our results highlight important changes on the ECM as well as the cancer cell lines. Cells growing in the TGFβ-derived ECM showed an increased malignant phenotype by undergoing EMT. Furthermore, this phenotype was further strengthened when cancer cells were cocultured with fibroblasts.

1.7 Exploration of cellular communication in cancer over different spatial contexts

Cells sense themselves as well as their environment to adapt their responses to a plethora of stimuli. Sensing the environment includes neighboring cells, both close and distant. Cell-to-cell communication is the cornerstone of tissue homeostasis. This still holds true even in a disease context like cancer. Corollary, the spatial context plays an critical role in cell response and communication. In this project, I analyzed spatially-resolved data from clonal cell spheroids generated from different cancer cell lines. Given the complexity of factors affecting cell communication, studying a single cell type allows us to better study these processes by eliminating one layer of intricacy.

I obtained spatially-resolved data from four cell lines that were analyzed

with imaging mass cytometry (IMC) as described in [46]. The original study aimed to analyze the environmental effects on four different cancer cell lines. The new IMC technology allows the analysis of spatially-resolved samples (slices) by direct laser ablation coupled to mass spectrometry. This allows for highly multiplexed measurements at single-cell resolution. In order to do so, spheroids were grown from different seeding concentrations (i.e. spheroid sizes) and different time points (growth times). Spheroids were then cryogenized and cut into slices. The slices were then labeled with a panel of antibodies for 33 different cellular markers including major pathways and cellular state. Finally the stained slices were ablated and analyzed using image mass cytometry (IMC).

The data set obtained, is comprised of intensity measurements for 33 different markers (Table S1) for each cell in each image as well as the cell position coordinates (x, y position of center of mass). The main aim of the analysis is to reveal the sources of variance of the different measured markers according to their spatial context. Therefore I applied a recently published explainable machine learning algorithm developed in our lab: the Multiview Intercellular SpaTial framework (MISTy) [39]. This framework allowed me to analyze the effect of different spatial contexts on cellular signaling variance. This tool builds a machine learning model for each spatial context (referred as a view) and image (spheroid slice). There are three main views that are considered: intraview, which considers the effects of intracellular markers or measurements within a given single cell, the juxtaview models the effects of the direct neighbour states on a cell and the paraview, which considers the broader or distant environment effects as a function of the distance. Each of the modeled views is then merged into a meta-model with standardized importances of the contribution to the variance explained. This allows for an easy aggregation of the models corresponding to a single condition by simple averaging. The approach allows us to extract information on cellular states based on the expression of different markers in the aforementioned spatial contexts.

MISTy was applied to the obtained sample data in order to find significant contributions on the marker expression based on different spatial contexts. Unfortunately, I was not able to obtain such results. After further investigation, the main reason behind the lack of meaningful results was found to be due to the high correlation of the measured markers across the cells. This is, lack of meaningful spatial differences between cells in this experimental system.

Chapter 2

Methods

2.1 Unveiling drug-resistance mechanisms in acute myeloid leukemia

I obtained the data from both patient *ex vivo* and cell line samples from our collaborators (Kristina B. Emdal, Caroline Wigerup, Kristina Masson). All credit from the experimental assays goes to the respective coauthors. For both sample types, I obtained cell viability measurements as well as phosphoproteomic data from Selinexor-treated and untreated samples as control. Cell viability was measured in both *ex vivo* and cell lines at increasing concentrations of the drug of interest as well as in combination with the AKT inhibitor MK2206 (in cell lines only). Viability was assessed at 48h and 72h post-treatment for single-treated and combination experiments respectively. For the phosphoproteomic data, cells were cultured with or without Selinexor (control) and analyzed using tandem mass tag (TMT) labeling. The pooled and labeled samples were then measured using liquid chromatography with tandem mass spectrometry (LC-MS/MS).

2.1.1 Dose-response model

Cell viability was evaluated from the patient *ex-vivo* samples after 48h post-treatment for nine increasing drug doses and DMSO as control. Each sample was measured in four independent replicates.

Viability was normalized against control for each replicate independently. Then a dose-response model was fitted using least squares method implemented in `scipy` [42]. The model used is a modified Hill function as follows:

$$R = \frac{mD^n}{k^n + D^n} \tag{2.1}$$

Where R is the relative response (i.e. relative cell survival, range [0, 1]) and D is the drug dose (in or case expressed in nM units). Parameters k, m and n were then fitted to the data as described above. Given that that the data in our model is relative to one, one can derive the formula to obtain the half-maximal effective concentration (EC_{50}) by considering $R = 0.5$ and the obtained parameters for a given sample then solving for D (which will correspond to the EC_{50}). Therefore:

$$EC_{50} = \sqrt[n]{\frac{k^n}{2m - 1}} \tag{2.2}$$

We then considered a threshold of 1000nM for the EC_{50} such that samples whose value was below are classified as responders and non-responders otherwise. The choice of this threshold value was based on the study by Crochiere et. al [11], which showed ¿80% XPO1 occucpancy by Selinexor in different cell lines. This corresponds to approximately the recommended phase two dose (RP2D) assessed during the Selinexor clinical trials.

2.1.2 Data normalization

Taking the raw intensity of all samples after the TMT-based MS analysis, I applied the variance stabilization normalization (VSN) method [19] on each batch of samples independently. The reason behind this is that if one were to apply the normalization over all the data set at once, the batch effects would become a confounding factor when applying the model-based normalization (explained below). Similarly, if one were to apply the batch correction before the normalization, the batch-correcting model transformation would be negatively affected by the non-normalized data.

VSN applies a general normalization based on a non-linear transformation of the data such that mean and variance of the features (measurements) are independent. The transformation is defined by the following function:

$$h(x_i) = \gamma \operatorname{arcsinh}(a + bx_i) \tag{2.3}$$

Where x_i is the vector of measurements from the feature i and γ, a and b are the model parameters which are learned from the data. The data transformation performed by VSN renders the data more suitable for multivariate analysis methods. The transformation defined in Equation (2.3) approximates to a logarithmic transformation for high intensity values, while low intensity values (close to the limit of detection) are contracted towards zero. For more details about the transformation and parametrization of the normalization model cf. Huber et al. 2002 [19]. Although the method was initially developed for microarray data, since phosphoproteomic data shares a similar structure, this approach is equally effective.

After normalization, the batch correction procedure implemented in the R package `limma` [34] was applied. This method basically fits a linear model to the data (by using least-squares method) considering the different batches and then removes that component from the data. The process described above was applied for both the *ex-vivo* and the cell line samples.

2.1.3 Differential phosphorylation analysis

Since the VSN method approximates to a logarithmic transformation (increasing with the raw intensity) one can consider the normalized values to be in logarithmic scale. I then computed the differential phosphorylation of our samples using the `lmFit` function from the `limma` R package [34] and computed the t- and p-values using the empirical Bayes method (function `eBayes` from the same package).

I compared the treated versus untreated samples for responders and non-responders independently in order to identify the phosphorylation effects of the treatment in both response groups. I also compared the responders against non-responders prior to treatment in order to identify differences in the baseline phosphorylation state between the response groups. The comparisons (or contrasts) between two given groups were computed by subtraction of their normalized $\log_2$ intensities. Considering the basic property of logarithms:

$$\log_n(\frac{a}{b}) = \log_n(a) - \log_n(b) \tag{2.4}$$

We therefore obtain the $\log_2(FC)$ values of a given condition a vs. another condition b (e.g. treated vs. untreated).

A phosphosite was considered to be significantly regulated if $|\log_2(FC)| > 1$ and uncorrected p-val ≤ 0.05.

Similarly, I contrasted the cell lines after treatment against untreated for each one individually as well as grouped by sensitive (GDM1 and MV411) or resistant (PL21 and NOMO1) cell line groups.

Biomarker analysis

Considering the sets of differentially regulated (up/down) phosphosites on the different response groups, we sought to search for potentially relevant biomarkers. This was achieved by comparing the different overlaps between the sets of significantly up- or down-regulated phosphosites in the *ex vivo* samples.

We then defined two major groups, namely predictive and excluder biomarkers regarding their capacity to identify responders and non-responders respectively. For each of the two major groups, two subgroups define each of the overlapping sets of significantly regulated phosphosites in the different computed contrasts of the differential phosphorylation analysis. Figure 4 represents the following subgroups:

- Group 1: Predictive biomarkers
  - Class 1: Phosphosites that are down-regulated in responders after treatment but not in the same case for the non-responders. While at the same time are up-regulated in responders before treatment when compared to the non-responders (i.e. are down-regulated in non-responders before treatment when compared to responders).
  - Class 2: Phosphosites that are up-regulated in responders after treatment but not in the same case for the non-responders. While at the same time are down-regulated in responders before treatment when compared to non-responders (i.e. are up-regulated in non-responders before treatment when compared to responders).
- Group 2: Excluder biomarkers
  - Class 3: Phosphosites that are down-regulated in non-responders after treatment but not in the same case for the responders. While at the same time are down-regulated in responders before treatment when compared to the non-responders (i.e. are up-regulated

in non-responders before treatment when compared to responders).

– Class 4: Phosphosites that are up-regulated in non-responders after treatment but not in the same case for the responders. While at the same time are up-regulated in responders before treatment when compared to non-responders (i.e. are down-regulated in non-responders before treatment when compared to responders).

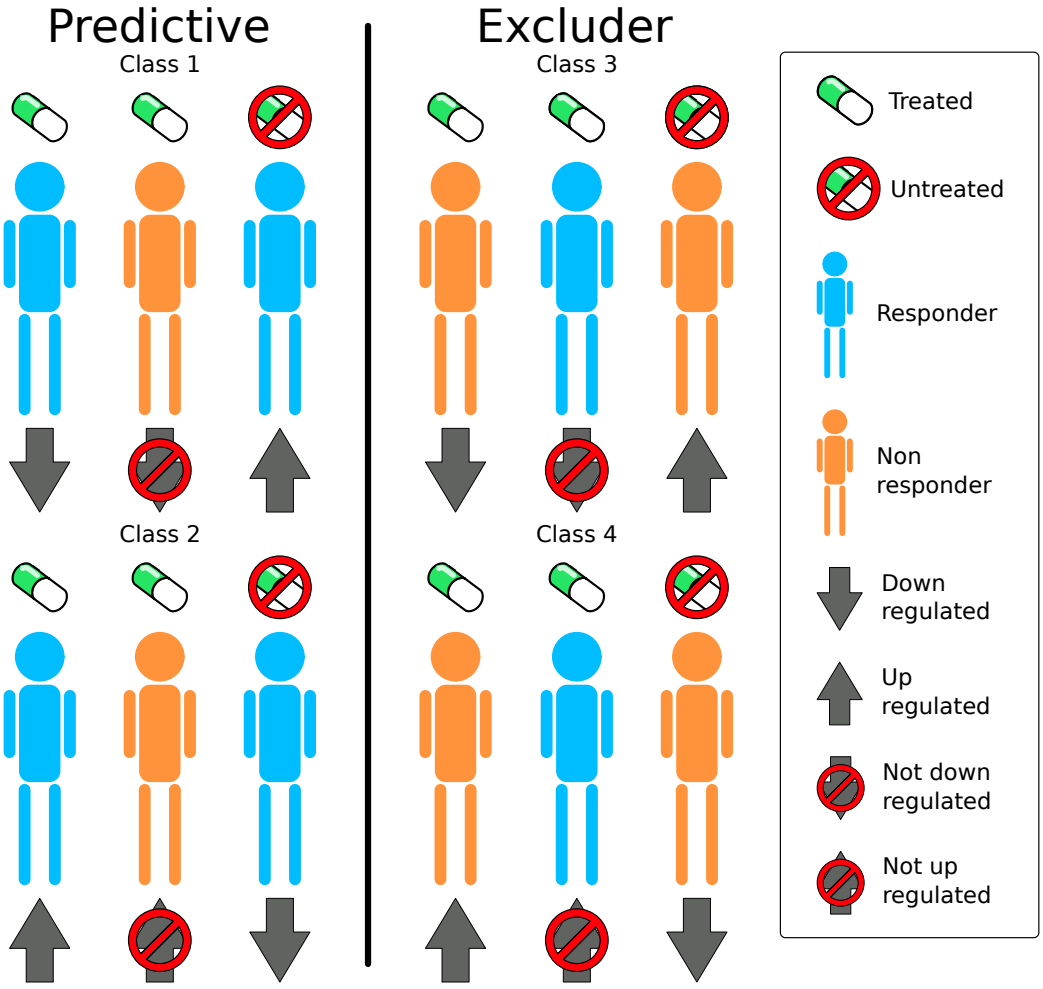

Figure 4: Overlapping group conditions used for the potential biomarker analysis of the differentially phosphorylated sites. On the left are the conditions specified for potential predictive biomarkers while on the right are for the potential excluder ones.

2.1.4 Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA) was computed over the different contrasts in order to obtain functional insights over the differentially phosphorylated sites. The enrichment scores were computed using the R package `piano` [41], which computes the enrichment scores based on different statistical methods and then ranks their results in order to obtain a comparable score for each gene set and compute their consensus (median). The gene sets used in this study were obtained from Molecular Signatures Database (MSigDB) [37]. More specifically the Canonical Pathways from the curated gene sets was used, which contains a total of 2232 gene sets curated from the major pathway databases BioCarta [31], KEGG [23], PID [35] and Reactome [21]. The differentially regulated phoshposites were mapped to their corresponding phosphoprotein and these in turn, to their corresponding member of the gene sets.

Based on the directionality of the phosphosite regulation, the enriched gene sets are classified into different classes: distinct directional (up or down), mixed directional (up or down) and non-directional. For our purposes, only the distinct directional classes were considered, since the enriched gene sets assigned to this class are those that show a clear consensual direction of regulation . The GSEA was applied in both *ex-vivo* and cell line samples results obtained in the differential phosphorylation analysis.

