Deciphering Cellular Signaling Pathways
and their Perturbations in Disease Using
Network-Based Strategies

THESIS SUBMITTED FOR THE DEGREE OF
"DOCTOR OF PHILOSOPHY"

by
Dana Silverbush

The work on this thesis has been carried out
under the supervision of Professor Roded Sharan

Submitted to the Senate of Tel-Aviv University

January 2018
ACKNOWLEDGEMENTS

First and foremost I would like to thank my supervisor, professor Roded Sharan, for his guidance, advise and encouragement throughout the work on this thesis. In particular I am grateful for his support in my academic future and long term visions, and his encouragement for independence and bold ideas.

I would like to thank Professor Danny Segev, Professor Jasmin Fisher, Professor Tamar Geiger and Professor Niko Beerenwinkel for their professional and personal mentoring.

Many friends and colleagues assisted in the path towards this thesis. I would like to thank Dr. Anthony Gitter for his help in comparing the SHORTEST method to his approach\(^1\), presented in Chapter 4. I would also like to thank Dr. Matan Hofree for providing the cancer driver gene lists used in Chapters 5 and 6, Mariya Mardamshina M.D. for her help to define the receptor-based classification used in Chapter 6, and Roni Wilentzik for her insightful comments on the figures in Chapters 5 and 6. I would like to thank Dr. Simona Cristea, my co-author to the "ModulOmics" paper, for her uncompromising professionalism.

Last but not least, I would like to thank my mother, Noemy Silverbush. While I sometimes went through ups and downs, her love and support were always steadfast.
Abstract

Disease relevant genetic and epigenetic differences between individuals are reflected in their protein levels, impacting protein functionality. As proteins work in concert and influence each other through protein-protein interactions (PPIs), the change in expression or function of a protein triggers a cascade of events leading to a changed function or a state of the cell. In particular, extracellular signals, including medical drugs, are transduced from stimulated surface proteins to proteins inside the cytoplasm, leading to a broad range of possible functional outcomes. Thus, interindividual differences in the genetic and epigenetic content may lead to substantial differences in the reaction and sensitivity to treatments that can only be understood by studying the cellular signaling circuits in a disease-specific context.

Constructing cellular signaling models is especially crucial in research on cancer, a disease which is caused by a chance accumulation of genetic alterations that trigger abnormal signaling circuits. The genetic alterations are often unique to a group of patients, resulting in a widely varied abnormal signaling cascades and diverse individual responses to therapy. One of the main efforts in cancer biology is to model the abnormal signaling cascades in patients and use the models to identify key protein candidates for drug targets so as to block the abnormal cascades. To this end, high-throughput technologies are used to map genetic alterations by DNA sequencing and to detect the resulting downstream effects by measuring gene expression, protein expression, protein phosphorylation events and cells count.
In this thesis, we rely on such high-throughput assays to facilitate the reconstruction of accurate cellular signaling models. We advance the capability of the PPI network to reflect the true nature of signaling cascades by inferring the directions of the physical links within the network. Using experimental causality data and building on an Integer Linear Programing (ILP) technique we provide an efficient orientation solution for a model organism. The method outperforms previous approaches and can orient considerable parts of the network, revealing its structure and function. We further devise a network diffusion technique to provide an orientation solution for the human PPI network, employing for the first time experimentally derived causality information to orient the human network. The oriented networks reflect more faithfully the true nature of the signal flow, benefiting not only signaling models, but also a variety of signaling based predictions, such as the inference of drug targets and the prediction of key cancer genes. Next, we focus on cellular signaling applications in cancer, and provide a tool that incorporates multiple high-throughput sources to identify gene modules which exhibit functional connectivity among their members and drives the disease. Finally, we design a framework to construct executable signaling models using a simulated annealing process and demonstrate its utility in tailoring optimal treatment for Acute Myeloid Leukemia (AML) patients.

Collectively, we provide a framework for elucidating cellular signaling pathways, and show its application to gain mechanistic understanding of cancer processes so as to suggest efficient therapeutic strategies.
CONTENTS

Acknowledgements .................................................................................................................. 2

1 Introduction .......................................................................................................................... 18
  1.1 Cellular signaling .............................................................................................................. 18
  1.2 PPI networks as a basis for signaling ............................................................................... 19
     1.2.1 PPI resources ........................................................................................................... 20
     1.2.2 PPI representation ................................................................................................... 21
  1.3 Orientation of PPI networks ............................................................................................... 22
  1.4 Orientation of the human PPI networks ............................................................................. 24
     1.4.1 Experimental resources for causality information .................................................... 25
     1.4.2 Network diffusion as a technique to explore network function and structure ........... 26
     1.4.3 Utilizing network diffusion to orient the human PPI network using causality information .................................................................................................................. 28
  1.5 Detecting cancer driver pathways in the PPI network ....................................................... 28
     1.5.1 Cancer driver genes .................................................................................................. 29
     1.5.2 Cancer driver modules ............................................................................................. 29
     1.5.3 Detection of gene modules in networks ...................................................................... 30
     1.5.4 Detection of gene modules in cancer ......................................................................... 33
  1.6 Executable computational models .................................................................................... 34
     1.6.1 Modeling node activity in a network ......................................................................... 34
     1.6.2 Finding stability in executable models using the Bio Model Analyzer tool .................. 36
  1.7 Summary of the papers included in this thesis ................................................................. 38
     1.7.1 Network orientation via shortest paths ....................................................................... 38
     1.7.2 A diffusion-based method for orienting the human protein-protein interaction network .................................................................................................................. 39
     1.7.3 ModulOomics: Integrating Multi-Omics Data to Identify Cancer Driver Modules ....... 39
     1.7.4 Towards Virtual Models of Cancer: Cell-Specific Computational Modeling of the PIM pathway in Acute Myeloid
2 Network orientation via shortest paths ............................................. 41
  2.1 Results ......................................................................................... 43
    2.1.1 Data Acquisition and Integration ........................................ 43
    2.1.2 Application and Performance Evaluation ........................... 44
    2.1.3 Comparison to Previous Work .......................................... 48
  2.2 Conclusions .............................................................................. 51
  2.3 Methods .................................................................................. 52
    2.3.1 Problem definition and the shortest-paths graph .......... 52
    2.3.2 The integer program ......................................................... 53
    2.3.3 A more efficient formulation .......................................... 55
    2.3.4 Computing orientation confidence ............................... 57
    2.3.5 Handling large networks ............................................... 58
    2.3.6 Iterative expansion ......................................................... 58

3 A diffusion-based method for orienting the human protein-protein
  interaction network ........................................................................... 60
  3.1 Results ...................................................................................... 61
  3.2 Conclusions ............................................................................. 72
  3.3 Methods .................................................................................. 74
    3.3.1 Diffusion-based orientation ........................................... 74
    3.3.2 The diffusion process and edge scoring ....................... 74
    3.3.3 Inference of directions .................................................... 75
    3.3.4 Drug response data ......................................................... 76
    3.3.5 Cancer genomics data ................................................. 76
    3.3.6 Implementation of previous orientation methods ....... 76
    3.3.7 Network and validation sets ....................................... 77

4 ModuLOmics: Integrating Multi-Omics Data to Identify Cancer
  Driver Modules ................................................................................ 79
  4.1 Results ...................................................................................... 82
    4.1.1 Driver modules are enriched with cancer drivers .......... 83
    4.1.2 Driver modules are functionally coherent ..................... 86
    4.1.3 Sensitivity analysis ......................................................... 90
    4.1.4 Driver modules in breast cancer subtypes recapitulate known
mechanisms and suggest novel functionalities

| 4.2 Conclusions                                                                 | 99 |
| 4.3 Methods                                                                     | 100 |
| 4.3.1 Model                                                                      | 100 |
| 4.3.2 PPI Connectivity                                                           | 101 |
| 4.3.3 Mutual exclusivity                                                         | 101 |
| 4.3.4 Co-regulation                                                             | 102 |
| 4.3.5 Co-expression                                                              | 103 |
| 4.3.6 Optimization procedure                                                     | 103 |
| 4.3.7 ILP                                                                        | 104 |
| 4.3.8 Stochastic search                                                          | 106 |
| 4.3.9 Comparison to HotNet2                                                      | 106 |
| 4.3.10 Comparison to TiMEx                                                       | 107 |
| 4.3.11 Evaluating module enrichment with known driver and non-driver genes       | 107 |
| 4.3.12 Evaluating module enrichment with known pathways                          | 108 |

5 Towards Virtual Models of Cancer: Cell-Specific Computational Modeling in Acute Myeloid Leukemia

| 5.1 Results                                                                     | 111 |
| 5.1.1 AML cell lines show differential sensitivity to PIM inhibition            | 111 |
| 5.1.2 Cell type-specific differences in PIM pathway signaling in response to treatment | 112 |
| 5.1.3 Building a generalized model of PIM signaling in AML                       | 114 |
| 5.1.4 Executable QN model validated by cell type-specific signaling behavior    | 116 |
| 5.1.5 Novel signaling components proposed through model refinement              | 120 |
| 5.1.6 In silico virtual experimentation with AML cell models can replicate independently reported data | 121 |
| 5.1.7 AML cell-specific model predicts synergistic drug combinations with the PIM inhibitor | 124 |
| 5.1.8 Executable QN model identify alternative susceptibilities in AZD1208 resistant cells | 127 |

| 5.2 Conclusions                                                                | 130 |
| 5.3 Methods                                                                    | 132 |
| 5.3.1 Framework for constructing executable models                              | 132 |
| 5.3.2 Introducing cell-specific context in QN models                           | 133 |
| 5.3.3 Comparison of boolean models, qualitative models and                      |  
qualitative AND/OR gates models .................................................. 134
5.3.4 Protein Array data transformation for qualitative models 138
5.3.5 Growth inhibition calculation ............................................ 139

6 Discussion ............................................................................... 140
7 References ............................................................................... 148

8 תיצות .................................................................................... 171
9 תקציר ..................................................................................... 173
List of Figures

Figure 2.1: Performance as a function of the number of expansion rounds used. The confidence cutoff in these experiments is set to 0…………………..46

Figure 2.2: Performance as a function of the confidence cutoff used. The number of expansion rounds for all experiments is 10 (11 iterations)…………………..46

Figure 2.3: Performance as a function of the percentage of cause-effect pairs guiding the orientation. The confidence cutoff in these experiments is set to 0 and there are no expansion rounds…………………………………………………………………………………47

Figure 2.4: Performance comparison of MIXED and SHORTEST…………………..47

Figure 2.5: Performance comparison of GITTER and SHORTEST using different number of expansion rounds. Labels denote the number of expansion rounds used by SHORTEST…………………………………………………………………………………50

Figure 2.6: Performance comparison of GITTER and SHORTEST using different confidence cutoff. Labels denote the confidence cutoff used by SHORTEST…………………………………………………………………………………47

Figure 2.7. Performance evaluation of SHORTEST against its extension to large networks (no expansion rounds). We scaled SHORTEST to large networks by changing confidence assignment. The extended method is comparable to the original SHORTEST, with 71% accuracy compared to 73%…………………..58

Figure 3.1. An overview of the Diffuse2Direct orientation algorithm. A) We start from an input network along with guiding information consisting of a collection of experiments, each of which induces a set of causal proteins and a set of effect proteins. As an example we show here drug response information, inducing a set of drug targets and their associated set of differentially expressed genes. B) For each experiment in the collection, we perform network diffusion once from the causal set and once from the effect set, and integrate the scores. C) The resulting scores are used as predictive features for the edge's direction. D) A classifier assigns a confidence to each directed edge, and using a cutoff we obtain a directed network…………………………………………………………………………………62

Figure 3.2. Application of Diffuse2Direct to VEGF signaling. We oriented the VEGF KEGG pathway from VEGF to the pathway's output nodes (VARP, Proliferation,
PGI2Production, Migration, Permeability and Survival). The figure depicts the directions chosen for each of the interactions, where the color of an interaction represents the confidence in its orientation………………………………………………..63

Figure 3.3. Performance evaluation. Shown are cross validation results comparing the directions inferred by Diffuse2Direct and Vinayagam et al. to subsets of interactions with known directionality: KPIs, PDIs, E3 ubiquitination interactions, interactions from the well-studied EGFR pathway, and signal-transduction interactions in mammals (STKE). The area under the precision-recall curve for each approach appears in parentheses………………………………………………………………………………..66

Figure 3.4. Comparison of orientation methods, limited to interactions covered by SHORTEST. Cross validation results comparing the inferred directions to subsets of interactions with known directionality: KPIs, PDIs, and the STKE interactions used for validation by Vinayagam et al. SHORTEST is limited in coverage and hence was not able to infer directions for smaller subsets such as EGFR and E3…………………………………………………………………………………………………………………………………………………..67

Figure 3.5. Performance as a function of the percentage of drugs guiding the orientation. In this setting only, we used a logistic regression classifier with no regularization in order to manually control the percentage of guiding data used………………………………………………………………………………………………………………………………………………………..68

Figure 3.6. Concurrence of orientations according to different sources of cause-effect information. A) Depicted are the agreement of directions assigned to the test edges when independently orienting the network using i) drug targets to differentially expressed genes, or ii) genetic perturbations to differentially expressed genes across different cancer patients (AML, breast, colon and ovarian cancer). B) The higher the agreement the higher is the orientation accuracy, which goes up to 96% when all sources agree…………………………………………………………………………………..69

Figure 3.7. Orientation improves the prediction of drug targets and cancer driver genes. A) The genes in the network are ranked according to their proximity to the differentially expressed genes of a given drug. We calculated the rank assigned to the correct drug targets by using the oriented network versus the unoriented one. B) The genes in the network are ranked by their proximity to the differentially expressed genes of the cancer patients. We calculated the percentage of genes out of the top 1% ranked genes matching known driver genes and known non-driver from multiple sources. The oriented network consistently reports more
known driver genes and less known non-drover genes than the unoriented one. 

Figure 4.1. Overview of ModulOmics. A) Four different data sources, corresponding to four different models M1,…,M4 (see Methods), contribute to the computation of the ModulOmics score: PPI connectivity (protein level), mutual exclusivity (DNA level), transcriptional co-regulation (regulatory connections and RNA level) and co-expression (RNA level). The four colors correspond to four different genes; full squares in the matrix for model M2 encode the presence of alterations, while empty squares encode their absence. In M3, genes 1 and 2 are regulated by a common transcription factor. In M4, the different color intensities depict different expression intensities. B) Potential mechanism leading to a driver module exhibiting patterns of PPI connectivity, mutual exclusivity, co-regulation and co-expression. C) The ILP optimization identifies modules with highest sum of pairwise ModulOmics scores, computed as the average of the four scores corresponding to models M1,…,M4, further z-scored and normalized to [0,1]. D) The stochastic search optimization uses the modules identified by ILP, depicted in panel C, as seeds, and aims to improve their scores and identify the global optimal solution. The space of initial solutions is clustered and genes are exchanged between clusters in order to identify modules with globally high scores. While the scores for models M1,…,M4 of the modules in panel C were approximated as the average pairwise scores, here they are computed explicitly for the entire module. E) The webserver tool computes the ModulOmics score of any chosen gene set, based on any of the TCGA datasets analyzed in this study. For each data source employed by ModulOmics, the tool plots the single omics scores of the top 50 modules, highlighting the score of the chosen gene set.

Figure 4.2. The driver modules inferred by ModulOmics are enriched with cancer driver genes. A) The average driver enrichment (red heatmaps) and non-driver enrichment (blue heatmaps) across the top 10 scoring modules inferred by each method in the three cancer types studied. The enrichment was calculated as the fraction of gene members in each module that are also part of each control list, averaged across the top 10 modules. The modules were ranked by their score, regardless of their sizes (the inferred groups consisted of two, three or four gene members). ME stands for mutual exclusivity, CoReg for co-regulation, and CoExp for co-expression single omics scores. NCG5, PosAG0, PosUnionAll, PosSomatic, PosUniprotKB, PosTextMine, PosTrans are the positive control lists, while NegAgoClean, NegAgoFull and NegEllidge are the negative control ones (see Methods). B) Detailed driver and non-driver enrichment scores for the positive
driver list PosUnionAll and the negative driver list NegAGOClean for the seven methods assessed, across the three cancer types, for the top scoring 5, 10, 15, 20 and 30 modules. C) Detailed driver and non-driver enrichment scores for the positive driver list PosUnionAll and the negative driver list NegAGOClean for the seven methods assessed, across the three cancer types, for the top scoring 5, 10, 15, 20 and 30 modules.

Figure 4.3. The driver modules inferred by ModulOmics are enriched with cancer driver pathways. A) Mean hyper-geometric p-value of the KEGG pathways significantly enriched in the top 10 modules identified by ModulOmics (red), HotNet2 (purple) and TiMEx (light blue). B) Average number of KEGG pathways significantly enriched in the top modules, indicated above the bars. The opaque bars indicate cancer-related pathways only. C) Average enrichment factors for top modules. The numbers displayed in panels B and C are normalized per module. Enrichment p-values and factors were computed with Expander.

Figure 4.4. Robustness of modules inferred by ModulOmics. A) Intersection of gene pairs co-residing in the same module, for various parameter choices. B) Intersection of the gene pool collectively reported by the top modules, for various parameter choices.

Figure 4.5. Modules inferred in mRNA-classified breast cancer subtypes reflect various levels of subtype aggressiveness and separate cancerous form healthy tissues. A) For each mRNA-based subtype and for the pooled set of genes in the top 20 modules, we computed their occurrence frequency in the top 20 modules, as well as their SNV and CNA alteration frequencies across the patient cohort. These genes are enriched with known cancer drivers and pathways, and could not have been identified if relying on SNV and CNA alteration frequencies alone. White corresponds to absent genes (0% frequency). B) Detailed PPI network view of the subset of genes in panel A that are either known drivers, or part of KEGG pathways. The displayed physical protein interactions underline cancer-related functional associations, such as the role of the PIK3A pathway in Luminal A tumors. C) Selected list of significantly enriched GO pathways across the top 20 modules, reflecting the aggressiveness of the Basal and Her2 subtypes, compared to Luminal A and Luminal B. Enrichment hyper-geometric p-values were computed with Expander. White corresponds to absent pathways. D) Module scores for top Luminal A modules suggesting two different biological roles of the tumor suppressor PTEN. E) The highest ranking Luminal A module in an independent mass spectrometry proteomics dataset separates cancerous from healthy patient
tissues. TP53 loss is measured by its downstream regulated protein CDC2, CDH1 loss is measured by its downstream regulated protein CTNNB1, and FOXA1 gain is measured directly……………………………………………………………………………….95

Figure 4.6. Modules inferred in receptors-classified breast cancer subtypes highlight differences among subtypes. A) For each receptors-based subtype and for the pooled set of genes in the top 20 modules, we computed their occurrence frequency in the top 20 modules, as well as their SNV and CNA alteration frequencies across the patient cohort. These genes are enriched with known cancer drivers and pathways, and could not have been identified if relying on SNV and CAN alteration frequencies alone. White corresponds to absent genes (0% frequency). B) The TN module PTEN, BRCA1, RB1, and TP53 (ModulOmics score 0.52) shows both Pairwise and group mutual exclusivity scores of 0, unique to this subtype, potentially suggestive that the accumulating activity of multiple tumor suppressors drives TN aggressiveness. C) RPPA data supports the observation that TN patients harbor lower activity of the tumor suppressors PTEN, BRCA1, RB1, and TP53, compared to other subtypes. PTEN, RB1, phosphorylated RB1 (RB1 PS807-S811) and RB1-related CYCLIN D1 show significantly lower activity, and phosphorylated AKT (AKT PS308 and AKT PS473), which is negatively regulated by PTEN, shows higher activity. TP53 and BRCA1 were not captured directly, yet their suppressed targets, CDK1 and CYCLIN B1, show lower activity. D) The tumor suppressors PTEN, BRCA1, RB1, and TP53 can be used to separate TN tissues from the other subtypes in RPPA data. CDK1, CYCLIN B1 and AKT, tumor promoters negatively regulated by TP53, BRCA1 and PTEN respectively, show significantly higher activity as a group; PTEN, RB1 and the RB1-related CYCLIN D1, tumor suppressors, show significantly lower activity as a group……………………………………………………………………………….98

Figure 5.1. AML cell lines sensitive to the PIM inhibitor AZD1208 have diverse genotypes. A) GI50 (uM) waterfall plot and molecular oncoprint illustrating the diverse pharmacological response of AML cells after 72 hours AZD1208 treatment as well as PIM expression and AML disease relevant mutations. Boxed cell lines names indicate responding cell lines further investigated. B) Protein expression measured by RPPA in treated and untreated cell lines show heterogeneity in signaling responses through various pathways. C) MV411, with an active FLT3-ITD, shows varied responses to concentrations of AZD1208 and/or AC220 for 72 hours. The number of viable cells was determined by Alamar Blue measurements where the values represent percent growth inhibition………………………………………………………………………………113
Figure 5.2. Schematic workflow of cell-specific model construction in BMA. Motifs and interactions curated from the literature are used to build a Qualitative Network in the BMA tool. The model was calibrated with the results of RPPA experiments for two cell lines with different AML-driving mutations. The model is designed to represent the general AML pathways, and provide a cell-specific context by “turning-on” a specific set of mutations. The mutations impact outgoing interactions, thus activating the pathways in a mutation-specific manner, resulting in mutation-specific phosphorylation activity throughout the pathways leading to specific cellular behavior. The model is iteratively refined by testing and comparing to the cell behavior measured as a response to different perturbations for the two cell lines. The model robustness was tested against perturbations from the literature performed on the explored mutations, and unseen cell lines incorporated automatically into the model. The model is then used for in-silico experimentation in order to test novel drug combinations, infer the source and mechanism for drug resistance, and predict drug response in resistant cell lines and suggest treatment for resistance.

Figure 5.3. Generation of a predictive cell behavior model for AML training cell lines (MOLM16 and MV411) and unseen cell lines (EOL1 and KG1A) A) Cell-Specific AML regulatory network model incorporating knowledge from the literature and calibrated to phosphorylation activity measured by RPPA. Perturbations, driving mutations, and internal genes are depicted in grey, green, and red, respectively. To simulate specific cell (MOLM16, MV411, EOL1 or KG1A) the node for the protein with driver mutations (TYK2, FLT-ITD, PDGFRA, or FGFR1 respectively) is set to 1, while all other proteins with mutations are set to 0. B) Protein signaling activity (phosphorylation) levels inferred in-silico using the cell-specific contexts (laptop icon), and the generalized model (papers icon) capturing levels of phosphorylation activity as measured in-vitro (petri-dish). C) Cell apoptosis and proliferation as inferred in-silico by the executable model compared to levels as observed in-vitro, with the generalized model capturing partial abnormal cell behavior, and the cell-specific context model recapitulating measured levels. D) Unseen cell lines EOL1 and KG1A are incorporated to the executable network model. The robustness of the model is tested via the ability of the model to capture the phosphorylation activity unseen at the time of model construction, and E) Cell behavior as a result of different perturbations.

Figure 5.4. Validation of predicted synergistic combinations of drugs reveals new effective treatment strategies. A) Cell-Specific AML model is used to test combinations of drugs and predict cell behavior in a cell specific manner. A drug
or a combination of drugs is simulated by partially or fully nullifying the target functions of their targets, and can be done automatically and efficiently with a large number of candidates. B) Predicted cell behavior of apoptosis and proliferations is validated via growth inhibition of AML cell lines cultured with the indicated concentration ranges of AZD1208 and/or tested combined inhibitor after 72 hours. Predicted synergic effect, as seen for EOL1 cell line with PIM and AKT and PIM and PI3K inhibitors, is used to prioritize combinations…………………………………………………………..126

Figure 5.5. Origin of resistance to AZD1208 in MOLM16 is computationally inferred and validated via whole exome DNA-seq, revealing signaling mechanism validated via western blots and offers combination to combat resistance which successfully induces apoptosis. A) Network model of MOLM16 resistant cell populations (R1-4). Perturbations (lightning bolts) were automatically predicted at specific nodes to simulate possible resistance mechanisms that would attenuate signaling down a specific pathway (shaded red and blue). B) Whole exome DNA-seq was performed on the 1uM AZD1208 resistant pool population to identify protein altering variants from variant calling as significantly different from the parental MOLM16 cell line. Genes whose proteins are involved with epigenetic machinery are underlined with dashes. C) Inferred signaling activity from the parental MOLM16 executable model is compared to activity from western blots for parental and resistant cell populations. D) Predictions of signaling activity and cell apoptosis for AZD1208 treated alone and in combination with AKT inhibitor AZD5363 is compared to activity from western blots. Prediction of induced apoptosis is supported by the increase in PARP Cleaved with AZD5363……………………………………………………129

Figure 5.6. Generation of a Boolean model to predict phosphorylation events, responses to single treatment and synergistic combinations of treatments. Activity of proteins was relaxed to binary (ON or OFF) and a new model was generated using the same methodology as described for the qualitative model. A) Protein signaling activity (phosphorylation) levels inferred in-silico using the Boolean model (laptop icon) capturing levels of phosphorylation activity as measured in-vitro (petri-dish). B) Cell apoptosis and proliferation as inferred in-silico by the Boolean model compared to levels as observed in-vitro. C) Unseen cell lines EOL1 and KG1A are incorporated to the executable Boolean model, and D) Cell behavior as a result of different perturbations. E) Predicted cell behavior of apoptosis vs. the Boolean model predictions…………………………………………………………..136

Figure 5.7. Generation of an AND/OR model to predict phosphorylation events, responses to single treatment and synergistic combinations of treatments. Activity
of relations is limited to AND/OR/NOT gates only and a new model was generated using the same methodology as described for the qualitative model. A) Protein signaling activity (phosphorylation) levels inferred in-silico using the AND/OR model (laptop icon) capturing levels of phosphorylation activity as measured in-vitro (petri-dish). B) Cell apoptosis and proliferation as inferred in-silico by the AND/OR model compared to levels as observed in-vitro. C) Unseen cell lines EOL1 and KG1A are incorporated to the executable AND/OR model, and D) Cell behavior as a result of different perturbations. E) Predicted cell behavior of apoptosis vs. the AND/OR model predictions.

Figure 5.8. A schematic representation for the method of transforming the log2 linear RPPA values to a 5 point categorical scale suitable for BMA modeling. The RPPA data was categorized by two methods 1) within cell and treatment dynamics of response and 2) within protein antibody measure across cell and treatment. This 5 point scale was also used to categorize the cell line fate (phenotypic response) at each cell line, treatment, and time point condition.
List of Tables

Table 2.1. A comparison of GITTER and SHORTEST for different numbers of expansion round........................................................................................................49

Table 2.2. A comparison of GITTER and SHORTEST for different confidence cutoffs ........................................................................................................50

Table 2.3. Algorithm run times in seconds.................................................................................................................................51

Table 3.1. AUPR results for the 5 unbiased test sets.........................................................................................................................64

Table 4.1. Genes in top modules participating in highly enriched pathways......................................................................................89

Table 5.1. Target functions for AML cell specific calibrated AML model.....................................................................................119

Table 5.2. BMA network replicates response to treatments reported in external publication........................................................................................................122
1 INTRODUCTION

1.1 CELLULAR SIGNALING

Cells respond to outer and inner cues via cellular signaling. Many types of cues exist: natural environmental cues such as nutrients help prokaryotic organisms navigate toward food sources; growth factors, hormones, neurotransmitters, and extracellular matrix components help multicellular organisms orchestrate cell behavior; human-made cues such as drugs and small molecules are used to influence cellular behavior. In cancer, both outer and inner cues are abnormally activated, such as the VEGF secreted from tumor cells to encourage angiogenesis and the abnormal expression of proteins caused by a mutation, which may trigger malignant signaling processes to initiate the hallmarks of cancer.

