Alizée Diatchenko

2022

Master's thesis

Master's thesis

NTNU
Norwegian University of
Science and Technology
Faculty of Natural Sciences
Department of Biology

NTNU
Norwegian University of
Science and Technology

Alizée Diatchenko

Logical Modelling of Triple Negative Breast Cancer and *In Silico* Drug Synergy Predictions

May 2022

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of
Science and Technology

Logical Modelling of Triple Negative Breast Cancer and *In Silico* Drug Synergy Predictions

Alizée Diatchenko

MSBIOTECH

Submission date:	May 2022
Supervisor:	Martin Kuiper
Co-supervisor:	Berit Johansen Kaisa Lehti Eirini Tsirvouli

Norwegian University of Science and Technology
Department of Biology

Department of Biology

Master's Thesis

Logical Modelling of Triple Negative Breast Cancer and *In Silico* Drug Synergy Predictions

Alizée Diatchenko

Supervisor:
Martin Kuiper

Co-supervisor:
Berit Johansen and Kaisa Lehti

Technical supervisor:
Eirini Tsirvouli

March, 2022

NTNU
Kunnskap for en bedre verden

Preface

This Master's thesis was conducted within the DrugLogics group and aimed to investigate whether a logical model could be extended in order to describe the main signaling processes responsible for the development and progression of triple negative breast cancer.

I want to thank my main supervisor, Professor Martin Kuiper, for always bringing his experimented scientist's eye on my work, which helped me understand the principles and subtleties of academic research. I am also grateful to Eirini Tsirvouli, my technical supervisor and PhD student, who was always willing to help me circumvent the technical difficulties I faced, and who also brought her scientist's point of view to my work. My thanks also go to the DrugLogics and cPLA2 group for the discussions about this project and many others.

I am grateful to the Ecole Centrale de Nantes and NTNU for allowing students to discover and graduate in different fields of study through their TIME partnership. I will never forget the two years spent as a student at NTNU, and all the amazing opportunities that Norway offered me.

And last but not least, I would like to thank my family and friends for their continuous support in this journey.

Alizée Diatchenko

Abstract

Breast cancer is the most commonly diagnosed cancer worldwide with more than 2 millions of new cases each year, and this number is only expected to rise. Breast cancers lacking the expression of the estrogen and progesterone receptors, and lacking amplification of the human epidermal growth factor receptor 2 are called "triple negative". Triple negative breast cancer is considered as the most aggressive form of breast cancer, and is extremely difficult to cure. While other forms of breast cancer can be treated quite efficiently by hormonal therapy, chemotherapy, surgery, or radiotherapy, triple negative breast cancers do not respond well to treatments and have poor prognoses. With the emergence of high-throughput sequencing techniques, the high molecular heterogeneity among triple negative cases can be more precisely characterised, setting the stage for the development of personalized treatments. Scientific studies recently suggested the use of new therapies including combinations of existing drugs, therefore, clinical trials are underway. However, exploring more and more possibilities for drug combinations can hardly be achieved through clinical trials, which are highly time-consuming and costly. *In silico* simulations are promising techniques to identify new target candidates in diseases with unmet clinical needs and relieve the burden on clinical trials. In this project, a Boolean network representing the altered signalling processes of triple negative breast cancer was built following a middle-out modelling approach, and was calibrated on mRNA expression data of four different cell lines. Multi-omics and pathway enrichment analyses were performed on triple negative breast cancer data to reveal the individual biological entities that are altered in triple negative cases compared to normal samples, and link them to underlying altered functions by pathway over-representation analysis. *In silico* drug synergy predictions were performed on the network to assess its ability to predict observed synergies and eventually find new promising drug combinations.

We found that the model had a limited predictive power when calibrated on mRNA expression data of three triple negative cell lines. Despite multiple steps to find the best parameters to run the synergy predictions, the results remained poor and led us to think that topological issues could have been introduced during the building of the model, and that the calibration data were not adapted to translate the dynamics of TNBC.

We suggest that those areas can be further investigated to improve the TNBC model. Notably, the refinement of the pathways that were modified during the calibration phase could be of considerable help.

Table of Contents

Preface i

Abstract iii

List of Figures vii

List of Tables viii

1 Theoretical Background 1

1.1 Cancers and Their Mechanisms 1

1.2 Cancers Worldwide 2

1.3 Triple Negative Breast Cancer 2

1.3.1 Molecular Heterogeneity of TNBC 3

1.3.2 Current Treatments and Chemotherapy Resistance 4

1.4 Multi-Omics Integration Towards Precision and Personalized Medicine 4

1.5 Systems Biology 5

1.6 The Role of P4 Medicine in Cancer Treatment 6

1.7 Biological Networks 6

1.7.1 Network Science and Biological Networks 6

1.7.2 Logical Modelling of Biological Networks 7

1.7.3 CASCADE 2.0 as a Starting Point for the TNBC Model 7

1.8 Objectives 8

2 Methods 10

2.1 Workflow Overview 10

2.2 Omics Data Resources 10

2.3 Bioinformatics Resources and Statistical Testing 11

2.4 Multi-Omics Analysis 11

2.4.1 TCGA Data Retrieval 11

2.4.2 Differential Expression Analysis 12

2.4.3 Somatic Copy Number-Alterations 12

2.4.4 Frequently Mutated Genes 13

2.4.5 Differentially Methylated Genes 13

2.5 List of Affected Genes 14

2.6 Pathway Enrichment Analysis 14

2.7 Literature Mining 14

2.8 Gene Regulatory Network Building - The TNBC Model . . . 15

2.9 Generation of Boolean Models and Drug Synergy Predictions . . . 16

2.9.1 Training Data Retrieval . . . 17

2.9.2 Generation of Logical Equations with Gitsbe . . . 19

2.9.3 Synergy Data Retrieval and Drug Targets . . . 19

2.9.4 Drug Synergy Analysis with Drabme . . . 20

3 Results and Discussion **21**

3.1 Multi-Omics Data Analyses . . . 21

3.1.1 Differentially Expressed Genes . . . 21

3.1.2 Somatic Copy-Number Alterations . . . 23

3.1.3 Frequently Mutated Genes . . . 25

3.1.4 Differentially Methylated Regions . . . 25

3.2 Composition of the Gene Lists Used for the Enrichment Analysis . . . 26

3.3 Identification of Over-Represented Pathways . . . 26

3.3.1 Results of the Enrichment Analysis against KEGG . . . 27

3.3.2 Results of the Enrichment Analysis against Reactome . . . 28

3.4 Identification of Pathways in the Literature . . . 29

3.5 Composition of the Prior Knowledge Network - the TNBC Model . . . 32

3.6 Conversion of the Prior Knowledge Network into a Boolean Network . . . 35

3.6.1 Training of the Models on Data from TNBC Cell Lines . . . 35

3.6.2 Training of the Models on TCGA Data . . . 37

3.7 Drug Synergy Predictions . . . 38

3.7.1 Tested Drugs and Their Targets . . . 38

3.7.2 Relevance of Using TCGA Samples for Drug Synergy Predictions . . . 39

3.7.3 Specificity of the Drug Panels . . . 40

3.8 Optimization of Gitsbe and Drabme Results . . . 40

3.8.1 Evolution of the Performance with an Increasing Number of Simulations . . 41

3.8.2 Evolution of the Performances and Calibration on Best Determinative Power Nodes . . . 44

3.8.3 Comparison of Prediction Performances With TCGA Calibration Data . . . 46

3.8.4 Best Predictive Results . . . 47

3.8.5 Prediction Performances with CASCADE 2.0 . . . 51

3.8.6 Predictive Performances of the TNBC Model . . . 51

4 Conclusion **53**

Bibliography **55**

Appendix **69**

A Versions of the Bioinformatics Tools Used 69

B TCGA Barcodes of the TNBC Patients 69

C Model Outputs 70

D Drug Panels of the CCLE Cell Lines 70

E Chromplot of the Somatic Copy Number Alterations 71

F Most Over-Represented Pathways Against KEGG 72

G Most Over-Represented Pathways Against Reactome 75

H Nodes of the TNBC Model and Their Topological Measures 78

I Fitness Evolution of the Cell Line-Specific Models 84

J Prediction Performances of MDA-MB-231 Models with the TNBC-Receptors Forced Inactive 85

K Prediction Performances of HS-578T Models Using the HSA Synergy Metric 85

L Prediction Performances of BT549 Models Using the HSA Synergy Metric 86

M Prediction Performances of MDA-MB-231 Models Using the HSA Synergy Metric and a Combination of Topology and Balance Mutations. 87

List of Figures

1 The ten hallmarks of cancer 2

2 Array-Array Intensity Correlation Matrix 21

3 Boxplot of mRNA Transcripts Before and After normalisation 22

4 Volcano plot of the Differentially Expressed Genes 22

5 Bubble plot of the significantly amplified or deleted chromosomal regions 24

6 Volcano plot of the Differentially Methylated Regions 25

7 Prior Knowledge Network visualisation 35

8 Heatmap CCLE 36

9 Heatmap of RNA expression of the TCGA samples used for training 38

10 Evolution of the prediction performances in terms of ROC-AUC for an increasing number of simulations using the three CCLE cell lines calibration and synergy data 43

11 Evolution of the prediction performances on a varying number of calibrated nodes 45

12 Evolution of the prediction performances of models trained on TCGA-BRCA data with an increasing number of simulations 46

13 Evolution of the prediction performances in terms of ROC-AUC of models trained on TCGA-BRCA data with an increasing number of calibrated nodes 47

14 Best predictive results on the TNBC cell lines 50

15 Prediction performances of CASCADE 2.0 on the TNBC cell lines 51

List of Tables

1 Availability of TNBC cases from the TCGA-BRCA project in each data type of the GDC Data Portal 12

2 Parameters used to run GISTIC 2.0 algorithm 13

3 Parameters used to run Gitsbe simulations 19

4 Number of genes identified in each of the omics analysis performed 26

5 Number of significantly over-represented pathways identified in each of the gene lists (FDR≤0.05). 26

6 Pathways identified by the Enrichment Analyses of the different gene lists against KEGG 27

7 Pathways identified by the Enrichment Analyses of the different gene lists against Reactome 28

8 Modules present in the TNBC model 34

9 Topological properties of the PKN 34

11 Expression status of key nodes 37

12 Synergistic drug combinations observed for the three CCLE cell lines 39

13 Activity of key nodes across the best models 48

1 Theoretical Background

1.1 Cancers and Their Mechanisms

Cancer development is a highly complex process. Animal bodies are composed of thousands of billions of cells which live in harmony and cooperation to sustain the basic biological functions necessary to the survival of the whole organism. However, in their lifetime, these cells undergo mutations in their genome, and even though a large part of these mutations are harmless to the balance of the organism, some of them can alter the behaviour of one cell, and lead to the development of cancer by giving it a selective advantage over normal cells. Cancer cells are usually defined as cells that are able to grow and divide despite the absence of growth or proliferation signals, and are able to invade other tissues than the tissue in which they arose [1]. These two properties are facilitated by what were characterised as "the hallmarks of cancer".