2.1.5 Kinase-Substrate Enrichment Analysis

Kinase-Substrate Enrichment Analysis works in a similar fashion as GSEA but in this case the enrichment is performed over a given set of kinases and their downstream substrates (i.e. phosphoproteins). The KSEA was computed over the differential phosphorylation thanks to the Python package `kinact` [44] which implements the KSEA method proposed in [9]. The kinase-substrate network was obtained from OmniPath [40], a comprehensive database that integrates more than a hundred different biological knowledge resources, including post-translational relationships. Similarly to the GSEA, I also applied this analysis to both *ex-vivo* and cell line samples. An enrichment score was considered significant if p-value ≤ 0.1 (uncorrected).

2.1.6 Drug combination synergy

To quantify the effects of the combination therapy in the cell lines, I computed synergy scores based on our dose response model (Equation (2.1)) by using the isobole method [38] originally presented by Loewe. This method allows us to compute the theoretical response of a combination of drugs as if their effects were independent and purely additive. Therefore, considering two drugs a and b, one can equate their response equations to find the theoretical dose of b that would match the same response level of a given dose of a:

$$R_a = R_b = \frac{m_a a^p}{k_a^p + a^p} = \frac{m_b b^q}{k_b^q + b^q} \tag{2.5}$$

The parameters correspond to those described previously for Equation (2.1). Solving for b as a function of a:

$$b(a) = \sqrt[q]{\frac{m_a a^p k_b^q}{k_a^p m_b - m_a a^p + m_b a^p}} \tag{2.6}$$

Considering purely additive effects, the equation above gives us the equivalent dose of b that achieves the same effect as the given a. Therefore, the expected response of a given combination of a and b can be assessed as the response computed from the model of R_b fitted to the single treatment data of b and then using a concentration of $b + b(a)$ for any combination of b and a. Finally, for each drug combination, I subtracted the expected response from the actual response measurement to obtain the differential response (i.e. synergy score). Note that the response variable R denotes the relative cell survival (between zero and one) and that one can consider as a positive effect its reduction (cell death). Therefore, a positive value denotes drug synergy (superadditivity) while a negative value shows antagonism (subadditivity).

2.2 Discerning cancer cell communication with associated fibroblasts and the extracellular matrix

I obtained peptidomics data from the different samples described above (Figure 3) from Magdalena Szczygiel (who performed all the experiments). The

samples were generated starting by growing a lung fibroblast cell line which was cultivated with and without (control) TGFβ to generate ECM. The cultures were then decellularized and lung cancer cell lines where seeded alone or in combination with the fibroblasts. Furthermore, cancer cell lines were labeled with Cell Type specific labeling with Amino acid Precursors (CTAP) in order to differentiate peptides belonging to the cancer cell lines from those of the fibroblasts or ECM. From the different cultures, samples were taken for three consecutive days and separated into supernatant (SN) and cell lysates (solid phase). The different samples were then measured using TMT-based proteomics with LC-MS/MS.

2.2.1 Data processing

Due to the nature of the experimental design, removal of batch effect would prove ill-advised since one would be removing sources of variance that could hold biological relevance. To circumvent this, I sought to compare the different samples based on their rates of production on the assumption that the main differences across batches affect the scale of the intensities and that these are constant within the same batch. Therefore, as will be described below in more detail, linear models were fitted for each protein across time and compared between conditions based on their slope, disregarding the intercept. Also, to minimize noise in the data and ease interpretability, we resolved to merge the peptide intensities by summing the raw intensities of all peptides pertaining to the same protein. This was done independently for the peptides containing CTAP and the ones without. Prior to this step, those peptides without Lysine were removed from the data set. This step was critical to avoid any bias towards unlabeled peptides since CTAP binds only to this specific amino acid. After merging the peptide intensities into our protein abundance proxy, data was normalized using the Variance Satbilization Normalization method (VSN) [19]. The normalization was applied on each batch of samples independently to avoid introducing biases due to batch effects.

2.2.2 Regressing protein dynamics

In order to be able to compare the effect of the different conditions in the protein expression in our samples, a linear regression was computed for each protein (labeled and unlabeled independently), for every replicate and condi-

tion. The regression was computed with the least squares method with time and $\log_2$ intensity as the independent and dependent variables respectively into a general linear model $y = mx + b$. Where y is the $\log_2$ raw proxy intensity of a protein, x the time point and m and b the slope and intercept to be fitted respectively. The slope of the model is obtained from the division of the covariance of x and y by the the variance of x:

$$m = \frac{SS_{xy}}{SS_{xx}} \tag{2.7}$$

Which are defined as:

$$SS_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n} \tag{2.8}$$

$$SS_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} \tag{2.9}$$

The intercept is then calculated as:

$$b = \frac{\sum_{i=1}^{n}(y_i + mx_i)}{n} \tag{2.10}$$

To assess the quality of the models, I also calculated the root mean square error (RMSE) as well as the coefficient of determination (R^2) as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - p - 1}} \tag{2.11}$$

Where $\hat{y}_i$ is the predicted value of y and p is the number of model parameters related to the independent variable (m, one in this case). The coefficient of determination is defined as the squared value of Pearson's correlation coefficient R. Models whose $RMSE \geq 1.25$ were disregarded for further analyses. The choice behind this threshold was deemed proper given the scale of our dependent variable ($\log_2$ intensity) for discarding those models with poor fit.

$$R^2 = \frac{SS_{xy}^2}{SS_{xx}SS_{yy}} \tag{2.12}$$

The variance of y (SS_{yy}) is computed in a similar fashion as for x (2.9).

Finally, I compared the slopes of the models between TGFβ-treated ECM samples against the slope of the models of the normal ECM using the Welch's t-test for independent samples with unequal variance. The t-statistic is therefore computed as follows:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{2.13}$$

Where the subindex 1 refers to the TGFβ-treated ECM samples and 2 to the normal ECM samples, $\bar{x}$ refers to the mean slope, σ^2 to the standard deviation and n the number of samples (replicates).

Therefore, the null hypothesis (H_0) is that both trends are not significantly different (i.e. the treatment did not affect the dynamics of the protein). The alternative hypothesis (H_1) is that the slopes are different and therefore the TGFβ-induced ECM has affected the protein dynamics.

The proteins whose absolute slope difference was below the 10% quantile (0.013) or whose uncorrected p-value was greater or equal to 0.1 were regarded as non-significant and therefore excluded from further analysis.

2.2.3 Reconstructing the ligand-receptor network

The ligand-receptor interaction network was downloaded by using pypath, the Python interface for the OmniPath database [40]. Pypath directly provides the subset of ligand-receptor interactions from the complete protein-protein interaction network curated from the integration of more than a hundred resources. From the list of all interaction pairs (edges), I filtered those interactions where either the ligand or receptor were present in our set of significantly changing proteins obtained in the previous step. The resulting network therefore contained both our significantly changing proteins as well as any other ligands or receptors interacting with them. I refer to these nodes not pertaining to our set of significantly changing proteins as intermediate nodes. This procedure was done separately for the mono- and coculture experiments. To reduce the network to the most relevant interactions, only the

intermediate nodes with degree (number of edges linked to a node) above one were kept. By using this filtering, I only keep those intermediate nodes that connect two or more of our significant proteins. Furthermore, I removed any of the intermediate nodes involving proteins that were not detected in the corresponding experiments (either mono- or coculture respectively). The reason behind this reduction was made on the basis that if any of the intermediate proteins was not detected in the experiment, none of the interactions involving this protein could not take place. The resulting network was then visualized with Cytoscape software [36].

2.3 Exploration of cellular communication in cancer over different spatial contexts

Data was obtained from Vito Zanotelli, who performed the experiments and presented in the original research study [46]. In this study, three-dimensional spheroids were grown from clonal cell lines in different conditions described below. Slices of these speroids were analyzed using a metal-based barcoding method with antibody-based multiplexed IMC [48, 8].

2.3.1 Experimental setup

Individual sphereoids were grown from four different cell lines, namely 293T (aka HEK293), DLD1, HT29 and T47D. These cell lines were derived from human embryonic kidney cells (293T), colorectal cancer (DLD1 and HT29) and breast cancer (T47D) respectively. These cell lines were grown in spheroid microplates in growth medium. In order to study whether spheroid size and/or growth time affects the results, three different cell concentrations (relatively 1.0, 0.5 and 0.25) and two time points (72 and 96h) were grown. Once grown, spheroids were cryogenized and sliced. The slices were then conjugated with a panel of isotope-labelled antibodies for different markers. In total, 459 slice images were analyzed for 33 different markers, including key signaling proteins and phosphosites (see Table S1).

2.3.2 Computational analysis

I applied Multiview Intercellular SpaTial framework (MISTy) [39] to the intensity measurements of each image. MISTy models interactions between

markers coming from different spatial contexts based on their contribution to the variance explained of the data. For example, interactions within the local cellular niche, or interactions within the broader tissue. Each spatial context is captured by a view. Each MISTy view is considered as a potential source of variability in the measured marker expressions. Each view is then analyzed for its contribution to the total expression of each marker and is explained in terms of the interactions with other markers that led to the observed contribution.

Considering $\mathbf{Y}_{u,v}$ as a matrix of cells $u = 1...n$ with (associated positions X_u) and marker measurements $v = 1...m$ the model corresponding to the vector of measurement of a given marker k follows the expression:

$$Y_{u,k} = \beta_I + \beta_0 f_0(Y_{u,\forall p \neq k}) + \sum_w (\beta_w f_w(g_w(X_u, Y_{u,\forall p \neq k}, T))) \tag{2.14}$$

Where f are the functions that define the models of the different views (random forests in this case) and g are functions defining the spatial contribution of measurements Y at the corresponding positions X and which can depend on other properties T (like cell types or other features). The β parameters control the weights of each view in the model.

The models are built for the following views:

- **Intraview (f_0):** The expression of each marker is modeled as a function of the expressions of the other markers within each cell.
- **Juxtaview (f_1):** The expression of each marker within a cell is modeled as a function of the expression of other markers in the neighboring cells. This is defined by:

  $$g_1(X_u, Y_{u,\forall p \neq k}) = \sum_{j \in N_i} Y_{j,\forall p \neq k} \tag{2.15}$$

  Where N_i is the set of neighboring cells of cell i. Neighboring cells are inferred based on a Delaunay triangulation of the cell positions of a given image, followed by removal of edges with length larger than the q percentile of all pairwise cell distances (where q is user-defined).
- **Paraview (f_2):** Expression of a marker in each cell is fitted to the weighted expression of other markers from of the surrounding cells in

that sample. The weighting is based on a radial basis function (RBF) as:

$$g_2(X_u, Y_{u,\forall p\neq k}) = \sum_{j=1}^{u} \exp(-\frac{d_{ij}^2}{l^2}) Y_{j,\forall p\neq k} \tag{2.16}$$

where d_{ij} is the Euclidean distance between cells i and j and Y_j is the expression vector of cell j. The parameter l controls the shape of the RBF and corresponds to the radius around the cell were the weighted expression have a substantial contribution (user-defined).

These three views are then merged into the global model for a given sample (image) as defined in (2.14). These are then aggregated based on their corresponding condition (cell line, concentration and time point) into a meta-model. Since the importances of each model are standardized, these can be aggregated by averaging. For more details regarding the importance weighting and aggregation of meta-models refer to [39].

To assess the performance of the models, I computed the improvement in the Root Mean Square Error (RMSE) of the multiview models when compared to the intraview model alone of each marker.

In order to explore different parameter settings, I tested different combinations of q and l in order to find the optimal predictive improvement based on feasible values. The set of parameters tested are $q \in \{8, 9, 10, 11, 12, 13, 14, 15\}$ and $l \in \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$.

Chapter 3

Results

3.1 Unveiling drug-resistance mechanisms in acute myeloid leukemia

3.1.1 Drug response classification

After fitting the dose-response models for each sample (see Supplementary Figure S2), EC_{50} values of the samples were computed as defined in (2.2). After considering the threshold of 1000nM to distinguish responders from non-responders, I obtained nine responders and eleven non-responders. Figure 5 shows the different EC_{50} values inferred for the different samples.

3.1.2 Potential response biomarkers

I first computed the differential phosphorylation between treated and untreated samples in responders and non responders as well as responders vs. non-responders prior to treatment (see Supplementary Figure S1). The analysis revealed 31 significantly phosphorylated sites in responders, 21 in non-responders and 163 when comparing responders with non-responders prior to treatment.

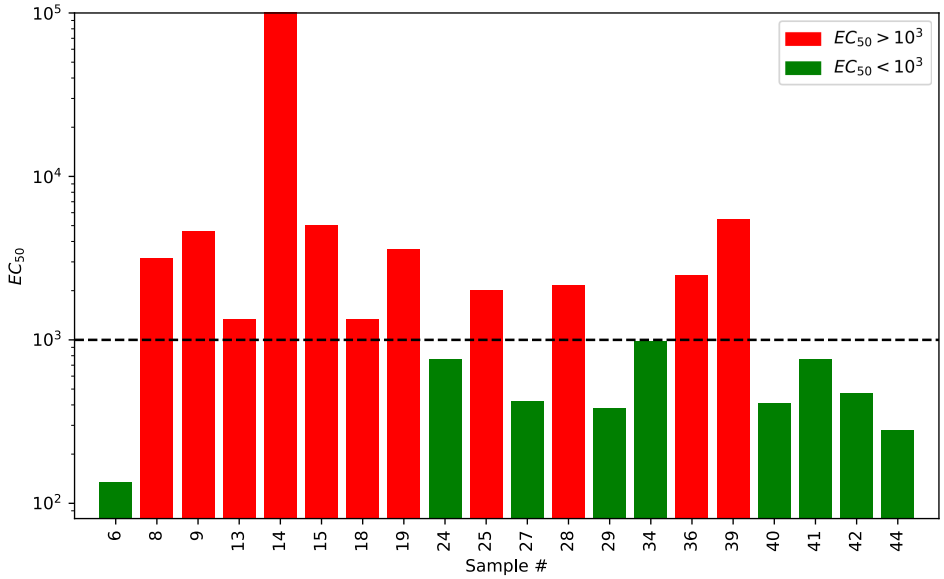

Figure 5: Inferred EC_{50} values from the dose-response models of the different samples defined in equation (2.2). Dashed line denotes the response threshold (1000nM) to separate responders (in green) from non-responders (in red). Note that the y axis is in log scale.