Cells use proteins called receptors that bind to signaling molecules, i.e. outer-cell cues. Different receptors are specific for different molecules, and there are hundreds of receptor types found in cells. Once activated by their respected signaling molecules outside the cell they transmit the signal to internal signaling pathways. The signal is typically amplified, activating multiple intracellular cascades for every one receptor that is bound. A signaling cascade includes several types of interactions, such as the addition or removal of a phosphate group, formation of complexes, conformational changes, initiation of protein degradation and more. At any one time, a cell is receiving and responding to numerous inner and outer signaling cues, and multiple cascade pathways are activated, with many overlaps among these pathways.
In the past two decades, high-throughput techniques have been developed with the goal of identifying in great detail the landscape of signaling activation and resulting gene expression. To name a few, genetic alterations which may trigger abnormal expression and signaling are measured using Sanger sequencing\(^4\) or Next Generation Sequencing\(^5\) techniques. The functional effects of these alterations and the site-specific signaling activation of proteins is measured using reverse phase protein array\(^6\), while total abundance is measured in small scale using western blot\(^7\) and in large scale via mass-spectrometry\(^8\). Gene expression differences that can stem either from genetic alterations or from signaling cascade effects are measured using microarrays\(^9\) or RNA sequencing\(^10\). These techniques provide partial portraits of the signaling initiating proteins, the downstream activated proteins, and the resulting gene-level changes. All these are linked via physical interactions of the PPI network, the common context on which they may be cast to expose the underlying molecular mechanisms.

### 1.2 PPI NETWORKS AS A BASIS FOR SIGNALING

Proteins work in concert by physically interacting to form cellular machineries, act on other molecular species, and transmit molecular signals. A PPI network model can be used to understand healthy common mechanisms and also as a basis to infer interindivial differences, disease related changes, and their effect on healthy signaling behavior. In the context of signaling, PPI networks reflect the potential of the cell signaling, where an interaction may imply that when a protein at the head of an interaction receives a signal (i.e. activated), it will pass the signal to the protein at its tail, which may trigger a cascade of activations.
1.2.1 PPI resources

At first, PPIs were measured at small scale. However, the development of automated technologies such as the yeast two hybrid system (Y2H)\textsuperscript{11} or protein complex isolation and mass spectrometry approaches\textsuperscript{8} has enabled the mapping of PPIs in large scale in multiple species. Nowadays large scale PPI assays are available for yeast\textsuperscript{12}, bacteria\textsuperscript{13}, worm\textsuperscript{14}, fly\textsuperscript{15}, and human\textsuperscript{16,17}. There are currently several large resources for PPIs. The BioGRID\textsuperscript{18} database contains curated sets of physical interactions, chemical associations, and post-translational modifications of 49 model organisms, 12 human viruses, and human, with a total of 1,168,521 interactions from 63,959 publications. The STRING\textsuperscript{19} database contains both known and predicted PPIs of over 2,000 model organisms and human with a total of 25,914,692 high confidence interactions (1,380,838,440 total interactions). The Hippie\textsuperscript{20} resource for human PPIs contains about 270,000 experimentally derived interactions. PPI interactions can also be extracted from the KEGG\textsuperscript{21} database, containing molecular interaction and reaction networks including curated biochemical reactions, enzyme reactions, and PPI pathways of more than 5,000 organisms. We orient the KEGG yeast network to compare to a previous approach in Chapter 2. The PPI collections of Hippie and BioGRID are large and well curated and we use them in Chapter 2, 3 and 4.

Many databases also provide pathway-views, grouping interactions into signaling pathways associated with a phenotype of interest, such as a disease or a defined biological process. Reactome Pathway Database\textsuperscript{22}, a free, open-source, curated and peer-reviewed pathway database, contains 23,419 pathways in 19 organisms. WikiPathways\textsuperscript{23} is an open, collaborative platform dedicated to the curation of
biological pathways, based on the Wikipedia platform, thus encouraging massive contribution of individual researchers; it contains 2,673 pathways for humans. Finally, KEGG provides the largest collection of pathways, containing 572,669 curated pathways which are very useful for exploring in-depth, well-studied processes, and we therefore use KEGG pathways for validation in Chapters 3 and 4.

As a pathway-view reflects the association of genes with a particular process, enrichment of groups of genes with known pathways is a common evaluation metric for shared functionality. An additional important and commonly used metric for shared functionality is the use of Gene Ontology Enrichment Analysis. The Gene Ontology\textsuperscript{24} (GO) provides the most comprehensive resource currently available for the function of genes, defining both the gene functions (‘GO-terms’) and how they relate to each other (‘relations’). The GO-terms focus on three aspects: molecular function, biological process, and cellular compartment. Shared functionality among genes is often reflected in shared GO relations or a small distance in the GO relations tree. In Chapter 4 we use both enrichment with KEGG pathways and GO enrichment analysis to evaluate the shared functionality of groups of genes.

1.2.2 PPI representation

Throughout this thesis, we represent a PPI network as a graph, consisting of nodes, representing proteins, and edges, representing physical protein-protein interactions. We focus on simple graphs, with no self-loops and parallel edges. A mixed graph is a triple $G = (V, E_U, E_D)$ that consists of a set of nodes $V$, a set of undirected edges $E_U \subseteq \{e \subseteq V : |e| = 2\}$ and a set of directed edges $E_D \subseteq (V \times V) \setminus E_U$. An undirected graph is a mixed graph in which $E_D$ is the empty set. Each edge $(i, j) \in E_U \cup E_D$ of the
network has an associated confidence score $weight(i,j)$, which may be equal to 1 for an unweighted network.

An orientation of $G$ is a directed graph $\vec{G}$ on the same vertex set $V$ whose edge set contains all the directed edges of $G$ and a single directed instance of every undirected edge, but nothing more.

### 1.3 Orientation of PPI Networks

The links in the PPI network indicate the potential presence of an interaction, yet lack information about its directionality\textsuperscript{25}. While a small subset of interactions are naturally directed (such as protein to DNA interactions), most of the directions in the PPI are unknown, calling for their computational inference. Once revealed, the oriented network could allow the construction of executable models of the cell with the ability to predict cellular behavior in response to diverse signals. It can further improve signaling-based predictions such as gene functions and connectivity and provide information of causality, such as predicting regulators leading to observed signaling events.

Most methods to orient a PPI at large scale have thus far focused on yeast. These methods rely on information from perturbation experiments, in which a gene is perturbed (cause) and as a result other genes change their expression levels (effects) to guide the assignment of directions\textsuperscript{26}. The common assumption is that for an effect to take place there must be a directed path in the network from the causal gene to the affected genes.
The resulting combinatorial problem can be formalized as follows: given cause-effect pairs represented by nodes in \( G \) the goal is to orient (assign single directions to) the undirected edges so that a maximum number of source-target pairs admit a directed path from the source to the target.

Previous work on this and related problems can be classified into theoretical and applied work. On the theoretical side, Arkin and Hassin\textsuperscript{27} studied the decision problem of orienting a mixed graph to admit directed paths for a given set of source-target vertex pairs and showed that this problem is NP-complete. We previously showed that the problem admits a sub-linear polynomial-time approximation\textsuperscript{28}. For the special case of an undirected network (with no pre-directed edges), the orientation problem was shown to be NP-complete and hard to approximate to within a constant factor of 11/12\textsuperscript{29}. On the positive side, Medvedovsky et al.\textsuperscript{29} provided an Integer Linear Program (ILP)-based algorithm, and showed that the problem is approximable to within a ratio of \( \Omega\left(\frac{1}{\log n}\right) \), where \( n \) is the number of nodes in the network. The approximation ratio was later improved to \( \Omega\left(\frac{\log \log n}{\log n}\right) \). Dorn et al.\textsuperscript{31} studied the parameterized complexity of orienting undirected networks.

On the practical side, several authors studied the orientation problem and related annotation problems. Yeang et al.\textsuperscript{26} were the first to use perturbation experiments to annotate protein networks. They proposed a probabilistic model and an accompanying inference approach to predict edge directions and signs of activation and repression from cause-effect data. Gitter et al.\textsuperscript{1} used SAT-based approximations to tackle the orientation problem. We previously provided an ILP-based solution which
avoids the need to enumerate all possible paths between a pair of genes\textsuperscript{32}. These methods are characterized by high precision but low recall.

In Chapter 1 we present a novel method, SHORTEST, which is based on the observation that signaling pathways tend to use relatively short paths. We orient a network so as to maximize the number of connected cause-effect pairs via shortest paths. The problem was shown to be NP-hard\textsuperscript{33}. We provide an ILP solution and orient the yeast PPI network. The SHORTEST method outperforms state-of-the-art methods, increasing the coverage by 8 fold while preserving a high accuracy rate.

1.4 Orientation of the Human PPI Networks

Orientation of the human PPI network poses additional challenges. First, in human we lack comprehensive causality information that was available to us for model organisms in the form of knockout experiments. Second, the size of the human PPI network (3-fold more interactions) calls for either extensive causality data to guide the orientation, or a new orientation technique that requires less training data. The only previous method to orient the human PPI network was developed by Vinayagam et al.\textsuperscript{34}. As guiding pairs for the orientation, Vinayagam et al. calculated Cartesian products of a) any membrane receptor to any transcription factor, and b) any member of a known family of membrane receptors to any member of a known family of transcription factors. They used the shortest paths induced by these pairs to calculate the probability of an edge and of its neighboring edges to participate in shortest paths and used these features to train a classifier to predict the directions of unseen interactions. However, as the correspondence between receptors and the
transcription factors they effect is only partially known, the pairs they have used represent only an approximation of the true signaling directions.

In our work, we utilize two types of experimental data on causality to orient the human PPI network: one derived from drug response experiments and the other derived from cancer omics. We review these resources below.

1.4.1 Experimental resources for causality information

In drug response experiments, a cell line is treated with a drug or small molecule, and the resulting gene expression is measured. We use such data to orient the network to connect drug targets to the genes that significantly changed their expression as a result of introducing the drug. This orientation is based on the assumption that there is a directed path from drug targets to the induced expression changes. Drug information to guide the orientation was obtained from the Connectivity Map project (CMAP, build235). CMAP contains 6,100 gene expression measurements in response to the administration of 1,309 drugs and small molecules. These measurements were taken under different drug concentrations and on different cell-line types using the Affymetrix HG-U133A and HT-HG-U133A Array. In order to form drug-specific signatures, we followed the normalization and filtering procedures described in Iskar et al.36.

The second resource for causality data is taken from cancer studies. In cancer, mutations often initiate a signaling cascade as they cause gain- or loss-of-function that affects the downstream genes and leads to changes in gene expression (see also Section 1.1). We used data generated by the TCGA Research Network to orient
the network from genetic alterations detected in the cancerous tissue, such as mutations, fusions, and copy number alterations, to the genes that significantly changed their expression in this tissue compared to a healthy tissue. The TCGA Research Network (http://cancergenome.nih.gov/) is a community resource project making data available rapidly after generation. TCGA contains patient data spanning 32 types of cancer, including DNA sequencing, miRNA sequencing, protein expression, mRNA sequencing, total RNA sequencing, array-based expression, DNA methylation, copy number alterations and clinical information.

TCGA data is used throughout this thesis. In Chapter 3 we also use RNA sequencing that was excluded from the orientation to predict key cancer genes. In Chapter 4 we use TCGA DNA sequencing, copy number alterations and RNA sequencing to predict groups of genes that drive a malignant process. We then use TCGA protein expression of the same patients to evaluate the predicted groups. Finally we use TCGA clinical data to separate the samples according to breast cancer subtypes, and analyze our results in light of the differences among the subtypes.

1.4.2 Network diffusion as a technique to explore network function and structure

Network diffusion, also termed network propagation or random walk with restart, is a computational technique used to diffuse a signal from a set of prior nodes over a network. It has been used in graph theory, statistical physics, electric engineering and machine learning\(^{37}\). In computational biology, the technique emerged as an efficient strategy to explore biological networks. It was used to amplify a biological signal such
as protein function or process membership from a group of proteins to its surroundings\textsuperscript{38}. In the context of cancer, it was used to amplify the signal of mutated genes so as to identify densely interacting mutated sub-networks\textsuperscript{39}, elucidate pathways connecting mutations to differential expression\textsuperscript{40}, and simulate abnormal signaling cascades so as to infer novel drug targets to reverse this effect\textsuperscript{41}.

The diffusion process computes a score for each protein which is the sum of a network term and a prior knowledge term. Formally, the score $F(v)$ of a node $v$ with a set of network neighbors $N(v)$ is:

$$F(v) = \alpha \left[ \sum_{u \in N(v)} F(u) \text{weight}(v,u) \right] + (1 - \alpha) Y(v)$$

where $\alpha$ is a smoothing parameter that balances between the network and the prior terms, $\text{weight}$ is normalized by the sum of outgoing edges\textsuperscript{37}:

$$\text{weight}_{v,u} = \frac{\text{weight}(v,u)}{\sum_{k \in N(v)} \text{weight}(v,k)}$$

and $Y(v)$ is a prior score to a protein, which is set to $\frac{1}{|\text{Priors}|}$ if it is part of the prior set, and to 0 otherwise. The diffusion score is computed in an iterative manner for a fixed number of steps or until convergence, as described by Cowen et al.\textsuperscript{37}. The process can be shown to converge to a steady-state distribution when the network is connected and the eigenvalues of its adjacency matrix are at most 1 in absolute value.
1.4.3 Utilizing network diffusion to orient the human PPI network using causality information

In Chapter 3 we present our novel method Diffuse2Direct, the first method to orient the human PPI network using experimentally derived guiding data. The low ratio of guiding information to network size hampers the use of previous causality-based orientation methods that rely on the network coverage induced by the causal pairs. We tackled this challenge by extending the network diffusion technique to reflect the signal traversing the network edges. For each experiment, we simulate the signal on each network edge going from a set of causal genes and the signal on each edge going into a set of affected genes. We score each direction based on its likelihood to participate in each of these signaling paths. We feed the resulting scores into a classifier to orient the network. Diffuse2Direct achieves both high coverage and high accuracy. The oriented PPI network is then utilized to improve the predictions of drug targets and cancer driver genes.

1.5 Detecting cancer driver pathways in the PPI network

The signaling activity of a cell is of particular interest in the context of cancer, where a chance accumulation of genetic alterations transforms the normal signaling pathways\(^3\). To elucidate cancer progression mechanisms it is often necessary to identify the molecular pathway transformed by the genetic alterations and the genetic alterations that drive this transformation.
1.5.1 Cancer driver genes

The first studies in the field focused on the identification of individual key genes which played a pivotal role in the progression of cancer. These genes are termed *driver* genes, indicating that a genetic alteration to these genes may drive cancer, unlike *passenger* genes where a genetic alteration is unlikely to drive the disease. One of the hallmarks of cancer is the perturbation of DNA stability, which is most likely driven by the driver genes and enables the sporadic creation of the passenger genes\(^3\). At first, the main methodology has been to filter putative driver genes according to their alteration frequency in a cohort of patients with respect to an expected background frequency\(^{42-44}\). Other prominent methods filter the candidate genes by studying their product protein structure or function. For example, CLUST\(^{45}\) studies the location of the mutation in the protein sequence, basing their ranking on the assumption that while passenger genes are distributed along the sequence, gain-of-function mutations tend to occur in specific residues or domains. ACTIVE\(^{46}\) studies the location in search of mutations in known functional residues such as phosphorylation sites. MutSig\(^{44}\) integrates three independent signals: (i) high mutational burden relative to an expected background frequency; (ii) clustering of mutations within the gene; and (iii) enrichment of mutations in evolutionarily conserved sites.

1.5.2 Cancer driver modules

Recent studies highlight the limitations of methods that focus on individual driver genes\(^{47}\). Most of all, these methods overlook the pathway level, which is crucial to characterizing malignant processes\(^{48-50}\). It is widely appreciated today that cancer is a disease not of individual mutations or genes, but of combinations of genes acting
in molecular networks corresponding to hallmark processes such as cell proliferation and apoptosis\textsuperscript{3,50,51}. Hence, an alternative view focuses on detecting cancer driver modules, a group of functionally related genes that drives a malignant phenotype.

An example of a driver module is (PI3K, S6, AKT, 4EBP1) that drives the inhibition of apoptosis in AML. Patients harboring genetic alterations in any one of the gene producing these proteins exhibit similar abnormal signaling\textsuperscript{52–55}. Furthermore their abnormal signaling may be targeted using the same inhibitors, demonstrating the importance of patient’s stratification on the pathway level\textsuperscript{56}.

The identification of cancer driver modules is important for drug development as well as patient stratification. Hofree et al. showed that the pathways perturbed by different patients can be used to stratify patients, and predict clinical outcomes such as patient survival, response to therapy or tumor histology\textsuperscript{50}. Dimitrakopoulos et al. discussed using gene modules to reposition drugs, hypothesizing that as the module members are functionally related and share a common subnetwork, they would be affected by the same drug which manipulates this subnetwork\textsuperscript{57}.

\subsection*{1.5.3 Detection of gene modules in networks}

Statistical approaches which do not rely on network strategies often search for a set of genes whose correlated expression in a cancerous tissue differs significantly from a healthy one. Their main focus is the detection of gene pairs\textsuperscript{58,59}, or integrating pairwise scores to rank a gene module\textsuperscript{60}. Alternatively, other prominent approaches search for driver modules in networks, relying on the observation that related functionality among genes can be quantified using the network structure. To detect
modules these approaches search the network for the densely connected groups of nodes, with only sparser connections between groups. The approaches are diverse in their detection method and in their definition for what constitutes a functional network (for example a PPI, coexpression, coregulation, or genetic interaction). Bar-Joseph et al. detected modules in yeast optimizing both the expression correlation among its members and the number of shared transcription factors regulating them. Tarnow et al. provided a method which can be used to integrate any two network-based data sources to find gene modules: they first detect a module using one network (a coexpression network for example) and then calculate the correlation strength of this group in another network (a PPI network for example). Their goal is to find modules with correlation strength that was unlikely to be found by chance. Tanay et al. suggested to rely on multiple sources of supporting information and integrated a collection of genome wide datasets, including gene expression, protein interactions, growth phenotype data, and transcription factor binding, to reveal the modular organization of the yeast system. Their goal is to find maximal gene modules that manifest a unique, common behavior across a significant set of different data sources.

Other detection methods utilize clustering approaches over the PPI network. The goal of the Markov Cluster algorithm (MCL) is to detect dense regions in a network. MCL performs iterative steps of expansion and inflation, both based on random walk. They assume that when a random walk starts from a gene within a module it is more likely for the walk to stay within the module than to walk to another one. This effect is boosted using the expansion step, which gives preference to revisiting nodes over
seeing new nodes. MCL converges quickly to an idempotent Markov matrix, whose structural properties are interpreted as a clustering of the given network. Nepusz et al.\textsuperscript{66} devised ClusterOne, a method for finding dense regions in a PPI. ClusterOne starts from single node and greedily adds and removes nodes to find groups with high cohesiveness. This process forms multiple, possibly overlapping groups. In the next step ClusterOne merges groups with high overlap. In the last step ClusterOne discards very small or sparse groups. ClusterOne may leave overlaps, reflecting the fact that proteins often have multiple functions and may belong to more than one module. They showed that the detected modules were highly enriched with known complexes. Newman et al.\textsuperscript{61} formulated the optimization task as searching for the maximal value of modularity over possible divisions of a network, where modularity is defined as the number of edges falling within groups minus the expected number in an equivalent network with edges placed at random. They showed that the modularity can be expressed in terms of the eigenvectors of a characteristic matrix for the network, termed the modularity matrix, and exploited this transformation to create a spectral algorithm for community detection in the network.

The approaches reviewed above aim to optimize the relations among genes, such as the connectivity of the group. Alternatively, a gene module may be defined as a connected subnetwork maximizing the individual scores of its members. Ideker et al.\textsuperscript{67} defined the problem as finding optimal node-scoring subgraphs and designed a method to detect connected regions of the PPI network that show significant changes in expression over particular subsets of conditions.
1.5.4 Detection of gene modules in cancer

Detecting modules in cancer provides additional layers of data that may be exploited. The first cancer-related layer is differential expression. Shrama et al. detected modules from a gene-gene interaction network, and then analyzed the expression patterns of the module members distinguishing between non-metastasis and metastasis tissue. Akavia et al. predicted driver genes in melanoma, and expanded from the driver to the differentially expressed genes likely governed by it. The second layer which focuses on the connectivity of mutated genes relies on the assumption that driver genes will be close to one another in the PPI network, while passenger genes will be spread randomly in the network. One of the leading approaches in this field is HotNet2, discussed in detail in Chapter 4, which seeks connected subnetworks with high proximity to mutated genes. The third layer relies on a pattern of mutual exclusivity, quantifying the observation that sets of genes that are commonly involved in the same cancer pathways tend not to be mutated together in the same patient. As powerful as each of these observations is individually, integrating them has the potential of improving cancer module detection significantly.

In Chapter 4 we describe ModulOmics, our method for the de-novo identification of cancer driver modules based on the integration of PPI networks, mutual exclusivity of DNA alterations (SNVs and CNAs), activity and co-expression at the RNA level, and co-regulation by shared transcription factor which are active in patients.
1.6 EXECUTABLE COMPUTATIONAL MODELS

Executable computational models of malignant cell signaling utilize the knowledge of PPIs and key disease-driving genes. They use the PPI network as a scaffold for the potential signaling cascade, and key genes as focal points to trigger abnormal signaling. A computational model in this context is a formal description of cellular behavior in which each cell component is translated into a computational component. The unique genetic alterations are represented by variables, indicating the presence and level of each alteration. The genetic alterations then activate their direct downstream proteins according to specific rules that can be represented by mathematical functions. The output of each function can then affect the downstream proteins, creating a cascade of functional alterations. In the cell, the activation cascade may lead to the nucleus where it changes gene expression and thus controls the cell phenotype, represented in the program output as the activity level of different cellular processes (apoptosis, proliferation, etc.).

1.6.1 Modeling node activity in a network

Computational models of biological processes were successfully used to analyze mammalian signaling pathways, biochemical processes and facilitate drug discovery. The scale of a computational model determines its level of detail, and therefore balances the complexity of the model and the faithful representation of signaling activity. A quantitative computational model translates a signaling cascade into a set of nonlinear ordinary differential equations (ODEs) that describe how all the interacting components are changing with time. These equations offer a mechanistic, chemically-based view on the change in the level of cellular species as a function of
the levels of their interactors\textsuperscript{82}. However, the dependency of such detailed models on parameter estimation requires a particularly rich experimental data, which is unattainable at the moment for moderate or large scale modeling\textsuperscript{83}. Hence, in this thesis we will not focus on ODE models, however they can be used for follow-up work to further refine the mechanisms identified by our larger scale computational models.

Recently, \textit{Boolean networks}, pioneered by Kauffman in 1969\textsuperscript{84}, are emerging as a successful platform for large scale modeling of signaling pathways\textsuperscript{85–87}. In these models, proteins are still translated to variables and interactions to functions, with the difference that proteins are either active or inactive \{0=off, 1=on\}. For every substance, a target function reads the values of other substances that affect it and calculates the change, if any, to the current substance. To further relax the model, target functions may be restricted to pre-specified types of functions.

Later, Kauffman\textsuperscript{84} extended Boolean networks to \textit{Qualitative Networks} (QN) by allowing variables to range over larger discrete domains and replacing Boolean functions by algebraic functions\textsuperscript{88}. Intuitively, a Qualitative Network associates a discrete variable with every substance the model follows. The variable ranges over a small discrete domain where values represent expression levels of the substance such as \{0=off, 1=low, 2=medium, 3=high, 4=maximum observed\}. The substance changes gradually to attain this target.
1.6.2 Finding stability in executable models using the Bio Model Analyzer tool

In this section we will review an executable model that is used to perform *in-silico* experiments testing different contexts (such as healthy state, or a cell altered by disease state) under different perturbations (such as an inhibition of a protein simulating the administration of a drug). This model is based on a qualitative network of substances, each containing a target function, which receives as input the value of the upstream substances. A state of the model is a specific value assignment for each of the substances, and a step in the model is an update of the assignments, based on the computation of the target functions. The Bio Model Analyzer tool (BMA)\textsuperscript{89} tool provides infrastructure to construct executable models and to explore the state-space, searching for either a steady state of the model or a counter example for convergence. The exploration is conducted not on the entire system as a whole but rather locally on small pieces of the system. The key to efficient exploration is that the model can be decomposed into restricted components from which the generalized behavior of the model can be inferred. Specifically, if a component diverges than the entire model diverges, while if a component converges to a steady state then this component can be fixed onto the steady state thus restricting the state-space.

In Chapter 5 we use the BMA tool to construct an executable network model for AML, and then use the executable model to predict drug efficacy for the four major subtypes of AML. To reconstruct the AML network we infer the key genes active in the AML cohort of patients, and then link them through interactions detected by text mining. We further infer the activation function of each protein using a simulated annealing
process striving to minimize the gap between the observed expression to the modeled one. We show that the resulting model is extremely effective in predicting drug synergy for first-line treatment, and can be automatically updated with the progression of the tumor to suggest effective second-line in cases where relapse occurs.
1.7 Summary of the papers included in this thesis

This thesis is based on the following papers:

1.7.1 Network orientation via shortest paths\textsuperscript{90}

In this paper, presented in Chapter 2, we studied the graph orientation problem, which calls for orienting the edges of a graph so as to maximize the number of pre-specified source-target vertex pairs that admit a directed path from the source to the target. Most algorithmic approaches to this problem share a common preprocessing step, in which the input graph is reduced to a tree by repeatedly contracting its cycles. While this reduction is valid from an algorithmic perspective, the assignment of directions to the edges of the contracted cycles becomes arbitrary, and the connecting source-target paths may be arbitrarily long. In the context of biological networks, the connection of vertex pairs via shortest paths is highly motivated, leading to the following problem variant: given a graph and a collection of source-target vertex pairs, assign directions to the edges so as to maximize the number of pairs that are connected by a shortest (in the original graph) directed path. This problem is NP-complete and hard to approximate to within sub-polynomial factors. Here we provide a first polynomial-size integer linear program formulation for this problem which allows its exact solution in seconds on current networks. We applied our algorithm to orient PPI networks in yeast and compared it to two state-of-the-art algorithms. We found that our algorithm outperforms previous approaches and can orient considerable parts of the network thus revealing its structure and function.
1.7.2 A diffusion-based method for orienting the human protein-protein interaction network\textsuperscript{91}

As our end goal is to gain insight into the human system, following the orientation of the yeast PPI network we turned to orient the human PPI network. Revealing the orientation of the human PPI network can enhance the modeling process of human diseases. In this paper, presented in Chapter 3, we introduce a systematic approach to orient the human PPI network using drug response and cancer genomics data. We provide a novel diffusion-based method for the orientation task that significantly outperforms existing methods. The oriented network leads to improved predictions of cancer driver genes and drug targets.