The hallmarks of cancer correspond to capabilities attributed to cancer cells that, for example, enable them to proliferate uncontrollably whereas their division and proliferation should just balance other cell's death. Figure 1 shows the well-known representation of the ten hallmarks of cancer as defined by Hanahan et al. (2011) [2]. In 2000, Hanahan et al. already defined six cancer hallmarks which are now well accepted and further characterised [3]. The sustained proliferation of cancer cells is due to their capacity of triggering growth signals which will act in a positive-regulatory loop to trigger the growth and proliferation of the cancer cells themselves. Cancer cells are also able to bypass the mechanisms that are supposed to negatively regulate their proliferation (with molecules such as growth suppressors) and the mechanisms that lead to apoptosis by disrupting the activity of molecules involved in different steps of these processes. Furthermore, the shortening of the telomeres, the extremities of the DNA strands, is crucial to impose a limit to the proliferation capacity of a cell: cancer cells express a high level of telomerase, the enzyme responsible for the addition of telomere segments at the end of the DNA strands, which makes them able to infinitely replicate. The angiogenic capacity of tumours is also remarkably enhanced compared to normal situation where the angiogenic process in adults is only used in cases of wound healing or child bearing. This increased vascular activity participates to the feeding of cancer cells which need oxygen and nutrients and are thus, able to be autonomous. As a sixth hallmark, Hanahan et al. defined the capacity of cancer cells to change their shape and lose their bindings to the extracellular matrix (ECM), enabling them to invade other tissues and develop metastasis. In 2011, two new hallmarks have been added to this list, namely the deregulation of cellular energetics, enabling the modification of the cell's metabolism to sustain the tumour's needs, and the evasion of the immune destruction (cancer cells are able to bypass the effect of immune cells). Two additional features of cancer cells were added to this map of hallmarks, for being facilitators of the development of the eight other hallmarks: genome instability and mutations to which cancer cells are prone, and the tumour-inflammatory response which promotes tumorigenesis by infiltrating tumours and providing essential molecules to it [2].

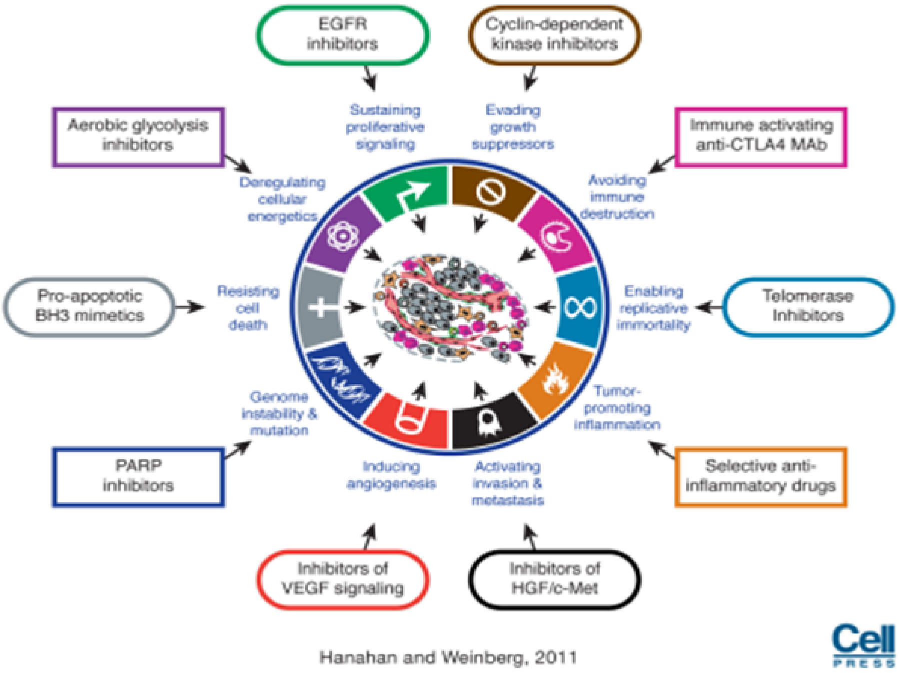

EGFR inhibitors
Cyclin-dependent kinase inhibitors
Aerobic glycolysis inhibitors
Sustaining proliferative signaling
Evading growth suppressors
Immune activating anti-CTLA4 MAb
Deregulating cellular energetics
Avoiding immune destruction
Pro-apoptotic BH3 mimetics
Resisting cell death
Enabling replicative immortality
Telomerase Inhibitors
Genome instability & mutation
Tumor-promoting inflammation
PARP inhibitors
Inducing angiogenesis
Activating invasion & metastasis
Selective anti-inflammatory drugs
Inhibitors of VEGF signaling
Inhibitors of HGF/c-Met
Hanahan and Weinberg, 2011
Cell PRESS

Figure 1: The ten hallmarks of cancer and drugs commonly used to target them. Retrieved from [2].

1.2 Cancers Worldwide

It is not only the complexity, but also the incidence of cancers that make it a global concern. Cancer is one of the leading causes of mortality in most countries worldwide. In 2020, it was estimated that 19.3 million new cancer cases arose, against 12 millions in 2012, and this number is expected to increase by 47% in 2040 compared to 2020 (for a total of 28.4 million new cases per year) [4, 5]. The expected increasing incidence of cancers worldwide, but especially in developing countries, is explained by the economic development of these countries which will lead to the emergence of new lifestyle and social habits, and also by the growing global population. In 2020, it was estimated that almost ten millions of persons died because of cancer, the most deadly type being lung cancer. Additionally, 22.8% of cancer cases and 19.6% of cancer deaths occur in Europe, while it only represents 9.7% of the world's population [4].

In 2020, female breast cancer represented 11.7% (2.3 million new cases) of all the diagnosed cancer cases, and was therefore the most commonly diagnosed cancer type, followed by lung cancer (11.4% of total cases) [4]. However, the mortality rate of lung cancer is higher than for breast cancer (18% and 6.9%, respectively). The incidence rate of breast cancer is 88% higher in developed countries than in developing countries, but the mortality rate is 17% higher in developing countries. These figures can be explained by the prevalence of risk factors in developed countries, such as a higher alcohol intake and the lack of physical activity, the advanced age at the first child's birth, or the intake of oral contraceptives. Also, the number of cases in developing countries can be underestimated due to a lower rate of mammography screening. The high rate of mortality in developing countries is mainly due to later-stage detection and more difficult access to treatment [4, 6].

1.3 Triple Negative Breast Cancer

Breast cancers can be classified according to the expression status of three protein receptors: the estrogen receptor (ER), the progesterone receptor (PR), and the human epidermal growth factor receptor 2 (HER2 or ERBB2) [7]. Additionally, differences in the gene expression patterns of breast cancer cases enabled to define several subtypes [8]: First, the basal-like which are ER and PR negative (ER- and PR-), and positive either for keratins 5/6 or 17 (or both), by immunohistochemical (IHC) staining. These markers are characteristic of basal cells, which explains the name of this subtype. We also have the luminal-like which are ER+ or PR+ (or both) and positive for keratin 8/18. Luminal-like can be either HER2 positive of negative (Luminal A or Luminal B), and are named like this since they exert a luminal phenotype. Basal and luminal cells are epithelial cells, in opposition to mesenchymal cells. Some tumours can be classified into an ERBB2-enriched subtype and show high expression of specific genes of the ERBB2 family and low expression of ER and PR, and ER associated genes. The last type described by Perou et al. (2000) is the normal breast like subtype, gathering gene expression patterns from basal epithelial, adipose, and luminal epithelial cells. This gave rise to the widely-known PAM50 classification [8].

Triple negative breast cancers (TNBC) are defined by the absence of ER and PR, and the lack of amplification of ERBB2 by IHC or fluorescent in situ hybridization (FISH) [9]. TNBC represent from 10% to 20% of all breast cancer cases [10], and if a majority of TNBC can be classified as basal-like (50% to 75%), they can also belong to the Luminal A or B subtypes, to the newly defined Claudin-low subtype, or, surprisingly, to the ERBB2-enriched subtype (which is not defined by a high expression of ERBB2, but of a subset of specific genes). The Claudin-low subtype is characterised by low expression of claudin genes, which are responsible for cells' tight junctions (for example the cadherins), and a triple negative IHC status of the ER, PR, and ERBB2 [11].
According to several studies, TNBC is more frequent in black or hispanic women under the age of 40 than non-TNBC [10, 12]. Also, TNBC patients have worse overall-survival (OS) than patient with non-TNBC with a median OS of 18 months according to some studies [12, 13].

1.3.1 Molecular Heterogeneity of TNBC

The mutational profile of breast cancer has been extensively studied, and it has been shown that the prevalence of BRCA (Breast Cancer genes) mutations is higher in TNBC than in non-TNBC, with up to 15.6% of patients with BRCA1 mutations [14, 15, 16]. BRCA genes are well known for their involvement in hereditary breast cancers and are classified as risk factors as women carrying a mutation in one of those genes have an increased chance of developing breast cancer in their lifetime [10]. Furthermore, women carrying germline mutations in BRCA genes and BARD1, PALB2, and RAD51 would have an increased risk of developing TNBC [17]. Additionally, copy number alterations (CNAs) have been found in PARK2, RB1, PTEN, and EGFR genes, and frequently mutated genes have been identified, including TP53, PIK3CA, USH2A, MYO3D, PTEN, and RB1 [18]. However, Shah et al. (2012) and many others also found a high mutational and expression heterogeneity in TNBC and emphasize the need of considering further subtypes in order to develop treatments that would increase the prognosis of TNBC patients [10, 19].

Indeed, to be able to further characterise the molecular biology of TNBC, classifications criteria have been defined. A well-known classification has been made by Lehmann et al., implying the existence of six transcriptional subtypes first [20], and the refinement of these six subtypes into four subtypes later [21]. TNBC cases could thus be classified into the basal-like 1 (BL1), basal-like 2 (BL2), luminal androgen receptor (LAR), or mesenchymal (M) subtypes, depending on several features such as the age at diagnosis, the histopathology, the tumour grade, and the spatial progression of the disease. High expression of cell cycle, cell division, proliferation, and DNA damage response genes, mRNA and pathways is characteristic of BL1 cases. They were found to have the highest pathological complete response (pCR) to neoadjuvant therapy (41%) and the highest OS compared to the other TNBC subtypes. On the other hand, BL2 uniquely show high activity of growth factor signalling pathways such as the epidermal growth factor (EGF) or Wnt/β-catenin. The LAR subtype is characterised by a high expression of genes involved in hormonally regulated pathways such as the androgen and estrogen metabolism. Analysis of the androgen receptor (AR) revealed that its level was high both at the trancriptomic and proteomic level in this subtype compared to non-LAR TNBC. Also, the LAR TNBC show a tendency to metastasize to bones and have unique globular histology. Finally, the M subtypes uniquely express higher levels of epithelial-to-mesenchymal transition (EMT) pathways such as the Rho signalling, and differentiation pathways, and have a tendency to metastasize to the lungs. Overall, it seems that the BL1, BL2, and M cases are a majority of basal-like breast cancer as described above, while LAR cases are more of the ERBB2-enriched subtype [20, 21].