After obtaining the significantly phosphorylated sites, I sought to find which of them could be potentially interesting biomarkers (cf. Biomarker analysis in Methods section). The following phosphosites were retrieved from the results of this analysis on the *ex-vivo* samples: For the predictive biomarkers (Class 1): CDK12_S423, FUS_S221, ZC3H13_S1265, MYC_S344 and AKAP12_S286. For the excluder biomarkers (Class 3): TNKS1BP1_S1545; (Class 4) SUN2_T107, PRRC2C_S633, RREB1_S1238, CWC22_S903, RBBP6_S780, MLLT1_S475 and SRRM2_S1831. None was found for Class 2 as well as for the cell line samples.

3.1.3 Functional analysis of drug response

Following the differential phosphorylation analyses, I applied Gene Set Enrichment Analysis (GSEA) in order to functionally contextualize these results. A summary of the top enriched gene sets can be seen in Figure 6.

As expected from the effects of a SINE, both response groups show a consistent positive enrichment of transcriptional and post-transcriptional processes (e.g. splicing). Also common in both response groups is protein ubiquitination (not shown in responders in Figure 6 as "REACTOME E3 UBIQ-

UITIN LIGASES UBIQUITINATE TARGET PROTEINS" appears in 11th position of the up-regulated enrichment). Interestingly, the responder group shows several enriched gene sets related to Zn^{+2} influx up-regulated that did not appear in the non-responder group. Similarly, also shows activation of caspase pathway (15th position in the enrichment). When looking to the exclusively up-regulated gene sets in non-responders, it shows several processes involved in $p75^{NTR}$ signaling (via NFκB) [18] as well as MTOR and NFκB triggering survival signals (not shown). This suggests a potential signaling axis that could explain the resistance of these cells to the treatment. Furthermore, are also enriched several processes related to senescence and cell-cycle arrest. The latter also appearing in responders in lower ranks, although they show more enriched processes related to mitosis.

Looking at the down-regulated gene sets, both groups show ERBB and MAPK signaling pathways as well as circadian clock and mitochondrial biogenesis. Similarly, several processes related to deubiquitination (including UCH proteinases) are also down-regulated in both groups, supporting the idea that protein degradation events are taking place in both groups as a result of drug response. There appears also consistent down-regulation of interleukin signaling (IL17 specially) in both groups. Among the down-regulated gene sets appearing only in responders, stands out the TGFβ pathway and interestingly "KEGG ACUTE MYELOID LEUKEMIA" gene set. In non-responders only, shows down-regulation of JAK-STAT and VEGF pathways.

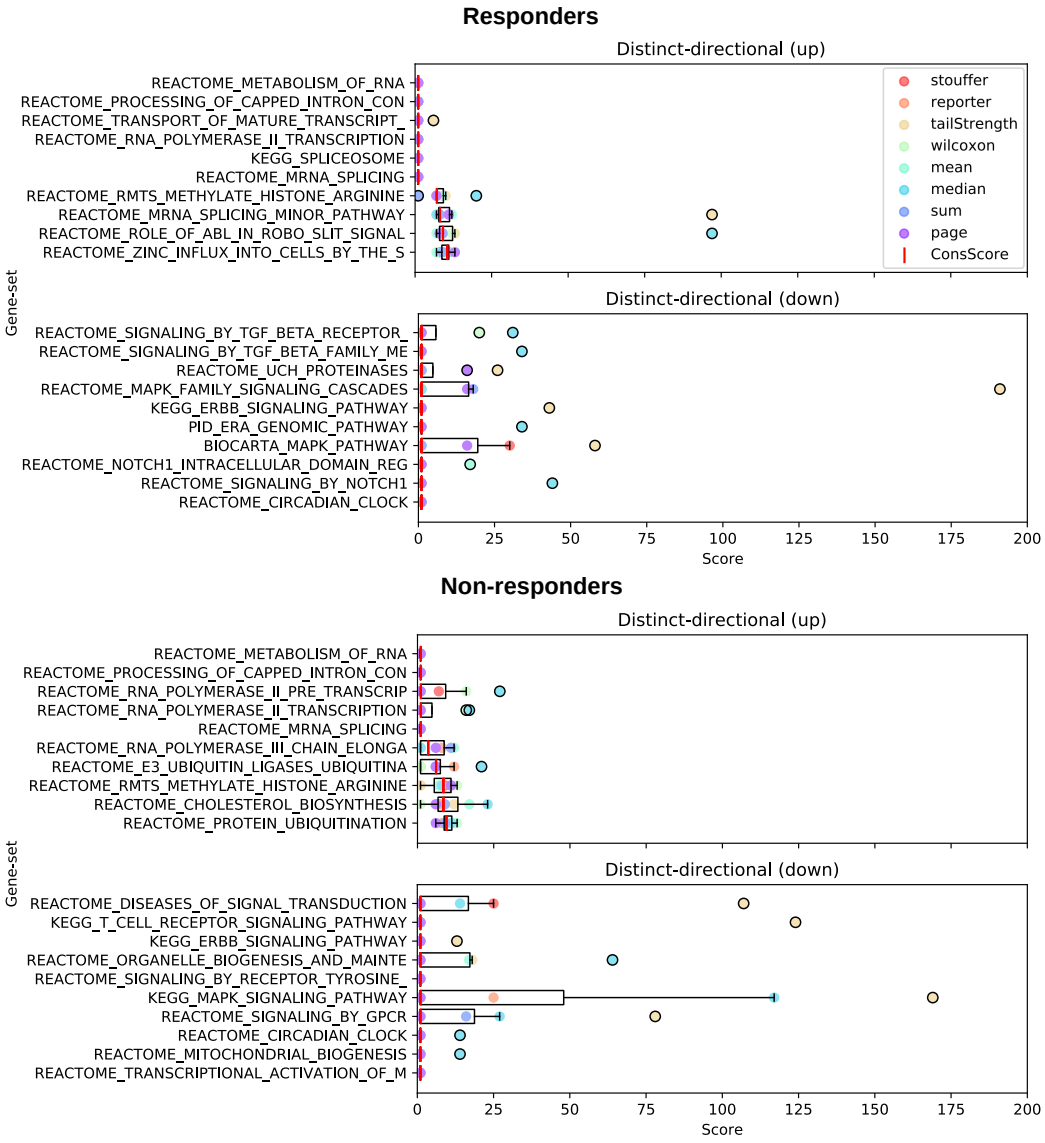

Figure 6: Top enriched gene sets in *ex vivo* samples. From top to bottom: Up-regulated and down-regulated in responders and up-regulated and down-regulated in in non-responders. The colored dots represent the rank of enrichment score of the different methods. The consensus score is represented by the median rank (red vertical bar).

When looking at the GSEA results of the cell lines (Figure 7), first observation is that the enrichment of the different gene sets is less consensual across methods than the observed in the *ex-vivo* samples. This is proba-

bly due to the lower number of samples and differences between the cell lineages. Nevertheless there is still a strong significant consensus. Among the up-regulated gene sets one can find several related to the innate immune system and pro-inflammatory signals (like IL-1) which are likely to be related to NFκB pathway activation. Can also be observed some gene sets related top75$^{\mathrm{NTR}}$ signaling that also appeared in the *ex-vivo* samples. We also observed several enriched processes related to autophagy in both response groups which is a known survival mechanism in AML [5]. Other common processes in both groups seem to be related to viral infection which is likely to be an artifact from immunity-related signaling.

Looking into the down-regulated gene sets also appear several commonalities like protein translation (ribosome, signal recognition particle SRP translation), cell cycle progression or nonsense-mediated mRNA decay. Both sensitive and resistant also show down-regulation of the gene set "REACTOME TP53 REGULATES TRANSCRIPTION OF CELL CYCLE GENES" but only in the sensitive cell lines one can find up-regulation of "REACTOME TP53 REGULATES TRANSCRIPTION OF DNA REPAIR GENES". Interestingly, are also found down-regulation in resistant cell lines, the MTOR pathway gene sets (not shown, enrichment positions 11-13), more specifically MTORC1. Furthermore, IL2-PI3K pathway also appears down-regulated in both groups. This findings further support the results obtained in *ex-vivo* samples, suggesting that AKT-MTOR signaling may play a role in the Selinexor resistance mechanism in AML samples, although the exact mechanism is yet unclear.

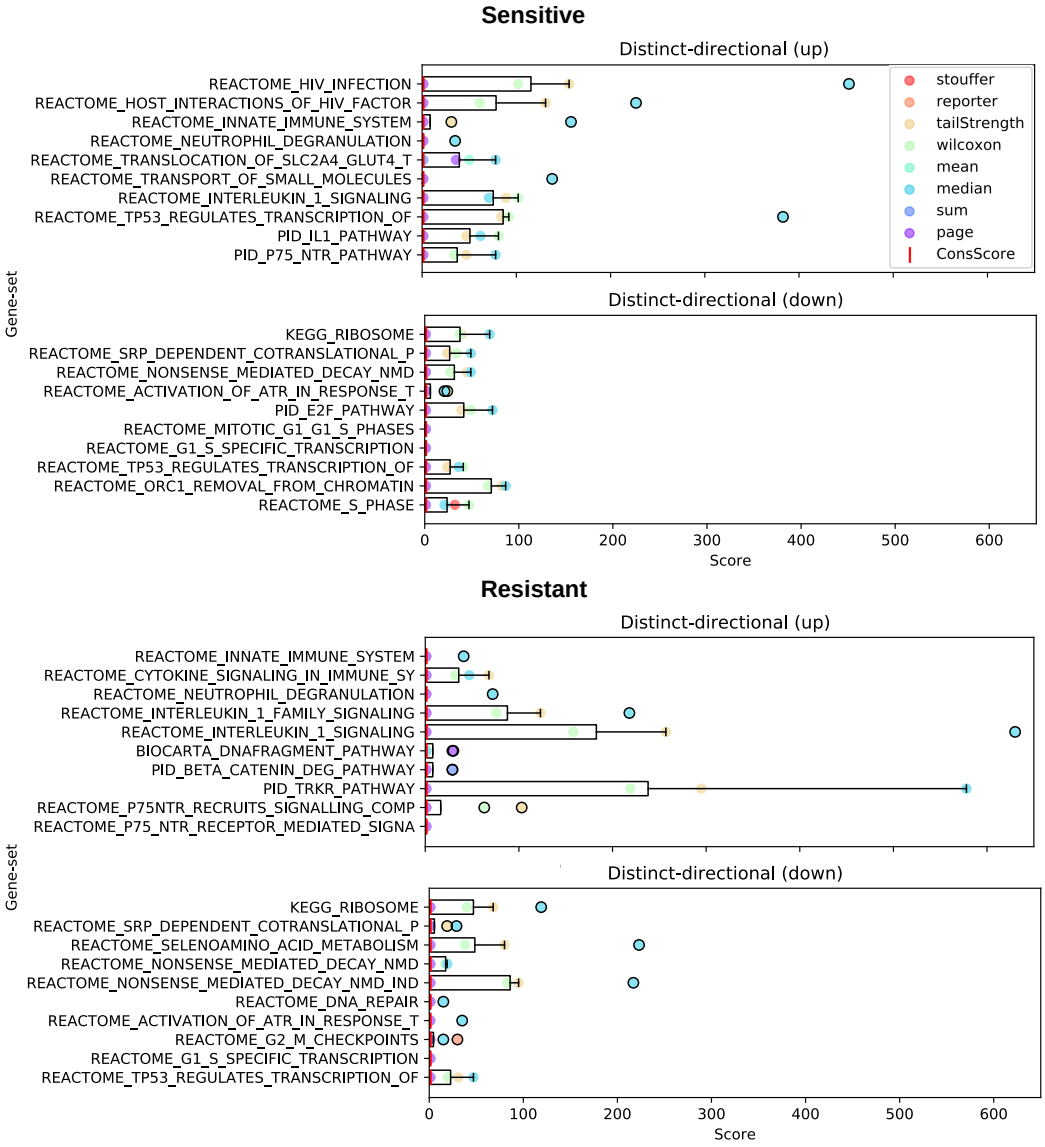

Figure 7: Top enriched gene sets in cell line samples. From top to bottom: Up-regulated and down-regulated in sensitive and up-regulated and down-regulated in in resistant. The colored dots represent the rank of enrichment score of the different methods. The consensus score is represented by the median rank (red vertical bar).

The enrichment of kinases shows few enriched kinases due to drug treatment, as can be seen in Figure 8. Nevertheless, the results are consistent with previous findings on GSEA and literature. Among the negatively enriched

kinases in non-responders can be found MELK which also appears negatively enriched when compared to responders against non-responders prior to treatment (i.e. positively enriched in non-responders). This is consistent with previous findings showing that MELK is overexpressed in AML [4] yet, despite being down-regulated by Selinexor in non-responders, these results indicate that there exists an alternative mechanism of resistance. When looking at the positively enriched kinases we find two members of the casein kinases family, namely CSNK2A1 and CSNK1A1. These have been shown to drive AML survival and progression [29, 10]. In the group of responders, we see negatively enriched kinases HRAS and MAPK1 (aka ERK2) which is consistent with the down-regulation of MAPK pathway observed in the GSEA and is an expected signal for cells halting growth and proliferation and undergoing apoptosis. Among the positively enriched kinases in responders after treatment, we found TNNI3K and CTDSPL. While the former is generally found in heart and associated with cardiac diseases, its role in AML is unknown. On the other hand, CTDSPL is a known tumor suppressor and has been found to be down-regulated in AML via miRNA [47].