1.7.3 ModulOmics: Integrating Multi-Omics Data to Identify Cancer Driver Modules\textsuperscript{92}

The identification of pathways driving cancer progression is a fundamental problem in tumorigenesis. Most current approaches to address this problem use primarily somatic mutations, not fully exploiting additional layers of biological information. In this paper, presented in Chapter 4, we describe ModulOmics, a method to identify cancer driver pathways \textit{de novo} by integrating multiple data types (protein-protein interactions, mutual exclusivity of mutations or copy number alterations, transcriptional co-regulation, and RNA co-expression) into a single probabilistic model. Across several cancer types, ModulOmics identifies highly functionally connected modules enriched with cancer driver genes, outperforming state-of-the-art methods. For breast cancer subtypes, the inferred modules recapitulate known molecular mechanisms and suggest novel subtype-specific functionalities.
1.7.4 Towards Virtual Models of Cancer: Cell-Specific Computational Modeling of the PIM pathway in Acute Myeloid Leukemia

Personalized therapy is a major goal of modern oncology, as patient responses vary greatly even within a histologically defined cancer subtype. This is especially true in AML, which exhibits striking heterogeneity in molecular segmentation. When calibrated to cell-specific data, executable network models can reveal subtle differences in signaling that help explain differences in drug response. Furthermore, they can suggest drug combinations to increase efficacy and combat acquired resistance.

In this paper, presented in Chapter 5, we constructed an executable AML model. The model is based on experimental testing of diverse AML cell lines, each representing an AML subtype, treated with multiple inhibitors while recording both the induced protein expression and resulting cell behavior. The model is based on general AML pathways and can be made subtype specific by switching-on the genetic alterations present in one of the studied cell lines. The model captures key differences in signaling among the cell lines, enabling to later predict unique cell line responses in expression and cell behavior. Furthermore, using cell-specific models we tailored combination therapies to individual cell lines, and successfully validated their efficacy experimentally. In particular, we show that cells with mild sensitivity to PIM inhibition show increased sensitivity in combination with PIK3CA inhibition. We also used the model to infer the origin of PIM resistance engineered through prolonged drug treatment of MOLM16 cell lines, and successfully validated experimentally our prediction that this resistance can be overcome with AKT1/2 inhibition.
2 Network orientation via shortest paths

Models that simulate protein activity and the signaling cascades traversing them require first and foremost a confident skeleton of PPIs to reflect the potential of physical interactions among proteins. The presence of PPIs is detectable at large scale, yet their directionality of signal flow must be computationally inferred. While many of these interactions carry directed signaling information, current PPI measurement technologies, such as yeast two hybrid and co-immunoprecipitation, reveal the presence of a signal flow but not its directionality. Some interactions are naturally directed, such as kinase-substrate and phosphatase-substrate interactions (KPIs), yet the directions of the vast majority of PPIs remain unknown. Identifying this directionality is fundamental to our understanding of how signaling networks function and was thus our first and main effort.

From a theoretical point of view, graph orientation problems have been studied by several authors. In Section 1.3 we presented theoretical results for orienting a network by connecting cause-effect pairs via any path. In the following we emphasize algorithmic results attending also to the length of the connecting paths, which are the focus of this work. Block et al. studied the problem of orienting the edges of an undirected graph so as to maximize the number of pre-specified source-target pairs that admit a directed shortest path between them (i.e., a directed path whose length is equal to the distance between them in the unoriented graph). They showed that for an undirected graph with a vertex set of size n, edge set of size m, source-target pair set of size P and paths whose length is bounded by k the problem is hard to approximate to within a factor of $O\left(\min\{\sqrt{m}, k\}^{1-\varepsilon}\right)$ for any $\varepsilon > 0$ and provided an
approximation algorithm with a ratio of \( \Omega\left(\frac{1}{\max \{n,|P|\}^{1/2}}\right) \). For the k-length-bounded orientations in a weighted undirected graphs, Gitter et al.\(^1\) proposed an \( O\left(\frac{2^k}{k}\right)\)-approximation to maximize the weight of all satisfied paths between sources and targets with length at most k, and proved this variant to be NP-hard.

In the biological domain, previous orientation studies has mostly used unsupervised methods, with the exception of Vinayagam et al.\(^34\) (see detailed description of the method in Section 1.4). As described in Section 1.3, previous unsupervised methods relied on information from perturbation experiments, yielding pairs of cause and effect to guide the inference\(^26\). They assumed that for an effect to take place there must be a directed path in the network from the causal gene to the affected gene. As there are many paths that can link two proteins in the interaction network, previous solution methods relied on either length-bounded paths\(^1,26,96\), or parameterized and integer programming techniques\(^29,31,32\). The former are limited by a fixed length of the considered paths (at most 5). The latter, while producing highly precise predictions, are very limited in their coverage as they start by contracting cycles of the input network, eliminating the vast majority of the interactions (>92% on yeast network) from further consideration.

It is likely that biological responses are controlled by relatively short signaling cascades\(^1,34,97\); however, in a large scale network, enumerating all possible paths between two nodes can still be computationally intractable, even when considering only paths of limited length. In this chapter, we propose an orientation method that overcomes the limitations of previous approaches by focusing attention on shortest
paths which allows the efficient representation of all such paths and avoids the contraction problem. At the same time it does not rely on a fixed path length. The method aims to orient the edges of an input graph $G$ (see Section 1.2 for definition) so that a maximum number of input source-target pairs admit a directed source-to-target path whose length is the shortest possible (i.e., equal to distance between the two nodes in $G$). Our algorithm builds on a first efficient (for realistic instances) ILP formulation of the problem, allowing the computation of optimal solutions in seconds on current networks. We applied our method to large scale data sets of yeast physical interactions and evaluated it using cross-validation experiments. Our method outperformed two previous state-of-the-art orientation methods $^{1,32}$ by a significant margin. The source code is available at www.cs.tau.ac.il/~roded/shortest.zip

The results in this chapter are based on Network orientation via shortest paths by Silverbush et al.$^{90}$

### 2.1 Results

We applied our method to orient the yeast network relying first on the information obtained from the shortest paths connecting causality pairs, and then iteratively expanding to rely on the next shortest paths (see Section 2.3).

#### 2.1.1 Data Acquisition and Integration

We gathered physical interactions and cause-effect pair information for the yeast *Saccharomyces Cerevisiae* from different sources. We used the PPI data set 'Y2H-union' from Yu et al.$^{98}$, which contains 2,930 highly-reliable undirected interactions
between 2,018 proteins. Protein DNA interactions (PDIs), which are directed by nature, were taken from Maclsaac et al.\textsuperscript{99}, an update of which can be found at (http://fraenkel.mit.edu/improved_map/). We used the collection of PDIs with $p<0.001$ conserved over at least two other yeast species, which consists of 4,113 unique PDIs spanning 2,079 proteins. KPIs were collected from Breitkreutz et al.\textsuperscript{100} by taking the directed kinase-substrate interactions out of their data set. This results in 1,361 KPIs among 802 proteins. We used a set of 14,427 knockout pairs between 2,870 genes and proteins from\textsuperscript{101} by taking their set of unrefined and unfiltered knockout pairs and filtering all pairs with $p<0.001$.

We integrated the data to obtain a physical network of undirected and directed interactions. We removed self-loops and parallel interactions; for the latter, whenever both a directed and an undirected edge were present between the same pair of nodes, we maintained the former only. Pairs of edges that are directed in opposite directions were integrated as an undirected edge. The resulting physical network spans 3,686 proteins, 2,655 PPIs, 4,091 PDIs and 1,359 KPIs.

2.1.2 Application and Performance Evaluation

We implemented our algorithm, which we call SHORTEST, in C++ using BOOST C++ libraries (version number 1.43.0) and the commercial IBM ILOG CPLEX optimizer (version number 12.5) to solve ILPs.

To evaluate the orientations suggested by our algorithm we ran the algorithm in a cross-validation setting, hiding the directions of the larger subset of known directed interactions, the PDIs. This subset was considered as undirected test edges. Guided
by the set of knockout pairs, our program computes orientations for all undirected edges, including the test edges.

We tested our algorithm using the efficient version of Section 2.3.3, taking advantage of the fact that the knockout pairs were derived from a small set of shared sources. When employing 10 expansion rounds the algorithm oriented with confidence 3,379 (82.6%) of the test edges, orienting correctly 2,283 (67.6%) of them (hypergeometric $p$-value $< 2.7^{-159}$). When restricting the confidence cutoff to 2 ($c_e \geq 2$) and performing a single iteration (i.e., no expansion rounds), the algorithm oriented with confidence 902 (22%) of the test edges with confidence, orienting correctly 714 (79.2%) of them (hypergeometric $p$-value = 0). To further evaluate our algorithm we define precision and recall as defined by Vinayagam et al.$^{34}$: we considered each interaction as two different instances, where the interaction from A to B is defined twice, as $A \rightarrow B$ representing its positive instance, and $B \rightarrow A$ representing its negative instance. If an orientation $A \rightarrow B$ exceeds the confidence cutoff then it is classified as positive. As expected, a higher confidence cutoff yields a higher precision [$TP/(TP + FP)$], and lower recall [$TP/(TP + FN)$]. More rounds of expansion increase recall and lower the precision. The performance of the algorithm for different choices of a confidence cutoff and number of expansion rounds is summarized in Figures 2.1 and 2.2.
As discussed earlier, co-complex interactions are not likely to have a preferred orientation. To test how frequently our algorithm assigns directions to such interactions, we evaluated our results against a list of 393 known complexes as annotated by GO\textsuperscript{24} and downloaded from SGD\textsuperscript{102} (June 2011). We found that 9% (720) of the interactions in the network lie within known complexes. Using a confidence cutoff of 0, 50% of all interactions in the network are assigned a direction, but only 3.4% of those (139 interactions) lie within known complexes. This number is significantly small compared to the random expectation (hypergeometric $p$-value $<1.17^{-70}$) testifying to the quality of our predictions.
To study the effect of the amount of cause-effect pairs on the orientation, we applied the algorithm with increasing fractions (chosen at random) of pairs. As evident from Figure 2.3, the more pairs the higher are the measured recall and precision, supporting our use of the cause-effect pairs to guide the orientation. It is interesting to note that a high percentage (approximately 93%) of the knockout pairs are satisfied throughout the experiments. It is also of note that our results seem more robust to variations in the percent of directed interactions in the input network. Even when eliminating the KPIs, an F-measure of 0.26 was attained (cutoff = 0, no expansion rounds).

To evaluate the scalability of the method we downloaded the full set of PPIs from BioGRID\textsuperscript{18} (October 2013), containing 89,512 unique interactions. While the preprocessing time increased to 5 minutes, the ILP solution was still obtained in under a second. The orientation obtained had very high quality: 2,061 (50%) out of the 4091
PDIs were oriented with confidence, and 1691 (82%) of the orientations were accurate (cutoff = 0, no expansion rounds). Expectedly, using a higher confidence cutoff yielded higher precision values (e.g., 90% for a cutoff of 1).

### 2.1.3 Comparison to Previous Work

We compared our approach to two previous state-of-the-art methods: the MIXED algorithm\(^{102}\) and the random orientation followed by local search algorithm of Gitter et al.\(^1\). Below we provide precision, recall and F-measure values for each approach, where the latter combines both precision and recall into a single value of their harmonic mean: 

\[
F\text{-measure}=2\left(\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}\right).
\]

The results of the comparison to MIXED are shown in Figure 2.4, where SHORTEST was applied in its maximum coverage setting. The 8-fold recall increase by SHORTEST is dramatic, albeit at the price of lower precision. To investigate the precision differences further, we compared both algorithms performances over the subset oriented by both. In this subset both approaches inferred identical orientation, achieving 94.8% precision over this evidently easier subset.

We employed an analogous procedure to infer the directions of the smaller subset of KPIs, hiding its directions while keeping the PDIs as a directed subset and guided by the same set of knockout pairs as in the previous experiment. We were able to orient with confidence 1,077 (79.2%) of them, orienting correctly 600 (56%) with an F-measure of 0.49. In comparison, the MIXED approach oriented 52 (3.8%) of the edges, orienting correctly 46 (88.4%) with an F-measure of 0.065. As before, within
the easier subset of edges oriented by both methods, a much higher precision of 85.4% was obtained.

To compare to the method of Gitter et al.\(^1\) we used the same data and network used in their paper (Anthony Gitter, private communication). The unweighted PPI network was extracted from BioGRID\(^1^8\) spanning 3,446 proteins and 10,945 undirected interactions. To guide our orientation we used the same source-target pair set manually chosen by Gitter et al.\(^1\). A test set was extracted from KEGG\(^2^1\) and the Science Signaling Database of Cell Signaling. It contains 91 directed interactions among 69 proteins. Gitter et al. oriented 2,447 interactions of the initial network. Their orientation oriented 55 of the 91 tested directed interactions (60.4%), out of which 37 orientations were accurate (67.3%). Applying our algorithm in its maximum recall setting, resulted in orienting 5,458 interactions among 2,221 proteins, more than a 2-fold increase compared to Gitter et al. On the test set our method oriented with confidence 79 interactions (86.8%), out of which 61 orientations were accurate (77.2%), providing both higher recall (67.0% vs. 40.7%), higher precision (77.2% vs. 67.3%) and achieving a higher F-measure of 0.71 compared to 0.51. The comparison is summarized in Tables 2.1 and 2.2 and depicted in Figures 2.5 and 2.6.

<table>
<thead>
<tr>
<th>Table 2.1. A comparison of GITTER and SHORTEST for different numbers of expansion rounds.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHORTEST</td>
</tr>
<tr>
<td>#0</td>
</tr>
<tr>
<td>Precision</td>
</tr>
<tr>
<td>Recall</td>
</tr>
<tr>
<td>F-measure</td>
</tr>
</tbody>
</table>
Table 2.2. A comparison of GITTER and SHORTEST for different confidence cutoffs.

<table>
<thead>
<tr>
<th></th>
<th>SHORTEST $conf &gt; 0$</th>
<th>SHORTEST $conf \geq 1$</th>
<th>SHORTEST $conf \geq 2$</th>
<th>SHORTEST $conf \geq 3$</th>
<th>GITTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>77.2%</td>
<td>74.6%</td>
<td>82.0%</td>
<td>90.9%</td>
<td>67.3%</td>
</tr>
<tr>
<td>Recall</td>
<td>67.0%</td>
<td>58.2%</td>
<td>45.1%</td>
<td>33.0%</td>
<td>40.7%</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.72</td>
<td>0.65</td>
<td>0.58</td>
<td>0.48</td>
<td>0.51</td>
</tr>
</tbody>
</table>

Figure 2.5: Performance comparison of GITTER and SHORTEST using different number of expansion rounds. Labels denote the number of expansion rounds used by SHORTEST.

Figure 2.6: Performance comparison of GITTER and SHORTEST using different confidence cutoff. Labels denote the confidence cutoff used by SHORTEST.

As pointed out by Gitter et al.\textsuperscript{1}, scalability is an important issue for methods analyzing high-throughput data sets, especially because current data is incomplete and networks for other organisms may be larger than those for yeast. Thus Gitter et al.\textsuperscript{1} had examined their running time using the different algorithms suggested in their paper. We compared our running time when using the same data set. Running our
algorithm one iteration only (which includes running the ILP for each test edge to determine confidence), reaching next to identical recall as Gitter et al.\textsuperscript{1} and much higher precision, took 8 seconds in average, faster than all algorithms used in Gitter et al.\textsuperscript{1}.

Running our approach when allowing the maximum number of expansion rounds up to exhaustion (reaching 5 iterations in total), reaching higher recall and higher precision, took 75 seconds, challenged only by the random approach. Running time is presented in Table 2.3.

Table 2.3. Algorithm run times in seconds. The full data set was used, with 256 source-target pairs.

<table>
<thead>
<tr>
<th></th>
<th>SHORTEST</th>
<th>Gitter et al.</th>
</tr>
</thead>
<tbody>
<tr>
<td>no expansion unlimited</td>
<td>8.0</td>
<td>75.0</td>
</tr>
<tr>
<td>unlimited expansion</td>
<td>16.2</td>
<td>2742.5</td>
</tr>
<tr>
<td>RANDOM</td>
<td></td>
<td>10806.7</td>
</tr>
<tr>
<td>MIN-SAT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAX-CSP</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.2 Conclusions

The orientation of a network is key to understanding its function. Here we have presented the SHORTEST approach, which allows us for the first time to confidently orient the majority of the edges in a network. The most natural use of such an orientation is in enhancing methods for pathway inference. Specifically, current pathway inference algorithms, like Scott et al.\textsuperscript{104}, receive as input an undirected PPI network and search for likely paths that start and end at specific proteins. The search space of such methods can be greatly reduced by the orientation information.
To test the potential utility of our orientation method in a pathway inference context, we checked its power in filtering candidate pathways by their agreement with the predicted orientations. To this end, we inspected the source-target pairs connected by more than one possible pathway within the networks in Section 2.2.3: (i) In the orientation instance due to Gitter et al. there are 740 possible shortest paths from source to target (6.2 paths per source-target pair). However, when filtering these paths against the confident orientation predictions, only 589 (4.9 on average per pair) remain. (ii) In the orientation instance due to Silverbush et. al. there are 46,782 shortest paths from source to target (11.9 on average per pair), of which only 28,273 (7.2 on average per pair) agree with the confident orientations. Thus, the use of our method can potentially reduce the search space by up to 40%.

### 2.3 Methods

#### 2.3.1 Problem definition and the shortest-paths graph

A path of length $\ell$ in $G$ is a sequence of distinct nodes $\langle v_1, \ldots , v_{\ell+1} \rangle$ such that $(v_i, v_{i+1}) \in E_U \cup E_D$ for every $1 \leq i \leq \ell$. We denote by $d_G(s, t)$ the distance between nodes $s$ and $t$ in $G$. We say that a pair of nodes $(s,t)$ is satisfied by an orientation $\vec{G}$ when the latter graph contains at least one directed path from $s$ to $t$ whose length is $d_G(s,t)$, i.e., equal to the length of a shortest $s$-to-$t$ path in $G$.

The Maximum Shortest-path Orientation (MSPO) problem is defined as follows:

**Input:** A mixed graph $G = (V, E_U, E_D)$ with non-negative edge weights $w(e)$ for every $e \in E_U \cup E_D$, and a collection of source-target vertex pairs $P = \{(s_1, t_1), \ldots , (s_k, t_k)\}$. 

**Objective:** An orientation of $G$ that satisfies (via shortest paths) a maximum number of pairs from $P$.

MSPO was shown to be NP-hard by Block et al. Here we provide a polynomial size ILP for it that allows solving it to optimality in seconds on current networks. For ease of presentation, we focus on unweighted graphs but the algorithm can be easily generalized to the weighted case. A key component of the algorithm is an efficient representation of shortest paths. For an ordered pair $(s,t)$ of nodes, we define their *shortest paths graph* $G_{(s,t)} = (V, E_{(s,t)})$ to be a directed graph on $V$ which consists of all edges that reside on a shortest path from $s$ to $t$ in $G$. $G_{(s,t)}$ can be efficiently constructed by employing breadth-first-searches from $s$ and $t$ (the latter, after reversing all the directed edges). Now for every edge $(u,v)$ such that $u$ is reachable from $s$ and such that $t$ is reachable from $v$, we include it in $G_{(s,t)}$ if and only if $d_{G_{(s,t)}} = d_{G_{(s,u)}} + d_{G_{(v,t)}} + 1$. It is easy to see that each shortest path in $G$ is a shortest path in $G_{(s,t)}$. Furthermore, each path in $G_{(s,t)}$ is a shortest path in $G$ by definition of its edges.

### 2.3.2 The integer program

The ILP formulation is based on checking pair connectivities via flow based computations. The basic observation is that a pair $(s,t) \in P$ is satisfied by a given orientation if and only if $t$ is reachable from $s$ in $G_{(s,t)}$. The latter condition can be rephrased as allowing us to send one unit of flow from $s$ to $t$. Our ILP consists of a set of binary orientation variables, describing the edge orientations, flow variables
describing the flow on edges of $G_{(s,t)}$ for every $(s, t) \in P$ and binary closure variables describing reachability of every pair in $P$. Formally, the variables are:

1. $\{o_{(u,v)}, o_{(v,u)} \in \{0,1\} \mid (u, v) \in E_U \cup E_D\}$

2. $\{f^{(s,t)}_{(u,v)} \in [0,1] \mid (s, t) \in P \land (u, v) \in E_{(s,t)}\}$

3. $\{c_{(s,t)} \in \{0,1\} \mid (s, t) \in P\}$

The orientation variables in (1) are used to encode orientations of the edges: an assignment of 1 to $o_{(u,v)}$ means that the undirected edge $(u, v)$ is oriented from $u$ to $v$. We set $o_{(u,v)} = 1$ and $o_{(u,v)} = 0$ for every directed edge in $E_D$. The flow variables in (2) are used to measure the existence of flow on every edge of $G_{(s,t)}$ for every pair $(s, t) \in P$. Unlike traditional flow problems, the amount of flow does not concern us, but rather its existence. The closure variables in (3) are used to represent which vertex pairs of the graph are satisfied: an assignment of 1 to $c_{(s,t)}$ will imply that the orientation satisfies $(s, t)$.

The objective is:

4. maximize $\sum_{(s,t)\in P} c_{(s,t)}$

The constraints are:

5. $o_{(u,v)} + o_{(v,u)} = 1$ for all $(u, v) \in E_U$

6. $o_{(u,v)} = 1, o_{(v,u)} = 0$ for all $(u, v) \in E_D$

7. $f^{(s,t)}_{(u,v)} \leq o_{(u,v)}$ for all $(s, t) \in P, (u, v) \in E_{(s,t)}$

8. $f^{(s,t)}_{(u,v)} \leq \sum_{w:(u,w)\in E_{(s,t)}} f^{(s,t)}_{(w,u)}$
9. for all \((s, t) \in P, (u, v) \in E(s,t), u \neq s\)

\[ c_{(s,t)} \leq \sum_{(w,t) \in E(s,t)} f_{(w,u)}^{(s,t)} \text{ for all } (s,t) \in P \]

Constraints in (5) ensure that each undirected edge is oriented in exactly one direction. We note that in case an interaction is unlikely to be directed (e.g., a co-complex interaction), the constraints in (5) can be modified to allow both directions. As we show in the sequel, the assignment of confidence scores to edge orientations allows us to (automatically) refrain from assigning directions to the vast majority of such interactions. Constraints in (6) ensure that the chosen orientations are consistent with the directed edges of \(G\). Constraints in (7)–(9) ensure that for every pair \((s, t) \in P\) the closure variable \(c_{(s,t)}\) can be set to 1 only if there is a flow from \(s\) to \(t\) in \(G(s,t)\). In detail, the constraints in (7) ensure that the flow respect the edge directions; the constraints in (8) ensure that no edge carries a flow from a vertex \(u\) if there is no incoming flow to \(u\); and the constraints in (9) ensure that the pair \((s, t) \in P\) is not satisfied if there is no flow from \(s\) to \(t\) in \(G(s,t)\). The overall size of the ILP is \(O(|P||E_D \cup E_U|)\).

2.3.3 A more efficient formulation

The ILP formulation above can be made more efficient by observing that in a biological knockout experiment one measures simultaneously all the effects of a certain knockout (cause), thus many pairs in \(P\) share the same source. We show below how to unify all the flow computations of a given source thus significantly reducing the number of variables and subsequently the time of solving the ILP.
Let $S(P)$ be the set of source nodes in $P$ and denote by $M_s = \{(s, t_1), \ldots, (s, t_i)\}$ the set of pairs with source $s$. Let $G_s = \bigcup_{i=1}^{t} G(s, t_i)$ be the union of all shortest path graphs with $s$ as a source. Denote its set of edges by $E_s$. We introduce the following updates to the flow variables and constraints of the program:

10. $\{f^s_{(u,v)} \in \mathcal{R}^+ \mid s \in S(P) \text{ and } (u, v) \in E_s\}$

11. $f^s_{(u,v)} \leq o_{(u,v)}$ for all $s \in S(P), (u, v) \in E_s$

12. $f^s_{(u,v)} \leq \sum_{(w,u) \in E_s} f^s_{(w,u)}$

13. for all $s \in S(P), (u, v) \in E_s, u \neq s$

$$c_{(s,t)} \leq \sum_{(w,t) \in E_s} f^s_{(w,t)} \text{ for all } (s, t) \in P$$

Clearly, any $(s, t)$-path identified by the previous ILP can be used in this formulation as well. Thus, correctness of this formulation depends on showing that any $(s, t)$-path it identifies must be a shortest path. Suppose to the contrary that there exists an $(s, t)$-path $\mathcal{L} \in G_s$ such that $\mathcal{L} \notin G_s(s,t)$. Let $p$ be the maximal prefix of $\mathcal{L}$ that is a prefix of some shortest path in $G(s,t)$ and let $(u, v)$ be the first edge on $\mathcal{L}$ that is not included in $p$ (i.e., $p$ diverges at $u$). Since $(u, v)$ belongs to some $G(s,t')$ (by definition of $G_s$), $p$ is also a prefix of a shortest $(s, t')$-path that traverses $(u, v)$. By our assumption, $(u, v)$ is not on a shortest path to $t$ and, hence, the same property holds for all the remaining edges on $\mathcal{L}$ including the last one – $(w, t)$. This results in a contradiction since the presence of $(w, t)$ in $G_s$ implies it must be on a shortest $(s, t)$ path.
2.3.4 Computing orientation confidence

In principle, there could be many optimal solutions to the orientation problem. Hence, some of the edges may be arbitrarily oriented in the sense that both of their directions can be used in some optimal solutions. Let $s_{opt}$ be the value of an optimal solution. To compute a measure of confidence in a given orientation of an edge $e = (v, w)$ we rerun the ILP while forcing $e$ to carry the opposite orientation $(w, v)$. We set its confidence value to $c_e = s_{opt} - s_e$ where $s_e$ is the maximum number of satisfied pairs for the modified instance. If $c_e > 0$ then the direction of $e$ is the same in all optimal solutions; thus, we keep $c_e$ as the edge’s confidence. If $c_e = 0$ and $e = (v, w)$ is on a shortest $s$-$t$ path then there are two cases to consider: (i) the opposite edge is on a shortest $s'$-$t'$ path; in this case an orientation of $e$ may be arbitrary and both $(v, w)$ and $(w, v)$ will be assigned with a confidence of 0. (ii) The opposite edge $(w, v)$ does not participate in any shortest path; thus, there must be some parallel path from $s$ to $t$ hat does not visit $e$, allowing it to be oppositely oriented without altering the objective. In this case $e$ is assigned a positive confidence according to its contribution to satisfying pairs. Precisely, each pair $p$ that $e = (v, w)$ is used to satisfy contributes $\frac{1}{n_p}$ to its confidence, where $n_p$ is the number of nodes $u \in G_p$ such that $d(s, u) = d(s, w)$ (reflecting the number of alternatives to using $e$). Now a cutoff may be defined, and an edge $e$ is said to be oriented with confidence if and only if its confidence exceeds the cutoff.
2.3.5 Handling large networks

The repeated solution of many ILP instances may be very costly for large network. We therefore extended SHORTEST to include weights on the pairs guiding the orientation. When all the weights are assigned 1, SHORTEST computes $S_{opt}$. To each weight we subsequently add random noise by sampling from a Gaussian distribution with mean 0 and variance 0.1. We repeat this process 1000 times and rerun the ILP. The final per-pair score assigned to a direction of an interaction is the fraction of times (out of 1000 repeats) it was oriented in that direction. The two methods for computing confidence are comparable (Figure 2.7).