This classification further proves the high heterogeneity of this disease and gives tools to study TNBC in more details.

1.3.2 Current Treatments and Chemotherapy Resistance

Since TNBC do not express ER, PR and lack amplification of ERBB2, they cannot benefit from the hormonal and anti-HER2 therapies commonly used in other breast cancers [10]. Currently, the standard of care (SOC) for TNBC patients is chemotherapy. For newly diagnosed cases, a combination of chemotherapy with neoadjuvant therapy and surgery is commonly used. However, for advanced cases, chemotherapy remains largely inefficient, with a median progression-free survival (PFS) of 1.7 to 3.7 months and an OS of ten to 13 months [22]. Advanced cases can benefit from single agent platinum or taxane-based chemotherapy, but overall, the rate of relapse and metastasized TNBC is higher than in non-TNBC, recalling once again the urgent need of finding more personalized and targeted therapies.

Recently, new treatments have been approved by the Food and Drug Administration (FDA), namely atezolizumab and nab-paclitaxel (paclitaxel combined with albumin) in the case of untreated metastatic TNBC positive for PD-L1 [23]. Indeed, some TNBC cases show a positive expression of the programmed cell death ligand PD-L1 [21], and the combination of an anti-PD-L1 antibody and a mitotic inhibitor significantly improved their median PFS and OS [23]. Another drug, pembrolizumab, is still undergoing clinical trials, but shows promising results as a PD-L1 inhibitor [24].

Additionally, two PARP inhibitors are FDA-approved and used as single-agents in the case of patients with germline mutation of BRCA. On one side, Olaparib is used for patient with ERBB2 loss, at a metastatic stage, and with germline BRCA mutation. It improves the median PFS and reduces the risk of disease progression or death [25]. On the other side, Talazoparib increases the PFS for patients with advanced stage and a germline mutation in BRCA [26].

All-in-all, these results suggest that targeted therapies could improve the prognosis of TNBC patients, given that we are able to classify them in relevant subtypes, according to their molecular profiles or other clinical features. This is the aim of precision and personalized medicine.

1.4 Multi-Omics Integration Towards Precision and Personalized Medicine

It has been almost two decades now since scientists have been able to fully sequence the human genome [27]. Since then, huge advances in sequencing techniques, and especially the emergence of high-throughput new generation sequencing (HT-NGS) techniques have led to the production of massive amounts of biological data, the so-called ”omics”. As a consequence, bioinformatic tools and strategies have been developed to handle these data and detect genomic variations, which among other applications, helped scientists further understand the complexity of human diseases [28]. HT-NGS enable to perform accurate genome sequencing much faster and at a much lower cost than the first generation sequencing techniques [29]. High-throughput technologies have not only been able to produce genomic data to help understand the genetic variants, but also new types of omics covering almost all the layers of compounds of an organism.

While DNA-sequencing provides information on the genome, and especially on single nucleotide variations (SNVs) and structural variations (SVs) present in a sample, RNA-sequencing provides information on the transcriptome, and therefore, on the genes that are transcribed into RNA. This represents the transition from DNA to RNA in the central dogma of molecular biology [1]. Transcriptomics comprise quantitative information on both coding and non-coding RNAs, such as long non-coding RNAs and microRNAs, which are now widely studied for their involvement in many abnormally functioning biological processes. Epigenomics data are tightly correlated to transcriptomics, since they relate the epigenetic variations affecting DNA (such as methylation), playing a key role in the transcription of genes. Proteomics, as the name suggests, offer an overview of the proteins that are finally translated (i.e. the translation of mRNA into proteins in the central

dogma), how they interact, and how they are modified after their translation (phosphorylation or ubiquitination for instance). This is key information to understand how the proteins' activity impacts biological phenotypes through their signalling, transport, and interaction with their environment [30].
This addition of layers to the knowledge of scientists has brought an important challenge to tackle: if it is important to understand the individual role of genes, proteins, or events driving their expression or activity, it is fundamental to handle the problem at a system's level as well, and combine the data types to leverage more knowledge on their interactions and role in diseases. Major discoveries have been done by combining multiple omics, and the scientific community is continuously developing new tools and methods to improve the use of these data [31, 32, 33, 34, 35].
Focusing on breast cancer, studies showed the evident need of integrating multi-omics to identify cancer-driver genes, or new therapeutic targets. For instance, Mertins et al. (2016) showed the correlation between recurrent somatic mutations in breast cancer and the elevated expression of several proteins, and they observed clusters of cases sharing similar proteomic expression patterns [36]. Also, the methylation of CpG islands is known to be involved in the repression of cancer genes, especially when located in the promoter regions of tumour suppressors [37]. Esteller et al. (2000) showed the correlation between the hypermethylation of the promoter regions of the BRCA1 gene and its loss in breast cancers [38].

1.5 Systems Biology

Systems level understanding has existed for many years before the emergence of omics data and is commonly used in biological sciences, and applied to ecology, population biology, and evolutionary studies. Systems biology comes as a powerful method to understand the behaviour of an entire system, each part of this system being able to influence the others. As an example, large-scale metabolic networks already existed in the 1970s [39]. This method is somehow opposed to reductionism, that claims that a complex system can be understood by studying its different parts separately. Both methods have their advantages, and using them both can be complementary, that is why they are not totally opposed. René Descartes (1596-1650) was a strong defender of reductionism, contrary to Ludwig von Bertallanfy (1901-1972), who is considered as the father of general systems theory in the 1940s. In molecular biology, reductionism tackles the genes and not the pathways for example, but it is also important to understand that an organism needs to be studied as a whole in order to draw conclusions on its dynamic behaviour [40]. Now that systems biology theories are well-established and progress in molecular biology has been made, especially in recombinant technologies and high-throughput sequencing, systems biology is emerging as a remarkably interesting tool in this particular field. [39]. Indeed, biological processes at the origin of life can be described as subparts of a network regulated by cis-regulatory elements and transcription factors acting together on the control of gene expression and signalling events. Genetic or environmental perturbations can modify these interactions, and eventually lead to a disease state if the network is not robust enough to circumvent these modifications [41]. Theories about systems biology are widely established and studied all around the world by many scientists such as Hiroaki Kitano, the head of the Systems Biology Institute. According to H. Kitano, systems biology research is composed of four major domains: Genomics, computation, analyses, and technologies. These domains complete each other for the creation of a realistic model, the simulation of its behaviours and predictions based on this model [42]. In the same perspective, T. Ideker defines the four steps of systems biology to build a robust, reliable, and realistic model. The first step is to define all the components of the system, their interactions, the general dynamic of the system, and to build a first model. The second step is about designing an experiment expected to emphasize the role of the components and their functions when they are put together. From this experiment, one should be able to compare what was expected and what was observed, and then to design new experiments on the model, which will respectively compose the third and fourth steps of systems biology [43]. Those four steps will be followed in this project.

1.6 The Role of P4 Medicine in Cancer Treatment

Systems biology unveils new key components at all levels of an organism: in diseases, it helps uncover the dissimilarities between the individuals, and before being able to implement individual medicine, it helps decomposing the diseases into subgroups that are characterised in more details than before, and share a larger part of their molecular profile [41]. New biomarkers are discovered by clustering patients together in smaller groups, as for example in TNBC (see section 1.3.1), which helps finding more targetable or responsive elements than the ones currently targeted by treatments. Further mapping of the diversity of diseases such as cancer brings us closer to being able of predicting the exact outcome of a treatment on a patient, only by knowing its genetic profile (which is becoming easier with HT-NGS). Adapting the treatment to each patient to minimise the risk of relapse or progression of the disease is one of the main goals of predictive, preventive, personalized, and participatory (P4) medicine. Some scientists predicted already ten years ago that the P4 medicine will be possible through the improvement of systems biology theories, integrative tools using all levels of biological data, and new measurements techniques, that will be used on a regular basis on each individual [44]. Today, if the efforts are largely focused on the treatment of diseases, the main goal of P4 medicine is still to completely change the structure of the medical system, which consists in treating the disease once symptoms are present, into a system that enables the ”maintenance of wellness” [45].

1.7 Biological Networks

1.7.1 Network Science and Biological Networks

Network science is a relatively new field of research accounting for the complexity of diverse types of systems. In theory, the interacting components of any system can be wired to each other in what we call a network, and can be described by mathematical tools and used to study the behaviour the system. The emergence of this scientific field has been eased by the Internet which is a powerful tool to store all the data relative to a network. Network types can range from social interactions (e.g. friendships in a high school), to the mapping of the power grid of a country, including neural and gene interactions networks. All these networks are made of components (the nodes) interacting through wires (the edges), and can simply be translated to graphs. The graph theory is then used to describe the topological properties of the nodes or of the network as a whole. For example, the degree of a node is its number of neighbours, and this property can be generalized to the network through its average degree, or the degree distribution. The latter is an essential measure to assess the robustness of a network, that is, its ability to circumvent perturbations without losing its general dynamic. Indeed, if most nodes of a network have a low degree, and that only a few (which we call hubs) have a high degree, the network's integrity would be threatened by a perturbation targeting his hubs, since they interact with most of the other nodes. On the other hand, a network with a normal degree distribution (following a Poisson's distribution, for which most of the nodes have the same degree) would be affected to a lower extent by a perturbation, but it would be equally affected by any perturbation. These two types of networks are respectively called scale-free and random [46].

The robustness of a cellular network is particularly crucial to maintain the basic biological functions of life, and it has been observed that most of the cellular networks are scale-free (e.g., the transcription regulatory networks of *Saccharomyces cerevisiae* and of *Escherichia coli*). In these networks, the hubs were also found to be essential genes [47].

Today, many scientists focus on developing biological networks to answer the challenges of P4 medicine. Many types of network can be mapped according to the nature of their nodes and edges. For example, the nodes can represent genes, proteins, complexes or metabolites, and the edges can represent protein-protein interactions (PPI), protein-metabolite interactions, or any type of regulatory relationship into a gene regulatory network [35, 47]. The integration of multi-omics data into biological networks is now commonly used to cover a wide range of genomic events involved in the emergence and the development of diseases [33, 48]. Several methods and tools have been developed to then study these networks and uncover valuable biological information, with the aims of discovering new molecular mechanisms, characterise further diseases, and help relieve the pre-clinical screening burden by finding new potential drugs [33, 35]. In particular, methods such as pathway enrichment analysis, module identification, or marker prioritization have been developed based on the network topology or on the significance of the nodes, in order to highlight key regulatory processes [35, 49]. Other measures based on the static or dynamic properties of biological networks such as the speed of information relayed by a node, the quantity of information flowing through an edge [50], or the relevance of a node in the dynamic of the network [51] showed that the most important features of a complex model can be identified and eventually given more weight in the calibration of simulation steps.