Finally, comparing the differences between responders and non-responders prior to treatment, we can see a high number of CDK family members to be enriched in non-responders, as well as other other cell cycle progression-related kinases like PIK3C3, WEE1 or MELK. If one looks to the enriched kinases in responders, one can find several kinases related to cell differentiation, survival and proliferation as well as several immune-related processes like PAK2, SYK and HCK. Interestingly, we also find TRPM7, a known widely expressed ion channel that has shown to permeate zinc (Zn^{+2}) ions [30]. This is consistent with our findings in the GSEA, although the effect of increased Zn^{+2} in Selinexor response is unclear.

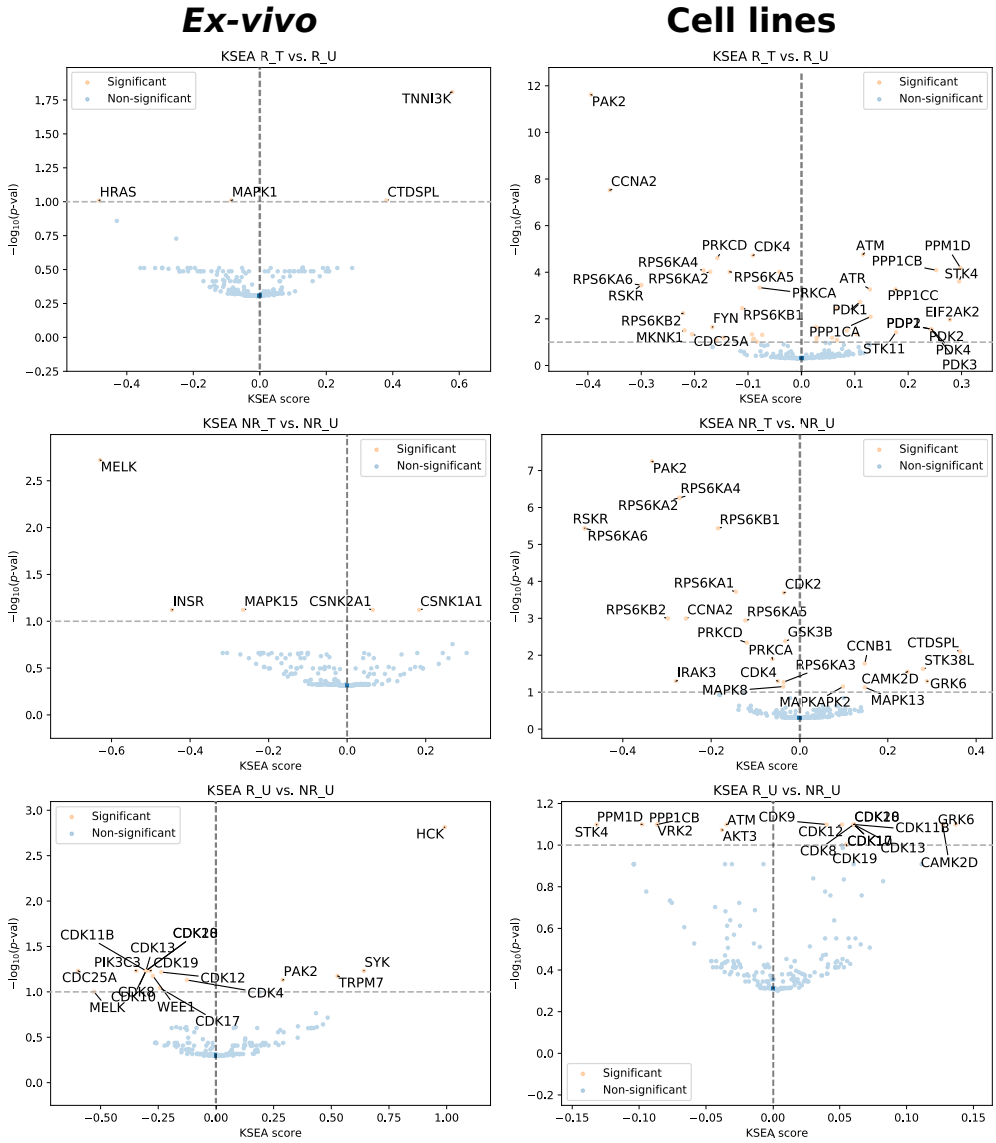

Figure 8: KSEA scores for the *ex vivo* (left) and cell line samples (right) against their $-\log_{10}(p\text{-value})$. In blue are shown the non significant kinases, in orange the significantly enriched kinases (p-val ≤ 0.1). Top left plots show the enrichment in responders/sensitive samples and middle plots show the enrichment in non-responders/resistant samples after treatment (compared to control). The bottom plots show the enriched kinases when comparing the phosphorylation of responders/sensitive against non-responders/resistant prior to treatment. Positive enrichment scores denote up-regulation of a kinase while a negative score corresponds to down-regulation. Only the top 30 enriched kinases display the label names due to space constraints.

When looking at the KSEA results of the cell line samples (Figure 8) three major common down-regulated kinases standing out. First is a group of RPS6 kinases family members, followed by CDK4. Both RPS6 and CDK4 are known to be activated by MTORC1 [6] which was shown to be down regulated in resistant cell lines in the GSEA. Third is PAK2 which is responsible for the internalization of IL2 [16], which is consistent with the findings of the GSEA (down-regulation of IL2-PI3K signaling). Futhermore, there seems to be higher levels of activation of AKT3 in resistant cell lines prior to treatment.

Overall, considering all the presented results, there is strong evidence that the AKT-MTOR signaling axis plays an important role in Selinexor resistance. More specifically, this seems to be achieved via MTORC2 as MTORC1 shows clear down-regulation, which is consistent with increased autophagy, down-regulation of RPS6 and CDK4 [1]. Furthermore, RPS6 is known to inhibit MTORC2, therefore is to be expected that the latter is up-regulated, given the clear down-regulation of RPS6. Nevertheless, there also seems to be down-regulation of PI3K, the canonical activator of AKT, at least via IL2, also consistent with the down-regulation of JAK in non-responder samples. Therefore, the activation of AKT could be achieved by a different mechanism [27]. Other group of well known activators of AKT are PDK family members. Other AKT activators include IKKϵ members, which are also related to viral infection processes, which could explain the enrichment of these terms in the GSEA. Nevertheless, PDK members appear clearly up-regulated in sensitive cell lines (see Figure 8) and also viral infection-related terms in GSEA (Figure 7), which suggests that the resistance mechanism lays downstream of PI3K.

3.1.4 Rational drug combination

After the analyses described above along other results obtained from collaborators, we hypothesized that resistance to Selinexor and therefore avoidance of apoptosis was achieved via AKT-FOXO3 signaling axis triggering survival response. Therefore we sought to investigate whether combination of Selinexor with the AKT inhibitor MK2206 would improve the effectivity in non-responders.

After considering Loewe additivity, I computed the expected survival of the different cell lines for each combination of drug doses. These were then compared to the mean measured survival by subtraction, such that a positive value represents the superadditivity (synergy) of that combination. The

results are shown in Figure 9.

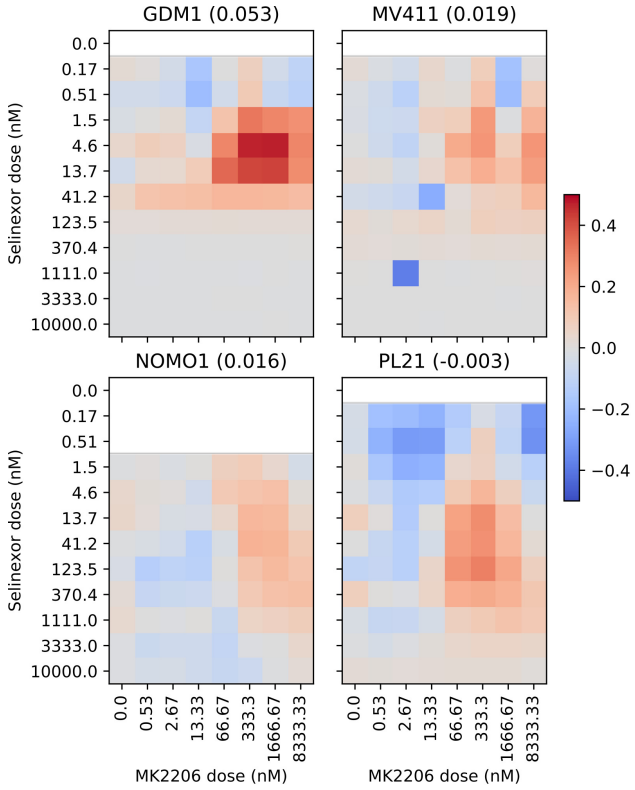

Figure 9: Predicted response (survival) difference with observed response for the different cell lines and combination doses. In parenthesis is shown the mean difference across all the dose combinations. Positive differences denote synergy (predicted survival > observed), negative difference shows subadditivity (observed survival > predicted).

The combination of MK2206 shows better synergy with Selinexor (increased response between 0.2 and 0.4 with respect to the predicted) for medium doses around the 333.3nM for all cell lines. Despite being GDM1 the only cell line showing high synergy, there is a clear effect in all cell lines. More importantly, the effect of this drug combination rendered the resistant cell lines into sensitive. This can be clearly seen in Figure 10, from 333.3nM of MK2206 on, the EC_{50} of the resistant cell lines (NOMO1 and PL21) fall to the same levels as the sensitive cell lines.

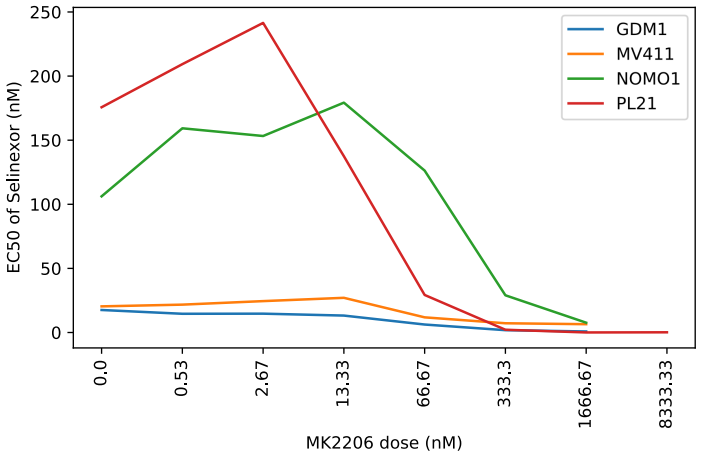

Figure 10: Inferred Slinexor EC_{50} values for the different cell lines and increasing MK2206 doses.

3.2 Discerning cancer cell communication with associated fibroblasts and the extracellular matrix

3.2.1 Differential regulation TGFβ-treated ECM

After processing the data, I computed a linear model for each protein in each condition over the three days. Once the models were computed, I discarded those whose RMSE was above 1.25. Overall, a total of 54.615 models were computed (mean of 3413 models for each condition replicate).Figure 11 shows a comparison of the mean slopes (across replicates) of the protein models between the treated ECM samples (ET, y-axis) against the normal ECM (x-axis).

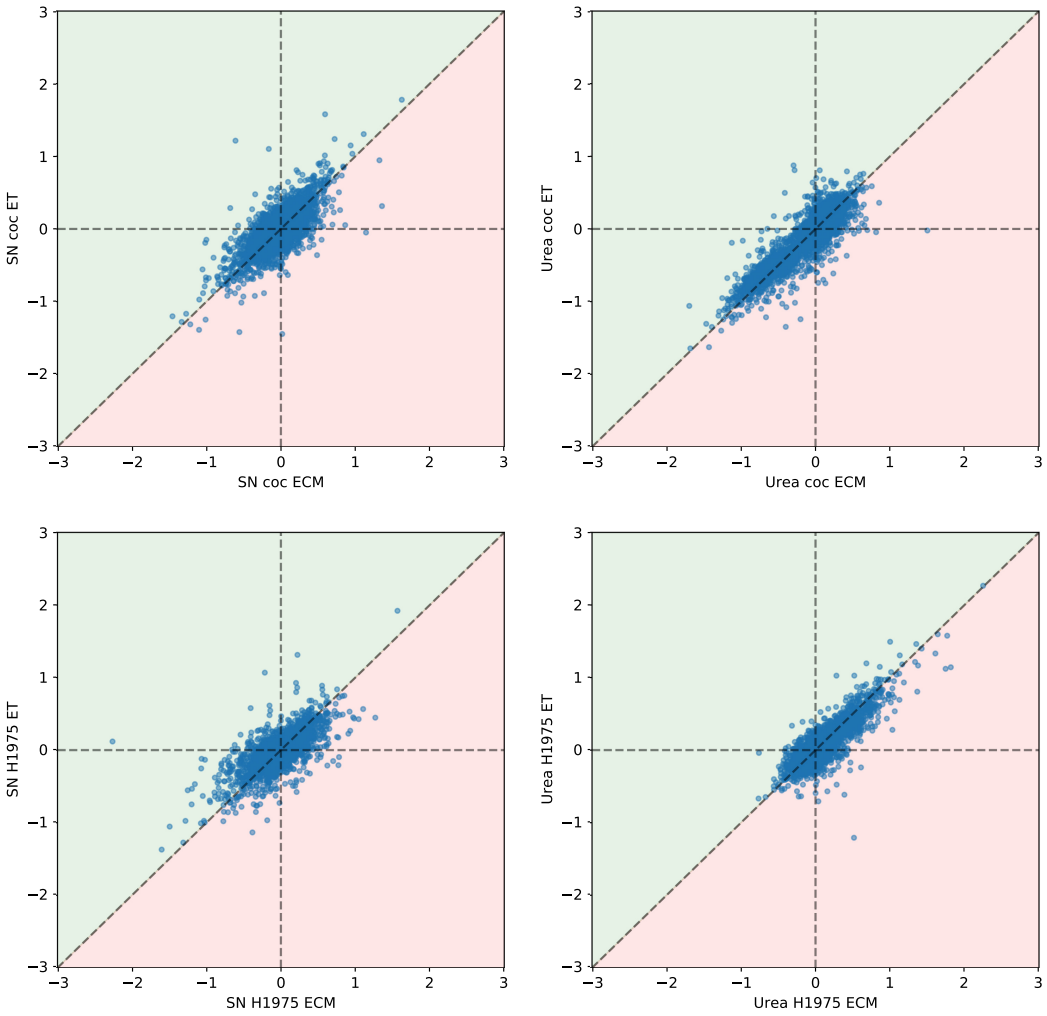

Figure 11: Comparison of the mean slope across replicates between the TGFβ-treated ECM (ET, y-axis) and normal ECM samples. Figures on top row show the comparisons of the coculture while H1975 monoculture on the bottom. Plots on the left show the SN samples while lysate samples are on the right. Dashed lines plotted for reference at slope zero and diagonal for no difference between ET and ECM. Green area denotes slopes which are higher in ET than ECM, red area for the opposite.