Figure 2.7. Performance evaluation of SHORTEST against its extension to large networks (no expansion rounds). We scaled SHORTEST to large networks by changing confidence assignment. The extended method is comparable to the original SHORTEST, with 71% accuracy compared to 73%.

2.3.6 Iterative expansion

In an application of our orientation algorithm, edges that are assigned zero confidence remain undirected. To expand our orientation to include some of those edges, we run several iterations of the algorithm. In each iteration, the directions of edges that were confidently oriented in the previous iteration are held fixed, while the criteria for satisfying the source-target pairs are modified to allow additional orientations. Specifically, let $G_i = (V_i, E_{U_i}, E_{D_i})$ be the input graph at iteration $i$. For a given source-
target pair \((s, t)\), let \(E_{i(s,t)} \subseteq E_{D_i}\) be the set of all directed edges occurring on directed paths from \(s\) to \(t\) in \(G_i\). Then we define \((s, t)\) to be satisfied by a given orientation of \(G_i\) if and only if it admits a directed path under this orientation such that: (i) the path does not intersect \(E_{i(s,t)}\); and (ii) its length is equal to the length of a shortest \(s\)-to- \(t\) path in \((V_i, E_{U_i}, E_{D_i} \setminus E_{i(s,t)})\). The algorithm terminates when no new edges are oriented with confidence.
3 A DIFFUSION-BASED METHOD FOR ORIENTING THE HUMAN PROTEIN-PROTEIN INTERACTION NETWORK

Signaling models in human are used to predict common properties of signaling pathways\textsuperscript{105}, study the progression of genome-related diseases\textsuperscript{106}, aid in the development of drugs\textsuperscript{107} and tailoring treatment combinations\textsuperscript{108}, investigate the effects of a chemical inhibitor in diverse disease settings\textsuperscript{109}, and many more. Inferring the orientation of the human network can greatly improve these predictions. Indeed, Vinayagam et al.\textsuperscript{34} showed that a directed network improves the prediction of previously unknown modulators in the human ERK signaling pathway. Nevertheless, to date, direction information in human is available for only a small percentage of the interactions. For example, KEGG\textsuperscript{21} contains 5,769 directed interactions in human out of the current 311,962 interactions present in BioGRID\textsuperscript{18}.

Existing orientation methods have been mainly applied to yeast\textsuperscript{1,26,90,96,103} (see also Chapter 2 in this thesis). The generalization of these methods to human is hampered by two main obstacles. First, there is a lack of cause-effect information to guide the orientation. Second, the scale of the problem is much bigger as the size of the human PPI network is almost 3-fold bigger than the yeast one (BioGRID, January 2016, non-redundant Interactions)\textsuperscript{18}. The only attempt to orient a human PPI network was made by Vinayagam et al.\textsuperscript{34}. They overcame the lack of a comprehensive collection of cause-effect pairs by focusing on (shortest) paths from membrane receptors to transcription factors (TFs). As the correspondence between receptors and TFs is only partially known, they assumed that all receptors should be connected to all TFs, thus obtaining only an approximation to the true signaling directions.
In this chapter we report on a novel method for network orientation, Diffuse2Direct, which is based on diffusing signals from causal proteins to affected proteins. To generate cause-effect information in human we utilize two independent resources: (i) drug response data, which captures the effect of drugs, represented by their protein targets, on gene expression; and (ii) cancer genomic data, which captures the effect of a patient’s somatic mutations on gene expression. We show that the assigned directions are highly accurate, outperforming the state-of-the-art methods by a wide margin. Moreover, we show that the predicted directions are robustly predicted and are consistent across the data source used. We then integrate all available data sources to construct a consensus oriented human network. We demonstrate its utility in the prediction of drug targets and cancer driver genes.

The results in this chapter are based on the paper *Orientation of the human protein-protein interaction network improves the prediction of drug targets and cancer driver genes* by Silverbush et al.91.

### 3.1 RESULTS

Diffuse2Direct uses causal proteins and paired effects to orient an undirected or a partially directed network (Figure 3.1). The input to Diffuse2Direct consists of a collection of experiments, each of which induces a set of causal proteins and a set of affected proteins, and a physical network of protein-protein and protein-DNA interactions (Figure 3.1A). Each experiment is used to score the network's proteins according to their proximity to the causal proteins of that experiment on the one hand and to the affected set on the other hand (Figure 3.1B). These proximities are
combined to score the likelihood of each of the two possible directions of an edge, and these scores are used as predictive features for the edge's true direction (Figure 3.1C). A classifier is then applied to the computed features to predict the direction of the interactions (Figure 3.1D). Each interaction is assigned with the most likely direction and a confidence estimate for it. For any given confidence threshold, the result is a partially directed network whose directed part can be viewed as representing signaling interactions and its undirected part can be viewed as representing intra-complex or other undirected interactions.

**Figure 3.1. An overview of the Diffuse2Direct orientation algorithm.** A) We start from an input network along with guiding information consisting of a collection of experiments, each of which induces a set of causal proteins and a set of effect proteins. As an example we show here drug response information, inducing a set of drug targets and their associated set of differentially expressed genes. B) For each experiment in the collection, we perform network diffusion once from the causal set and once from the effect set, and integrate the scores. C) The resulting scores are used as predictive features for the edge's direction. D) A classifier assigns a confidence to each directed edge, and using a cutoff we obtain a directed network.

To exemplify the workings of Diffuse2Direct, we apply it to a small scale example – the VEGF pathway from KEGG – stripping it off its orientation information. Diffuse2Direct orients the pathway based on given sets of input and output nodes.
one direction received a higher score than its opposite direction, then this direction is chosen; otherwise, the edge is left undirected. The output directed network is depicted in Figure 3.2. Out of 52 interactions, 48 are oriented as cataloged in KEGG, including the complex FAX-Paxillim whose interaction remains undirected.

Figure 3.2. Application of Diffuse2Direct to VEGF signaling. We oriented the VEGF KEGG pathway from VEGF to the pathway's output nodes (VARP, Proliferation, PGI2Production, Migration, Permeability and Survival). The figure depicts the directions chosen for each of the interactions, where the color of an interaction represents the confidence in its orientation.

As a first large scale application of the method, we used drug response data to orient a human PPI network. Specifically, in this application we assumed that the effect of each drug can be described as emanating from its set of targets and affecting the genes that were observed to be differentially expressed in response to the drug. As test sets for the orientation we used five subsets of directed interactions (see Table 3.1): interactions from large scale databases, including KPIs, PDIs and E3 ubiquitination interactions – all down sampled to avoid degree bias (see details in
Section 3.3.7). In addition, we used two sources of small-scale experiments: directed interactions from the well-studied EGFR pathway\textsuperscript{110}, and a collection of signal-transduction interactions in mammals compiled by Vinayagam et al.\textsuperscript{34}.

To evaluate the predicted orientation we computed precision-recall curves under a 3-fold cross validation scheme (Figure 3.3). Areas under these curves are given in Table 3.1, ranging from 0.74 to 0.92 depending on the test data used. To compare our results to those obtained using the network topology alone, we applied a variant of the diffusion suggested by Erten et al.\textsuperscript{111} which does not depend on prior cause-effect information. The benchmark results in an average of 0.56 (Table 3.1), demonstrating the importance of the causality information in guiding the orientation.

**Table 3.1. AUPR results for the 5 unbiased test sets.**

<table>
<thead>
<tr>
<th>Instances (50% correct orientations + 50% incorrect orientations)</th>
<th>Source</th>
<th>Drug response</th>
<th>Breast</th>
<th>Colon</th>
<th>AML</th>
<th>Ovarian</th>
<th>Benchmark orientation</th>
</tr>
</thead>
<tbody>
<tr>
<td>3,596 KPIs, filtered to avoid degree bias</td>
<td><a href="http://www.phosphosite.org%5Ctextsuperscript%7B112%7D">www.phosphosite.org\textsuperscript{112}</a></td>
<td>0.90</td>
<td>0.89</td>
<td>0.85</td>
<td>0.90</td>
<td>0.89</td>
<td>0.60</td>
</tr>
<tr>
<td>342 PDIs, filtered to avoid degree bias</td>
<td>ChEA database: integrating genome-wide ChIP-X experiments \textsuperscript{113}</td>
<td>0.92</td>
<td>0.90</td>
<td>0.84</td>
<td>0.92</td>
<td>0.88</td>
<td>0.69</td>
</tr>
</tbody>
</table>
We compared Diffuse2Direct to two previous orientation methods. The first method by Vinayagam et al.\textsuperscript{34} tackles the lack of experimental guiding data by using the functional annotation of proteins (Section 1.3). Vinayagam et al. computed features describing the probability of each direction of an interaction to participate in a shortest path from any membrane receptor to any transcription factor (see detailed description of the algorithm in Section 3.3.6). We also compared to SHORTEST\textsuperscript{90} presented in chapter 2, using the large-network variant presented in Section 2.3.5. We adapted SHORTEST to orient the human network by using the drug response data. Notably, the fraction of edges assigned with a direction by SHORTEST depends on the edge coverage by the shortest paths connecting the input pairs. For the drug response data, such paths cover only 44,647 of the edges, thus greatly limiting the method's
coverage. For comparison purpose, when including the SHORTEST orientation we restricted the evaluation to the interactions covered by SHORTEST (Figure 3.4). Both Diffuse2Direct and Vinayagam et al. perform similarly on the restricted set, with Vinayagam et al. outperforming SHORTEST, and Diffuse2Direct outperforming both. On the full network, Diffuse2Direct outperforms Vinayagam et al. on all test sets (Figure 3.3).

**Figure 3.3. Performance evaluation.** Shown are cross validation results comparing the directions inferred by Diffuse2Direct and Vinayagam et al. to subsets of interactions with known directionality: KPIs, PDIs, E3 ubiquitination interactions, interactions from the well-studied EGFR pathway, and signal-transduction interactions in mammals (STKE). The area under the precision-recall curve for each approach appears in parentheses.
Figure 3.4. Comparison of orientation methods, limited to interactions covered by SHORTTEST. Cross validation results comparing the inferred directions to subsets of interactions with known directionality: KPIs, PDIs, and the STKE interactions used for validation by Vinayagam et al. SHORTTEST is limited in coverage and hence was not able to infer directions for smaller subsets such as EGFR and E3.

To evaluate the effect of the size of the cause-effect input data, we used growing numbers of drugs to orient the network. As evident from Figure 3.5, increasing the number of drugs increases both the recall and the precision of the predictions, supporting our use of drug information for the orientation.
Figure 3.5. Performance as a function of the percentage of drugs guiding the orientation. In this setting only, we used a logistic regression classifier with no regularization in order to manually control the percentage of guiding data used.

As an alternative to the drug-based pairs, we used gene pairs derived from genomic mutation data and the resulting expression changes from TCGA. In this setting, mutations, fusions and copy number alterations of a patient are translated to causal proteins, and genes that are differentially expressed in a tumor with respect to a
matched healthy tissue serve as the effects (Section 3.3.4). We oriented the human network using guiding information from four sources: 200 AML, 960 breast cancer, 631 Colon cancer, and 316 Ovarian cancer patient samples. To evaluate the results we used the same cross validation scheme as above (Table 3.1). Remarkably, we observed wide agreement on the orientation derived from each of the guiding sources (Figure 3.6A). Furthermore, the higher the agreement the higher is the orientation accuracy, with 18.2% accuracy for directions supported by only one oriented network, 39.6% for directions supported by two sources, 60.4% for three, 81.8% for four, and 96.3% for directions supported by all five oriented networks (Figure 3.6B, chi-square p<1e^{-40}).

Figure 3.6. Concurrence of orientations according to different sources of cause-effect information. A) Depicted are the agreement of directions assigned to the test edges when independently orienting the network using i) drug targets to differentially expressed genes, or ii) genetic perturbations to differentially expressed genes across different cancer patients (AML, breast, colon and ovarian cancer). B) The higher the agreement the higher is the orientation accuracy, which goes up to 96% when all sources agree.

Reassured by the good performance and high agreement on the test sets, we turned to learn an orientation by using all available training data (i.e., using the five test sets
amounting to 33,756 directed interactions) and integrating all available guiding sources. We used a single guiding source at a time (drug response, AML, breast, colon or ovarian patients) training a classifier with the directed 33,756 interactions so as to assign direction estimates to the remaining 145,631 undirected interactions. Whenever the ratio $\text{Diffuse2Direct}(u, v)$ to $\text{Diffuse2Direct}(v, u)$ exceeded $1 + \epsilon$ for some information source, we included the edge $(u, v)$ in its directed network. Finally, we constructed a consensus oriented network by counting the votes for each direction from the five source-specific directed networks.

We reasoned that the resulting orientation can assist in predicting the functional roles of proteins based on their network location, as the edge directions greatly limit the number of paths in the network. As an example, we applied the orientation to predict drug targets. We oriented the network using only the cancer data sets. Then for each drug, we ranked the genes as potential targets of the drug based on their proximity to the drug induced expression. To this end, we flipped all directions in the network and computed a network diffusion score using the differentially expressed genes induced by the drug as a prior set. We found that the orientation-based scores outperform the unoriented ones (Figure 3.7A), with a mean rank for a true drug target of 6,045 (out of 15,501 network genes) compared to a mean rank of 8,837 obtained by the unoriented network ($p$-value = $1e^{-12}$, Wilcoxon signed-rank test).

Similarly, we hypothesized that the oriented network could assist in predicting cancer driver genes. We oriented the network leaving the AML data set out. Then, we ranked the genes based on their proximity to the AML differentially expressed genes. As before we flipped the network, and computed the network diffusion scores for each
patient separately, diffusing from the patient differentially expressed genes. We aggregated the ranking from all the patients to create one comprehensive list of putative driver genes. Focusing on the top 1% scoring genes (155 genes), we calculated their enrichment against multiple sources of driver (positive controls) and non-driver genes (negative controls). To this end, we used the gene lists introduced by Hofree et al.\textsuperscript{47}, compiled from different sources: Cancer Gene Census version 73 (CGC)\textsuperscript{120}, the Atlas of Genetics and Cytogenetics in Oncology and Hematology (AGO)\textsuperscript{121}, UniprotKB\textsuperscript{122}, DISEASES\textsuperscript{123} and MSigDB\textsuperscript{124}. The oriented network based computation consistently reported higher fractions of driver genes and lower fractions of non-driver genes compared to the unoriented one (Figure 3.7B). In particular, the orientation based computation reports 49\% of the AGO positive list consisting of 1429 genes (hyper geometric p-value = $7.14e^{-38}$); 30\% of the CGC positive list, consisting of CGC copy number variations, single nucleotide variants, somatic mutations and translocation of 531 genes (hyper geometric p-value = $1.13e^{-30}$); 39\% of the text mining list derived from the DISEASES database (711 genes, hyper geometric p-value = $2.88e^{-41}$), and 63\% of the genes found in the comprehensive list unifying CGC, UniprotKB, DISEASES and MSigDB\textsuperscript{47} and consisting of 2,045 genes (hyper geometric p-value = $4.86e^{-48}$). Importantly, known non-driver genes are underrepresented in the predictions, with 1.3\% genes in the curated negative AGO list (NegAGOClean) consisting of 3,272 known non-driver genes (hyper geometric p-value = $8.18e^{-14}$) and 12.9\% genes in the negative list introduced by Davoli et al.\textsuperscript{125} (NegDavoli), consisting of 10,271 known non-driver genes (hyper geometric p-value = $1.55e^{-43}$).
3.2 CONCLUSIONS

We have developed a novel method for orienting a network that is based on diffusion. We applied it to multiple drug response and cancer genomics data sets to infer a comprehensive and highly accurate orientation of the human protein-protein interaction network, significantly outperforming previous work. Key to the power of the oriented network is the great reduction in the number of possible paths in the network, guiding any subsequent analysis to the more plausible ones. To exemplify the power of this reduction, we applied the oriented network to the inference of drug targets and cancer driver genes. In both tasks, it significantly outperformed an application that is based on the unoriented network.
The network we have constructed can be incorporated with minimum to no change into a variety of existing solutions: as a skeleton for inferring process-specific subnetworks\textsuperscript{126–128}, propagating functional and disease-related information\textsuperscript{37,39,40,129}, and learning logical models of signaling pathways with the potential to greatly expand our understanding of their inner workings\textsuperscript{93,130,131}. Thus, we believe that the oriented network advances us an important step toward mechanistic understanding of cellular processes.

Diffuse2Direct may be extended to utilize recent experimental efforts on an even larger scale. A pivotal open problem is the orientation of tissue-specific networks, for which large scale perturbation information from different tissues is necessary. The Achilles project\textsuperscript{132} performed genome-wide perturbations using both RNAi and CRISPR-Cas9 in 216 cancer cell lines from over 20 different tissues. While currently providing only functional assays of survival and proliferation, in the future it may also include expression profiles needed to identify the affected genes in each cell line and perturbation. Such large scale data from diverse tissue will provide enough statistical power to allow a tissue-specific orientation, and therefore an intriguing comparison of the different orientations. Additionally, achieving higher coverage and accuracy may be feasible through the careful utilization of LINCS\textsuperscript{133}, the second generation of CMAP, containing the staggering number of 1.3 million profiles composed of 27,927 perturbagen and 476,251 expression signatures. While this volume of information cannot be readily utilized by Diffuse2Direct, it would be interesting to investigate even more efficient forms of network diffusion, with the potential to employ the increasingly rising volumes of perturbation experiments.
3.3 Methods

3.3.1 Diffusion-based orientation

We devised a novel orientation algorithm, Diffuse2Direct (D2D), which is based on propagating signals from causal to affected proteins and observing the direction of signal flow in the network. Unlike previous approaches, the algorithm does not rely on shortest paths only and accounts for all possible paths from source to target proteins. Diffuse2Direct receives as input a collection of experiments inducing paired sets of causes and effects, and a network of physical interactions between proteins for which an orientation is desired. The network may be undirected or mixed (containing both undirected interactions and directed ones). To ensure that the diffusion process converges, the network must be connected if undirected, and strongly connected if mixed (as is, e.g., the case for the BioGRID network we use here). Diffuse2Direct assigns directions to the edges using the following steps: (i) diffusing once from the causal proteins and once from the paired effect proteins; (ii) combining the two diffusion scores into a single score for each possible edge direction; and (iii) using the combined scores across multiple pairs as predictive features for inferring the direction of each interaction. These steps are described in detail below.

3.3.2 The diffusion process and edge scoring

To assign a score for each edge direction, we compute a propagation (see detailed description of propagation in Section 1.5) $F^C$ from the set of causes (i.e., the set of causal proteins induced by an experiment serves as the prior set in this propagation...
computation) and a propagation $F^E$ from the set of effects. If an edge $(u,v)$ is directed from $u$ to $v$ then we expect $u$ to be closer to the causal proteins than $v$, thus the ratios $\frac{F^C(u)}{F^C(v)}$ and $\frac{F^E(v)}{F^E(u)}$ should be greater than 1. We take the score of the directed edge $(u,v)$ to be the product of these ratios, i.e., the larger the score the more likely it is that the edge is directed from $u$ to $v$ (Figure 3.1B):

$$score(u,v) = \frac{F^C(u) \cdot F^E(v)}{F^C(v) \cdot F^E(u)}$$

### 3.3.3 Inference of directions

The above process can be recorded in a vector in which each entry corresponds to a candidate directed edge in the network (for example $(x,y)$, $(y,x)$, $(y,z)$, $(z,y)$ in Figure 3.1C) and contains its score. The process may be repeated for multiple input experiments inducing paired sets of causes and effects, resulting in a matrix of edge scores, where rows correspond to edges and columns to an experiment. Each row reflects the contribution of its directed edge to connect the corresponding paired sets of causes and effects induced by an experiment. These features are fed to a logistic regression classifier which assigns a confidence score to each directed edge (Figure 3.1D). To direct an edge, we choose the highest scoring direction unless the confidence ratio of both directions is below $1 + \epsilon$, where we set $\epsilon$ to be 0.1. To avoid over fitting and restrict the number of features we use L1 regularization where the balancing parameter is chosen by 3-fold cross-validation within the training set, tested in the range of $10^{-4}$ to $10^4$. 
3.3.4 Drug response data

For each drug, we consider its known targets as causal proteins and genes whose expression changes in response to treatment by the drug as effects. The data was downloaded from CMAP, build2\textsuperscript{35}. In total we extracted 551 drugs with 1,915 known targets (702 unique targets, 3.48 targets per drug on average). We extracted 21,424 differentially expressed genes (3,487 unique genes, 38.88 genes per drug on average).

3.3.5 Cancer genomics data

We utilized data generated by the TCGA Research Network downloaded on April 2016. Samples were taken from 200 AML patients, 631 Colon cancer patients, 960 Breast cancer patients, and 316 Ovarian cancer. Per sample, we labeled a gene as causal if it was either called as mutated or had copy number variation by TCGA. We labeled a gene as an effect if it was called as significantly differentially expressed by Cosmic\textsuperscript{120}.

3.3.6 Implementation of previous orientation methods

We reimplemented the method of Vinayagam et al.\textsuperscript{34}. To this end, we counted for each directed edge the number of shortest paths traversing it to connect a) any membrane receptor to any transcription factor, and b) any member of a known family of membrane receptors to any member of a known family of transcription factors. These counts were used to derive classification features for predicting edge directions as done by Vinayagam et al. In addition, we tried a variant of the method guided by drug response data, yet this led to a low recall.
A second method we compared to is the SHORTEST\textsuperscript{90} method presented in chapter 2, together with its extension in Section 2.3.5.

Last, we used the DADA approach\textsuperscript{111} to compute a benchmark orientation which is based solely on the network topology. Specifically, we computed the Diffuse2Direct diffusion scores by setting $\alpha$ to 1, thus eliminating the effect of the chosen priors. For each edge $(u, v)$ we calculated three such (DADA) scores $F(u)$, $F(v)$ and $F(u)/F(v)$, and used them as features for a logistic regression classifier.

### 3.3.7 Network and validation sets

We used the Homo Sapiens network from BioGRID (release 3.4.126)\textsuperscript{18} of 147,753 PPIs integrated with known collections of directed interactions with proteins in the BioGRID network: (i) 450 signal-transduction interactions in mammalian cells from the Database of Cell Signaling (http://stke.sciencemag.org/cm/ April 23, 2009 version) used for validation by Vinayagam et al.

(ii) 117 interactions of the EGFR signaling pathway from Samaga et al.\textsuperscript{110}.

(iii) 4,293 KPIs from Phosphositeplus (www.phosphosite.org)\textsuperscript{112}, filtered from an original set of 4293 interactions.

(iv) 28,566 PDIs downloaded from ChEA database: integrating genome-wide ChIP-X experiments\textsuperscript{113} and filtered from 28,566 interactions.

(iiv) 330 E3 ubiquitination interactions, downloaded from hUbiquitome\textsuperscript{115}, a database of experimentally verified human ubiquitination enzymes and substrates, and filtered from 330 interactions.

Extending the BioGRID network with these sets resulted in an integrated network with
179,487 interactions (358,974 bi-directed interactions). Confidence score estimates for interactions are taken from the ANAT\textsuperscript{39} software, which uses a logistic regression-based scheme based on the techniques by which an interaction was detected. Directed interactions added to the network are assigned a fixed confidence value of 0.8.

For validation purposes, we used the known collections of directed interactions. We note that although PDIs and KPIs are routinely used as test sets for network orientation, both are strongly biased toward having interactions directed from a high degree protein to a low degree protein. To remove this degree bias, we filtered the test sets by random down-sampling so that similar numbers of interactions from high to low degree and from low to high degree proteins remain. The filtered test sets included: (i) 450 STKE interactions; (ii) 117 interactions of the EGFR signaling pathway; (iii) 1798 KPIs down-sampled from an original set of 4,293 interactions; (iv) 171 PDIs down-sampled from 28,566 interactions; and (v) 104 E3 ubiquitination interactions down-sampled from 330 interactions.
4 ModulOmics: Integrating Multi-Omics Data to Identify Cancer Driver Modules

Rapid advancements in sequencing technologies led to an unprecedented increase in the generation and availability of high-resolution DNA, RNA, and protein cancer data. These large datasets are analyzed with mathematical and computational tools, unveiling mechanistic and predictive insights into cancer progression and treatment\textsuperscript{117,119,135}. We thus far focused on genome scale predictions, however, it is often desirable to constrain the scale of a model to a group of genes, or a pathway, governing the studied phenotype.

To this end we seek cancer driver modules that activate mechanisms driving tumorigenesis and gradually contribute to triggering the hallmarks of cancer, conferring fitness advantages to the tumors\textsuperscript{3,136}. Driver module elucidation can further our understanding of cancer initiation and progression, as well as inform the development of targeted therapies by focusing the scale of a model.

It has been observed that members of cancer pathways often display specific alteration patterns across tumor samples, most notably co-occurrence and mutual exclusivity\textsuperscript{70,71,137}. Even though finding mutually exclusive groups of altered genes is an efficient way to identify cancer modules, most existing methods rely only on DNA-level data, particularly SNVs. These methods centrally focus on DNA alterations to explore cancer progression, failing to fully exploit the complex interactions involving RNA or protein molecules potentially driving tumorigenesis\textsuperscript{138–140}. Thus, data integration strategies have the potential to unravel previously unknown cancer driver modules\textsuperscript{141}. 
An additional important data source for identifying interactions among cancer drivers is PPI networks (Section 1.2). Studies that exploit this data source include HotNet2\textsuperscript{39}, which uses PPI networks of genetic alterations to identify significantly altered subnetworks connecting recurrently mutated genes; EnrichNet\textsuperscript{142}, which identifies functional gene sets based on PPI proximity calculated similarly to HotNet2; and MEMO\textsuperscript{137}, which identifies mutually exclusive gene sets on the basis of PPI-filtered pairwise connections. Nonetheless, these methods do not include additional layers of biological information directly into their model, such as RNA regulation or gene expression.