1.7.2 Logical Modelling of Biological Networks

Biological networks can be formally represented as abstract circuits of components together playing on the activity of each others. In a logical model, at time t, a discrete value is assigned to each component, which corresponds to their state. The state of a node x is function of logical equations, which depends on other components' state, and on the nature of the interaction of these components with x. There can be positive or negative interactions, and they can be weighted depending on the level of regulation that its regulators exert on it [52]. In biological network, and especially in gene regulatory networks, the activity or presence of a molecular compound of the system is represented by its state.

A simplified logical formalism is the Boolean formalism. In Boolean Networks (BN), the states of the components can only take two values: 0 if they are inactive or not present in the system at time t, or 1 if they are active. The logical functions become Boolean equations using the Boolean algebra to describe the control that the regulators of node x exert on it. The Boolean algebra basically describes which combination of the positive or negative regulators should be active or not for x to be active, by using the logical connectives AND, OR, and NOT [53]. However, while Boolean models only consider binary states of activity, multi-level logical models in general offer the possibility of multiple discrete states which are particularly relevant to represent the influence of a regulatory component over another, and both formalisms are used to study biological networks [52].

The topology of a logical network is the same as the topology of the corresponding biological network: the difference resides in the dynamic of the model, which can be qualitatively studied in logical model. State transition graphs represent every possible global state that can be taken by the system (=2^n possible combinations for a system made of n components), and the possible transitions from a global state to another. Many properties of the system can be seen in this graph, such as stable states or attractors. In biology, the stable states of a system are usually reached when a lasting, stable situation is reached. They can be fix points, or cyclic attractors, where the system transits between a set of possible global states [52], and they are of particular interest in cancers. Indeed, multiple genomic events would modify the topology of a normally functioning cellular network to create new attractors, trapping the tumour cells into an unhealthy cycle [54].

Compelling examples of logical networks representing cancer systems indicate that it is a promising area of systems biology. Also, principles mentioned in Section 1.7.1, such as the identification of high-influence nodes to eventually reduce the size of the system of interest are often applied to logical models [50, 55, 56, 57]. Various contexts have been explored using logical modelling, such as the modelling of differentiation processes [55] or drug combination discovery [57, 58]. Breast cancer models have been developed too, and used to assess the sensitivity and robustness of the models to many drugs [50, 59].

1.7.3 CASCADE 2.0 as a Starting Point for the TNBC Model

Another compelling example of the use of the logical formalism in systems biology is CASCADE 2.0. CASCADE 2.0 is a manually-curated logical model representing cancer signalling pathways. It was itself extended from CASCADE 1.0, a gene regulatory network representative of adenocarcinoma, in the aim of representing the most important cancer signalling pathways [58, 60]. CASCADE 2.0 contains 11 pathways, namely NFκB signalling, PI3K-Akt signalling, signalling by Rho GTPases, signalling by RTKs, Apoptosis, Cell cycle, JAK-STAT signalling, MAPK signalling, TGFβ signalling, Wnt signalling, and mTOR signalling. 144 nodes and 367 edges were present in

the first version of this network, which was later refined based on the results of dynamical analyses that they performed.
The extension process of CASCADE 1.0 to CASCADE 2.0 aimed at constructing a model covering the signalling processes in which the targets of 18 small-molecule inhibitors are involved. Using baseline activity data of cell lines from different cancer types, they converted this Prior Knowledge Network (PKN) into a Boolean network that fits the observed activity of the nodes in each cell line, and perturbed this model by inhibiting the targets of the 18 drugs. It resulted in a significant number of true positive predictions and few false negatives, which is promising to reduce the panel of preclinical drug screening to discover new drug synergies. Later, they refined the network by improving the topology of the pathways responsible for the false negative predictions, and they ran new drug synergy predictions by calibrating only the most influential nodes of the network. The influence of a node was calculated based on its capacity to change the synergy predictions or create a complex attractor when switching its state to active or inactive. The predictions based on only a subset of calibrated nodes resulted in equal or higher number of true positives, and equal or lower number of false positives and false negatives [58].

In the wave of CASCADE 2.0, the CASCADE 3.0 model was built by extension to represent colorectal adenocarcinoma through major signalling pathways. The extension effort was based on a middle-out modelling method, combining both multi-omics data integration and literature mining. CASCADE 3.0 construction was based on the over-represented pathways in a list of genomically and transcriptionally altered genes, and on a large literature curation focused on the known altered pathways in colorectal adenocarcinoma [61]. This effort resulted in a network of 183 nodes and 603 edges, with the addition of signalling pathways such as Notch, Hedgehog, and Hippo, and the completion of some modules of CASCADE 2.0. Drug synergy predictions were improved using CASCADE 3.0 rather than CASCADE 2.0 for one colorectal adenocarcinoma cell line, and were more mitigated for another one. Surprisingly, CASCADE 3.0 performed well on non-colorectal, epithelial cell lines used as calibration data in Niederdofer et al. (2020). CASCADE 3.0 is thus considered as an interesting reference to extend existing models by middle-out modelling into more specific models.

1.8 Objectives

The general objectives of the project include:

1. The development of a logical model that allows accurate representation and system behaviour prediction for TNBC
2. The deployment of a middle-out modelling approach, starting from CASCADE 2.0, and extending it to a TNBC model
3. The calibration and testing of the model against existing data specific to TNBC
4. The prediction of new drug combinations which could be potentially considered as treatment candidates

Additionally, this project aimed to answer the following questions:

- Can multi-omics data leverage enough relevant biological information at the individual entity level to enable the discovery of key altered pathways in TNBC?
- Can we identify typical parameters for the pipeline which optimise the prediction power of the TNBC model?
- Do TNBC cell lines produce Boolean models which behave in different ways?

Objective 2. was motivated by the previous successful extension of CASCADE 2.0 into a colorectal cancer model by middle-out modelling. This project enabled the construction of a CASCADE

2.0-family model, extended for TNBC, which we will commonly designate as the TNBC model. Objective 3. was handled with the use of mRNA expression data of TNBC cell lines for the calibration, allowing the construction of cell line-specific models. Those models were expected to display similar behaviour to their respective cell line. The comparison of the steady states of the model with the experimental cell line data was done by computation of a fitness score, enabling us to assess the quality of the model topology. Additionally, the predictive power of the model was tested by the introduction of perturbations imitating the effects of drug combinations that were already experimentally assessed. Objective 4. would be valuable if the models were able to provide correct predictions on existing data specific to TNBC.

2 Methods

A Github repository was created to store the necessary files to reproduce the bioinformatics analyses performed throughout this Master's thesis. The repository can be found following this link: **TNBC repository**. The name of the files of interest will be mentioned along the report.

2.1 Workflow Overview

The goal of this Master's thesis was to develop a Boolean model representing the main signalling pathways involved in Triple Negative Breast Cancer (TNBC), and be able to correctly predict known drug synergies and possibly predict new ones, by *in silico* perturbation of the system. The challenge here was that TNBC cases are known to have highly heterogeneous molecular profiles, making it hard to find an appropriate treatment for most of the patients [1, 10]. Indeed, we built a gene regulatory Boolean model of TNBC, and the synergy predictions were then performed using cell lines-specific data, in order to calibrate the state of the genes accordingly to the molecular profile of different cell-lines.

The TNBC model was built by extension of an existing model: CASCADE 2.0 [58]. CASCADE 2.0 is a generic logical model representing the main signalling pathways in cancer. It contains 144 nodes and 367 edges, which represent 11 cancer signalling pathways. The list of nodes composing the pathways is not exhaustive, but they contain the most important signalling genes in order to respect the dynamic of the cancer processes, as identified by Niederdofer et al. (2020). The nodes of CASCADE 2.0 either represent the protein product of a gene and are named by their HUGO gene symbol, or they are genes, family of proteins, or complexes. These three types of nodes are annotated with "_g", "_f" or "_c" respectively. This model was efficiently used to predict drug synergies on different cell lines from gastric adenocarcinoma, colorectal, and prostate cancers, and was indeed considered as a basis for a CASCADE 2.0-family model for TNBC. The same convention of annotation as in CASCADE 2.0 were used in the TNBC model.

A middle-out modelling approach was used to identify the relevant pathways that needed to be added to CASCADE 2.0. This approach was used in Tsirvouli et al. (2020) [61], and was proven effective to extend CASCADE 2.0 to a colorectal cancer model. The middle-out approach included a top-down step during which we analysed different types of omics data to identify genes that were either differentially expressed, differentially methylated, frequently mutated, or abnormally duplicated or deleted. These genes were enriched against pathway databases to find out which processes were affected in TNBC cases, and add those which were not represented in CASCADE 2.0. Additionally, we identified the pathways that were already known to be abnormally regulated in TNBC, by literature curation. The bottom-up modelling step consisted in the completion and the detailed study of these general processes at the components and interactions level. To be able to handle the quantity of information contained in the model, and to run simulations in an acceptable time, the size of the network needed to be constrained, and only the most important pathways were added.

CASCADE 2.0 was then extended with new pathways composed of the main nodes and interactions retrieved from pathway and interaction databases, but also in the literature. At this step, the model was a gene regulatory network giving information about the nature of the causal interactions between different genes, proteins, or complexes, which we call a Prior Knowledge Network (PKN). The next step was to convert this PKN into a Boolean network, based on RNA expression data of TNBC cases. Several sets of Boolean equations ruling the network were generated so that the dynamic of the resulting networks would fit these RNA expression data. The final step was to predict drug synergies of two-drug combinations on these different models, and assess the performance of the models in predicting True Positive (TP) and True Negative (TN) synergies.

2.2 Omics Data Resources

The Genomic Data Commons (GDC) is a project from the National Cancer Institute (NCI) which aims to make cancer genomic data accessible for the scientists worldwide. Indeed, the development of high-throughput sequencing produces a massive range of data, and the GDC resources provide an easy access to large extent genomic consortia such as The Cancer Genome Atlas (TCGA) [62]. TCGA is a program orchestrated by the NCI and the National Human Genome Research Institute. 33 cancer types have been profiled and analysed by the TCGA Research Network. Breast cancer was covered in TCGA Pan-Cancer analysis, along with 11 other cancer types. six types of omics data were measured for these 12 cancer types: mutation, copy number, gene expression, DNA methylation, microRNA, reverse-phase proteomic arrays, and clinical data [63]. The omics data used in this project were mainly retrieved from the GDC data portal and belong to the TCGA-BRCA project, unless stated otherwise.