I then computed a t-test between TGFβ-treated ECM and normal ECM replicates in order to assess whether the protein rates were significantly different between these. From the 54.615 models, I was able to compute a total of 11.368 slope contrasts ($\frac{54615\text{models}}{2\text{replicates}2\text{treated/untreated ECM}} \simeq 13654$), the actual number of t-tests obtained is slightly lower than expected due to missing

values, which led to missing models in some replicates or ECM conditions. The results of these contrasts are shown in Figure 12 as volcano plots.

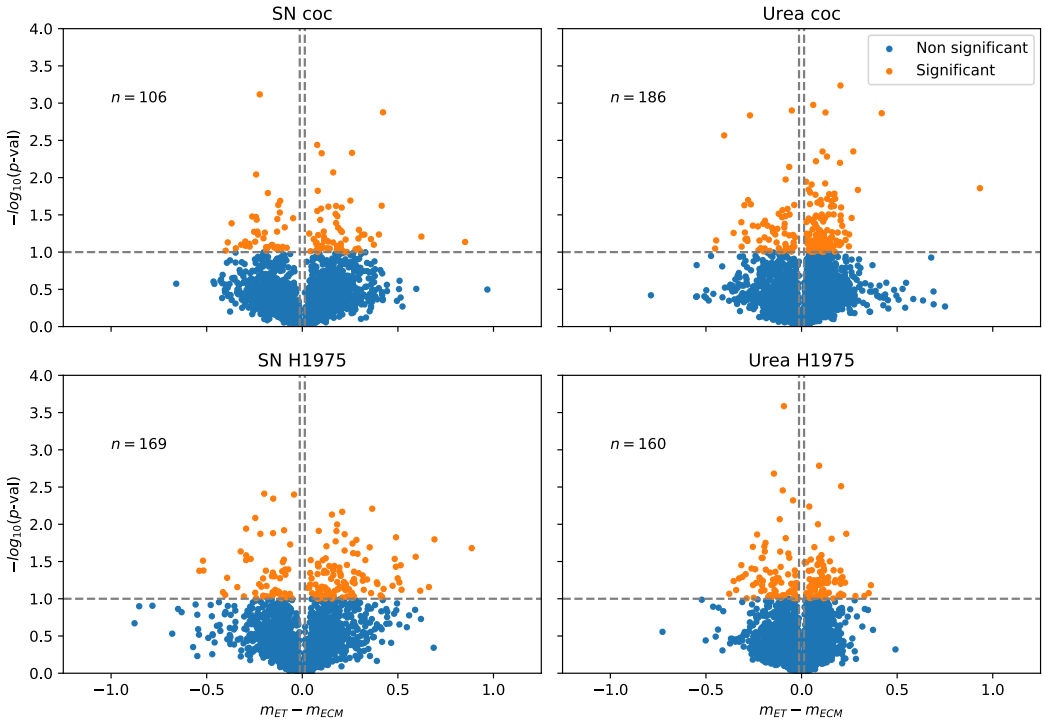

Figure 12: Volcano plots showing statistical significance vs. the slope difference between ECM+TGFβ and ECM. Upper plots show these results in the coculture samples and bottom plots for the monoculture. Supernatant samples and cell lysate samples are shown on the left and right respectively. "n" denotes the number of significantly changing proteins in the different sample types. These are considered as p-value ≤ 0.1 and $|m_{ET} - m_{ECM}|$ (absolute difference) $\geq$ than the global 10% absolute quantile (0.013).

3.2.2 Effects in cell-cell communication

In order to contextualize the previous results regarding cell-to-cell and cell-ECM communication, I extracted the ligand-receptor network based on the list of significantly changing proteins (as described in Methods section). The resulting networks from the H1975 monoculture and coculture with HFL1 are shown in Figure 13.

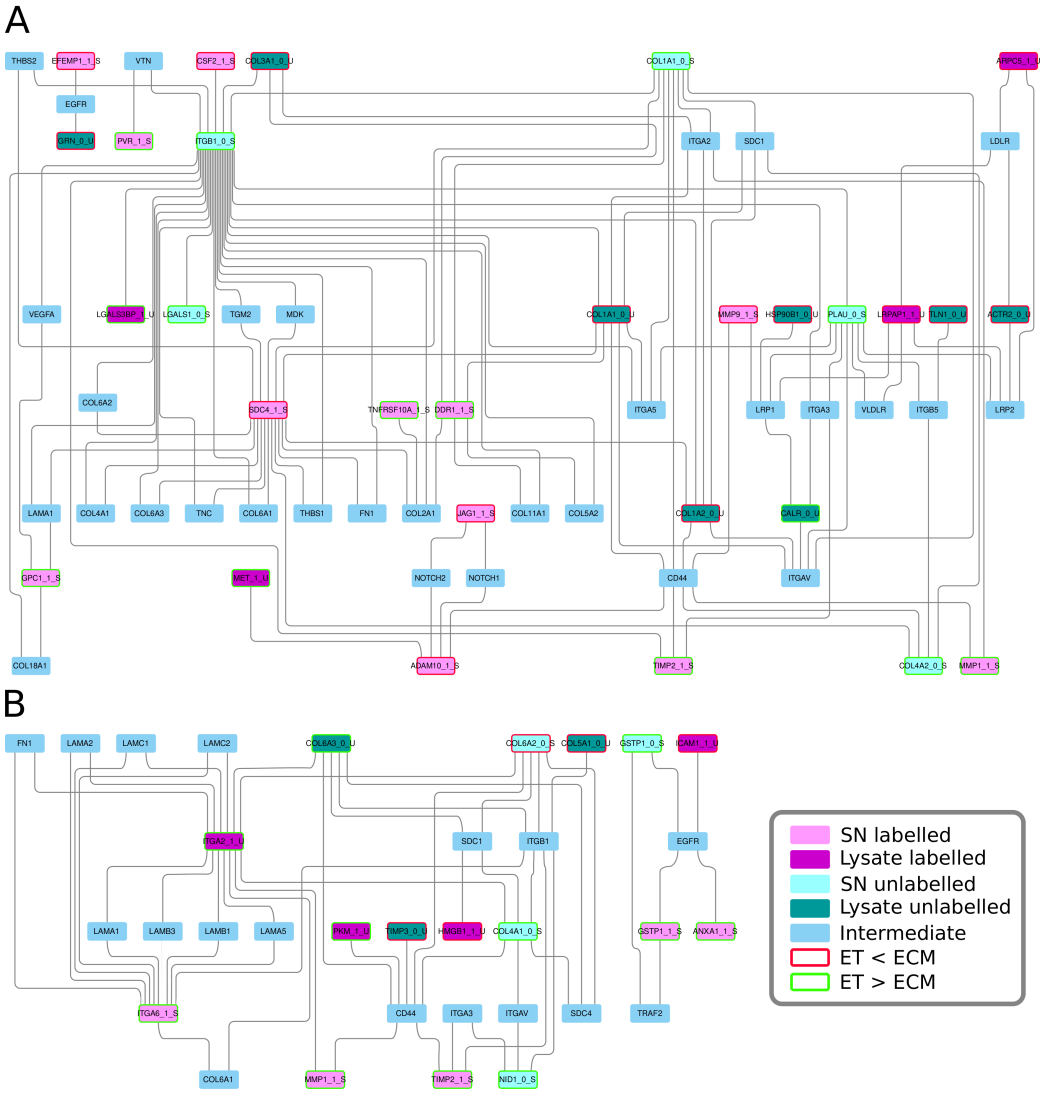

Figure 13: Reconstructed ligand-receptor networks of significantly changing proteins in (A) H1975 monoculture and (B) coculture. Colors represent the different conditions or intermediates (in blue): pink colors denote proteins that are labelled (coming from the H1975 cell line) while turquoise is used for unlabelled proteins (coming from HFL1 or ECM). Dark/saturated colors denote proteins coming from Urea samples (cell lysates) while their lighter version indicates proteins from SN (supernatant/secreted). Significantly changing proteins have their borders colored according to the sign of the contrast between TGFβ-treated ECM (ET) vs. control (ECM), in green when the slope in ET $>$ ECM and red otherwise.

Looking at the cancer cell monoculture ligand-receptor network (Figure 13A), we see increased rates of collagens COL1A1 and COL4A2 on the ECM supernatant while the lysate levels are decreasing (COL3A1, COL1A1 and COL1A2). This may suggest an increased degradation of the ECM in the TGFβ-treated cultures. This is further supported by the increased rate of PLAU, known to cleave plasminogen to its active form plasmin, which in turn is known for its proteolytic activity and remodeling of the ECM, activating growth factors as well as metalloproteinases [3, 26]. Futhermore we see changes in proteinases like increased rate of MMP1 and the inhibitor TIMP2 secreted by the cancer cells while ADAM10 and MMP9 show decreasing rates when compared to normal ECM samples. While several studies generally relate higher expression of MMPs to cancer progression and worse prognosis, expression of TIMP proteins seems to be more context-dependent [20, 26, 32]. Hepatocyte growth factor receptor MET, a well-known oncogene and linked to drug resistance, also shows increased rates in cancer cell lines grown in the TGFβ-treated ECM [12]. Other receptors linked to collagen and ECM interaction are also higher, namely ITGB1 and DDR1.

The ligand-receptor network for the coculture samples (Figure 13B) shows also several changes in collagens. COL6A3 and COL4A1 show higher rates when cells are grown in the TGFβ-treated ECM, while COL6A2 and COL5A1 show decreased rates. We also see increased ITGA2 and 6 from the cancer cell lines which have been related to increased metastatic potential in breast cancer [2]. Similarly as in the monoculture, we see an increase in MMP1 and TIMP2. Furthermore, the inhibitor TIMP3 appears in lower rate in the unlabelled samples, which is known to be the inhibitor of metalloproteinases with highest spectrum, including all MMPs and many ADAMs [15]. Interestingly, cancer cell lines also show increased intracellular levels of pyruvate kinase PKM which is linked to cancer metabolism and can induce EMT by enhancing cell survival and invasiveness. We also find lower levels of high mobility group protein B1 (HMGB1) in cancer cell lysates. As its name indicates this protein can be found in many cellular localizations, from the nucleus to the extracellular space. Its functions are also quite diverse, but among the most relevant are acting as a damage-associated molecular pattern (DAMP), triggering inflammation and immune responses, DNA repair and stability, autophagy and apoptosis [24]. Neverthelessits functions and implications in cancer seem to be highly context-dependent.

Overall, one can conclude that the effect of TGFβ-treated ECM greatly affects the structure and composition of the ECM via changes in the signaling

and secretion from both the cancer cell lines and the associated fibroblasts, therefore altering the cancer phenotype.

3.3 Exploration of cellular communication in cancer over different spatial contexts

3.3.1 Model performance and parametrization

For each pair of parameters, I executed MISTy and computed the median RMSE improvement of all markers in each given condition (set of images corresponding to the same cell line, seeding concentration and time point) after the corresponding aggregation of models in those images. As can be seen in Figures 14B and 15, there is no improvement at all in the models including the juxta- and para- views when compared to just modeling the intraview model. Actually, the performance is decreasing in all cases.

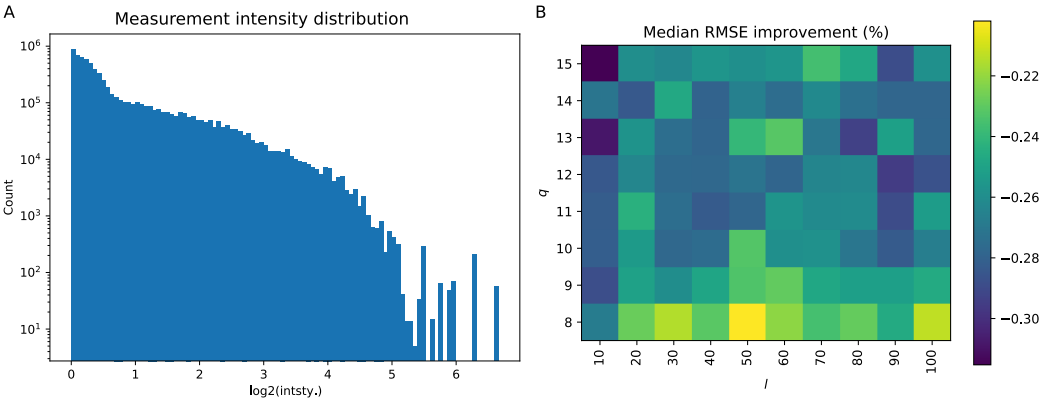

Figure 14: A) Distribution of measurement intensities across all samples and channels. B) Heatmap representing the median of the median RMSE improvement of the models with varying parameter combinations (i.e. the median of each box as shown in Figure 15).