In this chapter we describe ModulOmics, a method for the \textit{de novo} identification of cancer driver modules based on the integration of PPI networks, mutual exclusivity of DNA alterations (SNVs and CNAs), and RNA-level co-regulation and co-expression, into a single probabilistic score (Figure 4.1). We identify modules that maximize this score by performing a two-step optimization procedure that combines ILP with stochastic search. We apply ModulOmics on three large scale TCGA datasets of breast cancer\textsuperscript{143}, glioblastoma (GBM)\textsuperscript{117} and ovarian cancer \textsuperscript{119}, and show that it accurately identifies known cancer driver genes and pathways. Moreover, ModulOmics outperforms two state-of-the-art methods to detect cancer modules, namely the DNA-centric method TiMEx\textsuperscript{75} and the PPI-based method HotNet2\textsuperscript{71}.

We further use ModulOmics to identify modules that characterize breast cancer subtypes. The highest scoring modules are enriched with cancer drivers, and reliably separate cancerous from normal tissues in an independent patient cohort\textsuperscript{144,145}. Moreover, the modules characterizing aggressive subtypes, such as Her2 and Basal,
are further enriched with Gene Ontology (GO) terms related to cell proliferation. In the triple negative (TN) subtype, we identify functional connections among multiple down-regulated tumor suppressors, including \textit{TP53}, \textit{BRCA1}, \textit{RB1} and \textit{PTEN}. This pattern is also supported by reverse phase protein array (RPPA) data\textsuperscript{117}. In Luminal A, high scoring modules containing \textit{PTEN} suggest two potential functionalities of this protein: a canonical one as part of the PI3K pathway, and a non-canonical one as a regulator of cell proliferation.

ModulOmics is freely available in two forms, as an open-source R code for the identification of cancer driver modules from a cohort of cancer samples (https://github.com/danasilv/ModulOmics), and as a webserver for the evaluation of any set of genes of interest using the TCGA data processed in this study (http://anat.cs.tau.ac.il/ModulOmicsServer/).

\textbf{Figure 4.1. Overview of ModulOmics. A)} Four different data sources, corresponding to four different models M\textsubscript{1,…,4} (see Methods), contribute to the computation of the ModulOmics score: PPI connectivity (protein level), mutual exclusivity (DNA level), transcriptional co-regulation (regulatory connections and RNA level) and co-expression (RNA level). The four colors correspond to four different genes; full squares in the matrix for model M\textsubscript{2} encode the presence of alterations, while empty squares encode their absence. In M\textsubscript{3}, genes 1 and 2 are regulated by a common transcription factor. In M\textsubscript{4}, the different color intensities depict different expression.
intensities. B) Potential mechanism leading to a driver module exhibiting patterns of PPI connectivity, mutual exclusivity, co-regulation and co-expression. C) The ILP optimization identifies modules with highest sum of pairwise ModulOmics scores, computed as the average of the four scores corresponding to models M1,…,M4, further z-scored and normalized to [0,1]. D) The stochastic search optimization uses the modules identified by ILP, depicted in panel C, as seeds, and aims to improve their scores and identify the global optimal solution. The space of initial solutions is clustered and genes are exchanged between clusters in order to identify modules with globally high scores. While the scores for models M1,…,M4 of the modules in panel C were approximated as the average pairwise scores, here they are computed explicitly for the entire module. E) The webserver tool computes the ModulOmics score of any chosen gene set, based on any of the TCGA datasets analyzed in this study. For each data source employed by ModulOmics, the tool plots the single omics scores of the top 50 modules, highlighting the score of the chosen gene set.

The results in this chapter are based on the paper ModulOmics: Integrating Multi-Omics Data to Identify Cancer Driver Modules by Silverbush et al.92

4.1 RESULTS

ModulOmics identifies driver modules on the basis on DNA and RNA cancer patient data, integrated with PPI networks and known regulatory connections. Each candidate module is scored according to the degree of mutual exclusivity among DNA alterations in its members across the patient cohort, the correlation of the RNA expression of its members across the cohort, the probability that its gene members are connected in the PPI network, and the fraction of its members that are co-regulated by a common active transcription factor. As the number of candidate groups grows exponentially with maximal group size, ModulOmics uses a heuristic two-step optimization procedure to first find good initial solutions by linearly approximating the scoring function and then refining these solutions via stochastic search (see Section 4.3).
We applied ModulOmics on three large TCGA cancer datasets: breast cancer, glioblastoma, and ovarian cancer. On these datasets, we compared ModulOmics to four simplified similar approaches, in which the score of a group is computed using only single omics data sources, namely PPI connectivity, mutual exclusivity, co-regulation or co-expression, as well as to two state-of-the-art methods for the identification of driver modules, HotNet2 and TiMEx.

4.1.1 Driver modules are enriched with cancer drivers

To assess the performance of ModulOmics, we calculated the enrichment of the highest scoring drive modules with known driver genes (positive controls) and known non-driver genes (negative controls). To this end, we used the gene lists introduced by Hofree et al.\textsuperscript{47} (see Section 3.1). The enrichment was calculated as the fraction of gene members in each module that were also part of each control list, averaged across the top modules considered. The top 10 modules inferred by ModulOmics generally outperformed the top 10 modules identified with the four single omics approaches and with TiMEx and HotNet2 across the seven positive and three negative control lists tested (Figure 4.2A). Specifically, ModulOmics achieved enrichment score of close to 1 across all three cancer types in the three largest positive control lists: the manually curated resource NCG5, the positive AGO list (PosAGO), and the Union All list (PosUnionAll), consisting of between 1,429 and 2,144 known drivers. Importantly, the modules inferred by ModulOmics scored close to 0 in all three negative control list assessed, namely the complete negative AGO list (NegAgoFull), the curated negative AGO list (NegAGOClean), and the negative list introduced in Davoli et al.\textsuperscript{125} (NegEllidge),
consisting of between 3,272 and 9,457 known non-driver genes.

In addition, ModulOmics also outperformed the other methods when evaluating the highest scoring 5, 10, 15, 20, or 30 modules of any size (Figure 4.2B), or when separately evaluating modules of fixed sizes (Figure 4.2C). Among the competing methods, PPI-based and co-regulation-based scorings exhibited good performances, while co-expression, mutual exclusivity, TiMEx and HotNet2 generally performed poorly on both positive and negative control metrics.
Figure 4.2. The driver modules inferred by ModulOmics are enriched with cancer driver genes. A) The average driver enrichment (red heatmaps) and non-driver enrichment (blue heatmaps) across the top 10 scoring modules inferred by each method in the three cancer types studied. The enrichment was calculated as the fraction of gene members in each module that are also part of each control list, averaged across the top 10 modules. The modules were ranked by their score, regardless of their sizes (the inferred groups consisted of two, three or four gene members). ME stands for mutual exclusivity, CoReg for co-regulation, and CoExp for co-expression single omics scores. NCG5, PosAGO, PosUnionAll, PosSomatic, PosUniProtKB, PosTextMine, PosTrans are the positive control lists, while NegAgoClean, NegAgoFull and NegEllidge are the negative control ones (see Methods). B) Detailed driver and non-driver enrichment scores for the positive driver list PosUnionAll and the negative driver list NegAgoClean for the seven methods assessed, across the three cancer types, for the top scoring 5, 10, 15, 20 and 30 modules. C) Detailed driver and non-driver enrichment scores for the positive driver list PosUnionAll and the negative driver list NegAgoClean for the seven methods assessed, across the three cancer types, for the top scoring 5, 10, 15, 20 and 30 modules.

4.1.2 Driver modules are functionally coherent

An additional metric for evaluating the relevance of the inferred modules concerns their functional coherence, which we accessed via their enrichment with curated pathways from KEGG\textsuperscript{21} (see Section 4.3.12). ModulOmics identified key cancer-related pathways, such as pathways in cancer, across all three cancer types (Figure 4.3A). In contrast, HotNet2 identified this pathway only in the GBM and ovarian cancer datasets, and TiMEx did not identify it at all. Additional highly enriched pathways included apoptosis, cell cycle, TP53 signaling, mTOR signaling, and the angiogenesis-related VEGF pathway. Interestingly, the set of enriched pathways also included pathways characterizing other cancers types, indicating shared mechanisms among malignancies.

To quantify the pathway enrichment performance of ModulOmics, TiMEx, and HotNet2, we counted the number of pathways significantly enriched (Bonferroni-
corrected $p$-value $\leq 0.05$) in each of the top 5, 10 and 15 highest scoring modules (Figure 4.3B), and computed their average enrichment factor (Figure 4.3C). Enrichment $p$-values and factors were computed with Expander$^{146}$ (see Section 4.3.12). Overall, ModulOmics identified modules enriched with more general pathways and cancer-related pathways than the two competing methods, and was the only method for which all highest scoring 10 modules were enriched with at least one pathway. Specifically, ModulOmics identified an average of 7.8 enriched pathways per module in breast cancer, while HotNet2 and TiMEx only identified an average of 1 and 0.1 pathways, respectively. In GBM, ModulOmics slightly outperformed HotNet2, with averages of 8.8 and 4.3 enriched pathways per module, respectively, while the modules reported by TiMEx in these settings were not enriched with any pathways. In ovarian cancer, ModulOmics outperformed HotNet2 and TiMEx by an even wider margin, with an average of 7.4 enriched pathways per module, as compared to 2.7 and 2 pathways, respectively. When evaluating the enrichment factor, ModulOmics outperformed both HotNet2 and TiMEx across all three cancer types, by particularly wide margins in breast and ovarian cancers. A high percentage of the genes identified by ModulOmics participated in known KEGG pathways, reaching an average of 75% across all three cancer types, compared to 39%, as identified by HotNet2, and 13%, as identified by TiMEx (Table 4.1).
Figure 4.3. The driver modules inferred by ModulOmics are enriched with cancer driver pathways. A) Mean hyper-geometric p-value of the KEGG pathways significantly enriched in the top 10 modules identified by ModulOmics (red), HotNet2 (purple) and TiMEx (light blue). B) Average number of KEGG pathways significantly enriched in the top modules, indicated above the bars. The opaque bars indicate cancer-related pathways only. C) Average enrichment factors for top modules. The numbers displayed in panels B and C are normalized per module. Enrichment p-values and factors were computed with Expander\(^{146}\).
Table 4.1. Genes in top modules participating in highly enriched pathways

<table>
<thead>
<tr>
<th>Cancer</th>
<th># top modules</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ModulOmics</td>
<td>HotNet2</td>
<td>TiMEx</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td># genes in enriched pathways</td>
<td># total genes</td>
<td># genes in enriched pathways</td>
<td># total genes</td>
<td># genes in enriched pathways</td>
<td># total genes</td>
<td># genes in enriched pathways</td>
<td># total genes</td>
</tr>
<tr>
<td>Breast</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>0.8</td>
<td>6</td>
<td>20</td>
<td>0.3</td>
<td>2</td>
</tr>
<tr>
<td>Breast</td>
<td>10</td>
<td>4</td>
<td>7</td>
<td>0.571</td>
<td>8</td>
<td>30</td>
<td>0.267</td>
<td>2</td>
</tr>
<tr>
<td>Breast</td>
<td>15</td>
<td>5</td>
<td>10</td>
<td>0.5</td>
<td>8</td>
<td>40</td>
<td>0.2</td>
<td>3</td>
</tr>
<tr>
<td>GBM</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>0.833</td>
<td>9</td>
<td>18</td>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>GBM</td>
<td>10</td>
<td>7</td>
<td>9</td>
<td>0.778</td>
<td>15</td>
<td>26</td>
<td>0.577</td>
<td>0</td>
</tr>
<tr>
<td>GBM</td>
<td>15</td>
<td>8</td>
<td>11</td>
<td>0.727</td>
<td>17</td>
<td>30</td>
<td>0.567</td>
<td>0</td>
</tr>
<tr>
<td>Ovarian</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>9</td>
<td>0.333</td>
<td>2</td>
</tr>
<tr>
<td>Ovarian</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>0.917</td>
<td>3</td>
<td>9</td>
<td>0.333</td>
<td>3</td>
</tr>
<tr>
<td>Ovarian</td>
<td>15</td>
<td>13</td>
<td>15</td>
<td>0.867</td>
<td>3</td>
<td>9</td>
<td>0.333</td>
<td>3</td>
</tr>
</tbody>
</table>
4.1.3 Sensitivity analysis

We evaluated the sensitivity of the inferred modules to different parameter choices by running ModulOmics with changed parameter values and counting repetitions of results in the top 10 modules of size 4. We changed the parameters of the stochastic search as follows: 300 initial module seeds instead of the default 200, 15 clusters instead of the default 10, and 7 top results reported by each cluster instead of the default 5. We evaluated the following metrics: i) the repetition of gene connections, i.e. gene pairs co-residing in the same module and ii) the repetition of the gene pool reported by the top modules, regardless of which module they belonged to (Figure 4.4). Across the three cancer types, the majority of gene pairs remain connected: 24, 19 and 16 gene pairs in breast cancer, GBM, and ovarian cancer, respectively, are common to the top modules inferred with each parameter configuration. Additionally, the majority of genes collectively included in the top modules are robust to parameter changes: 21, 8 and 8 genes respectively.
Figure 4.4. Robustness of modules inferred by ModulOmics. A) Intersection of gene pairs co-residing in the same module, for various parameter choices. B) Intersection of the gene pool collectively reported by the top modules, for various parameter choices.

### 4.1.4 Driver modules in breast cancer subtypes recapitulate known mechanisms and suggest novel functionalities

Next, we applied ModulOmics on molecularly defined subtypes of breast cancer, classified using the mRNA PAM50 classification\textsuperscript{147} into Basal (125 patients), Her2 (61), Luminal A (364) and Luminal B (174). Across all subtypes, the genes in the top 20 modules (Figure 4.5A) were highly enriched with cancer drivers (66% were part of the NCG5 positive control list and 70% were part of the UnionAll positive list, while only 4% were part of the AGOClean Negative control list) and KEGG pathways (44 enriched pathways, 24 of which were directly related to cancer, average p-value
The top drivers identified by ModulOmics included TP53, AKT1, mTOR and PTEN, as well as subtype-signature genes such as BRCA1 and BRCA2 for Basal, CDH1 for Luminal A and B, MAP3K1 for Luminal B and EGFR for Her2. An alternative strategy to ModulOmics for identifying relevant drivers would have been selecting genes with highest SNV or CNA alteration frequencies. However, in that case, a substantial portion of the enriched gene list identified by ModulOmics would have been overlooked, as 34% fall below the SNV median frequency per gene and 40% fall below the CNA median frequency per gene (Figure 4.5A). Therefore, integrative approaches such as ModulOmics are essential to this end.

A detailed PPI network view of the genes identified by ModulOmics showed that TP53 is a key player in tumor progression for all subtypes, while subtype-specific key players included EGFR for Her2 and BRCA1 for Basal (Figure 4.5B). The network view highlighted the higher rate of PPI-connected established tumor suppressors in the Basal subtype, as compared to Luminal A, matching the aggressive nature of these tumors. In addition, Luminal A modules were characterized by a higher occurrence of PI3K pathway members, such as PIK3R1, AKT1, mTOR and PTEN, as previously observed. The top modules identified by ModulOmics were further highly enriched with functional relations, highlighting different GO annotations for each subtype (Figure 4.5C). These results captured the increased pathway activity of key pathways required for tumor progression, such as apoptosis, cell cycle process or cell proliferation, as well as the known aggressiveness of Basal and Her2 tumors, reflected in their higher pathway enrichment.
The highest ranking module for both Luminal A and Luminal B was TP53 and CDH1, two known functionally associated drivers in the Luminal subtypes\textsuperscript{117,152}. The top Her2 modules were characterized by the recurrent appearance of the nuclear export gene XPO1 together with TP53, which is one of its known targets\textsuperscript{153,154}. Interestingly, the mechanism of TP53 nuclear export by XPO1 is well characterized\textsuperscript{155}, yet no pivotal role was suggested specifically in Her2 breast cancer. The highest ranking module of the Basal subtype consisted of RB1, BRCA1, NF1, and CREBBP. CREBBP is a BRCA1 activator\textsuperscript{156} and, since both BRCA1 and CREBBP are involved in DNA repair, this module potentially reflects the altered DNA damage repair mechanism specific to Basal tumors\textsuperscript{157}.

One of the frequently occurring genes in the top Luminal A modules was the tumor suppressor PTEN, occurring both in modules reflecting its canonical PI3K pathway role, and in modules suggesting a non-canonical role (Figure 4.5D). The canonical PTEN module also included PIK3R1 and AKT, thus supporting the known mutual exclusivity pattern of mutations within the PI3K pathway\textsuperscript{158}. The module suggesting the non-canonical role of PTEN also included CDH1 and TP53, supporting the hypothesis that PTEN regulates cell proliferation by increasing the binding of CDH1 to APC\textbackslash C, a complex known for its tumor-suppressive function, and by increasing TP53 acetylation following DNA damage\textsuperscript{159}. Indeed, according to the database of Transcriptional Regulatory Relationships Unraveled by Sentence-based Text Mining (TRRUST)\textsuperscript{160}, PTEN and CDH1 are co-regulated by two common transcription factors, namely STAT3 and NFKB1.
In order to further explore the clinical relevance of the highest scoring driver modules, we examined how well they can distinguish healthy tissues from cancerous ones in an independent omic data source. To this end, we used a recently published proteomics dataset consisting of 62 samples of Luminal A and healthy tissues\textsuperscript{144,145}, and focused on the two highest scoring Luminal A modules: \textit{TP53} and \textit{CDH1}, and \textit{FOXA1} and \textit{TP53}, respectively. These top two modules significantly separated the Luminal A cancerous tissues from the healthy ones (p-value $1.6e^{-06}$ and p-value $6.2e^{-06}$ respectively, Kolmogorov–Smirnov (KS) test, Figure 4.5E). For comparison, neither \textit{GATA3} or \textit{PIK3CA}, the most frequently mutated genes in Luminal A, nor \textit{TP53}, the most frequently mutated gene in breast cancer, were able to significantly separate the two types of tissue (p-value 0.065, p-value 0.054, and p-value 0.69, respectively, KS test). Similarly, a random module of the same size did not significantly separate the tissues (p-value 0.14, averaged over 1,000 random modules, KS test).
Figure 4.5. Modules inferred in mRNA-classified breast cancer subtypes reflect various levels of subtype aggressiveness and separate cancerous from healthy tissues. A) For each mRNA-based subtype and for the pooled set of genes in the top 20 modules, we computed their occurrence frequency in the top 20 modules, as well as their SNV and CNA alteration frequencies across the patient cohort. These genes are enriched with known cancer drivers and pathways, and could not have been identified if relying on SNV and CNA alteration frequencies alone. White corresponds to absent genes (0% frequency). B) Detailed PPI network view of the subset of genes in panel A that are either known drivers, or part of KEGG pathways. The displayed physical protein interactions underline cancer-related functional associations, such as the role of the PIK3A pathway in Luminal A tumors. C) Selected list of significantly enriched GO pathways across the top 20 modules, reflecting the aggressiveness of the Basal and Her2 subtypes, compared to Luminal A and Luminal B. Enrichment hyper-geometric p-values were
computed with Expander [146]. White corresponds to absent pathways. **D)** Module scores for top Luminal A modules suggesting two different biological roles of the tumor suppressor PTEN. **E)** The highest ranking Luminal A module in an independent mass spectrometry proteomics dataset separates cancerous from healthy patient tissues. TP53 loss is measured by its downstream regulated protein CDC2, CDH1 loss is measured by its downstream regulated protein CTNNB1, and FOXA1 gain is measured directly.

An alternative way to study breast cancer progression is by stratifying patients according to immunohistochemistry results assessing the HER2, ER and PR receptors. To this end, we separated our patient cohort into the following subtypes: TN (116 patients), Her2-enriched (30), Luminal A (477) and Luminal B (88), and used ModulOmics to infer modules for each subtype. Similarly to the mRNA-based classification, the genes part of the highest scoring 20 modules were enriched with cancer drivers (67% were part of the NCG5 positive control list and 59% were part of the UnionAll positive list, while only 2% were part of the AGOClean negative control list) and with known cancer pathways (46 enriched pathways, 25 of which were directly related to cancer, average p-value 0.01, Figure 4.6A). Across subtypes, the highest scoring modules highlighted a unique alteration pattern for the tumor suppressor TP53. In Luminal A, Luminal B and Her2-enriched, TP53 was mutually exclusive with other tumor suppressors, such as PTEN and BRCA1 in Luminal B, or BRCA2 in Luminal A, which led to ModulOmics inferring these groups as high scoring modules. However, in TN, TP53 was mutually exclusive with BRCA2, but not with other key TN drivers, such as BRCA1, PTEN or RB1, as both the pairwise and the group mutual exclusivity scores of TP53 and these three drivers were 0 (Figure 4.6B). This suggests a TN-specific concerted down-regulation of multiple tumor suppressors, namely TP53, BRCA1, RB1 and PTEN, potentially contributing to the bad prognosis of this subtype. Taken together, these results imply that the level of
mutual exclusivity in tumor suppressors might reflect the aggressiveness of the tumor subtype\textsuperscript{161–164}.

Finally, we used an independent omic data source (RPPA) to further evaluate the functional connectivity among the tumor suppressors \textit{PTEN}, \textit{BRCA1}, \textit{RB1}, and \textit{TP53}. In general, evaluating protein measurements limits automatic and exhaustive analyses, since loss of function can lead to missing data, requiring the identification of downstream regulated proteins that can serve as surrogates. \textit{PTEN} was found to be downregulated, while phosphorylated \textit{AKT}, which is suppressed by \textit{PTEN}, was upregulated. \textit{BRCA1} loss was accounted for by its downstream regulated protein \textit{CYCLIN B1}, which was highly expressed. \textit{RB1} showed an overall low expression in most samples, while the \textit{RB1}-related \textit{CYCLIN-D1} was lowly expressed mostly in TN tumors (Figure 4.6C). \textit{TP53} loss was accounted for by its target \textit{CDK1}, which was also highly expressed. \textit{CDK1}, \textit{CYCLIN B1} and \textit{AKT}, the tumor promoters regulated by \textit{TP53}, \textit{BRCA1} and \textit{PTEN}, were significantly upregulated in TN tissues compared to the other subtypes (p-value $1.2e^{-16}$, KS test), while the tumor suppressors \textit{PTEN}, \textit{RB1}, and \textit{BRCA2} were significantly downregulated (p-value $2.9e^{-09}$, KS test, Figure 4.6D). These results suggest that these two groups of genes can be used to separate TN from the other subtypes.
Figure 4.6. Modules inferred in receptors-classified breast cancer subtypes highlight differences among subtypes. A) For each receptors-based subtype and for the pooled set of genes in the top 20 modules, we computed their occurrence frequency in the top 20 modules, as well as their SNV and CNA alteration frequencies across the patient cohort. These genes are enriched with known cancer drivers and pathways, and could not have been identified if relying on SNV and CAN alteration frequencies alone. White corresponds to absent genes (0% frequency). B) The TN module PTEN, BRCA1, RB1, and TP53 (ModulOomics score 0.52) shows both Pairwise and group mutual exclusivity scores of 0, unique to this subtype, potentially suggestive that the accumulating activity of multiple tumor suppressors drives TN aggressiveness. C) RPPA data supports the observation that TN patients harbor lower activity of the tumor suppressors PTEN, BRCA1, RB1, and TP53, compared to other subtypes. PTEN, RB1, phosphorylated RB1 (RB1 PS807-S811) and RB1-related CYCLIN D1 show significantly lower activity, and phosphorylated AKT (AKT PS308 and AKT PS473), which is negatively regulated by PTEN, shows higher activity. TP53 and BRCA1 were not captured directly, yet their suppressed targets, CDK1 and CYCLIN B1, show lower activity. D) The tumor suppressors PTEN, BRCA1, RB1, and TP53 can be used to separate TN tissues from the other subtypes in RPPA data. CDK1, CYCLIN B1 and AKT, tumor promoters...
negatively regulated by TP53, BRCA1 and PTEN respectively, show significantly higher activity as a group; PTEN, RB1 and the RB1-related CYCLIN D1, tumor suppressors, show significantly lower activity as a group.

4.2 CONCLUSIONS

ModulOmics is a novel method to de novo identify molecular cancer driver pathways, based on the integration of connectivity within protein-protein interaction networks, mutual exclusivity among SNV or CNA alterations, transcriptional co-regulation, and RNA co-expression, into a single probabilistic score. ModulOmics uses an efficient two-step optimization procedure to first find good initial solutions using linear approximation, and then refine these solutions with stochastic search. We demonstrate the performance of ModulOmics in identifying modules enriched with known cancer driver genes and pathways in three large scale multi-omics TCGA datasets: breast cancer, GBM, and ovarian cancer. We further investigate breast cancer subtypes and find that some of the highest scoring modules are known to be involved in cancer-related molecular mechanisms, while others suggest novel functionalities. We evaluate these results using an independent patient cohort and independent proteomic and phosphoproteomic datasets. In addition, we show that the top modules inferred by ModulOmics can be used to reliably separate cancerous from normal tissues in Luminal A samples, as well as to distinguish TN samples from the other subtypes. ModulOmics is a freely available open-source software. It is designed to aid cancer research in various stages: to identify driver modules de novo from a large cancer dataset, to integrate new omics data sources for improved modules identification, and to evaluate the functional connectivity of any group of genes identified by other means. With the rapid accumulation of high-
throughput cancer data of different types, ModulOmics will produce ever refined predictions that will help elucidate key molecular mechanisms driving cancer progression.

Some of the modules identified by ModulOmics may merit further experimental investigation. For example, on the basis of the highest ranking Basal module (RB1, BRCA1, NF1, and CREBBP), we propose that further validation experiments could evaluate the clinical implications of using the CREBBP inhibitor in BRCA1 patients, similarly to PARP1, another DNA repair agent successfully used in treatment\textsuperscript{165}. Based on the recurrent joint occurrence of XPO1 and TP53 in top Her2 modules, we propose to further evaluate the role of the export mechanism of XPO1 leading to TP53 depletion in the nucleus, thus decreasing its tumor suppression capability. The role of XPO1 in tumor progression was previously investigated in a preclinical context of TN treatment\textsuperscript{154,166}. Here, we suggest it may also play a role in the Her2 subtype.

4.3 METHODS

4.3.1 Model

Given a set \( G = \{G_1, \ldots, G_n\} \) of genes and a collection \( M = \{M_1, \ldots, M_m\} \) of models for different data types, we are interested in computing \( S_G \), the ModulOmics probabilistic score of the set \( G \), reflecting how likely are the genes in \( G \) to be functionally connected. \( S_G \) is computed as the mean of \( m \) probabilistic scores \( P(G \mid M_k) \). Each of these \( m \) scores represents how strongly functionally connected the genes in \( G \) are, under different models:

\[
S_G = \frac{1}{m} \sum_{k=1}^{m} P(G \mid M_k) \quad (1)
\]
The models we consider here are: connectivity among protein-protein interactions ($M_1$), mutual exclusivity among point mutations or copy number alterations ($M_2$), transcriptional co-regulation ($M_3$), and gene co-expression ($M_4$).