2.3 Bioinformatics Resources and Statistical Testing

In this project, R was mainly used for the bioinformatics analyses [64]. R is a programming language for statistical analyses which can be used on RStudio, an Integrated Development Environment (IDE) [65]. Bioconductor packages were used to analyse data [66]. The code used for the different omics data analyses, the enrichment analyses, and to trace most of the plots related to it can be found on the Github in the file $TNBC\_data\_analysis\_pipeline\_v2.R$.
When performing statistical tests, it is important to assess the relevance of the results. The "null hypothesis" is commonly defined as the absence of significant difference between two sets of observations that we wanted to prove different. For instance, a true null hypothesis in a clinical trial would mean that there is no difference between the drug tested and the control test [67]. The p-value of an observation enable to evaluate if the probability of the event described as "the null hypothesis is true" is low enough to reject it. p-value are necessary in findings that will lead to decision-making such as marketing of drugs that were deemed efficient, in the way that it is a "safety" threshold, and it can be taken really low to increase the reliability of the product. However, it is necessary to consider it carefully and keep in mind that this is a probability, and we cannot conclude from a p-value that a result is true or false. It is advised to consider other significance measures such as confidence interval or the confidence of the tools used in the study to assess the significance of statistical tests [68].
Furthermore, when performing multiple tests of the probability that the null hypothesis is true, the chances to get false results can be high. For instance, performing 100 tests and considering that a p-value lower than 0.05 is enough to reject the null hypothesis could lead to five false positive results, among the results that were supposed significant [69]. To take the multiple testing into account, it is relevant to adjust the p-value with the False Discovery Rate (FDR), which is defined by Glickman et al. as "the expected fraction of tests declared statistically significant in which the null hypothesis is actually true" [70]. A widely-used method to decrease the FDR in multiple testing is the Benjamini-Hochberg (BH) procedure [69], which uses the p-value ranking of the multiple tests and correct them with the ratio of the FDR and the total number of tests.
In this project, the significance of the statistical tests performed in the omics analyses, and in the enrichment analyses was assessed using the BH procedure to calculate the FDR-corrected p-value (also referred as q-value in some tools).
In parallel, other bioinformatics tools such as GISTIC and MutSigCV, and public knowledge databases have been used to conduct the analyses and will be described in more details in the following sections. A summary of all the tools used and their version can be found in Appendix A.

2.4 Multi-Omics Analysis

2.4.1 TCGA Data Retrieval

Triple Negative Breast Cancer (TNBC) sample data were retrieved from the BRCA project of TCGA using the list of barcodes provided by Jiang et al. [71]. These barcodes were selected based on the Immunohistochemistry (IHC) status of the estrogen receptor (ER), progesterone receptor (PR), and the human epidermal growth factor 2 (HER2). A total of 173 patient's barcodes corresponded to a TNBC in the GDC data portal repository (and are available in Appendix B and **here** in the *tnbc_id.csv* file). However, the number of available cases for each data type was varying, as shown in Table 1.

Table 1: Availability of TNBC cases from the TCGA-BRCA project in each data type of the GDC Data Portal. Gene expr.: gene expression; Seq. Reads: sequencing reads; Masked SNV: masked single nucleotide variation; Masked CNS: masked copy number segment; DNA Meth.: DNA methylation; Protein Expr.: protein expression; Bio: biospecimen.

Gene expr.	Seq. Reads	Masked SNV	Masked CNS	DNA Meth.	Protein Expr.	Clinical	Bio
173	173	157	173	173	144	173	173

2.4.2 Differential Expression Analysis

Gene expression data of the 173 TNBC cases were analysed using RStudio and the TCGAbiolinks Bioconductor, a package dedicated to the analysis of TCGA data [72]. A total of 286 gene expression samples were selected based on their TCGA sample barcode which matched the patients barcodes mentioned above, including 173 breast invasive carcinoma samples and 113 normal tissue samples. The samples barcodes for the gene expression are available **here** in the *samples_barcodes_TNBC_gene_expr.csv* file The samples were then pre-processed by removing samples with a Spearman correlation lower than 0.6, and the mRNA transcripts were normalised to adjust for within-lane gene-specific effects and between-lane distributional differences. Indeed, RNA-sequencing read counts are known to be biased by the genes' lengths, but also by the GC-content of the target sequence and by the sequencing depth, which is the amount of sequences generated from a lane. To overcome these biases, it is crucial to normalise the transcripts within the same lane (i.e., within-lane normalisation) and between replicate lanes (i.e., between-lane normalisation) [73]. After normalisation, we considered that the expression levels were correctly processed to be subjected to further analyses. A Differential Expression Analysis (DEA) was performed between the normal samples and the tumour samples groups, respectively named "Solid Tissue Normal" and "Primary solid tumour" in the TCGA database. An FDR-adjusted p-value of less than 0.05 and an absolute logFC higher than 2 were necessary for a gene to be considered significantly differentially expressed.

2.4.3 Somatic Copy Number-Alterations

Genes affected by Somatic Copy Number-Alterations (SCNAs) were also identified. SCNAs are very common in cancers, but they also are very complex processes and should be carefully taken into consideration in multi-omics analyses. In one hand, somatic alterations can be "passenger" alterations and are not responsible for the development of cancer. In the other hand, SCNAs can be "driver" alterations, and are thought to occur at a higher frequency among cancer samples. The driver mutations are powerful tools to identify underlying mechanisms of cancers. Furthermore, one or a few genes affected by SCNAs could counterbalance the effect of SCNAs observed on thousands of other genes, the total effect of these alterations becoming neutral [74, 75].

GISTIC is an algorithm from the GenePattern platform that has been designed to tackle these two challenges. The algorithm is divided in several steps, and was revised into a better version, named GISTIC 2.0. The first step defines the segmented SCNAs profile of each cancer sample, removing the copy number variations occurring in the germline before proceeding to the next steps. As the segments can describe overlapping regions of the genome, the algorithm successively divides them into the most likely set of individual SCNAs based on their probability of occurrence. Then, it estimates the background frequency of each aberration and identifies the most significantly frequently altered genomic regions compared to the background frequency. These identified regions can get a high score because they are close to an altered gene, thus, GISTIC 2.0 decomposes the SCNAs into its different peaks and assess their significance to identify the final set of regions with significant SCNAs.

The "masked" copy number segment of 173 TNBC tumor samples from the TCGA-BRCA cohort were directly retrieved using the TCGAbiolinks package, meaning that the germline CNVs were already removed. Indeed, the GDC Masked Somatic Aggregation workflow removes all columns and variants that could be germline mutations, in order to protect the anonymity of the data. A file containing the segmented data for all the samples was created, and given as an input to the GISTIC tool. The parameters used to run the algorithm were left to their default value, except a few that are given in Table 2. The results from the GISTIC analysis were read and visualised using the Maftools R bioconductor.

Table 2: Parameters used to run GISTIC 2.0 algorithm

Reference genome	GRCh38
Amplification threshold	0.1
Deletion threshold	0.1
q-value threshold	0.05

2.4.4 Frequently Mutated Genes

Single nucleotide variations are other genetic somatic alterations known to be responsible for the development of cancers through the disruption of biological processes, giving rise to the hallmarks of cancer [2]. The emergence of large genome sequencing projects such as TCGA provides access to more and more samples, which were expected to provide knowledge on new mutational cancer-driver events. However, it seems that screening the mutation profile of an increasing number of samples leads to the discovery of a high number of false positives. Lawrence et al. (2013) suggested that this was due to making the assumptions that the average mutation frequency is constant in all genes, whereas it can be in fact highly heterogeneous, with some genes carrying an average number of mutations much higher than others. MutSigCV was developed to take into account this mutational heterogeneity and was used to find the recurrently mutated genes in TNBC [76].

A Mutation Annotation Format (MAF) file containing mutation data from the TCGA-BRCA cohort harmonized against GRCh38 was queried using TCGAbiolinks and filtered based on their TNBC barcodes. This MAF file contained mutation data of 24,612 genes among 157 different tumour samples. MutSigCV was then used to identify genes with recurrent somatic mutations in TNBC patients from the TCGA-BRCA cohort. Analysis and visualisation modules from the Maftools R bioconductor were used to prepare the MAF file for the MutSigCV analysis, and to visualise the results given by the latter. MutSigCV was run locally on the GenePattern server. The full human exome coverage file, the covariates table file, and the mutation type dictionary file provided on the platform were used as input parameters along with the MAF file previously described.

2.4.5 Differentially Methylated Genes

Methylation data are becoming widely used in omics analyses since heterogeneous levels of DNA methylation have been observed in cancers cells compared to normal tissue cells [77]. Methylation is an epigenetic event consisting in the binding of a methyl group to the carbon 5 of cytosines most often located just before a guanine, called a CpG site. In mammals, approximately 70% of the CpG sites are methylated [78]. When promoter regions of genes are methylated, the expression of these genes is usually lowered, since it inhibits the binding of proteins required for the initiation of transcription [77, 79]. Many studies have found a clear correlation between the expression of cancer genes (e.g., low expression of tumour genes, or high expression of oncogenes) and the methylation level of their promoter region [80, 81].

Using the TCGAbiolinks bioconductor, we obtained 192 methylation data samples of 173 cases including 16 Solid Tissue Normal, one Metastatic, and 175 Primary solid tumour samples. The samples barcodes are available **here** in the *samples_barcodes_TNBC_met_data.csv* file The data were obtained from the Illumina Infinium HumanMethylation450 platform. A Differentially Methylated Regions (DMR) analysis was performed using an FDR-adjusted p-value lower than 0.05 and a DNA methylation difference threshold of 0.25. The level of methylation was calculated using the beta-values, a measure ranging from 0.0 to 1.0 and roughly corresponding to the ratio of methylation events in a CpG island [82].

2.5 List of Affected Genes

In order to perform an Enrichment Analysis (EA) on the results of the multi-omics analyses, a list of affected genes needed to be constituted. Each individual omics analysis produced a list of genes that were significantly affected in the context of TNBC. Several lists of genes were compiled to run different EA and compare their results. First, a gene list was compiled from the results of each omics analysis, except for the recurrently mutated genes. Then, a gene list was compiled using the union of the genes resulting from the DEA and from the DMR (later referred to as DEA ∪ DMR). As mentioned above, it is interesting to compare the methylation level of the promoter regions of the genes, and their expression level. A gene list containing the union of the genes from the DEA, the SCNA analysis, and the DMR was built and analysed (DEA ∪ SCNA ∪ DMR). Lastly, we added the genes found to be recurrently mutated to form a list containing all the genes that were found altered in at least one type of omics. The lists can be found **here** in the *gene_lists_EA* folder.

2.6 Pathway Enrichment Analysis

A pathway Enrichment Analysis was conducted to identify the molecular and biological processes affected in TNBC. Indeed, the affected genes highlighted by the multi-omics analyses and constituting our gene lists are involved in many biological functions. The goal of the EA is to find the most over-represented pathways among these lists, which would then be considered as abnormally regulated in the context of TNBC.

The genes lists were enriched against two different pathway databases, Reactome and the Kyoto Encyclopedia of Genes and Genomes (KEGG), in order to find the biological processes affected in TNBC [83, 84]. The R bioconductors ReactomePA and clusterProfiler provide functions for pathway over-representation analysis (ORA) against Reactome and KEGG databases, respectively [85, 86].