One potential source for the low performance may be related to technical noise in the data. The original study shows a clear relationship between lower model predictability and low-intensity measurement channels (cf. Results section 3.2.1 from [46]). The distribution of measurement values can be seen in Figure 14A, which clearly shows great amount of low intensity values.

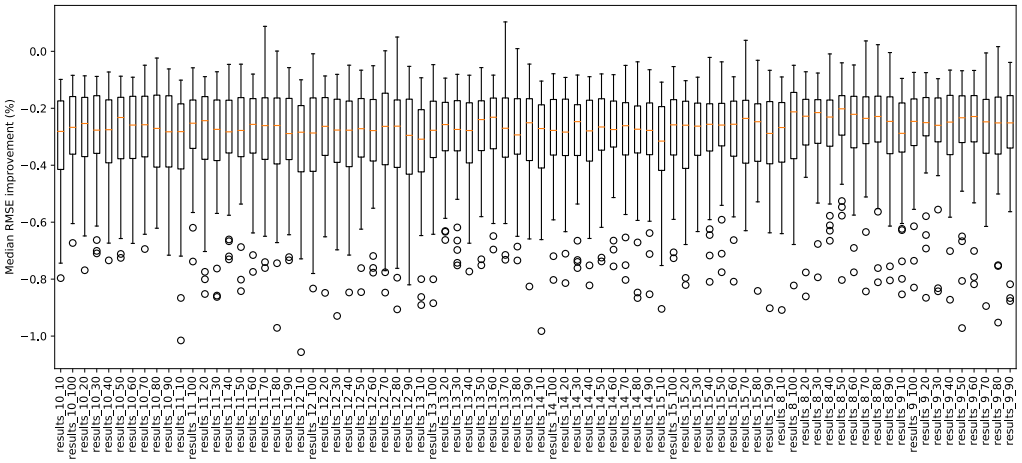

Figure 15: Box plot of median RMSE improvements for each set of parameters. Each column corresponds to the results corresponding to a given of a pair of parameters named as follows: `results_<`q `value>_<`l `value>`.

3.3.2 Performance examination

In order to pinpoint the source(s) of the performances obtained, I examined the variance explained in the models corresponding to the best performing parameters described above ($l = 50$ and $q = 8$). The results are shown in Figure 16. A t-test was performed between the explained variances of intra- and multiview models on each marker. Only three markers appeared significantly different ($p \leq 0.05$) between intraview models and the multiview models, namely STAT3, SAPK/JNK Y183/Y185 and FAK Y397 S1. Despite the significance, these improvements are not notable. In general, these results show that multivew models are not very different than the intraview models alone in terms of variance explained. Clearly, some predictors perform better than others but inclusion of spatially-resolved terms in the model did not improve efficiency.

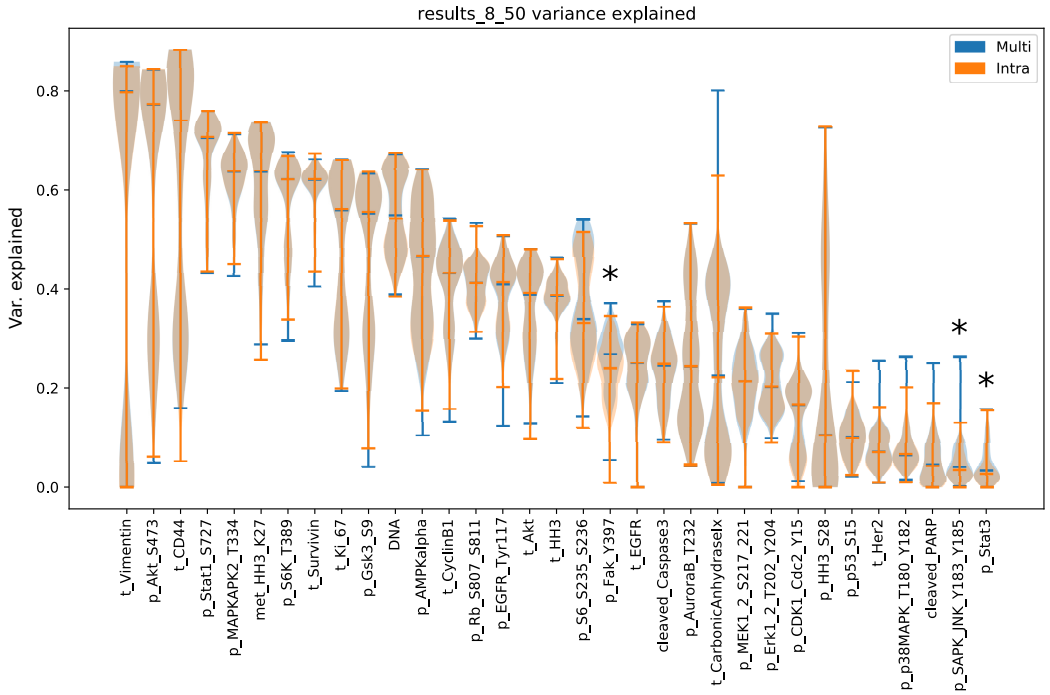

Figure 16: Violin plot showing the distribution of variance explained over the different samples (images) across the markers. In orange are shown the intraview models alone and in blue the whole multiview models. Asterisks denote those markers which are significantly different between intra- and multiview models ($p \leq 0.05$).

It is important to note that I am considering overall performance over the models in different conditions and cell lines. I therefore hypothesized that the approach did perform well in some cell lines or growth conditions but not in others. In order to test this, I computed the variance explained (R^2) difference between the multiview models against their corresponding intraview model alone. I then plotted these in a heatmap with hierarchical clustering and labeled according to the different conditions (cell line, seeding concentration and growth time), shown in Figure S3. Overall, there is no clear clustering of specific conditions towards increasing performance. Nevertheless, a few light trends can be appreciated. Most of the models corresponding to cell line HT29 gather on the top region of the heatmap which shows the highest differences between multi- and intraview models (including both positive and negative differences). On the other hand, the lowest seeding concentration and growth times tend to be more present in the lower part of the heatmap, indicating lower or absent differences between the models in

these conditions.

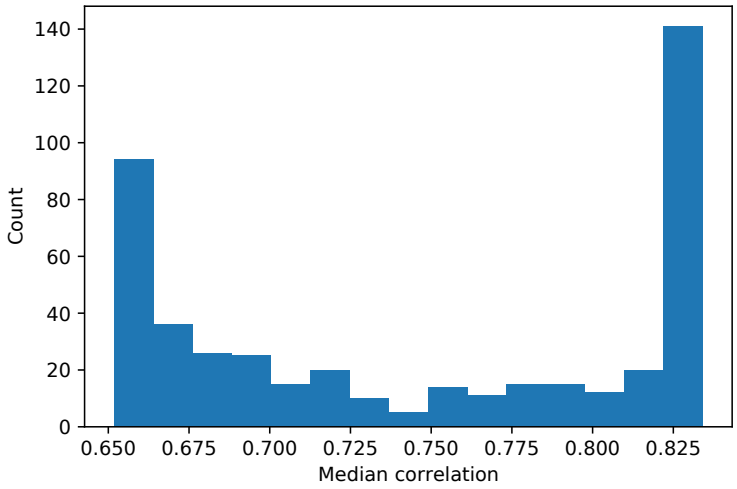

Figure 17: Histogram of correlations of marker measurements between cells and across all images.

Finally, I sought to examine the measurement correlations between cells on each of the samples. I therefore computed the Pearson correlation coefficient between the measurements of each cell in an given image. From the correlation matrix of a sample, I computed the overall median and plotted the distribution of these across all images, shown in Figure 17. Overall, one can observe that the median correlation between cells in an image is significantly high. The median correlation ranges between 0.65 and 0.83, which can be partially explained by the fact of the spheroids being grown from clonal cell lines. Nevertheless, one would expect higher variances due to the differences in spatial distribution of the cells, especially considering how cells are grown. For instance, cells in the inner spheroid should show signs of hypoxia, while cells on the surface would have easier access to nutrients and oxygen.

Chapter 4

Discussion

4.1 Unveiling drug-resistance mechanisms in acute myeloid leukemia

Overall, in this project I applied different modeling and omic data analysis tools in order to identify drug resistance mechanisms in AML. Thanks to this approach, we successfully identified another potential target for combination therapy with Selinexor. Based on the results obtained from the cell line models when combining Selinexor with the AKT inhibitor MK2206, it is clear that there exists synergy between both drugs that has potential to overcome the resistance mechanism described here. Nevertheless, these results need to be further validated *in vivo*. Follow-up experiments would also help pinpointing more exactly the underlying cause of the resistance mechanism proposed here.

Furthermore, the resistance mechanism presented may not be unique or applicable to all the AML cases, as the scope of this study is limited to the amount of samples that were available. The results obtained here are also limited by the experimental procedures, which may include (but not limited to) biological variance between the patient samples, collection, storage and/or processing methods, etc. Further limitations need also to be considered in the data analysis, including the classification of the response groups or choice of significance thresholds, the prior knowledge used or the algorithms applied on the different steps of the analysis. Overall, and to the extent of our knowledge, the choice of methods applied in this study have been based on a combination of state-of-the-art and well-known, widely-used and robust procedures.

In conclusion, the approach presented has proven to be useful to generate clinically relevant hypotheses in order to tackle drug resistance. Based on the results of the rational drug combination shown in this study, our strategy has already shown promising results. Furthermore, this approach can easily be extrapolated to other contexts apart from AML, increasing the value of the work presented here.

4.2 Discerning cancer cell communication with associated fibroblasts and the extracellular matrix

In this work, I present the study of protein dynamics of a lung cancer cell line and ECM both in mono- and coculture with the fibroblasts responsible for its deposition under TGFβ stimulation. By using CTAP labelling, we can easily differentiate between proteins synthesized by the H1975 lung cancer line or from the HFL1 lung fibroblast cell line. Furthermore, the separation of the cultures between the cell lysates (Urea) and supernatant (SN) allows us to differentiate between secreted proteins and those coming from the cells or the ECM.

After applying a linear model to each protein for each condition, I was able to assess the significantly changing proteins after TGFβ treatment when compared to the non-treated samples. Considering all sample types, I found a total of 621 differentially changing proteins (Figure 12). From the TGFβ-induced significantly changing proteins, I reconstructed a ligand-receptor network (Figure 13). Overall, both mono- and cocultures show a clear remodeling of the ECM driven by the cancer cell lines when growing in the TGFβ-treated ECM. Yet, these changes differ considerably between the mono- and coculture samples. On one hand, the cancer cell line monoculture shows a decreased ECM-degradation phenotype. On the other hand, when cocultured with fibroblasts, there is increased metalloproteinase activity. Our results suggest that TGFβ-treated ECM increases invasiveness and malignancy in this cell line model. Additionally, and considering the fact that H1975 has been proven to be a model of partial EMT [22], coculture with fibroblasts shows further promotion of the mesenchymal phenotype both in the cancer cell line and the associated fibroblasts.

In summary, I was able to obtain meaningful mechanistic insights regard-

ing how different tumor microenvironments can affect cancer malignancy and phenotype. The results presented here also suggest potential targets that may revert the enhanced EMT phenotype via disruption of specific cancer cell-ECM interactions. Nevertheless, further investigation will be required in order to pinpoint which of these interactions will provide the most beneficial outcomes.

4.3 Exploration of cellular communication in cancer over different spatial contexts

Given the results presented here, I was not able to identify significant marker interactions coming from the different spatial contexts. After exploring the sources of variance across markers and conditions, it is plausible that the main reason for the lack of improvement in the model lies in the high correlation of the measurements between the cells. This was also shown in the original study [46]. In their study, they also show that the model based in the intracellular measurements captured more variability than the model containing the environmental and spatial effects (97% of the cases). Therefore, and despite using a more sophisticated model than the original study, the data provided is not suitable to extract meaningful knowledge from the different spatial contexts.

Based on the fact that the data was collected from clonal cells, it is likely that their response to environmental cues is highly similar. As shown in the original study, the main differences in the cell responses was tightly linked to their distance from the spheroid border. Therefore the variance in their phenotype was mainly linked to oxygen and nutrient availability and not to intercellular signaling between the cells. Hence, our approach could not unveil further insights on this matter. Nevertheless, this approach would be suitable in different settings where actual spatial differences are present. Samples with more complex or heterogeneous structures like for instance, tissue samples or multiple cell type cultures would be more appropriate. Furthermore, thanks to the flexibility of MISTy, the design and application of other user-defined views could also provide further insights. Although MISTy is currently a young tool, as different settings and users explore its flexibility, further views and therefore applications will become available.

Chapter 5

Conclusions

In this thesis I aimed to apply data analysis and modeling to study cell communication in cancer. To do so, several approaches have been applied based on the available data and research questions. Furthermore, the analytical pipelines developed on each project have been designed to in order to be easily adaptable to similar systems. The development of the different pipelines allowed me to learn various techniques in the field of biological data analysis, from well-known, widely applied methods to state-of-the-art algorithms. These in turn, allowed me to contribute to different research projects on cancer signaling and obtain relevant biological and clinical insights as well as new hypothesis generation.

In the following sections, I will detail relevant conclusions obtained for each of the projects presented in this thesis.