### 4.3.2 PPI Connectivity

Model $M_1$ assesses the functional connectivity of the set $G$ at the protein level, by computing the probability of $G$ being connected in the PPI network. Starting with the HIPPIE network and its associated interaction probabilities, we define, for each pair of genes $(G_i, G_j)$, $con(G_i, G_j)$ as the probability of the shortest path connecting $G_i$ and $G_j$, i.e., the product of the probabilities of the path’s edges. The computation of $con(G_i, G_j)$ for all $G_i, G_j \in G$ yields a complete graph on $G$, denoted $\mathcal{G}(G)$. If we denote the edge set corresponding to any graph $H$ by $E(H)$, then the connectivity of the set $G$ is defined as the sum of the probabilities over all connected subgraphs spanning $G$, as follows:

$$
P(G \mid M_1) = \sum_{c \in C(G)} \prod_{G_i, G_j \in E(c)} con(G_i, G_j) \times \prod_{(G_i, G_j) \in E(G)} (1 - con(G_i, G_j)) \tag{2}
$$

where $C(G)$ is the collection of connected subgraphs spanning the genes in $G$.

### 4.3.3 Mutual exclusivity

Model $M_2$ estimates the degree with which DNA alterations support the functional connectivity of the genes in $G$. Following the mutual exclusivity framework defined in the context of waiting times to alteration introduced in TiMEx$^{75}$ and pathTiMEx$^{167}$, $P(G \mid M_2)$ is computed as the degree of mutual exclusivity of the set $G$, as follows:

$$
P(G \mid M_2) = \begin{cases} 
\mu_G & \text{if } p - \text{value} \leq 0.05 \\
0 & \text{otherwise}
\end{cases} \tag{3}
$$
where both $\mu_G$ and p-value are reported by TiMEx. The TiMEx probabilistic graphical model estimates $\mu_G$, which is the mutual exclusivity intensity of the group $G$, via a nested likelihood ratio test between an independence model and an alternative, mutual exclusivity model. The independence model assumes that the genes evolve independently during disease progression, whereas the mutual exclusivity model assumes that only the gene with the shortest waiting time in a functionally connected group of genes will fixate. The parameter $\mu_G$ represents the probability that a group of genes is perfectly mutually exclusive, i.e., that no two genes in $G$ share alterations in the same patient. Therefore, $\mu_G=1$ corresponds to perfect mutual exclusivity, and $\mu_G=0$ corresponds to independence. The p-value in Equation 3 is the probability of observing a given alteration pattern of the set $G$ under the null hypothesis of independence, as described by Constantinescu et al.\textsuperscript{75} and Cristea et al.\textsuperscript{167}.

### 4.3.4 Co-regulation

Model $M_3$ assesses the functional connectivity of the genes in $G$ on the basis of their transcriptional regulation. The co-regulation score $P(G \mid M_3)$ is defined as the fraction of genes in $G$ which are co-regulated by at least one common active transcription factor,

$$P(G \mid M_3) = \frac{|G_{\text{co-reg}}|}{|G|}$$  \hspace{1cm} (4)

where $G_{\text{co-reg}} \subseteq G$ is the maximal set in which all genes are regulated by at least one common active transcription factor. A transcription factor is considered active if it is
differentially expressed (z-score of fold change is either >1 or <-1) in at least 25% of samples.

### 4.3.5 Co-expression

Model $M_4$ evaluates the functional connectivity of the genes in $G$ based on their transcriptional profiles. Let a gene be defined as expressed if its average expression across all samples is in the $k$th $q$-quantile, and let $G_{co-exp} \subseteq G$ be the set of all expressed genes. Then, the co-expression score of $G$ is defined as the mean among all pairwise Spearman correlations of the expression profiles of the genes in $G_{co-exp}$, and 0 corresponding to the remaining pairs, in which at least one of the genes is not expressed,

$$P(G \mid M_4) = \sum_{\{G_i, G_j \in G_{exp}\}} \frac{\text{cor}(E_i, E_j)}{{|G| \choose 2}}$$

(5)

where $E_i$ is the continuous expression level of gene $G_i$ across all samples, and $\text{cor}(E_i, E_j)$ is the Spearman correlation among the expression profiles of $G_i$ and $G_j$.

For this application, we chose $k=3$ and $q=4$, i.e., the 3rd quartile.

### 4.3.6 Optimization procedure

Given a large cancer dataset, identifying groups of functionally connected genes is challenging, as the number of candidate groups increases exponentially with maximal group size. Therefore, we employ a two-step procedure to optimize the global ModulOmics score in equation 1. First, in order to identify a large set of good initial solutions, we formulate the optimization problem as an ILP, and optimize a linear
approximation of the global ModulOmics score. Second, we perform a stochastic search starting from these initial solutions and using the global score.

### 4.3.7 ILP

The first step of our optimization procedure linearly approximates the exact scores of the set $G$ under each of the four models $M_k$, by decomposing them into pairwise scores. For each model $M_k$, the score of each pair of genes $(G_i, G_j)$ is denoted by $w_{G_iG_j}^k$ and equals to $P \left( (G_i, G_j) \mid M_k \right)$, further z-scored and normalized to $[0,1]$. The goal of the optimization routine is to identify candidate subsets $G$ with high total scores $w_G$, computed as:

$$w_G = \sum_{k=1}^m \sum_{G_i \subseteq G, i < j \in G} w_{G_iG_j}^k \quad (6)$$

The ILP retrieves sets $G$ of fixed size $K$ with maximal $w_G$ score. Thus, $G$ is the maximal weight subgraph of size $K$ in a weighted complete graph with nodes $V$, corresponding to a large set of genes, and edges $E_{i,j} = \{w_{V_iV_j} \mid V_i, V_j \in V\}$. The ILP consists of the following set of binary vertex variables $V_{(i)}$ denoting the inclusion of vertex $V_i$ in a set $G$, and edge variables $E_{(i,j)}$, denoting the inclusion of edge $E_{i,j}$ in $G$:

$$V_{(i)} \in \{0,1\} \quad \forall V_i \in V \quad (7)$$

$$E_{(i,j)} \in \{0,1\} \quad \forall V_i, V_j \in V, i < j \quad (8)$$

and the objective function:

$$\text{maximize} \sum_{V_i \subseteq V, i < j} \sum_{V_i \subseteq V} w_{V_iV_j} \cdot E_{(i,j)} \quad (9)$$

under the constraints:
Constraints 10, 11, and 12 ensure that the retrieved set is a clique, and constraints 13 and 14 ensure that the clique is of size $K$. Let us note that identical solutions would be retrieved by discarding either constraint 13 or 14, yet we include both for efficiency considerations. With each candidate set $G$ found, we add constraint 15 to prevent the ILP to choose the entire set $G$ again:

$$
\sum_{i \in G} V(i) \leq K - 1
$$
4.3.8 Stochastic search

We use 200 high-ranking modules identified by the ILP as seeds for a stochastic search that expands the search space and optimizes directly the exact score of the modules, rather than their pairwise approximations. The stochastic search uses the seed modules as starting points and aims to find the modules with global optimal score. The seed modules are clustered into 10 clusters using $k$-means, and a search cycle starts independently from each cluster, in order to increase the chances of finding modules with global optimal scores. Each of these 10 cycles iterates among the modules in its cluster and tries to improve each one by suggesting 20 possible exchanges of a random module member with another random gene. If the score improves, then the exchange is accepted and the module is updated accordingly. Each cycle reports its 5 highest scoring modules. The modules reported by all 10 cycles are finally aggregated and re-ranked. Sensitivity analyses show that the performance of ModulOmics is robust under different parameter choices (Section 4.1.3).

4.3.9 Comparison to HotNet2

To identify driver modules, HotNet2\textsuperscript{71} simulates heat diffusion from a set of prior nodes throughout a PPI network, extracts hot modules, and assigns a p-value to each module based on the probability to detect a module of the same size using random permutations of the PPI network. As recommended by the authors, we used SNVs and CNAs as the prior set, and assigned the initial prior score of each genetic alteration to be its alteration frequency in the data. We applied HotNet2 on the same PPI network we used with ModulOmics. To assign a p-value, we used 100 permuted
networks as background. To calculate a hyper-geometric score for pathway enrichment with Expander\textsuperscript{146}, we used modules of up to size 7, since larger modules are more likely to be unspecific from a mechanistic perspective.

4.3.10 Comparison to TiMEx

We ran TiMEx\textsuperscript{75} with default parameters on the same binary dataset used as input for ModulOmics, consisting of binary SNVs and CNAs alterations. We considered as significant all resulting mutually exclusive groups with Bonferroni-corrected p-value $<0.1$.

4.3.11 Evaluating module enrichment with known driver and non-driver genes

To evaluate the enrichment of modules with known driver genes, we used the following lists introduced in Hofree et al.$^{47}$, as positive controls: i) The Cancer Gene Census (CGC) version 73 (PosSomatic and PosTrans), a set of 569 genes manually curated by The Sanger Institute, which have alterations in somatic and germline SNVs, CNVs and translocations; ii) UniprotKB\textsuperscript{122} (PosUniprotKB), a manually curated database of 412 functional proteins, classified as proto-oncogene, oncogene and tumor suppressor gene; iii) a query of DISEASES\textsuperscript{123} (PosTextMine), a database of disease-gene associations extracted mainly from text-mining, which consists of 711 genes associated with cancer; and iv) The Atlas of Genetics and Cytogenetics in Oncology and Hematology (PosAGO)\textsuperscript{121}, a list of 1,430 cancer genes manually curated by a collaborative effort spanning multiple centers. PosUnionAll is the union of all these positive control lists. In addition, we used the Network of Cancer Genes
(NCG5)\textsuperscript{168}, a manually curated list consisting of 1,571 protein-coding cancer driver genes compiled by The Sanger Institute. The gene members of the two shortest CGC lists, germline SNVs (38 genes) and CNVs (15 genes), were not identified in any high scoring module by any of the tested methods, hence are not shown here.

To evaluate the enrichment of modules with known non-driver genes, we used the following lists introduced by Hofree et al.\textsuperscript{47}, as negative controls: i) a list derived from AGO\textsuperscript{121} consisting of 9,457 genes that have no evidence of association with cancer (\textit{NegAgoFull}); ii) a conservative version of the negative AGO list (\textit{NegAGOClean}), created by filtering genes that are part of any cancer-related pathway from the MSigDB database\textsuperscript{124}, resulting in 3,272 genes, and iii) a list of known non-driver genes introduced by Davoli et al.\textsuperscript{125} (\textit{NegEllidge}).

### 4.3.12 Evaluating module enrichment with known pathways

To evaluate the enrichment of modules with known pathways, we used two statistical tests of module-pathway intersection, as proposed by the Expander software\textsuperscript{146}: i) a hyper-geometric enrichment test to calculate the occurrence probability of the intersection of a module with a pathway at random when drawing from all gene coding proteins, and ii) an enrichment factor designed to ease the bias towards larger modules. The enrichment factor is defined as the ratio between the sizes of the intersection of each module and each pathway and the intersection of that pathway and the set of all background genes (all protein coding genes), normalized by the sizes of the module and background respectively.
5 TOWARDS VIRTUAL MODELS OF CANCER: CELL-SPECIFIC COMPUTATIONAL MODELING IN ACUTE MYELOID LEUKEMIA

The potential of personalized medicine depends on our ability to translate the molecular context of patients’ tumors into interpretable clinical outcomes. Successful steps have been taken to accurately predict tumor progression and response to treatment from molecular disease markers\textsuperscript{169,170}. Using tumor cell line based compound screening, we can provide robust readouts of cellular responses to multiple compounds. This information can be used to systematically train computational models of the molecular signaling pathways contributing to drug sensitivity and resistance in various cancer settings, and to propose novel drug targets and combination approaches. Cell line screens have provided some success in explaining or predicting drug responses by driver gene mutations\textsuperscript{170–172}; however in many cases the true mechanism of resistance remains elusive or more complex. Most predictive methods routinely used today employ correlative statistics or feature-based learning techniques such as machine learning, while network methods remain scarce despite their potential for extracting mechanistic insights and actionable biomarkers.

The molecular heterogeneity within cancer types further complicates the prediction of tumor cell behavior determining a patient's drug response. Multiple somatic mutations, epigenetic events or otherwise deregulated gene/protein expression may contribute to driving the disease. This is true in AML, where patients may harbor somatic mutations in a number of potential oncogenes including FLT3, MLL, TYK2,
FGFR1, PDGFRA, IDH1, DNMT3A, impacting expression of downstream signaling for example through PIM kinases\textsuperscript{173,174}. FLT3 internal tandem duplications (FLT3-ITD) and PIM over-expression are associated with poor prognosis in AML patients, motivating the development of small molecule inhibitors targeting these proteins\textsuperscript{175,176}. Incomplete signaling inhibition or the presence of multiple molecular alterations that reduce a tumors dependency on any one target may result in drug resistance\textsuperscript{177,178}. This may be overcome through rational drug combinations; however optimal approaches are rarely obvious and high throughput combination screening is complex and expensive with limited success shown.

With an aging population, the incidence of AML is increasing, with the number of new cases per year approaching 20,000 in the USA alone. AML therefore presents a large unmet clinical need, with overall 5-year survival rates remaining at around 25%. Most patients will respond to initial cytoreductive therapy but a large proportion will relapse with emergence of drug-resistant clones. Given that bone marrow transplantation as the only curative therapy is not an option for many patients, a better understanding of the regulatory pathways causing leukemic transformation and in particular the emergence of resistance will be essential to improve treatment outcomes in AML.

Computational simulations of cancer cell signaling have the potential to overcome both the limitation of cell line diversity and \textit{in-vitro} screening throughput. Computational modeling approaches can be used to capture and integrate knowledge with molecular and phenotypic data to better understand the genetic and signaling dependencies determining a drug’s mechanism of action. The models should be unique to the tumor cell context, include key proteins and their interactions
whilst accounting for influential gene mutations, and would ideally extend to other molecules involved in cell signaling. Execution of such models should demonstrate the intracellular signaling activity as it is triggered by different mutations and different therapeutic modalities, resulting in different cell phenotypes.

In this chapter, we use QNs to model the protein signaling connecting genetic aberrations in FLT3, TYK2, PDGFRA or FGFR1 to cell proliferation/apoptosis via the PIM and PI3K kinases for four AML cell lines, accounting for their unique genetic and phenotypic diversity. Construction and analysis of the biological QN model was achieved in BMA\textsuperscript{89}. By incorporating cell-specific context switches in the model for four cell lines, we were able to accurately model response and resistance to a pan-PIM kinase inhibitor AZD1208 and the FLT3 inhibitor AC220 and validated experimentally our predictions. The model provides a useful tool for AML research, and the approach offers value to drug discovery and early development.

The results in this chapter are based on the paper *Towards Virtual Models of Cancer: Cell-Specific Computational Modeling of the PIM pathway in Acute Myeloid Leukemia* by Silverbush et al.\textsuperscript{93}.

### 5.1 Results

#### 5.1.1 AML cell lines show differential sensitivity to PIM inhibition

To identify potential genetic alterations associated with sensitivity to the pan-PIM kinase inhibitor AZD1208, we surveyed gene variants by whole exome DNA sequencing (\url{http://cancerres.aacrjournals.org/content/77/4/827.figures-only}: Data
set S4) and prioritized by AML disease relevance\textsuperscript{116}. Although cells sensitive to AZD1208 do harbor AML relevant PDGFRA, FGFR1, FLT3, and MLL genetic variants, only a small number of cell lines harbor the same variant, thereby failing to reach statistical significance in association to drug response (Figure 5.1A). Basal cell line PIM1 mRNA expression tends to be higher in sensitive lines, as previously shown at the protein level\textsuperscript{179}, underlying the importance of compound target expression alongside the interplay with genetic alterations for sensitivity. However cells harboring pathway relevant genetic alterations or over-expressing PIM exhibit varied response to treatment, calling for a deeper examination of the cell signaling relating genotype to phenotype to provide a better understanding of the molecular dependencies underlying PIM inhibitor sensitivity in AML cell lines.

5.1.2 Cell type-specific differences in PIM pathway signaling in response to treatment

Given the wide variability of response to therapeutic agents across AML cell lines, we explored the differences in phospho-protein signaling downstream of PIM for AML cell lines MOLM16, MV411, EOL1 and KG1A. RPPA measurements taken 24 hours post AZD1208 treatment reproduced published findings\textsuperscript{179} of reduced BAD phosphorylation in the MOLM16 cell line, and reduced S6 pS235/236 in EOL1 (Figure 5.1B). To estimate response of cell lines, growth inhibition was quantified according to the number of viable cells after culturing with different concentrations of AZD1208 and FLT3 inhibitor, AC220, in combination (Figure 5.1C). Directional de-phosphorylation signaling trends seen in RPPA for BRAF pS445, EIF4B pS406, MTOR pS2481 and global BAD phosphorylation were confirmed by PhosphoScan.
mass-spectrometry in MOLM16 cells after 3 hour treatment with AZD1208 (http://cancerres.aacrjournals.org/content/77/4/827.figures-only: Data set S9).

Figure 5.1. AML cell lines sensitive to the PIM inhibitor AZD1208 have diverse genotypes. A) GI50 (uM) waterfall plot and molecular oncoprint illustrating the diverse pharmacological response of AML cells after 72 hours AZD1208 treatment as well as PIM expression and AML disease relevant mutations. Boxed cell lines names indicate responding cell lines further investigated. B) Protein expression measured by RPPA in treated and untreated cell lines show heterogeneity in signaling responses through various pathways. C) MV411, with an
active FLT3-ITD, shows varied responses to concentrations of AZD1208 and/or AC220 for 72 hours. The number of viable cells was determined by Alamar Blue measurements where the values represent percent growth inhibition.

### 5.1.3 Building a generalized model of PIM signaling in AML

In order to model the observed genotypic and phenotypic differences between the AML cell lines, we proposed a workflow for developing a cell-specific context network model using the BMA tool from cell line molecular information (Figure 5.2). We generated an initial generalized model from the semi-supervised curation of 68 publications ([http://cancerres.aacrjournals.org/content/77/4/827.figures-only: Data set S5](http://cancerres.aacrjournals.org/content/77/4/827.figures-only) for AML cell line. The initial model contains a canonical set of 64 interactions among 32 interacting proteins connected to 2 cell phenotypes/behaviors of apoptosis and cell proliferation ([http://www.bioc.cam.ac.uk/fisher/aml - GeneralModel.json](http://www.bioc.cam.ac.uk/fisher/aml - GeneralModel.json), Table 5.1). All values at nodes range from 0 to 4 to represent the phosphorylation activity from the transformed RPPA data, with 0 representing low to no activity and 4 representing abnormal over activity. The cellular behavior outcome for each disease state is reflected by the two terminal downstream nodes, which model the outcome for cellular abnormal proliferation and apoptosis rates. The generalized model of AML signaling was able to capture only partial abnormal cell behavior for untreated cells, capturing the abnormal low apoptosis levels for both MOLM16 and MV411, and showing an increase in proliferation, yet not capturing the magnitude of the increase. In addition, known perturbations such as simulating inhibition of PIMs in the model, showed the expected trend line of decreased proliferation, yet did not exhibit the expected effect on apoptosis levels.
Figure 5.2. Schematic workflow of cell-specific model construction in BMA. Motifs and interactions curated from the literature are used to build a Qualitative Network in the BMA tool. The model was calibrated with the results of RPPA experiments for two cell lines with different AML-driving mutations. The model is designed to represent the general AML pathways, and provide a cell-specific context by "turning-on" a specific set of mutations. The mutations impact outgoing interactions, thus activating the pathways in a mutation-specific manner, resulting in mutation-specific phosphorylation activity throughout the pathways leading to specific cellular behavior. The model is iteratively refined by testing and comparing to the cell behavior measured as a response to different perturbations for the two cell lines. The model robustness was tested against perturbations from the literature performed on the explored mutations, and unseen cell lines incorporated automatically into the model. The model is then used for in-silico experimentation in order to test novel drug combinations, infer the source and mechanism for drug resistance, and predict drug response in resistant cell lines and suggest treatment for resistance.
5.1.4 Executable QN model validated by cell type-specific signaling behavior

We incorporated multiple gene mutation switches (http://www.bioc.cam.ac.uk/fisher/aml - CellSpecificAML.json, table 5.1) to construct cell-specific context model, and used an annealing technique (see Section 5.3.1) to iteratively refine the target function of the proteins. The executable QN model (Figure 5.3A) was built on the RPPA and growth inhibition of MOLM16 and MV411 cell lines, harboring TYK2 mutation and FLT3-ITD respectively (Figure 5.1B). For each cell line across each treatment (Figure 5.3B), the mean square error (MSE) observed between the transformed RPPA values and modeled signaling activity ranged from 0.15 to 0.21 and median of 0.17 (0.3 to 0.57 in the generalized model, median of 0.41), with the lowest seen for MV411 cell line treated with AZD1208 + AC220 combination and the highest for the MOLM16 cell line treated with AZD1208. Meanwhile across each protein, the MSE observed between the transformed RPPA values and each protein signaling activity ranged from 0 to 0.36 and median of 0.29 (0.41 to 0.88 in the generalized model, median of 0.58), with the lowest seen for BAD and BCR and the highest seen for 4EBP1.

Equally as important, the cell-specific context model performed well in predicting cellular response as measured both by growth inhibition and markers of reduced proliferation and increased apoptosis (Figure 5.1B-C). The model accurately predicted (Figure 5.3C) the reduction in proliferation as a result of treatment with AZD1208 single agent, AC220 single agent, and drug combinations in MV411 cells. Although under predicting the magnitude of increase in apoptosis for AC220 single
agent, the model accurately predicted the directional responses with increases in apoptosis for AZD1208 single agent, AC220 single agent, and combination treatments in MV411 cells.

In addition to predicting differential phenotypic responses in each cell line, the model highlights key signaling events that may underlie the mechanism for each. We validated the robustness of events suggested for MOLM16 using mass-spectrometry. Most importantly, the mass-spectrometry corroborated the decreased EIF4B pS406 phosphorylation after AZD1208 treatment, contributing to decrease in proliferation, as well as the decrease in BAD pS112 and pS155 after AZD1208 treatment, which increases apoptosis. A key differentiating feature of MOLM16 cell lines is the lack of hyperactivity from the MAPK (Ras Raf MEK ERK) and AKT-mTORC1 pathways post AZD1208 treatment, supported by dephosphorylation at downstream BRAF pS445 and MTOR pS2481 in the mass-spectrometry data.

Testing the adaptability of the model to new cell-specific contexts, we "turned on" the genetic alterations FIP1L1-PDGFRA fusion and FGFR1 fusion, matching EOL1 and KG1A cell lines respectively. The apoptosis range was expanded to span the full dynamic range seen in these cell lines, yet no further refinement of the model was performed. The model reflected the cellular signaling changes observed in RPPA data (Figure 5.3D) where the MSE ranged from 0.18 to 0.28 with the lowest seen for EOL1 cell line treated with single agent AZD1208 and the highest for the KG1A cell line treated in combination with AC220. Across each protein, the MSE observed between each protein signaling activity and the transformed RPPA values ranged from 0 to 0.58, with the lowest seen for AKT and the highest seen for p27. The model
also performed well in predicting cellular response (Figure 5.3E). For the proliferation and apoptosis cell behaviors, the model accurately predicted the cellular responses seen in KG1A for AZD1208, AC220, and combination treatments, as well as the cell behaviors for EOL1 with AZD1208 treatment (AC220 was not tested for EOL1).

The model also replicated variations in sensitivity, such as EOL1 reacting with reduced apoptosis to PIM inhibition when compared to MOLM16.
Figure 5.3. Generation of a predictive cell behavior model for AML training cell lines (MOLM16 and MV411) and unseen cell lines (EOL1 and KG1A)  

A) Cell-Specific AML regulatory network model incorporating knowledge from the literature and calibrated to phosphorylation activity measured by RPPA. Perturbations, driving mutations, and internal genes are depicted in grey, green, and red, respectively. To simulate specific cell (MOLM16, MV411, EOL1 or KG1A) the node for the protein with driver mutations (TYK2, FLT-ITD, PDGFRA, or FGFR1 respectively) is set to 1, while all other proteins with mutations are set to 0.  

B) Protein signaling activity (phosphorylation) levels inferred in-silico using the cell-specific contexts (laptop icon), and the generalized model (papers icon) capturing levels of phosphorylation activity as measured in-vitro (petri-dish).  

C) Cell apoptosis and proliferation as inferred in-silico by the executable model compared to levels as observed in-vitro, with the generalized model capturing partial abnormal cell behavior, and the cell-specific context model recapitulating measured levels.  

D) Unseen cell lines EOL1 and KG1A are incorporated to the executable network model. The robustness of the model is tested via the ability of the model to capture the phosphorylation activity unseen at the time of model construction, and  

E) Cell behavior as a result of different perturbations.  

Table 5.1. Target functions for AML cell specific calibrated AML model. Target functions are associated to the nodes and aim to capture biological relationships. MAX function corresponds to independent activation by upstream proteins, while MIN to dependent activation, such that the effect is governed by the lower expression of the two upstream proteins. + corresponds to an additive effect, and * is used to assign magnitude of effect.
ERK  |  max(MEK, 1/2*PDGFRA)
RSK  |  AVG(ERK)
AKT  |  max(Pi3K, mTORC2)
mTORC2 | AVG(Pi3K)
mTORC1 | 1/2*PRAS40 + TSC2
TSC2  |  1/2*((PIM2-1) + 1/2*Akt)
PRAS40 | 1/4*PIM1 + 5/4*Akt
CHK  |  max(PIM1, PIM2)
H3    |  AVG(CHK)
cMyc  |  max(3/4*FGFR1, max(1, 1/4*(max(PIM1, PIM2) + H3)))
P27   |  max(1, var(cMyc)*(var(cMyc)-2) + 1/2*max(var(PIM1), var(PIM2)))
Proliferation | (EIF4B-2) + 1/2*ERK + 2/3*p27 + 2/3*cMyc
Apoptosis | !MAX(BAD, S6, 1/2*BAD + cMyc + S6 + 2*EIF4E))

*range expanded by 2 for FGFR1

5.1.5 Novel signaling components proposed through model refinement

A by-product of refining the QN model to capture cell type-specific signaling is a graphical and descriptive representation of cell specific signaling dynamics between proteins in the network (Figure 5.3A). By simulating the QN model, we were able to test our assumptions regarding the signaling dependencies between proteins, as described by the target functions (Table 5.1). For instance, despite FLT3-ITD being upstream of PIM1, the effect revealed by the iterative optimization of the model was less than other interacting proteins, also suggested by the RPPA measurements (Figure 5.1A) leading to BAD over-activity in MOLM16 but not MV411. The target function of AKT shows that it is dependent on the activity of the FLT3-ITD and FGFR1 fusion, reflecting the accumulation of evidences for AKT/mTOR pathway role in AML (suggested previously by Keeton et al.179). The target function of S6 reflects the dominant over activation of it via AKT-mTOR pathway, additive to the activity of
MAPK pathway, and leading to anti-apoptotic cell behavior of MV411 and KG1A. At the same time the target function of BAD accumulates with activity of MAPK pathway and of PIM1 direct phosphorylation of all three sites of BAD\textsuperscript{180} leading to the anti-apoptotic behavior observed in MOLM16.