Pathway ORA are a common type of analyses which determine whether the gene list of interest contains more genes associated to a pathway than what would be expected by chance. To assess if a pathway P is over-represented in a gene list G, these tools calculate the ratio between the number of genes in G associated with P over the total number of genes in G. Then they calculate the ratio of genes in the reference genome associated with P over the total number of genes of the genome. If the first ratio is greater than the second one, then P is over-represented in the gene list. What differs between KEGG and Reactome is the association of the genes with the pathways contained in the database: since there are no clear definition of what a pathway is, the same gene can be associated to different pathways depending on the database.

The *enrichKEGG* and *enrichPathway* functions were applied on the different gene lists. Pathways with an FDR-adjusted p-value lower than 0.05 were considered significantly over-represented. An EA was performed on each of the lists described above and the results were manually checked on the Reactome Pathway Browser or on KEGG pathway database and compared to each other, in order to select the ones that were related to cancer. As we chose to keep CASCADE 2.0 as the basis of the model, and to extend it, the pathways found in the EA were also compared to CASCADE 2.0.

2.7 Literature Mining

Simultaneously to the Enrichment Analyses, literature related to TNBC was curated, in order to gain more knowledge about the relevant pathways to add in the model. The PubMed database [87] was used to retrieve scientific papers or reviews concerning TNBC-specific pathways. First, the literature was used to confirm that CASCADE 2.0 represented pathways that are altered in TNBC. The literature was also used to support the results from the EA: if a cancer-related pathway found to be over-represented in the EA was also known to be abnormally regulated in TNBC, it would very likely be selected for modelling. Finally, the literature was used to discover new pathways that were not necessarily over-represented in the altered gene lists, and that were not present in CASCADE 2.0. Pathways that were often mentioned to be abnormally functioning in the context of TNBC were automatically selected for the modelling part.

2.8 Gene Regulatory Network Building - The TNBC Model

The gene regulatory network was built on GINsim, using CASCADE 2.0 as a base. GINsim 3.0 (Gene Interaction Network simulation) is a software used to model and perform static and dynamic analyses on gene regulatory networks. The user can add new nodes and directed edges to a network, annotate it, and define logical rules for each node of the network. Many network analyses can be performed directly from GINsim, such as the computation of the state transition graph or a stable states analysis [88]. CASCADE 2.0 was imported to GINsim, and the model was extended from there. New pathways were added only if they were:

- Over-represented in the lists of altered genes AND documented as important/altered pathways in the context of TNBC
- Strongly documented as important/altered pathways in the context of TNBC

The nodes and interactions composing the new modules were curated using knowledge and pathway databases and the scientific literature. Reactome, KEGG, and SIGNOR were mainly used to retrieve information about the nodes composing the modules [84, 85]. The genes, proteins, and complexes mentioned in the literature as members of the newly added pathways were added to the network only if they respected one of these conditions:

- There was strong literature evidence mentioning the nodes as important/up-regulated/down-regulated/mutated in TNBC
- The nodes were highlighted in the pathway EA and at least in two papers related to TNBC
- The nodes were annotated in one of the pathway databases used, and there was strong evidence showing that they are involved in TNBC

The reader should be aware that the genes, proteins, and complexes composing the TNBC model will sometimes wrongly be called ”genes” for simplification.

SIGNOR 2.0, standing for SIGnaling Network Open Resource, is a manually-annotated database gathering information about the logical interactions between biological entities, in the form of positive or negative, and directed edges. SIGNOR also offers manually curated pathway graphs. The nature of each interaction is annotated with the mechanism of action of the regulator entity on the target entity, and with the effect of this mechanism on the target, if these information were given in the scientific evidence that they used to curate it. A confidence score is attributed to each reported interaction and is calculated based on four features. First, the number of publications in which they report this interaction. Second, the occurrence of this interaction in the SIGNOR annotated pathways. Third, the occurrence of this interaction in a Reactome file containing the protein-protein interactions for humans. And finally, the occurrence of the target entity in the UniProtKB webpage of the regulator entity. This score ranges between 0 and 1, and the closer it is to 1, the more evidences there are about this interactions [89]. SIGNOR is the main knowledge database that was used to curate the causal interactions between the added nodes.

Some general rules were used to consider that an edge should be added to the module:

- The edge is referenced on SIGNOR with a score above 0.5 AND is not reported in a specific cell type or disease (except for TNBC)
- The edge is referenced on SIGNOR AND clearly mentioned in at least two papers
- If the edge is not referenced on SIGNOR, the edge should be strongly supported by the literature AND be important in the dynamic of TNBC

- The edge should be relevant for the dynamic of the network (e.g. it should not be redundant with another existing edge)
- The edge should not be reported both as positive and negative on SIGNOR. Such edges were ignored

No logical rules were defined at this step for the added nodes, since they should be defined later in the project. The network was later visualised with Cytoscape, a software developed to enable the visualisation of biological networks, integrate a wide range of information from large interaction databases, and compute biological and mathematical features of the network [90]. Cytoscape was used to display the network into specific layouts, and to analyse the network's properties. The nodes were clustered by modules, using the Group Attributes Layout. Cytoscape offers an analysis tool named Network Analyzer, providing several network measures, especially relevant for the comparison and the characterisation of networks. This plugin allows the user to understand what type of network they are using, its general organization and structure, but also the properties inherent to each node of the network. The degree and betweenness centrality of each node were computer using the NetworkAnalyzer, as well as the average degree and the clustering coefficient of the network.

The degree of a node is the number of nearest neighbors it has, or in other words, the number of nodes it is interacting with. In oriented networks, we distinguish incoming edges and out-going edges of a node, which means that we can calculate an "Indegree" and an "Outdegree". By definition, the average degree of a network is the sum of the degrees of its nodes divided by the number of nodes it contains.

The betweenness centrality of node A denotes the number of shortest path going through node A, the shortest path being the minimum number of nodes one has to pass through to go from a node B to node C. A node with a high betweenness centrality is called "a bridge", refering to the fact that if we remove this node, a lot of paths are disrupted and longest paths will be used to go from one node to another.

Lastly, the clustering coefficient of node A is proportional to the number of interactions between the nearest neighbors of A. It reveals if A and its neighbors form a cluster or not. The clustering coefficient of a network ranges between 0 and 1, and the closer it is to 1, the better the nodes are forming clusters [47].

These measures were computed on the PKN considered as a directed network, and they were compared to the measures related to CASCADE 2.0.

2.9 Generation of Boolean Models and Drug Synergy Predictions

The gene regulatory network was further studied using the Druglogics pipeline [91]. This pipeline is composed of two modules, Gitsbe and Drabme.

Gitsbe is a module aiming to transform a PKN into a Boolean network, by generating logical rules that fit an expected behaviour. Indeed, the PKN is a model representing only the causal relationship between two genes/proteins/complexes. It does not include the fact that the activity of the nodes are inter-dependent to several other nodes, which can be translated through Boolean rules.

Gitsbe necessitates several input files to function. A SIF file containing the network's topology will be used as a base: this file is the one that one wants to transform into a Boolean network. At this point, it only contains simple activation or inhibitory edges. The model outputs file is also necessary to specify to the algorithm which nodes' states should be considered to calculate the resulting phenotype of the model. In this project, we considered that the dynamic of the model could either be "pro-apoptosis", or "anti-apoptosis", and the output nodes were genes that were leading either to apoptosis or to cell survival. Thus, the steady state of these nodes were used to calculate a global output response. Lastly, a configuration file specifying the options used to run the pipeline needs to be given as input. Some of these options are, for example, the number of simulations, generations, and the population per generation, or the tool used to compute the attractors of the models. They will be explained in further details later.

Gitsbe functions as a genetic algorithm, which means that it reproduces the principles of natural evolution: the best models of a generation will be selected to create the models of the next one. During the first generation, Gitsbe first creates several basic Boolean models where each node's state is ruled by a default equation (the number of models generated is specified by the "population" parameter):

$$A = (B \lor C) \land \neg(D \lor E) \tag{1}$$

Where B and C activate A, and D and E inhibit it. In other words, at least one of A's activators and none of its inhibitors should be active for A to be active. The algorithm modifies these rules with a certain number of mutations specified by the user, in order to find the best model's dynamic to fit "training" data. These training data correspond to the expected behaviour of the model under certain conditions. Under unperturbed conditions, one can either calibrate the steady states of some (or all) nodes of the network, or they can leave it unspecified and ask for a global output response between 0 and 1.
If one chooses to calibrate the state of some nodes, the algorithm will try to produce a model that has the closest steady states from these calibration data, which is quantified by the fitness score. In our case, the use of a global output response would be relevant to simulate the death of cancer cells (global output set to 0), or the proliferation of cancer cells (global output set to 1).
It is also possible to perturb the network's dynamic by forcing the state of some nodes (for example forcing it to 0 for a gene knock-out or to 1 for a knock-in).
After having calculated the fitness score of each model of the first generation, the algorithm selects the best ones and exchange their logical rules to generate models for the next generation. This is iteratively repeated until a fitness threshold or the maximum number of generations is reached, and called a "simulation". In the end, a given number of the best models of the last generation are saved.
The user can choose the number N of simulations to run, and the whole process described above is repeated for N simulations.
The resulting files from Gitsbe are all the best models (i.e. the ones with the best fitness score) from all the simulations.

After having produced several boolean models with Gitsbe, Drabme can be used to predict drug synergies on this set of "best" models. Drabme is the second module of the pipeline. It is made to test the effect of drug combinations on the dynamic of the models. It takes as input files the best models produced by Gitsbe (in .gitsbe format), along with a file containing all the drugs to be tested and their targets. Based on this file, the algorithm will be able to modify the state of the drug targets to 0 or 1 according to the action of the drugs, and assess their effects. In the configuration file mentioned above, the synergy score used to assess the effect of the drug combinations can be set either to "hsa" or to "bliss" depending on which synergy metrics is used. With these two metrics, a drug combination is considered synergistic when the viability of the model is lowered compared to the viability of the model when the effects of the separate drugs are added. The viability of the model is calculated using the global output response of the perturbed model, indeed the model outputs file mentioned above will be required. The combination size (i.e. the number of drugs combined) also needs to be specified in the configuration file.
In this project, only two-drugs combinations were tested. To calculate the synergy score of a combination of drug A and drug B, Drabme first perturbs the model using drug A (i.e. by forcing the state of the drug A's target(s) to 0 or 1 accordingly to the action of drug A) and calculates the global output response of all the best models provided by Gitsbe under perturbation A. It then perturbs the models with drug B, and calculates the global outputs. Finally, it perturbs the model with both drugs A and B (i.e. by forcing both the state of drug A's and drug B's targets to 0 or 1 accordingly to the action of drug A and B). Depending on the score metrics used, the synergy score from the combination of drug A and B is calculated based on the global output responses of the models under these three perturbations. All the files necessary to reproduce the simulations are available **here**.