5.1 Unveiling drug-resistance mechanisms in acute myeloid leukemia

In this project I successfully analyzed dose-response as well as phosphoproteomic data from AML cancer patient and cell line samples. Our main goal was to identify Selinexor resistance mechanism. Thanks to our analyses, we proposed a rational drug combination to overcome Selinexor resistance in AML by using an AKT inhibitor. Follow up experiments showed promising results of this combination in AML cell lines. These results surely prove relevant, although further investigation would be required for our findings to be applicable in the clinical context in humans.

5.2 Discerning cancer cell communication with associated fibroblasts and the extracellular matrix

I studied cell-to-cell and cell-ECM communication in lung cancer cell line-based on proteomic data. The experimental design allowed us to study both the changes within the cells and the secretome as well as distinguish between proteins synthesized by the cancer cells or of fibroblast/ECM origin. Thanks to the time-resolved data, I was able to identify which proteins were significantly changing between different tumor microenvironments and conditions. This allowed us to reconstruct the putative significant ligand-receptor interactions between the cancer cell lines and the microenvironment and fibroblasts based on the changes between TGFβ-induced ECM and normally produced ECM. The results obtained provided meaningful mechanistic insights on how tumor microenvironment can affect cancer cells phenotype and malignancy.

5.3 Exploration of cellular communication in cancer over different spatial contexts

In this project I had the opportunity to study spatially-resolved data generated with the recently-developed IMC system. I aimed to unveil novel contributions of the different spatial contexts in cell signaling. To do so, I applied a state-of-the-art algorithm developed in out lab which has already shown promising results in other IMC experiments. Nevertheless, given the nature of the data generated and the experimental design, I was not able to identify any novel or remarkable results. Further investigation on why I was not able to obtain such results showed very high correlation across the cells in a given image (spheroid slice). This fact did not render any significant contributions on the variance of the different spatial contexts studied. The main reason behind this was because all spheroids were grown from clonal cells and therefore no significant differences between the cells could be identified.

Bibliography

[1] D. A. Altomare and A. R. Khaled. Homeostasis and the importance for a balance between akt/mtor activity and intracellular signaling. *Current Medicinal Chemistry*, 19(22):3748–3762, 2012.

[2] Valery Adorno-Cruz, Andrew D. Hoffmann, Xia Liu, Nurmaa K. Dashzeveg, Rokana Taftaf, Brian Wray, Ruth A. Keri, and Huiping Liu. Itga2 promotes expression of acly and ccnd1 in enhancing breast cancer stemness and metastasis. *Genes Diseases*, 2020.

[3] Chao Ai, Jixin Zhang, Shenyi Lian, Jie Ma, Balázs Győrffy, Zhenyuan Qian, Yong Han, and Qin Feng. Foxm1 functions collaboratively with plau to promote gastric cancer progression. *Journal of Cancer*, 11:788–794, 2020.

[4] Houda Alachkar, Martin BG Mutonga, Klaus H Metzeler, Noreen Fulton, Gregory Malnassy, Tobias Herold, Karsten Spiekermann, Stefan K Bohlander, Wolfgang Hiddemann, Yo Matsuo, Wendy Stock, and Yusuke Nakamura. Preclinical efficacy of maternal embryonic leucine-zipper kinase (melk) inhibition in acute myeloid leukemia. *Oncotarget*, 5(23):12371, 2014.

[5] Jessica K. Altman, Amy Szilard, Dennis J. Goussetis, Antonella Sassano, Marco Colamonici, Elias Gounaris, Olga Frankfurt, Francis J. Giles, Elizabeth A. Eklund, Elspeth M. Beauchamp, and Leonidas C. Platanias. Autophagy is a survival mechanism of acute myelogenous leukemia precursors during dual mtorc2/mtorc1 targeting. *Clinical Cancer Research*, 20(9):2400–2409, 2014.

[6] Norman Balcazar, Aruna Sathyamurthy, Lynda Elghazi, Aaron Gould, Aaron Weiss, Ichiro Shiojima, Kenneth Walsh, and Ernesto Bernal-Mizrachi. mtorc1 activation regulates -cell mass and proliferation by

modulation of cyclin d2 synthesis and stability. *Journal of Biological Chemistry*, 284(12):7832–7842, 2009.

[7] Frances R. Balkwill, Melania Capasso, and Thorsten Hagemann. The tumor microenvironment at a glance. *Journal of Cell Science*, 125(23):5591–5596, 2012.

[8] Bernd Bodenmiller, Eli R Zunder, Rachel Finck, Tiffany J Chen, Erica S Savig, Robert V Bruggner, Erin F Simonds, Sean C Bendall, Karen Sachs, Peter O Krutzik, and Garry P Nolan. Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators. *Nature biotechnology*, 30(9):858–867, 2012.

[9] Pedro Casado, Juan-Carlos Rodriguez-Prados, Juan-Carlos Rodriguez-Prados, Sabina C Cosulich, Sylvie Guichard, Bart Vanhaesebroeck, Simon Joel, and Pedro R Cutillas. Kinase-substrate enrichment analysis provides insights into the heterogeneity of signaling pathway activation in leukemia cells. *Science signaling*, 6(268):rs6, 2013.

[10] Hae J Chon, Kyoung J Bae, Yura Lee, and Jiyeon Kim. The casein kinase 2 inhibitor, cx-4945, as an anti-cancer drug in treatment of human hematological malignancies. *Frontiers in pharmacology*, 6:70, 2015.

[11] Marsha L Crochiere, Stefan Hannus, Kerrin Hansen, Frank Becker, Erkan Baloglu, Margaret Lee, Michael Kauffman, Sharon Shacham, and Yosef Landesman. Xpo1 target occupancy measurements confirm the selinexor recommended phase 2 dose. *Oncotarget*, 8(66):110503, 2017.

[12] Alexander Drilon, Federico Cappuzzo, Sai-Hong Ignatius Ou, and D Ross Camidge. Targeting met in lung cancer: will expectations finally be met? *Journal of Thoracic Oncology*, 12(1):15–26, 2017.

[13] J Etchin, J Montero, A Berezovskaya, B T Le, A Kentsis, A L Christie, A S Conway, W C Chen, C Reed, M R Mansour, C E L Ng, S Adamia, S J Rodig, I A Galinsky, R M Stone, B Klebanov, Y Landesman, M Kauffman, S Shacham, A L Kung, J C Y Wang, A Letai, and A T Look. Activity of a selective inhibitor of nuclear export, selinexor (kpt-330), against aml-initiating cells engrafted into immunosuppressed nsg mice. *Leukemia*, 30(1):190–90, 2016.

[14] J Etchin, Q Sun, A Kentsis, A Farmer, Z C Zhang, T Sanda, M R Mansour, C Barcelo, D McCauley, M Kauffman, S Shacham, A L Christie, A L Kung, S J Rodig, Y M Chook, and A T Look. Antileukemic activity of nuclear export inhibitors that spare normal hematopoietic cells. *Leukemia*, 27(1):66–74, 2013.

[15] Dong Fan and Zamaneh Kassiri. Biology of tissue inhibitor of metalloproteinase 3 (timp3), and its therapeutic implications in cardiovascular pathology. *Frontiers in Physiology*, 11:661, 2020.

[16] Alexandre Grassart, Annick Dujeancourt, Paul B Lazarow, Alice Dautry-Varsat, and Nathalie Sauvonnet. Clathrin-independent endocytosis used by the il-2 receptor is regulated by rac1, pak1 and pak2. *EMBO reports*, 9(4):356–362, 2008.

[17] Douglas Hanahan and Robert A. Weinberg. Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674, 2011.

[18] Barbara L Hempstead. The many faces of p75ntr. *Current opinion in neurobiology*, 12(3):260–267, 2002.

[19] Wolfgang Huber, Anja von Heydebreck, Holger Sueltmann, Annemarie Poustka, and Martin Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl. 1:S96–S104, 2002.

[20] Pilar Iniesta, Alberto Morán, Carmen De Juan, Ana Gómez, Florentino Hernando, Cristina García-Aranda, Cristina Frías, Antonio Díaz-López, Francisco-Javier Rodríguez-Jiménez, Jose-Luis Balibrea, and Manuel Benito. Biological and clinical significance of mmp-2, mmp-9, timp-1 and timp-2 in non-small cell lung cancer. *Oncology reports*, 17(1):217–223, 2007.

[21] 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.

[22] Mohit Kumar Jolly, Satyendra C Tripathi, Dongya Jia, Steven M Mooney, Muge Celiktas, Samir M Hanash, Sendurai A Mani, Kenneth J Pienta, Eshel Ben-Jacob, and Herbert Levine. Stability of the hybrid epithelial/mesenchymal phenotype. *Oncotarget*, 7(19):27067–27084, 2016.

[23] Minoru Kanehisa and Susumu Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[24] Hyunkoo Kang, Hyunwoo Kim, Sungmin Lee, HyeSook Youn, and BuHyun Youn. Role of metabolic reprogramming in epithelial–mesenchymal transition (emt). *International Journal of Molecular Sciences*, 20(8):2042, 2019.

[25] Trinayan Kashyap, Christian Argueta, Thaddeus Unger, Boris Klebanov, Sophia Debler, William Senapedis, Marsha L Crochiere, Margaret S Lee, Michael Kauffman, Sharon Shacham, and Yosef Landesman. Selinexor reduces the expression of dna damage repair proteins and sensitizes cancer cells to dna damaging agents. *Oncotarget*, 9(56):30773–30786, 2018.

[26] Kai Kessenbrock, Vicki Plaks, and Zena Werb. Matrix metalloproteinases: Regulators of the tumor microenvironment. *Cell*, 141(1):52–67, 2010.

[27] Kiran Mahajan and Nupam P. Mahajan. Pi3k-independent akt activation in cancers: A treasure trove for novel therapeutics. *Journal of Cellular Physiology*, 227(9):3178–3184, 2012.

[28] Bruno C. Medeiros, Steven M. Chan, Naval G. Daver, Brian A. Jonas, and Daniel A. Pollyea. Optimizing survival outcomes with post-remission therapy in acute myeloid leukemia. *American Journal of Hematology*, 94(7):803–811, 2019.

[29] Waleed Minzel, Avanthika Venkatachalam, Avner Fink, Eric Hung, Guy Brachya, Ido Burstain, Maya Shaham, Amitai Rivlin, Itay Omer, Adar Zinger, Shlomo Elias, Eitan Winter, Paul E. Erdman, Robert W. Sullivan, Leah Fung, Frank Mercurio, Dansu Li, Joseph Vacca, Nathali Kaushansky, Liran Shlush, Moshe Oren, Ross Levine, Eli Pikarsky, Irit Snir-Alkalay, and Yinon Ben-Neriah. Small molecules co-targeting ckiα and the transcriptional kinases cdk7/9 control aml in preclinical models. *Cell*, 175(1):171–185, 2018.

[30] Mahealani K Monteilh-Zoller, Meredith C Hermosura, Monica JS Nadler, Andrew M Scharenberg, Reinhold Penner, and Andrea Fleig. Trpm7 provides an ion channel mechanism for cellular entry of trace metal ions. *The Journal of general physiology*, 121(1):49–60, 2003.

[31] Darryl Nishimura. Biocarta. *Biotech Software & Internet Report*, 2(3):117–120, 2001.

[32] Ferhat Ozden, Caner Saygin, Didem Uzunaslan, Bulent Onal, Haydar Durak, and Hilal Aki. Expression of mmp-1, mmp-9 and timp-2 in prostate carcinoma and their influence on prognosis and survival. *Journal of cancer research and clinical oncology*, 139(8):1373–1382, 2013.

[33] Elli Papaemmanuil, Moritz Gerstung, Lars Bullinger, Verena I. Gaidzik, Peter Paschka, Nicola D. Roberts, Nicola E. Potter, Michael Heuser, Felicitas Thol, Niccolo Bolli, Gunes Gundem, Peter Van Loo, Inigo Martincorena, Peter Ganly, Laura Mudie, Stuart McLaren, Sarah O'Meara, Keiran Raine, David R. Jones, Jon W. Teague, Adam P. Butler, Mel F. Greaves, Arnold Ganser, Konstanze Döhner, Richard F. Schlenk, Hartmut Döhner, and Peter J. Campbell. Genomic classification and prognosis in acute myeloid leukemia. *New England Journal of Medicine*, 374(23):2209–2221, 2016.

[34] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015.

[35] Carl F Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H Buetow. Pid: the pathway interaction database. *Nucleic Acids Research*, 37(suppl_1):D674–D679, 2009.

[36] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[37] 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*, 102(43):15545–15550, 2005.

[38] Ronald J. Tallarida. Quantitative methods for assessing drug synergism. *Genes & Cancer*, 2(11):1003–1008, 2011.

[39] Jovan Tanevski, Attila Gabor, Ricardo Omar Ramirez Flores, Denis Schapiro, and Julio Saez-Rodriguez. Explainable multi-view framework for dissecting inter-cellular signaling from highly multiplexed spatial data. *bioRxiv*, 2020.

[40] Dénes Türei, Alberto Valdeolivas, Lejla Gul, Nicolàs Palacio-Escat, Michal Klein, Olga Ivanova, Márton Ölbei, Attila Gábor, Fabian Theis, Dezső Módos, Tamás Korcsmáros, and Julio Saez-Rodriguez. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Molecular Systems Biology*, 17(3):e9923, 2021.

[41] Leif Väremo, Jens Nielsen, and Intawat Nookaew. Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Research*, 41(8):4378–4391, 2013.

[42] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake van der Plas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. Scipy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, 17:261–272, 2020.

[43] Amy Y. Wang and Hongtao Liu. The past, present, and future of crm1/xpo1 inhibitors. *Stem Cell Investigation*, 6, 2019.

[44] Jakob Wirbel, Pedro Cutillas, and Julio Saez-Rodriguez. Phosphoproteomics-based profiling of kinase activities in cancer cells. pages 103–132, 2018.