5.1.6 \textit{In silico} virtual experimentation with AML cell models can replicate independently reported data

As a first independent test of the AML cell-specific model, we assessed its ability to replicate \textit{in-silico} a sample of protein and phenotypic cell line responses to drug treatment reported in the literature but not used as part of model construction or refinement. We replicated each in-vitro experiment by turning on a respected set of mutations and adding the new examined inhibitor to the model, then observing the predicted protein expression. All eight protein changes were successfully predicted (Table 5.2). The model successfully predicted cell specific response to compounds including: failure of a MEK inhibitor to induce apoptosis in EOL1\textsuperscript{181}; insensitivity of KG1A to the combination of AKT, PDK1 and FLT3 inhibitors\textsuperscript{182}; and the growth inhibition induced on EOL1 by combining PIM and AKT inhibition\textsuperscript{177}. The decrease in cell proliferation of MV411 in response to mTORC\textsuperscript{183} inhibitor was not recapitulated; however Willems et al.\textsuperscript{183} attribute the decrease in proliferation to eIF4E decreased expression which was accurately replicated by the model.
Table 5.2. BMA network replicates response to treatments reported in external publications. Summary of experimental conclusions are given in \textit{in-vitro}, and model predictions are given in \textit{in-silico}.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Cell Line</th>
<th>Perturbations</th>
<th>In-Vitro</th>
<th>In-Silico</th>
<th>Check.</th>
</tr>
</thead>
<tbody>
<tr>
<td>184</td>
<td>MV411</td>
<td>FLT3-ITD inhibitor</td>
<td>Moderate effect on viable cell number (especially compared to MOLM13 and MOLM14 FLT-ITD positive)</td>
<td>Proliferation reduced to 0</td>
<td>V</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PI3K inhibitor</td>
<td>Neglectable effect on viable cell number (especially compared to MOLM13 and MOLM14 FLT-ITD positive)</td>
<td>Proliferation remains 4</td>
<td>V</td>
</tr>
<tr>
<td></td>
<td>MV411</td>
<td>FLT3-ITD + PI3K inhibitor</td>
<td>induced more apoptosis than each of the agents alone in all samples tested</td>
<td>Apoptosis increased to 4</td>
<td>V</td>
</tr>
<tr>
<td>182</td>
<td>KG1A</td>
<td>AKT, PDK1, and FLT3 inhibitor</td>
<td>showed resistant to the inhibitors</td>
<td>Proliferation remains 4</td>
<td>V</td>
</tr>
<tr>
<td>183</td>
<td>MV411</td>
<td>mTORC inhibitor</td>
<td>Decreased cell proliferation</td>
<td>Proliferation remains 4</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Induced apoptosis</td>
<td>Apoptosis increased to 4</td>
<td>V</td>
</tr>
<tr>
<td>EOL1</td>
<td>MEK inhibitor</td>
<td>AZD8055 fully inhibited multisite eIF4E-binding protein 1 phosphorylation</td>
<td>Phosphorylation activity of eIF4E (3-&gt;1)</td>
<td>V</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>---------------</td>
<td>------------------------------------------------------------------------------</td>
<td>------------------------------------------</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Induced a dephosphorylation of 4EBP1 on T37/46, S65 and T70 residues</td>
<td>Phosphorylation activity of 4EBP1 (3-&gt;1)</td>
<td>V</td>
<td></td>
</tr>
<tr>
<td>181</td>
<td>EOL1</td>
<td>Failed to induce apoptosis</td>
<td>Apoptosis remains 0</td>
<td>V</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Level of p-4E-BP1 were not down-regulated</td>
<td>≈ Phosphorylation activity of 4EBP1 remains 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>177</td>
<td>EOL1</td>
<td>4EBP1 is hardly effected</td>
<td>4EBP1 slightly affected (3 -&gt; 2)</td>
<td>V</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PIM inhibitor</td>
<td></td>
<td>S6 very slightly affected (3 -&gt; 2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Induced apoptosis, yet not as sensitively as MOLM16</td>
<td>Apoptosis induced (0 -&gt; 1), yet profoundly less sensitive than MOLM16 (0 -&gt; 4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>177</td>
<td>EOL1</td>
<td>Effect on S6 in more profound than with PIM inhibition</td>
<td>S6 (3 -&gt; 1)</td>
<td>~V</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AKT inhibitor</td>
<td></td>
<td>Phosphorylation of 4EBP1 decreased (3 -&gt; 1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>177</td>
<td>EOL1</td>
<td>PIM and AKT inhibitors</td>
<td>Significant effect on apoptosis</td>
<td>Significant effect on apoptosis (0 -&gt; 4)</td>
<td>V</td>
</tr>
</tbody>
</table>

### 5.1.7 AML cell-specific model predicts synergistic drug combinations with the PIM inhibitor

To assess the potential to prioritize synergistic combinations through *in silico* hypothesis testing with these models we assessed the PIM inhibitor AZD1208, the AKT inhibitor AZD5363, MEK inhibitor selumetinib (AZD6244, ARRY-142886), FLT3 inhibitor AC220, and PI3K inhibitor AZ2426 across the 4 AML cell lines (Figure 5.4A) also summarized in table S4. For each cell line we simulated inhibition of the drug targets first as single agents and then as combinations with PIM inhibition. We validated each combination in each cell line experimentally across a dose range for each agent (Figure 5.4B).

The MOLM16 cell line was correctly predicted to be hyper-sensitive to the PIM inhibitor resulting in almost complete cell kill, and no additional effect was predicted in combination with other inhibitors.

In contrast, the MV411 context model, which harbors a FLT3-ITD, correctly predicted a strong synergy between AZD1208 and AC220 combination attributed to apoptotic...
effect, evident even at lower dosage of combined treatments. Very weak synergy with mild apoptosis was correctly predicted in MV411 in combination with either MEK or PI3K inhibition.

Meanwhile, EOL1 was correctly predicted to gain apoptotic synergic effect with the PIM and AKT inhibitor combination, as well as the PIM and PI3K inhibitor combination. Surprisingly, and the only synergy of the 16 combinations not predicted by the model, EOL1 also exhibited a synergic effect with the AZD1208 and AC220 combination. AC220 efficacy has previously only been reported in FLT3 driven tumors, however these data suggest efficacy from AC220 in PDGFRA mutated tumors potentially through inhibition of PDGFRA driven AKT/PI3K and MAPK signaling.

Finally, in the KG1A model, which harbors an activating FGFR1 fusion, we did not see a co-occurrence of high apoptosis and high growth inhibition for any of the combination treatments, validated as well by the in-vitro assays. Our model suggests that the persistent insensitivity of KG1A may be derived by the high levels of cMyc, which is not directly targeted by any of the combinations.
Figure 5.4. Validation of predicted synergistic combinations of drugs reveals new effective treatment strategies. A) Cell-Specific AML model is used to test combinations of drugs and predict cell behavior in a cell specific manner. A drug or a combination of drugs is simulated by partially or fully nullifying the target functions of their targets, and can be done automatically and efficiently with a large number of candidates. B) Predicted cell behavior of apoptosis and proliferations is validated via growth inhibition of AML cell lines cultured with the indicated concentration ranges of AZD1208 and/or tested combined inhibitor after 72 hours. Predicted synergic effect, as seen for EOL1 cell line with PIM and AKT and PIM and PI3K inhibitors, is used to prioritize combinations.
5.1.8 Executable QN model identify alternative susceptibilities in AZD1208 resistant cells

Four separate populations of MOLM16 cells were made resistant to PIM inhibition by growth in the presence of increasing doses of the compound over a four month period until resulting cell populations were able to maintain growth at 1uM AZD1208. While the parental MOLM16 cell has a 50nM GI50 in a 3 day MTS proliferation assay, all four resistant populations had GI50s greater than 9uM to AZD1208 over the same 3 day growth period. RPPA measurements were taken for the parental and resistant cell lines.

We predicted candidate genetic causes of resistance by iteratively perturbing all individual and pairs of nodes in the parental MOLM16 model, and choosing those leading to similar signaling activity and phenotype as observed in the resistant populations, quantified by lower MSE. This resulted in four different resistant contexts, one for each resistant cell population (Figure 5.5A). All contexts show over-activation through RAS/PI3K as well as AKT/MTOR signaling, supported by RPPA. Interestingly, the different resistance contexts differ in their strength of altered signaling where resistant cell population R1 has a higher activity for both pathways and resistant cell population R3 has lower activity for the AKT pathway. The predicted and observed pathway signaling suggests increased signaling activity through 4EBP1, EIF4B, S6, and BAD contributing to resistance. In particular it highlights AKT-S6 pathway as a major cause for the decreased apoptosis compared to MOLM16 parental when treated with AZD1208.
Whole exome DNA-seq was performed to identify potential protein altering genetic variants that could be driving the AZD1208 resistance. All variant calls with significant differences from the parental line (http://cancerres.aacrjournals.org/content/77/4/827.figures-only: Data set S6) were further parsed to highlight genes encoding proteins that have BIOGRID interactions (http://cancerres.aacrconnectivity.org/content/77/4/827.figures-only: Data set S7) to the RAS/PI3K and/or the AKT/MTOR signaling pathways (Figure 5.5B).

Using the four resistant MOLM16 context models, we predicted possible treatments to overcome resistance by simulating inhibition at each point through systematic addition of an inhibitor node to the network. In line with signaling changes, introduction of an AKT inhibitor AZD5363 to the resistant populations was predicted to overcome the AZD1208 resistance by blocking the abnormal PRAS40, 4EBP1 and S6 activity (Figure 5.5D). To test this prediction, parental MOLM16 and AZD1208 resistant populations were treated with and without 1uM AZD5363 for 1 hour. The resistant populations responded to AKT inhibition with AZD5363 by decreased pS235/235 S6 ribosomal protein and pT246 PRAS40 (Figure 5.5D), providing strong evidence for inhibition of AKT/MTOR signaling. The decrease in AKT/MTOR signaling was accompanied with an increase in cleaved PARP, indicating increased apoptosis and highlighting the dependency on this signaling pathway during AZD1208 resistance in MOLM16 cells.
Figure 5.5. Origin of resistance to AZD1208 in MOLM16 is computationally inferred and validated via whole exome DNA-seq, revealing signaling mechanism validated via western blots and offers combination to combat resistance which successfully induces apoptosis. A) Network model of MOLM16 resistant cell populations (R1-4). Perturbations (lightning bolts) were automatically predicted at specific nodes to simulate possible resistance mechanisms that would attenuate signaling down a specific pathway (shaded red and blue). B) Whole exome DNA-seq was performed on the 1μM AZD1208 resistant pool population to identify protein altering variants from variant calling as significantly different from the parental MOLM16 cell line. Genes whose proteins are involved with epigenetic machinery are underlined with dashes. C) Inferred signaling activity from the
parental MOLM16 executable model is compared to activity from western blots for parental and resistant cell populations. D) Predictions of signaling activity and cell apoptosis for AZD1208 treated alone and in combination with AKT inhibitor AZD5363 is compared to activity from western blots. Prediction of induced apoptosis is supported by the increase in PARP Cleaved with AZD5363.

5.2 CONCLUSIONS

The success of personalizing treatments for AML patients by tailoring to respective genetic alterations that characterize cancer subtypes has so far been limited. Moreover, drug responses seen in genetically matched patients or representative cell lines show considerable diversity\textsuperscript{116,178}. By integrating both genomic and baseline proteomic data from AML cell lines with known tumor-driving genetic events we generated an AML network model capturing cell-context-specific signaling in the PIM kinase pathway. We developed a workflow methodology for constructing a network model with cell-specific context switches, which focuses on iterative refinement of the target function to reflect literature and experimental evidences. Users may also consider applying automated tools to decipher the target functions, such as CellNOpt-cFL tool developed by Morris et al.\textsuperscript{13}, and follow by manual refinement of the target functions.

The resulting cell-specific model captures cell specific signaling and response to cancer therapeutics, and provide virtual cell line models in which to test hypotheses for tailored therapy \textit{in silico}. The cell-specific model significantly reduced the prediction error for both the baseline training data and on-treatment changes in protein expression compared to the generalized model. This is unsurprising since a generalized AML model insufficiently explains the heterogeneity in the mutational
landscape and protein-signaling dynamics reported across different cell lines, for example a lack of signaling through AKT unique to cells with mutations in TYK2.

The cell-specific model accurately and directly recapitulated published experimental results for reported changes in expression in all 8 cases, and 9 out of 10 responses in cell behavior. These results are particularly remarkable when considering the potential variability in signaling and phenotypic output over time, and the focus of these models on the cells steady state reflected by model stability.

We progressed to experimentally validate predictions made with the cell-specific model. The MV411 context model captured the signaling impact of the FLT3-ITD to correctly predict induction of apoptosis after treatment with PIM and PI3K inhibitors, and no effect with PI3K inhibitor alone. For the cell line KG1A we identified contribution of high cMyc activity to cell proliferation, and correctly predicted insensitivity to inhibition of targets thought to be elevated by the FGFR fusion including AKT, PDK1 and FLT3. The EOL1 context model identified previously unreported combination synergy between PIM and PI3KCA inhibitors, validated through increased tumor growth inhibition. This could lead to patients treated with lower doses of the inhibitors if the same efficacy is achieved by combinations, and thereby, reducing the risk of toxicity.

Model discrepancies highlight potential gaps in the captured network knowledge, and hypotheses that warrant further investigation. For example, our model fails to capture BCR and ERK overexpression following treatment in EOL1 and KG1a cell lines. This cannot be resolved through simple optimization of the current network, suggesting a potential gap in our knowledge of how the MAPK pathway influences these
mechanisms (Figure 5.3D). We found that Siendones et al.\textsuperscript{185} had also previously hypothesized the coexistence of transduction signal event, triggering the MAPK pathway independent of the FLT-ITD event, and coupled with poor response to FLT3 inhibitor. Investigating this discrepancy may shed new light on the resistance mechanism of these patients to FLT inhibitors.

Furthermore, using the MOLM16 context model we were able to systematically explore genetic changes that may render the cell resistant to PIM inhibition. Exome sequencing and subsequent drug combination treatment of MOLM16 cell populations with acquired resistance to AZD1208 confirmed our predicted mechanistic dependency on AKT signaling and AKT inhibition as a second line therapy to overcome resistance.

5.3 METHODS

5.3.1 Framework for constructing executable models

The input to the framework is an experimental investigation of cell signaling responses measured by RPPA, yielding a quantitative measure of site-specific and total protein phosphorylation level. To establish the PPI relevant environment we filtered to proteins which exhibit statistically significant total protein and phosphorylation changes. To do so we normalized protein levels using quadrant median (QMN) (http://cancerres.aacrjournals.org/content/77/4/827.figures-only: Data set S2), and filtered by log2 QMN differences greater than or equal to 0.5 and Wilcoxon Rank Sum Tests p-values less than or equal to 0.1
This resulted in key proteins that serve as the skeleton of the model. We performed a text mining search using these proteins, extracting both full documents and sentence-level information. We extended the text mining filtering of Tudor et al. \(^\text{186}\) by adding an additional layer of filtering terms suited for AML. The aggregated information provides a text representation of physical connections, translatable to the BMA “generalized” model presented in Section 5.1.3. The generalized model contains all the proteins, interactions and cell behaviors of the cell-specific model, excluding the context of the specific cells studied and the tailored target function for each protein.

### 5.3.2 Introducing cell-specific context in QN models

We incorporated switches in the BMA tool used to add abnormal expression to cell-specific genetic alterations, thus simulating a cell-specific context. As a result of a different set of mutations “switched on”, the protein activity exhibited by the model will differ between cell lines.

We iteratively refined the target function for each internal node to reflect the information reported in the literature as well as the levels of phosphorylation activity as measured by the transformed RPPA data for each cell line (see Section 5.3.4 for data transformation). While we were able to fully rely on the literature and databases for the presence or absence of an interaction, information about its nature (such as maximization, addition, multiplication) was scarce. If such information existed we set the target function accordingly, yet for most interactions we inferred the target functions guided by the RPPA measurements.
To this end we utilized a simulated annealing process, with its objective set to minimize the MSE observed between the transformed RPPA values and the modeled ones. In each step of the annealing process, we randomly select an interaction and a modification to its target function, which is accepted or rejected based on its contribution to the optimization function and with a probability which is dependent on a decreasing temperature parameter.

We used 49 proteins measured across separate conditions in two cell-lines to train the model, and validated it using an equal number of proteins and conditions in two hidden cell-lines.

5.3.3 Comparison of boolean models, qualitative models and qualitative AND/OR gates models

Executable models focus on the activation of a protein or a cell behavior, rather than the accurate quantitative measurements produced by experimental studies such as we used in this chapter. Executable qualitative models provide coarse-grained descriptions useful for systems whose mechanistic underpinnings remain incomplete. The range of qualitative modeling approaches provide three major types: qualitative models relax the activity of biological entities to integer values within a specified range, Boolean models relax the activity of biological entities to binary (ON or OFF), AND/OR models relax the relation of entities to simple logic operators (AND, OR, NOT). We explored the properties of the three approaches, by changing the model once to a boolean model and once to an AND/OR qualitative model and reconstructing the model. The construction of the models was easier with the refinement time cut from 2 hours to 55 minutes for the Boolean model and 1.15 for
the AND/OR model. For single agent PIM-inhibitor treatment the Boolean model was able to reasonably predict the proliferation and apoptotic responses in MOLM16 and KG1a, partially predicted proliferation response in EOL1, but poorly predicted responses in MV411 (Figure 5.6A-D). The MV411 cell line was correctly predicted to response well to FLT3 inhibition. The model was not, however, able to predict treatment combination synergies (Figure 5.6E). Since the Boolean model is simpler and easier to construct than a qualitative model it offers a useful tool for investigating single agent treatment in larger networks.

The AND/OR gated model recaptured most of the responses to single treatments, as well as synergistic combinations, revalidating the predictions made by our model (Figure 5.7). The synergistic response of KG1a to the combination of AZD1208 and AC220 was the only response not recaptured. This phenotype is likely derived by the protein S6 additive activity from the MAPK pathway and AKT- mTORC1, which cannot be accurately described using AND/OR gates. AND/OR models may be generated by automated tools\textsuperscript{13}, and can serve well as an initial model scaffold. However more complex relationships such as those in our model between BAD, S6, 4EBP1, TSC2 and EIF4B in AML need to be further refined.
Figure 5.6. Generation of a Boolean model to predict phosphorylation events, responses to single treatment and synergistic combinations of treatments. Activity of proteins was relaxed to binary (ON or OFF) and a new model was generated using the same methodology as described for the qualitative model. 

A) Protein signaling activity (phosphorylation) levels inferred in-silico using the Boolean model (laptop icon) capturing levels of phosphorylation activity as measured in-vitro (petri-dish). 

B) Cell apoptosis and proliferation as inferred in-silico by the Boolean model compared to levels as observed in-vitro. 

C) Unseen cell lines EOL1 and KG1A are incorporated to the executable Boolean model, and 

D) Cell behavior as a result of different perturbations. 

E) Predicted cell behavior of apoptosis vs. the Boolean model predictions.
Figure 5.7. Generation of an AND/OR model to predict phosphorylation events, responses to single treatment and synergistic combinations of treatments. Activity of relations is limited to AND/OR/NOT gates only and a new model was generated using the same methodology as described for the qualitative model. A) Protein signaling activity (phosphorylation) levels inferred \textit{in-silico} using the AND/OR model (laptop icon) capturing levels of phosphorylation activity as measured \textit{in-vitro} (petri-dish). B) Cell apoptosis and proliferation as inferred \textit{in-silico} by the AND/OR model compared to levels as observed \textit{in-vitro}. C) Unseen cell lines EOL1 and KG1A are incorporated to the executable AND/OR model, and D) Cell behavior as a result of different perturbations. E) Predicted cell behavior of apoptosis vs. the AND/OR model predictions.
5.3.4 Protein Array data transformation for qualitative models

The relative linear log2 RPPA values were categorized for use in executable network modelling (Figure 5.8). For each column which represents a single treatment and cell spot replicate across the paneled antibodies, a distribution is generated and the values are categorized on a 5 point scale from 0-4 for that treatment and cell line. The median is taken across replicate spots for a treatment and cell categorized values for a single value. For each row which represents a single protein antibody readout across the treatments, a distribution is generated and the values are categorized on a 5 point scale for that single protein antibody (http://cancerres.aacrjournals.org/content/77/4/827.figures-only: Data set S1).

Figure 5.8. A schematic representation for the method of transforming the log2 linear RPPA values to a 5 point categorical scale suitable for BMA modeling. The RPPA data was categorized by two methods 1) within cell and treatment dynamics of response and 2) within protein antibody measure across cell and treatment. This 5 point scale was also used to categorize the cell line fate (phenotypic response) at each cell line, treatment, and time point condition.
5.3.5 Growth inhibition calculation

For single agent, GI\textsubscript{50} were calculated from the ratio of the 72 hour treatment to 72 hour DMSO control, after subtraction of the Day 0 data from each measurement. The dose-response data were fitted using Xlfit (Microsoft Excel). For combination, percent growth inhibition was determined using the Chalice software with values of 0 to 100% indicating anti-proliferation (fewer number of cells than the vehicle control but greater than or equal to the number of cells at the start of treatment) and values of 101 to 200% indicating cell death (fewer cells than at the start of treatment). Day 0 values were subtracted from the Day 3 treatments. The combination Indexes (CI) and Synergy scores were determined using the software program Chalice (Zalicus) and CI determination was made at the ED50 value. Synergy was determined by the Loewe additivity model.
6 DISCUSSION

In this thesis we presented our novel tools designed to facilitate reconstruction of cellular signaling models in health and disease, and demonstrated the utility of such models in tailoring optimal drug combinations for AML patients. It is our vision to provide a full framework for reconstructing cellular signaling models that may accompany future studies. The studies may use the framework firstly to project multiple large scale omics measurements onto a common context of the PPI network, secondly to identify key modules activated in a cohort of patients under study, and finally to reconstruct an executable model of the pathway spanned by the members of the module and linked by the oriented PPI network. The model can then be used to test options to manipulate its pathways to trigger the desired response.

We presented two methods for orienting a PPI network. We devised SHORTEST, an ILP-based method that orients the network so as to maximize the number of shortest and almost-shortest paths connecting pairs of knocked out genes and their associated differentially expressed genes. The ILP formulation allowed us to account for all shortest paths without explicitly calculating them, and in practice converged to an optimum solution in seconds. Unique to our approach was the possibility to extend the solution by inferring parallel shortest paths and increasingly longer paths, reflecting the natural pathway redundancy in the cell. We applied SHORTEST to orient the yeast PPI network, as yeast has abundant causality data in the form of knockout experiments. SHORTEST outperformed state-of-the-art methods, increasing the orientation coverage by 8 fold while preserving a high accuracy rate. We demonstrated the utility of the resulting oriented network by its ability to filter
candidate pathways connecting new pairs, showing that the oriented network reduces the search space by about 40%.

In contrast to yeast, orienting the human network has remained a long standing challenge. A previous approach oriented the network from all receptors to all transcription factors, thus using only an approximation to the true cause-effect pairs. We pioneered the use of experimental causality data to orient the human network, exploiting drug response information to guide the orientation from drug targets to their induced expression change, and using cancer genomic information to guide the orientation from genetic alterations to their associated differentially expressed genes. Extending SHORTEST to orient the human network using these guiding data did not yield satisfactory results, and we concluded that this is due to the ratio of guiding data size to the network size. At the same time, the SHORTEST application hinted at the need to extend beyond strictly shortest paths. Thus, we designed a novel method called Diffuse2Direct, which uses network diffusion techniques to account for all possible paths from source to target proteins. Diffuse2Direct is able to orient the human network with high accuracy, and provides a robust orientation even when switching to different guiding sources.

The orientation methods can be further improved by including additional guiding layers. With the increase of quality proteomics and particularly phosphoproteomic data in human it will soon be possible to incorporate it, into the orientation process. The two orientation methods that we presented here are easily extensible to incorporate such data, by adding cause-effect pairs going from stimulus to phosphorylated proteins. Further improvement of the orientation algorithm efficiency
will also enable the use of the recently published CMAP newest addition\textsuperscript{187}, featuring an incredible 1,000-fold scale-up of the last CMAP addition. A further possible extension of the orientation methods is by requiring an orientation to admit directed paths going first from stimulus to phosphorylated proteins and only then to differentially expressed genes, mimicking the natural signal flow in the cell. Although large scale dynamic time-course data is not yet available, they can greatly improve the orientation accuracy. Network diffusion techniques are particularly fitting for this data, as early time points are expected to be closer to the causal genes, and respectively receive higher diffusion scores. Such time-course data will allow to reliably account for feedback loops, again by enforcing order. For example, suppose that protein A activates protein B, secondly protein B activates protein C, and finally protein C deactivates protein A. Using static data we will observe that when A is overexpressed, both B and C are affected. Hence we will orient the network assigning A->B and A->C, leaving B-C unoriented. Using dynamic data we will be able to observe that when A is overexpressed, B is affected - revealing the direction A->B, only then C - revealing the direction B->C. Following C activation we may observe that A is deactivated, revealing the direction C->A. Further oscillation will increase our confidence in such an orientation.

Focusing on a key pathway in the network may enable a more detailed model, and a closer examination of its underlying biology mechanism. With that intent we designed ModulOmics, a novel method for the identification of molecular pathways that drive cancer progression, integrating protein interactions, mutual exclusivity of DNA alterations, RNA expression, and gene regulation data. ModulOmics is designed to
be flexible and user-friendly, such that each set of data can be excluded or replaced with a new one without changing the ModulOmics framework. In addition, since ModulOmics integrates different sources of information independently, single omics datasets from other cancer studies can be readily integrated. The framework can also be extended to include calibrating the weight of each data source according to the reliability and estimated error rates of the relevant measurement technologies.

Importantly, we presented a proof of concept for detecting cancer driver modules from large multi-omics datasets, by using ModulOmics to investigate the functional connectivity of breast, GBM and ovarian cancer multi-omics TCGA datasets. Additionally, the ModulOmics webserver can be used to evaluate the ModulOmics score of any user-defined gene set, on the basis of any of the TCGA datasets used in this study. The webserver can be very useful in situations where gene sets are deduced from independent biological or computational analyses. In the future, once high-throughput single-cell profiling of tumors becomes routinely performed in the clinic, ModulOmics can be used to identify functional connections derived from multiple tumor cells of single cancer patients, rather than a patient cohort. In this way, the design of personalized cancer therapies based on the tumor heterogeneity of each patient can be facilitated.