2.9.1 Training Data Retrieval

Four well-studied TNBC cell-lines data were used to train the algorithm. These data were RNA-sequencing data obtained from the Cancer Cell-Line Encyclopedia (CCLE). The expression profile of 41,717 gene probes was found for the following cell-lines: MDA-MB-231, MDA-MB-453, HS-578T, and BT-549. Following the PAM50 subtypes, MDA-MB-231, HS-578T, and BT549 belong to the basal-like subtype, meaning that they exert a basal phenotype, while MDA-MB-453 belong to the luminal-B subtype [8, 92]. In the TNBC subtypes defined by Lehmann et al. (2011), MDA-MB-231 and HS-578T are together classified as mesenchymal-stem-like whereas BT549 is classified as mesenchymal, and MDA-MB-453 as luminal-androgen-receptor. Differences in their RNA expression profiles are expected considering the heterogeneity of TNBC.

We only kept the expression of genes that were present in the network as training data, and averaged each gene expression among their probes, which ended in four lists of expression data of 184 genes. By cell-line, their expression values were then scaled from 0 to 1 using the equation:

$$x_{i;scaled} = \frac{x_{max} - x_i}{x_{max} - x_{min}} \tag{2}$$

Where $x_{i;scaled}$ is the scaled expression of gene x_i, and x_{max} and x_{min} are respectively the maximum and the minimum expression of all the genes in the cell line.

In a first step, all the nodes for which we had an expression value were calibrated on these scaled expression data (184 nodes). The training data of the 184 nodes for the four CCLE cell lines are available **here**: *training_files → training_CCLE_full_name.tab.*

Another set of expression data was used to investigate the influence of the training file on the results of the pipeline. These data were RNA-sequencing data from the TNBC cases of the TCGA-BRCA project. TCGAbiolinks was used to perform hierarchical clustering of these samples with the RNA expression of the four cell lines previously used, based on the Euclidean distance of the expression profile of 184 genes of our model. The aim of this step was to select the TCGA samples that clustered with the CCLE cell lines, use these samples as training data, and predict drug combinations that were observed synergistic for the CCLE cell lines.
Separately, the TCGA and CCLE data were normalised using counts per million, they were then scaled between 0 and 1, and finally, in order to get closer to Boolean states and investigating if the fitting score of the models could be improved, the values strictly higher than 0.5 were set to 1, and the values equal or below 0.5 were set to 0. The expression values of the genes were averaged across the samples of a same cluster, and then scaled between 0 and 1. Then, the variance of each gene across all the samples was computed, and we clustered the samples based on the discretised expression of the top 1,000 variable genes. In this case as well, 184 genes of the model were calibrated using the binary TCGA expression data, which are available **here**: *training_files → training_TCGA_full.tab.*

Previous observations on several models developed by the Druglogics group, including CASCADE 3.0 [61], showed that calibrating the model with a smaller number of nodes improves the predictive performance. Also, there is ongoing work from the group on optimizing the calibration parameters, and especially on the use of a metric, the Determinative Power (DP) of the nodes, to assess the topological importance of the nodes of a network. The DP has been shown very promising to improve the prediction performance of some models from the group, and we decided to investigate whether it could have the same effect on the TNBC model.
The DP is a metric based on the topology of Boolean networks, and does not account for the state of the nodes, but for the quantity of information flowing through the nodes. The DP can be calculated for each node, based on the information provided by its regulators, and it unveils the quantity of information that can be deduced from its state regarding the other nodes of the network. Then, this is a static measure giving information on the dynamic of the network [56, 93]. The DP of the nodes of the network was calculated using the script provided by Weidner et al. (2021) [51], and using the topology of the first model generated by Gitsbe. This model uses the equation 1 for all the nodes.

Finally, rather than using calibration data to train the models, it is possible to train them on what is called a random proliferative training data profile. This situation corresponds to an unperturbed model (we are not calibrating the state of any node), left in a proliferative dynamic.

The global output is expected to be equal to 1, which translates the fact that if we leave cancer cells unperturbed, they keep proliferate. Following the workflow shown in the **tutorial** relative to the DrugLogics pipeline [91], the prediction results from the random models are then used to normalise the calibrated models predictions. It is expected that the normalization of the calibrated predictions with the random proliferative predictions increases the prediction performances in terms of ROC-AUC.

2.9.2 Generation of Logical Equations with Gitsbe

In this project, some parameters remained unchanged and the most important ones are summarized in Table 3. Other parameters such as the number of simulations and the type and number of mutations were modified and their influence will be discussed in the Results and Discussion part (3).

Table 3: Parameters used to run Gitsbe simulations

Attractor tool	Nr of generations	Population	Nr of best models saved
BioLQM	20	20	3

The configuration file allows the user to run the simulation on the raw PKN, but also on a network where the "input" or "output" nodes were removed. The input nodes are the ones which do not have incoming edges, and the output nodes do not have outgoing edges. Input nodes do not have a defined logical function since they have no regulators and represent external signals [52]. We removed them from the PKN to run the simulations.

The model outputs file used in this project was always the same. CASP8, CASP9, and FOXO_f were the nodes contributing to the "pro-apoptosis" phenotype, thus they all had a weight of -1, whereas RSK_f, CCND1, and MYC were the nodes contributing to "anti-apoptosis" with a positive weight of one each (see Appendix C). The state of these nodes were therefore used to calculate the global output response of the model and assess, for example, the effect of a perturbation.

2.9.3 Synergy Data Retrieval and Drug Targets

Synergy data were used to generate a list of drugs to test two-drugs combinations on the generated models. These synergies were obtained from several drug synergy repositories depending on the availability of such data on the four used cell lines. As explained above, Drabme can calculate both an HSA synergy score, or a BLISS synergy score, based on the global response of the perturbed model. Indeed, both HSA scores and BLISS scores were retrieved from SynergxDB, DrugComb, and DrugCombDB, when they were available [94, 95, 96]. These databases contain drug combination information on different datasets, therefore, we needed to combine them to find drug synergies for all of our cell lines.

To create the drug panel file for Drabme predictions for each of the cell line, we needed to specify the drugs and targets that we wanted to test. One file per cell line needed to be created, since the combinations found in the synergy databases were specifically tested on one cell line. The tested drugs to include in this file should act on nodes of the model, indeed, some of the synergies found could possibly be excluded of the analysis if it was not targeting our nodes. To create this file, it was necessary to know the targets of each drug involved in the synergistic combinations tested on the four cell lines.
DrugBank is a database gathering extensive information about drugs, their molecular properties (structure, weight, formula), their targets, and the nature and mechanism of their action [97]. These information are available for drugs approved by the Food and Drugs Administration (FDA-approved), and for drugs that are in the process of approval as well. The complete database was downloaded and searched using R and the drugbankR Bioconductor. The tested drugs were queried along with their targets and their action (inhibition or activation), and only the drugs acting on nodes of our network were selected for synergy predictions. The drug panel files were then created

and contain the drugs acting synergistically for their respective cell line and targeting nodes of the models. They are displayed in Appendix D.

To predict drug synergies when using TCGA-calibration RNA expression data as mentioned in Section 2.9.1, we needed to adjust the drug panel. Indeed, the drug panels contained drugs that were observed synergistic for the four CCLE cell lines. The observed synergies differ from one CCLE cell line to another, indeed we could not assume that they would be observed in TCGA samples. The choice of the composition of the drug panel used in this case will be detailed in the Results and Discussion part 3. All the files necessary to reproduce the simulations are available **here**.

2.9.4 Drug Synergy Analysis with Drabme

Drabme produces several output files, including a so-called "Ensemble-wise synergies" file which provides the average synergy score of each drug combination tested, across all the best models provided by Gitsbe (or in a .gitsbe format). This score is calculated either as an HSA score or a BLISS score. Other files are also produced by Drabme, but were not used for further analyses and will not be described here.
To assess the performance of the predictions, we calculated the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). This metric is used to assess the performance of a binary test (here, Drabme classifies the combinations as synergistic or not), by comparing its True Positive Rate (TPR) and its False Positive Rate (FPR) at different values. The TPR, also called sensitivity, can be calculated with the following equation:

$$TPR = \frac{TP}{TP + FN} \tag{3}$$

Where TP stands for True Positive, and FN for False Negative. The sensitivity highlights how accurate is the model for predicting true synergies (it is the number of true synergies that were correctly predicted divided by the total number of synergies observed). On the other hand, the FPR can be calculated as follows:

$$FPR = \frac{FP}{FP + TN} \tag{4}$$

Where FP stands for False Positive, and TN for True Negative. This is the number of wrongly predicted synergies divided by the total number of non-synergistic combinations. The ROC curve is obtained when plotting the TPR on the y-axis, and the FPR on the x-axis, and will give information on the usefulness of the prediction: if the curve is of equation y=x, then the prediction is random, and does not predict more TP than FP. If the prediction is efficient and Drabme is able to predict more TP than FP, then the AUC of the ROC curve will be higher than 0.5.

In this project, the AUC of the ROC curves of all the predictions were traced and compared to each other, in order to find the best performing cell lines or parameters used to generate and calibrate the models. The code used to trace the ROC plots is available **here** in the $Drabme\_performances.R$ file.

3 Results and Discussion

3.1 Multi-Omics Data Analyses

As highlighted by multiple studies, the integration of multi-omics in the study of complex diseases such as TNBC is necessary to understand the interplay of different types of molecules in the regulation of the cancer processes, and be able to represent it at the systems biology level [80, 48]. In this study, we integrated transcriptomics, epigenomics, and genomics data from TNBC patients, and we attempted to link the results provided by the analyses to general biological processes (or pathways), in order to draw the molecular story of this disease into a biological network.
Among the 271 different TNBC patients included in the TCGA-BRCA study, 270 were females and only one was male, 228 were below 70 years old, 196 were white, 57 were black or african-american, ten were asian, and seven were of unknown ethnicity [62].

3.1.1 Differentially Expressed Genes

The first omics analysis performed was a Differential Expression Analysis (DEA) between solid tissue tumour samples and normal tissue samples. All the gene expression samples from the TCGA-BRCA project were first queried and prepared into an R object, and were then explored, pre-processed, and analysed using functions of the TCGAbiolinks package. The pre-processing step was performed using the *TCGAanalyze-Preprocessing* function which plots the Array-Array Intensity Correlation (AAIC) matrix along with a boxplot of correlation samples by samples (fig.2). A Spearman correlation threshold of 0.6 was applied, but no sample with a lower correlation than 0.6 was detected, which means that no outliers needed to be removed from the data. This step involved the 1222 BRCA samples, since the filtering based on the TCGA barcodes was done at a later stage.