[45] Li Yang, Yanli Pang, and Harold L. Moses. Tgf- and immune cells: an important regulatory axis in the tumor microenvironment and progression. *Trends in Immunology*, 31(6):220–227, 2010.

[46] Vito RT Zanotelli, Matthias Leutenegger, Xiao-Kang Lun, Fanny Georgi, Natalie de Souza, and Bernd Bodenmiller. A quantitative analysis of the interplay of environment, neighborhood and cell state in 3d spheroids. *bioRxiv*, 2020.

[47] Lidan Zhang, Xiaojuan Li, Zhiyong Ke, Libin Huang, Yanni Liang, Jun Wu, Xiaoli Zhang, Yueqin Chen, Hua Zhang, and Xuequn Luo. Mir-99a may serve as a potential oncogene in pediatric myeloid leukemia. *Cancer cell international*, 13(1):1–11, 2013.

[48] Eli R Zunder, Rachel Finck, Gregory K Behbehani, D Amir El-ad, Smita Krishnaswamy, Veronica D Gonzalez, Cynthia G Lorang, Zach Bjornson, Matthew H Spitzer, Bernd Bodenmiller, Wendy J Fantl, Dana Pe'er, and Garry P Nolan. Palladium-based mass tag cell barcoding with a doublet-filtering scheme and single-cell deconvolution algorithm. *Nature protocols*, 10(2):316, 2015.

Appendix A

Appendix

A.1 Supplementary figures

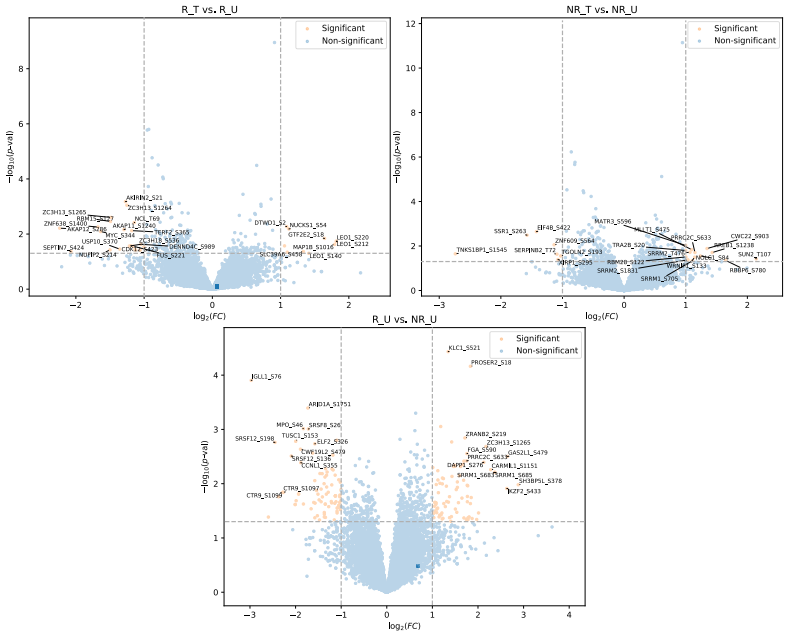

Figure S1: Volcano plots of the *ex vivo* patient samples displaying the $\log_2(FC)$ versus the $-\log_{10}(p\text{-value})$. In orange are shown the significant differentially phosphorylated sites (p-value ≤ 0.05 and $|\log_2(FC)| > 1$) and in blue the non-significant. Only the top 25 sites are labeled due to space limitations. The top left volcano shows the differential phosphorylation of responders (treated vs. untreated), top right the non-responders (treated vs. untreated) and the bottom volcano shows the responders vs. non-responders (untreated). Only the top 25 differentially phosphorylated sites display the label names due to space constraints.

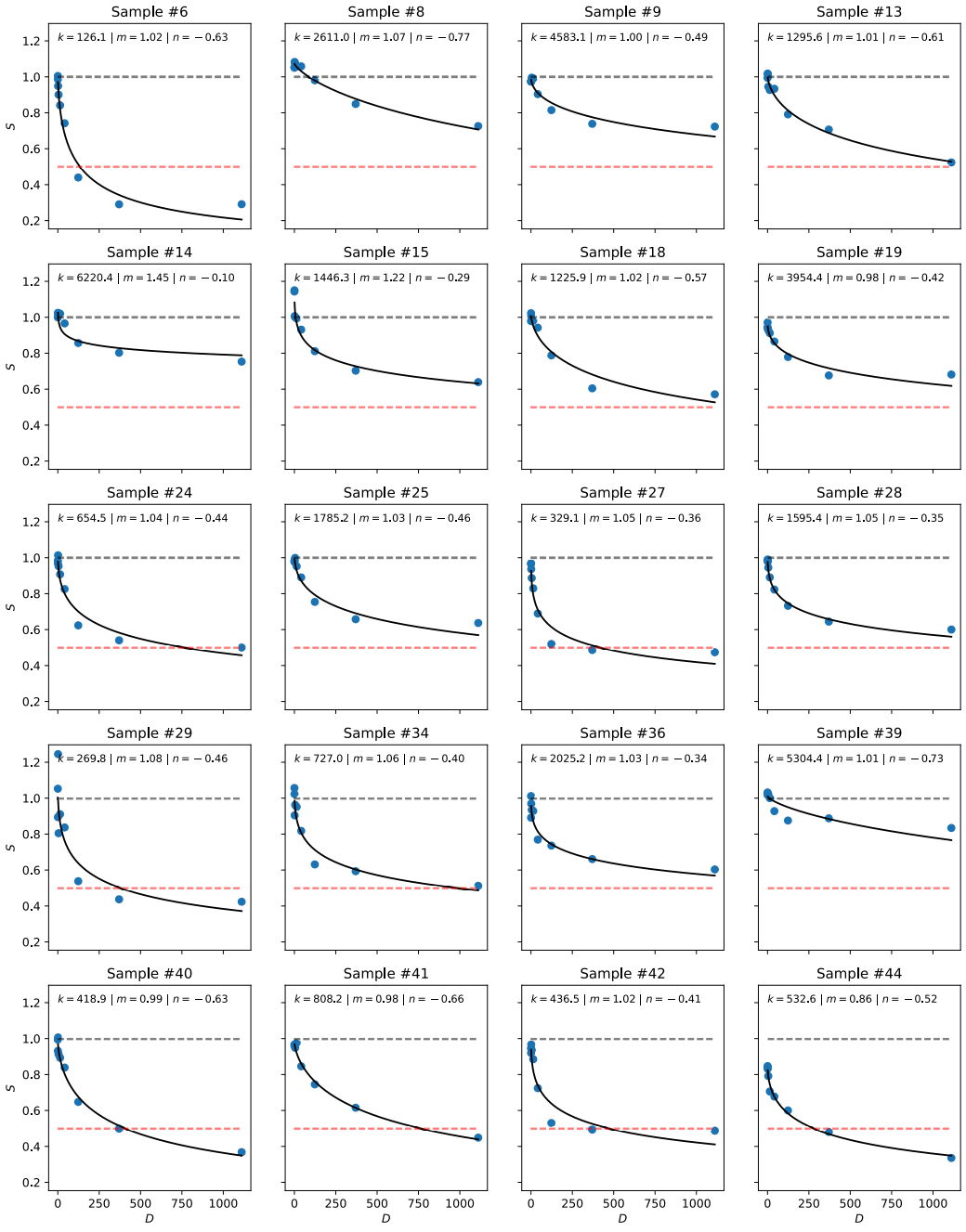

Figure S2: Fitted dose-response models for each *ex vivo* patient sample. The measured relative survival (blue dots) was used to fit the model described in equation (2.1) (black line). The parameter values are shown at the top of each model. Dotted lines are added for reference, in gray 100% survival and in red 50%.

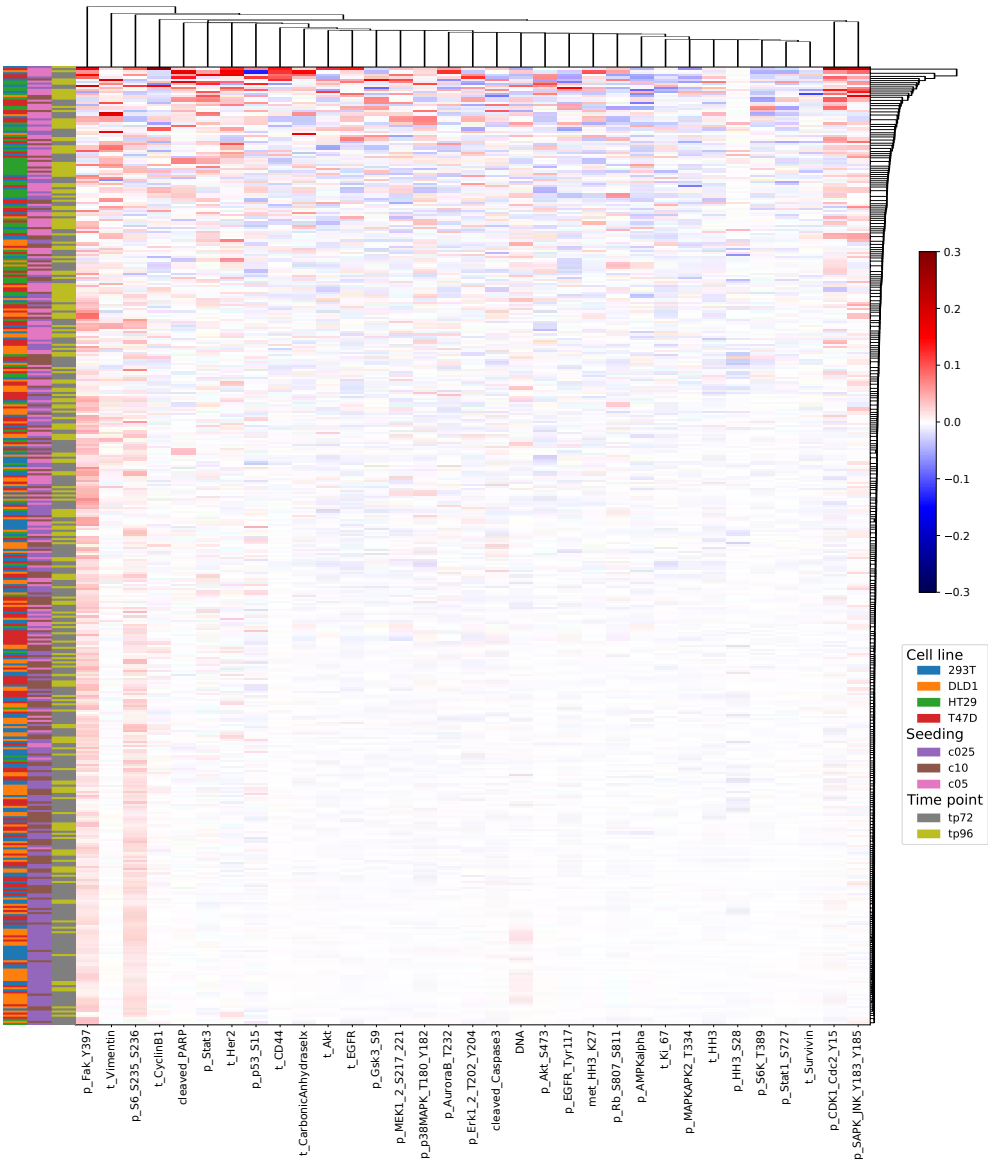

Figure S3: Variance explained (R^2) difference between full and intracellular models for the selected parameters (q=8 and l=50). Markers are shown in the x-axis and samples on the y-axis. Samples and markers are ordered according to hierarchical clustering. Red indicates an increase in variance explained in the multiview model with respect to the intraview model alone, blue otherwise (white means no difference. Color codes on the left denote the different conditions of each image: left column corresponds to cell lines (293T in blue, DLD1 in orange, HT29 in green and T47D in red), central column corresponds to the seeding concentration (purple for 0.25, pink for 0.5 and brown for 1.0 in relative scale) and the right column denotes the growth time (72h in silver and 96h in gold).

A.2 Supplementary tables

Table S1: List of markers measured in the IMC experiments. Class defines the major process each marker is involved in and the antibody clone used for that marker.

Marker	**Class**	**Antibody clone**
Vimentin	Structural	EPR3776
p-AKT_S473	AKT	D9E
CD44	Receptor	IM7
p-STAT1_S727	JAK/STAT	polyclonal pStat1
p-MAPKAPK2_T334	EGF/Stress	27B7
met-HH3_K27	Structural	C36B11
p-S6K_T389	Nutrient	1A5
Survivin	Apoptosis	71G4B7
KI-67	Cell Cycle	B56
p-GSK3_S9	Nutrient	D85E12
p-AMPKα	Nutrient	40H9
CyclinB1	Cell Cycle	GNS-11
p-RB_S807/S811	Cell Cycle	D20B12
p-EGFR_T117	EGF	3G3.2
AKT	AKT	C67E7
HH3	Structural	D1H2
p-S6_S235/S236	Nutrient	D57.2.2E
p-FAK_Y397	EGF	Polyclonal FAK
EGFR	EGF	D38B1
Cleaved CASP3	Apoptosis	C92-605
p-AURKB_T232	Cell Cycle	D13A11
CA9	Nutrient	Polyclonal CAIX
p-MEK1/2_S217/221	EGF	41G9
p-ERK1/2_T202/Y204	EGF	20A
p-CDK1/CDC2_Y15	Cell Cycle	44/CDK1/CDC2(pY15)
p-HH3_S28	Cell Cycle	HTA28
p-TP53_S15	Stress	16G8
HER2	EGF	D8F12
p-MAPK14_T180/Y182	Stress	36/p38(pT180/pY182)
Cleaved PARP	Apoptosis	F21-852
p-SAPK/JNK_Y183/Y185	Stress	G9
p-STAT3	JAK/STAT	4/pSTAT3