To reconstruct the functional network connecting the members of a driver module and extend it to relevant neighboring proteins we developed ANAT, a tool for the inference of protein networks that underlie a process of interest\textsuperscript{188}. As input ANAT can receive a collection of genes governing a process, such as the members of a module, or causality sets, such as the members of a driver module as causes and the associated
differentially expressed genes as effects. ANAT will carve out from the PPI network a subnetwork that underlies the molecular process triggered by the causal genes. We recently published an extension of ANAT\textsuperscript{188} (not included in this thesis) including an exact inference of the underlying network and the option to refine the inferred network by predicting the transcription factors that govern the process under study.

Finally, we demonstrated the utility of cellular signaling models and suggested a scheme for the inference of the network logic. We designed a cell-specific computational model to optimally tailor treatments for AML patients, which was successfully validated in collaboration with the Astra-Zeneca pharmaceutical company. Following our success in first-line treatment, we extended our program to infer new mechanisms of resistance for second-line treatment. We used it to suggest a therapy to overcome the resistance, which was also successfully validated in-vitro. By accurately predicting drug responses and combination synergies, and providing the mechanistic insight on the proteins driving the response, we highlighted the ability of simulated models and virtual experimentation to prioritize effective therapies accompanied with associated predictive and dynamic biomarkers. Successful drug combinations could significantly augment therapy options for AML patients by overcoming innate and acquired resistance to drugs. Simulated qualitative models potentially offer a virtual platform to screen, discover and prioritize drug combinations \textit{in silico}, focusing experimental approaches to validation. Comprehensive genetic diagnosis using targeted exome sequencing is already entering the clinic in major teaching hospitals. When coupled with emerging mass cytometry analysis, all the biological information to build patient specific qualitative networks models may soon
be available from frontline diagnostics data, paving the way for improved personalized treatments and enhanced precision medicine choices.

In a different study\textsuperscript{41} (not included in this thesis) we present a possible large-scale application to simulate drug treatment in a personalized setting. In this study we employed network diffusion to simulate the impact of patient mutations on their associated differentially expressed genes. Our goal was to predict efficient drug targets, and to do so we simulated a drug inhibition by deleting the corresponding drug target from the network and recalculating the impact of the mutations on the differentially expressed genes. We assumed that potential drug targets will maximize such impact. Applying the method to AML patient data from TCGA, we found that our predictions are enriched with known driver genes from COSMIC\textsuperscript{120} (p-value $< 10^{-10}$) as well as known drug targets from DrugBank\textsuperscript{189} (p-value $< 10^{-5}$). Notably, the method also recaptured the efficacy of the three treatments that were found in the small scale application presented in Chapter 5 for the MV411 cell line.

The tools we designed form an initial framework for the construction of cellular signaling models and the study of their perturbation. A typical workflow of the tools presented here, as we envision it, may include using ModulOmics to identify driver modules triggering a process of interest, then using ANAT to infer the functional network connecting the members of the module. The functional network will provide a broader view of the pathway activated in the studied process, forming a cellular signaling model. The oriented networks can be integrated into a variety of existing approaches, specifically it can be integrated into ANAT, providing even more accurate signaling models. For medium size networks, the logical functions of the proteins in
the network may be deduced using the simulated annealing process we have presented. The resulting signaling model could be used for simulating perturbations and inferring novel drug targets, optimal drug combinations, and personalized drug treatments.

As more large-scale data accumulates, further tuning will be possible. For example, several studies showed that using tissue-specific PPI networks improved the prediction of disease causing genes\textsuperscript{190–192}. In cancer, even the host tissue expression is slightly changed from a healthy one, due to enhanced secretion of cancer related ligands\textsuperscript{193}. Similarly to the way that large-scale gene-expression data was used to infer healthy tissue-specific networks\textsuperscript{189}, cancer gene-expression can be used to infer host-tissue-specific networks. Finally, it is becoming increasingly clear that an important reason for treatment failure in cancer is due to intra-tumor heterogeneity, in which tumors may harbor several populations of cells that react differently to treatment. Completing the framework by designing a solution to model population-based networks and their inter-populations connections is desirable. Such a tool will receive as input large-scale omics for different populations, for example in the form of single-cell (sc) genetic alterations and differentially expressed genes. Similarly to our assumptions regarding orientation we may assume that there exists a path in each cell from the sc-genetic alterations to the sc-differentially expressed genes. The tool will simultaneously infer oriented networks and assign each cell to a network. The goal will be to: i) minimize the number of different networks corresponding different populations; and ii) maximize the number of satisfied sc-genomic alterations and sc-differentially expressed genes pairs. Such a solution can model population-specific
networks, and provide a basis for executable network models accounting for the intra-tumor heterogeneity.
7 REFERENCES


symbolic approach to analyze biological signaling networks. *BMC Syst. Biol.* 1,

89. Benque, D. *et al.* BMA: Visual Tool for Modeling and Analyzing Biological
Networks. in *Computer Aided Verification* 686–692 (Springer, Berlin,
Heidelberg, 2012).

90. Silverbush, D. & Sharan, R. Network orientation via shortest paths.

interaction network improves the prediction of drug targets and cancer driver
genes. *Submitted*

92. Silverbush, D. *et al.* ModulOmics: Integrating Multi-Omics Data to Identify
Cancer Driver Modules. in *The Annual International Conference on Research
in Computational Molecular Biology (RECOMB 2018), to appear*

93. Silverbush, D. *et al.* Cell-specific computational modeling of the PIM

94. Fields, S. & Song, O. A novel genetic system to detect protein-protein

95. Gavin, A.-C. *et al.* Functional organization of the yeast proteome by

96. Ourfali, O., Shlomi, T., Ideker, T., Ruppin, E. & Sharan, R. SPINE: a
framework for signaling-regulatory pathway inference from cause-effect


182. Zeng, Z. et al. Simultaneous inhibition of PDK1/AKT and Fms-like tyrosine kinase 3 signaling by a small-molecule KP372-1 induces mitochondrial


הבדלים נטייה々ו-נטייה々ו אנשי מישחקים או בורמת הבנויים של החלבונים וה止损ים הפונקزينואליות שלדם. הבדלים אלה ידועים ולהביזיל שלשיות מורכבים ושפכים תקדומים חוה, חות השלבונים עובד בתיאום ומעכירים אתıyoruz לע ידי הפסקת החלה, בקצרה,apkאריאים.

אותראקציה החלבון-חלבון. ההזג תזוניות תלו, בכבלס תורמות, מועבר אמצעים
אותראקציה החלבון-חלבון מمنتجת החוה (ممברנה) של החלבון בתוך החוה (ביוסטפליימה), ומייסים
לעמוס רוח לש העצאות הפיזיולוגיות נפרדו.בס, הבדלים אופייניים בהזג תזוניות-גנרט ו)<=
לחבים בתבלונים מעשיים ביניהם בוחנים ועל ידי החיתוך, אחור רשים במדידת קרקע על ידי החיתוך.

אחר רשם אתראקציה החלבון-חלבון האופייניים של המוטל.

בניו מוטל של מוסלמיה אוחזים בתאו החוניים במחושב לש חקר משל噀ה, או חות השמלה מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת מחוזקת המחוזק התصناعת השלב של החלבונים

וית את ביצים המושנות של החלבונים והגוז, או רוחבי בין החלבונים, סכום התאונות

הדלילות אסなくて בתרבויות חומשיות של התרבות.

בותים ובנות סוסמיסי על שניים בתקופה בהברה בחלקית לאפח את תגיית השלב ויושבים
המтяרים מסוכל אוחזים בתאו. גו נפרדים את החיתוך עליה גזם מוסלמיה, לע די חוחה
הכינוים של הקורמים פיטיסיים ביו השלום במנור, היינו אתראקציות החלבון-חלבון. באומנויות מדע
מניסיון המרדים שונים סיביות והتصمמות של סכינקה חתנית בלעימונים ובניי שיטות עלולות
ByUsername החלבון והanshipונות בבדיוライフ את הכינוים ברחת של גזם מוסלמיה לצדי. השית
הצלחת ולהורר ולקחיים מוניצים יישומי המוסלמים בכרות במדידות החלבונים בכרות בשיטה
באים השמשות במדידות מוניצים יישומי בתרבויות הבינלאומיים בכרות במדידות החלבונים בכרות בשיטה

Perfil סיביה-נוחה. בס, כולנו את רשת האמצעים של מודל החיתוך את מסכם בכרות במדידות החלבונים בכרות בשיטה.
הרושת המכותות אנק סיפקה תשתיות מכלולות יوتر שיכולה לשכין בכבוד של מדלי גיוסים והמחוקים.
ואף יוכל לסיע ב鼙ונות נסף של התורות המבוקשות על מדלי היותם בחרת, כמו חורי תכלים
המחוקים מעריקות לתורמת חורי מוטיצים וממתכת מצ potrà. הציע הבה היה להटמד בשכימיים
מדלי יוחט בכרות. בגני שיח להאר משכלב ממסר מחזורות בצור י訓וד ובהנה בכידי לאת
מדליות, كبוצע של גני בלע פונקציונליים משותפים או ביד מחזור גורו חשב לחפחתות
הטרון. בכרות, הצעות דרכ פעולות שנעודה לגביע מדלי גיוסים והדגמת את הת настояות הרהב של מדלי
מסווג הזה לשים התיאום מיטבה של טיפי חולייל לקומית מיילבלטיטי חרמה
ל솥ים, בוחת והני שית ל母校 התuttle מדלי גיוסים ויומיים נתון בnonnullי היה הבה התפורטות התולות
במסולות ב倌 התוקדונות מחולות הסטרו, ואת בכידי התהלות אסרטרניים טיפי עילוטי יותר.
תקציר

הבלדונות שבבירושלים פסימיות, יзоיריים מסלול אלוהים והשלטים על תחנות התחנה. הפורטלים מעיפונים על ידי שיניים בבוכבין, בני אשר חיבורו כנון כת החור, בני און שפתי פומפר
וכן מתכתי אẻ השלום את רון החניך של הלבון בה. השינוי מתחלף אתון העובד ב' הלבונים
ביציופלותה ונמיגה לירע החנה, והאתת גורל שלני בהפעתה התשא, על ידי שוני של תרוכוביה

פנימיית בתה וא על ידי שוני במרת הביתות שלàngוים.

מודלים שונים שלא מסלול האות אפשהו לוחות את השינויים מסלולות בראית ביתאות
תולות, בצבר רגיל ולאחר מתת חורמה, והיאפרים בטון של תרפה כשתכלה שלבחון מתפח
_bt產業 מיסול האותות והقبول את החירה מסלולות הגרמה.ekt יירה תשתיות המיתנה לתייכ
מסלול האותות היה החוי. בעד שניים לונDisposed בתופקה בגובה מסלולות למדוד את הגחון של
כוקר פיסי, והא יומנויות לד今まで את הכינון של הקושר. בחקק קוס מקשיזיורים הכורותיה ייו,ככמ הבר יישריה של בלדונים קונ הל랩יני המשררה שלחת, כל ברו י кнопкיפי הפתורהים בבדיר
הלסיק את הכורותיה של מרב יתרושיה בתה.

cדי חלון את רשת הלוחון-הלבון בין סמיכה על ניסיונים המספסים מגדיל של סירה והתחנה ברחת.
אוח הינטושיים המורכבים המושמשים למלתיים והבראריות הדגตน כי ניסי השתקת גנים. בנסייה כל
ן אוח מושחו (ה שתנה) וחותפה בכל גין אוריים משני את רמת הביצי שלדח (גיני ההתחנה).
ון אוחים מסתמכים על הגחוןشدد בשניים התתחנה ישן את רמת הביצי שלדח מקרך מסלול
מכוך של כשירותי מיסים בו גן הסביבה צלום. כל גין להומשטים יבשימי סיסבה-חותפת מסוב זה
cדי חלון את הכורותיה מסלולות ממקל הסביבהEndPoint התתחנה.

בפרקخش יבשות אשר מתאיר את הפרזון שבגון להקשיק כווית ברחת. ביעיט התשק הכורותיה ברחת
 العبשות לכלון את הרשת, לכלר מעבר כו נבר כל חאח מקושיותו הפיסים שבת, כה שומסר
מקסימלי של גון סיבר-חותפת יושר מסלולות ממקוכם. שיטות קדומות בצל מתשק הכורותיה לכל
לעיע מקודם לחיישו שמידותיה חלים מעליות בברחת המקורות והלא ממקות וממציה לכל
קודק, כה שומיץ ער. איסנטים קולנועים שכר לשキーוריה ברחת, היזין כל נבנית אילוגית ו
אף אפוא lngול של כלدور מהרשת. התשק הכורותיה מקושיות שיצחקם שמצימתיו הפקת
לحيا אתא אריאול ושמשע בוליונית שיקול שוליות לולית רשת מהרותה. בנסף כל, הפרשות
הלוחה א┢וארת וכל שמשע בוליונית שיקול שוליות לולית רשת מהרותה. בנסף כל, הפרשות
אלל מסלולות שבין כל סיבת לה תצא עידל הלוחה ארוך באמפ בלבבי, בעד שורשיה....
ביוולוקית מוסללים יאוזו היה והסכמה מגוונות שהרי קצרים כל האפשרויות. על ביסס האבחנה בדרכ
דיאוה המוסללים הקצרים ביוור גית ולא הדירigos של בניית הסקט המהויים לפיכך
לתשא את המقوا יברך אם מ subparagraphי של גוון ספה-זנצה-זנצה קיורר דר מוסללים זכר
ביוור (ברשת המוקדחת). לעם ויהי שמלות.

ואו מיצוגי אננואים חס לעסרונ עבעה, זקנה, המתחטטת על ברכות בשמלום עולעילה,
ולככל מתוך החברים, SHORTEST המתחטטת על ברכות בשמלום עולעילה.
לחומר את החברים בכיר for längerמה. בחר את יוצאת הפתחים בשמש יאוצים yet לדס – 
שינה היה אורכומגคง דמג נועה על עליל נטסי סיביה-זנצה-זנצה ריבים שלשם חלום שבעים בכרום לכל
את החבר. הערוך את סיב ירט החבר השבר לפיכך על חיתות חבר וחבר המהויים
שהלשה לועז ההשתר הלך הערכה: ניב בין קינו שלחה ובין הלאם המנהר
ולכן. השווונאי את יוצאת הפתחים של פלארטמקים יחכים ההרגה את העזרי יד יגוגו. הגדיל בה דע כה
כינו החבר עלית חלה השמות וסансקטיונים שלוי, ומצענו כי החבר מקדישים ליזן מוסללים
אופיריס בין גוון ספה-זנצה חדישים ידע כלפי מצות מברך יהו כה בזכר
 الساد מדעי קימונים מוסר שיתוף המכמה את הרט השמיר, ניסיון יינו רשת הלאם-חלבנה של
בוע שחרים קימונים מוסר שיתוף המכמה את הרט השמיר, ניסיון יינו רשת הלאם-חלבנה של
א造血 לא נגוע חבלשה הממחק. כינו רשת האני היה עריך לと思っていた שיפוע חוויים המובססים עלי
הרט השמיעים לאיכי עד ית מחלחל המשמשים מוסללים יאוזו חוכמן וחבר המקדישים שיפוע
על מוסללים אלא. כינו רשת האני היה אך כמאתורי העיך, אך כינו רשת ומייצר קימיים המסוג
הדות. ראשים, בנדづ או נטסי סיביה-זנצה-זנצה-זנצה בנדづ 발표.شعار, גרל רשת האני, יפו שלוש מרשוע
המשמר, יוצ עקרב שב יוצר ב₪מסטרים רעי פלא שלי טסיי סיביה-זנצה-זנצה – בדוכ יזביה יומי
ניסיון לוגל רשת – או עיזוב של שיתוף החוויים כשמח התנגב על ייז מנוק לים שלי
לגלל רשת. המלטכי מיקסים עכ חוויה רק שיחה את חבלן רשת אדום ואינו מתבנסת על החור
של חלבון שפגוסו על חלבון הנקשורים לדר מוסללים זכר ביוור. התוים שלıyla מTouchEvent
המסזור בין חלבון שפגוסו ספגוסי חלבון הנקשורים לדר שוחלו פרק💡, והרי שוחליו וחבר
בי כל רפסטיים כל חלבון דוא, ולכן היא מתבנסת על שירוט של גוון ספה-זנצה-זנצה.
בפק דנל בונה עם פציות את Diffuse2Direct, יישון הראות הממנה רשת חלבון-חלבנה
באמדת בתבונת על מידו נטסיי ספה-זנצה. לושר, כינו זה אדום, ואנו מפיקים ומסי של ססי

174
הביס-האצות ינשמ תורוקמ עדימםילודג. תישאר ונא םישמתשמ יוסינב הבוגת תופורתל ושענש תולשושב תיאתךותמ טקיורפ

Connectivity Map, המבורה וא הפורהו וא הלוקלומ הנטקר תירשומםע תולשושה תוייאתה האצותכו םינגםימיוסמ םינשמ תא תמר יוטיבה םהלש.

םדוקמכ, ןתינ חינהל ידכש םינגהש וнести תא תמר יוטיבה םהלשםייקש תשרב לולסמ קنوוכמה הנורה לש הפורה תרשוקםינגל ולא.

םא ךכ ינובלחה הרטמה לש הפורה םישמשמ התוא טסכ הביס, םינגהו םינשמש תא תמר יוטיבה םהלש.

תינש, ונא םישמתשמ עדימב ףסאנש לע ילוחןטרסךותמדסמ המותנה לש טס ההובג.

 TASR, תוברת, גנטיזוטילו געטיוור גירעתי הלועפי אלפקט הדוהיה הלשתקת, ובשקבותיתשתהמ מטסלה תיאתות הנג מסומיוו מייסנים את רמ היוביי תשלוח. ב막רה ושהמשישו

המוטיעות חסב סיבה הווהים שסיים את רמ הבוכי תשלוח חסויוּתָה.

ועזוב נייטשה חדות ארשר מתמטכ בר şekilde פעמון אחזור ברש בכה למלד את האיתות המחברת וסהק של טס סיבה ע לטוּתָהוּת. טסניקית הפוע甯ה גאוי מוגבלת טססלים קרפרים יביר, יא מאמפרותת הלוהיטי לכל המטסליים האפורים יברה. תמרשת Diffuse2Direct, מוטעתי ו𝛽אף גנטיסטי תנייה ליתוי מובס תרש. זא מראים ייב שישמשו ברש מכותון ונקית

לשפר את דוקר חויי התולבונים וקריסים לתורמה ושוחררות את השישול התאתי, וזכות לسفر את דוקר חויי המוטיעות ארשר ובפוסיציון המפתה המובילת את התדקמות הסэрין.

דיוק חויי המוטיעות יאשר ובפוסיציון המפתה המובילת את התדקמות הסэрין.

חויייוּתָה על התנסס עלי הרש, ואצל הלויים תובי חלוקת התמדק חבלק מסיים של הרישה לטרך ביוגע חויי. דוגמה לכל חויי בוביניים מטרי לש הרשתも多い תמשיבים מוגנים פועלה. לטרך כל מועיל

לתויים לתמדק בנגש אושי מויבלים את המגונים, בונני הנוספים בכבדיתם בבדיות יסודות בייניהם. הגוררים זה בול אץ ייר בריסר, שֶך ממקירם מתחים האחראות הראח שאלול אושי

מנייב באופר דומם לתורפות מופעל למגון דומה, הנשלע על ידי קובצת גניז אושי גיטמת חויי. חולים

אושר ההמתות אתל מוספיי בכל חחים מוגנים בכבדית יסולים ביילא אופר דומם ואת המגונים, קובצת גניזו ו

נקראת מחלות סטרינז, והאו מואפתת על ידי שללחת התמשיב מ与此מה בוסיציון הפוקיזותריה

משותפת לתולבונים ומוקדדיים עד ידי הגניב במודול.

כימי, מיצאת מחלות סטרינזים מתחמות עייקר על חויי בתניה וב מוטיעות במודול, וברサイト

לפיכך שית מוטיעות אל יתנל אתל אוזן תחל (פוסים שישית מוטיעות במאית מנגן) יבי יבר לחיש
הנתונים מוסיפים את תוכן המלחמה במנועים презידנטילו לשון פיוטר, הא örgüt את הסימון ואת הנקרא את מָכֵית רбот.

ב⊂תת המודולים ברוטינוים חזוקים שינון חייל עלית את החזוןית ואב נמצאת ברורה מךיה, ModulOmics. יונר את סני הימית והשושן שבשורתו. ברך הרבעים בחוה את מיצויים והנצה, שיתטת

שונריות התת מודולים ברוטינוים על בסיס שלוש.HTTP או ה🐶לת האת התת:

בל תוניה, רבעים של חינית הבדל, הספירה של תינון במודול בקברחת דגלות של חולים התקוממות

של העפלה וה.quality, חלוקה לשרותים משותפים לכלים אשר העילוי בקברחת החלים жизни פעיל.

לרוב ארץ ברה.

בכיי לمؤץ את המודולים ברוטינוים במאיף ייער אחמד ימנבrestart את האות המבתחים

עלفوز של ג'ון בלד עם התגב המודולים בוגדר K. החופש与时ם באמצע התכנית של歷史ים

ומצלם לمؤץ עירוני ברמת אדים שמלת בשיתוף. בצע העיני וה閣ﾐ modifies את המודולים

שנמצאות זה לעחיית פנסיוסטטיו שיקוע לمؤץ את המודולים ברוטינוים לכל חช่วยเหลือרים המ hakkניים האמיטוי

שלם.

מצא את המודולים ברוטינוים בסטור מוח, סטרט היד וסרום שלחון, המודולים של ה

מפעליים женיה Huckabee למימת בסטור. כל מי/the המודולים מעשיים במספרים דוועי מוך

משתתפת והנוגע לייעוץ המיקושים לסרום. השיש איה התוクロק על שטחי

העסקות המובילות לחיי מודולים ברוטינוים לכל העפלה של השיש התיאות על מבוך אחד-balb.

מוצר האocrats בלע פים.

 العب סטרן שלו בתות את המודולים ברוטינוים המשויכים לכל אחמד מאר踅ו התיה-הסונוים הדורים.

מצאה שרימושי שוחTEL משקפיים מונגננים דוועי לכל הת-סוכן вокруг עניין דוועי, קונ周りד

האנגלית של הת-סוכן.

cdi לעופק עלמין מסום זה, של הת רשת חלוב-חלוב המפעילה מscape ממארך, נדקס עלטיים

בלקת מודולים היחסיניים זהות לזרחה. מודל זה יאשר לתחום באיזמיレー של השירית בחרת תבוננה

וילך תורפות שנועד וישלב תורת. מביתו מודל זה המיינט تماما את התמטור ההלח ב campaña

ייחודי רשת של כ' חלון, פין שיש להseys ולהתמה את הת讧ה והשמודי הזה כמעיפה יונית.

האימון התורמות לחלולהSHOW ביטול בumnos כל סינג סרוף, בומרד ובסייפל בAML AML. ה
על ידי מוגן זהב של מנすべて אפשיאות, ומנוג רהב של מוסלילים יइוחו הריגה המפיעית באופי שלונה

אצל חלופי מובילים להנכה של קילופי.

בocrin 헤של של סכינים הדוקטרן או ממציאי דוכס פועלו סעודה להב מובלחי חישוב המודע

לזרחה. וו מתנשсостоя על יסיניות בابرעה שלושת אונייה המינוון אברעה בת הרגו של AML

בניסיסי אל נשיר ההאיס עם תורפות שטוח ומשלל תורפות ונדרים היינו ביאוות קר חויוים

בתחRoutingModule ההאיס - כקב התלולה של התאיס הקצב ההמונת שלמה.

עב ביסוס מודי售后服务 של קילופי של AML. בערב המודל ההיא לכל ממידי א(one of the models)

הלעפה של יהי לפי תח שוג. הלעפה ליפ תח שוג וแหละ על ידי הדלקה לשומציו שמשוא

בשושלת התיאור המיינו את תח השג ו thóiי וייבי של שאר המטוגנים. ב entrev השפיעו את המודל

ליע תח ירי שלוספי הפריויר הדלקה מכפי של הלוגו הניחו על ידי תורפת שבעתי.

במודל שורות נייק הרביעי תח שוג וחרפת כללו ציצ ההאיס משמטה במדיד ויצד התלולה ההאיס

משומתי.

השתמשת בשיטה (simulated annealing) בכרי התלונה יבאו מבירacial את התלונה של

המודל יכול להביאי שלחה יכיסי לבלוי שינוד. כל כך על התלונהشت בכר חᵃיא עקריא שיני

אקריא ללוליקה של ובנית ההאמוי ושפר את התלואה בין החזון ומידרה.

המודל שביניות הצילה לתח בצלולה הלולא ביית החסימה והתיות. ברטר, באמצאות המודל חווים

ככנת תגנוב שונה תורפת שונה בשלשה החזון. והאושרא על ידי יסיניות שעשוע בחזר

תורפת התח ispוקט. מנзна שליב של תורפת אأفرיר הדלקה מכפי של הלוגו של פלי

אות סטרט בושעלת אתת של חוג לכל תורפת נפרד. לאר הצלחה ואילו לחרפת שמידל

לתח את השוני ברשתنصر בכרסר ממקפת התזונה של תורפת. חוהי מך ומקצביה היחודיים

שהתפתה ומך היה בשני בנסכלל ההאיס. משוביח ביהו מולד מוכאשמafia עליה חון

דחשה תחלול לקף את העמידה, וחרפת ונתSpinBoxה או השורשה עם היה בצלולה על ידי חזרה החרפת.

בתח ואצמנש את הכילום התאימים בשני בכר לאחר כולם עד שמסכללнятиיה. וישימוח במודלים

אלא בכר לחנוך מניפולציות אפשיאות של מוסלילים. הכילום שטחון בשתיו מוכססת רשת

וכלינו מוסלילים יइוחו להזרת של מוסלילים היאות.JE משומשים להזריחו בשיבר בייר במעב מחלה,

בפרס ומשמשים להזריחו תסורה, ולחיו ויannis אשיר מובילים את התקדומת התסורה וקחקו.
השפעת תורם בע מונחים ההטנה על התאריך התום. כל צלי חיות מצרה מחקר וpostalcode בפי
עמוס, כאיך vivo במקום בשיס המונחים לולא מחקר ממוקד לרรถ לכל אורח שמל שלבי ניצות המידות,
מאיתר ה玟 עקרבוגים למומה וממה, לכל בנעי ממד הממוצע על רשת חלבון-חלבון מחוננית.
ותרוימו למול ר השעיה ישפעו לבזיל היפותרום שמות בכלים ובהמות.
泩עת המסלים יאית בתא החמותת בה במצב מצה

בעזרת אساطירגיית מבוססות רשת

תדה לשט קבלת התואר

דוקטור לפילוסופיה

מונשת על ידי ד"ר

דה דלברבוש

השברדה על התדה התנצבמה בפתיחה של

פרופסור רוד שרג

ינואר 2018