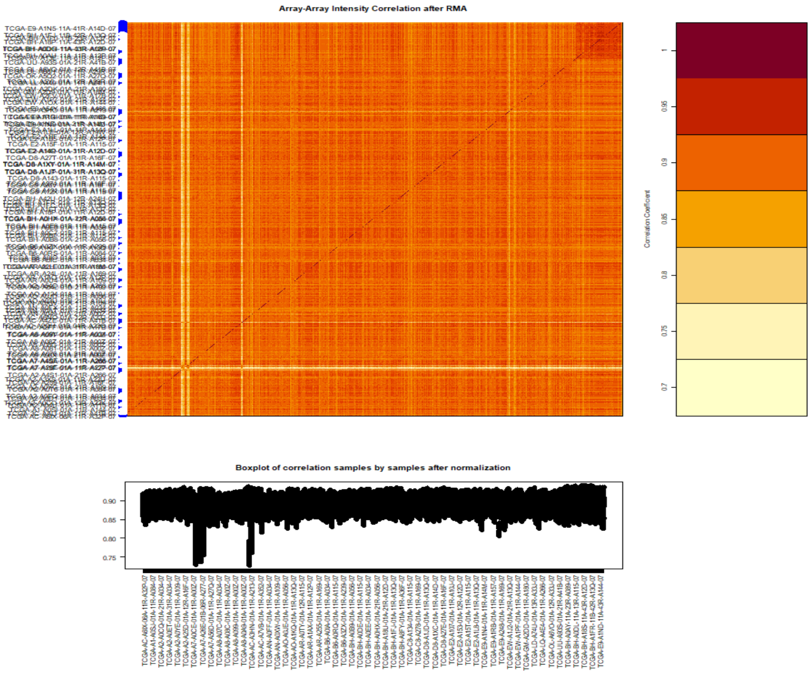


Figure 2: Array-Array Intensity Correlation (AAIC) matrix and boxplot of correlation samples by samples of the TCGA-BRCA gene expression samples.

Before transforming further these BRCA samples, we filtered them based on their barcode. A list of TNBC patients' barcodes from the TCGA-BRCA study was provided in Jiang et al. (2016), thus we only selected these patients [71]. To ensure the quality of the analysis, the data were normalised using within-lane normalisation and between-lane normalisation, as mentioned in the Methods. The result of the normalisation is displayed as boxplots in Figure 3.

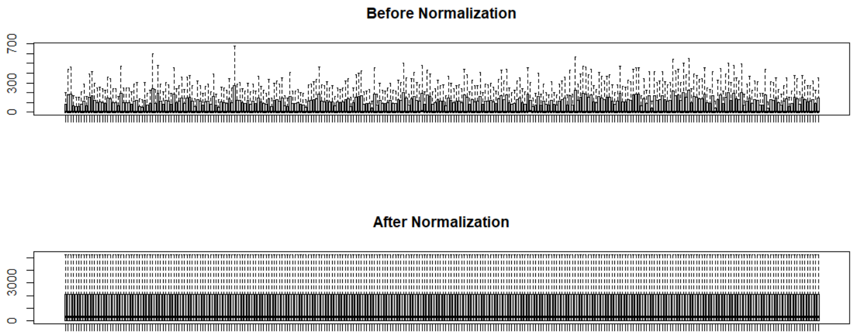


Figure 3: Boxplot of the mRNA transcripts read counts before and after within-lane and between-lane normalisation. These boxplots show the dispersion of the reads for each sample before and after normalisation. The box represents the interquartile range, meaning that half of the data are comprised in this range (i.e. from the first quartile to the third). Outside this box, the dotted line represent the lowest and highest value for each sample.

The DEA was performed on 17,394 genes among the 286 tumour and normal samples. The list of the samples' barcodes can be found on the **Github** in the *samples_barcodes_TNBC_gene_expr.csv* file. Using an absolute logFC lower than 2 and a FDR threshold of 0.05, we identified 2,262 DEGs, which were represented in a volcano plot (fig.4). We found 1,520 over-expressed genes (i.e., with a logFC≥ -2) and 753 under-expressed genes (i.e., with a logFC≤ -2), respectively. The list of these genes along with their logFC and FDR value can be found in the **Github** in the *DEGs_with_logFC_value.csv* file.

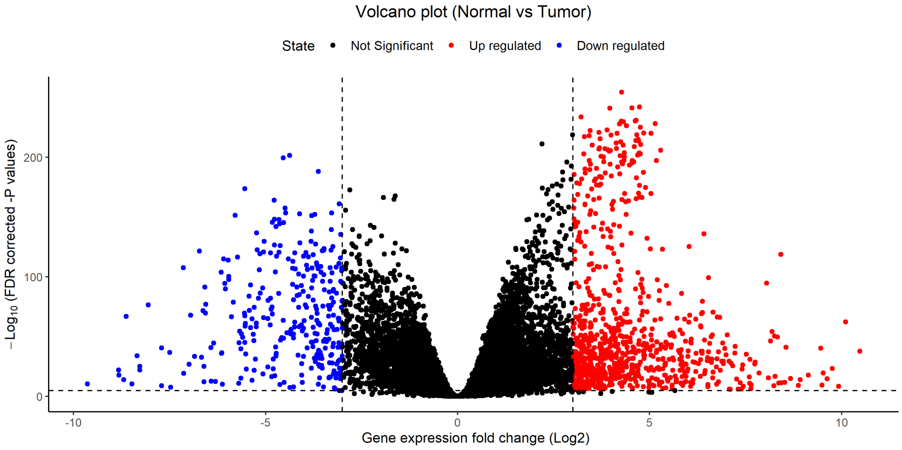


Figure 4: Volcano plot of the Differentially Expressed Genes (DEGs). The dashed lines represent the FDR threshold of 0.05 and the logFC threshold of 2. The blue and the red dots represent the down-regulated genes (logFC≤ -2) and the up-regulated (logFC≥ 2) genes, while the black plots represent the non-significantly differentially expressed genes.

To verify our findings, we compared the 2,262 DEGs from our analysis to the findings of Wang et Guda (2016). They analysed the gene expression profile of 55 TNBC samples from TCGA microarrays and found 1,800 significantly over-expressed genes compared to the paired normal samples using the same logFC and FDR thresholds as we did [80]. Among the 2,262 DEGs that we found, 1,520 were over-expressed genes (i.e., with a logFC$\leq$0), and 196 were common to the list of over-expressed genes of Wang et Guda (2016). This represents an overlap of approximately 13%, which is low, but not surprising since the two analyses were run on a different number of samples, and that most comparisons of gene expressions obtained through RNA sequencing and microarrays are usually done on subsets of the same size [98].

The total number of DEGs that we found is quite large and could not be analysed by checking all the genes one by one. To reduce this number, we could have chosen more stringent thresholds for the significance (FDR value) and the differential measure (logFC). However, the next steps of the project involved enrichment analysis of these genes, which can be performed on a large gene list, thus it was not necessary to be too stringent. Furthermore the logFC and FDR thresholds were chosen based on the value that we found in most scientific studies that we encountered involving DEA. Furthermore, Wang et Guda (2016) classified the genes resulting from their omics analyses into levels of hyperactivation, the first level being the DEGs, and the following levels being the genes belonging to level one and being hyperactivated in another omics type (i.e a gene with high copy number, low methylation, or target of miRNA found with low expression in TNBC). This is translating the fact that they based the hyperactivated status of a gene on its RNA expression before considering further omics types [80]. Thus, based on the proven relevance of gene expression data integration into many multi-omics-based models, we decided to keep all those DEGs [99] .

In our results, nine micro RNAs (miRNAs) were found differentially expressed (three were under-expressed, six were over-expressed), including three that were earlier reported in the literature: miR-155, miR-497, and miR-497 [100, 101]. miRNAs are small non-coding RNAs which target specific mRNA and control their expression in the genome of eukaryotic organisms [1]. Except the fact that we included the miRNAs in the next step of our multi-omics analysis, no specific attention has been brought to it. However, specific miRNA expression patterns have been highlighted in different types of cancer, and it is now clear that some of them function as oncogenes or tumour suppressor through their regulatory activity, and could be interesting prognostic factors or even treatment targets [102].

3.1.2 Somatic Copy-Number Alterations

As a second omics analysis, we searched for significantly amplified or deleted regions of the genome, referred to as SCNAs. Copy Nunmber Variation (CNV) profiles of 174 tumour samples from TCGA-BRCA were available and retrieved with the TCGAbiolinks bioconductor. GISTIC provided several results files. The first file of interest contains all the data about the aberrant regions, and the name of the samples that are significantly amplified or deleted in the corresponding regions. Two other important files contain the amplified and the deletion peaks, respectively with the genes contained in these regions. Lastly, it provided a file with the scores of the identified amplifications and deletions. FDR-corrected p-value and information about the amplitude of the peaks and the frequency of aberrations are provided in this file. Functions of Maftools were used to read the files of interest and to display the significantly altered cytobands as a function of the number of samples in which this region is altered, and the number of genes it contains (Fig.5). A chromplot of the altered regions is represented in Appendix E.

1,062 genes were found located in significantly amplified or deleted regions and were considered for further analysis. The full results are available on the Github **repository** in the *GISTIC_results* folder.

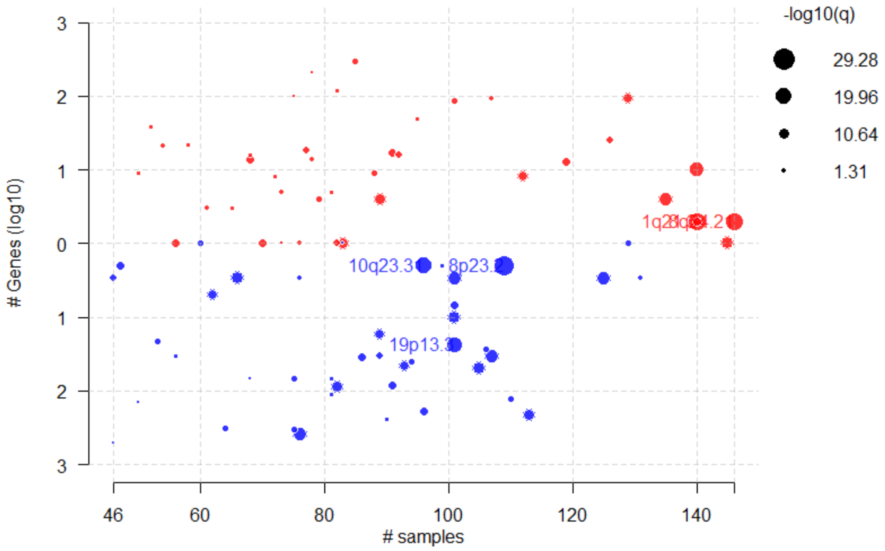


Figure 5: Bubble plot of the significantly amplified or deleted chromosomal regions (cytobands) in the TCGA-BRCA TNBC samples. The red and the blue dots represent the amplified and deleted regions, respectively, and their size is inversely proportional to the FDR. The five named dots represent the five locus which had the most significant alteration in terms of q-value.

Among these 1,062 genes, RB1 and PTEN were found in amplified regions, and EGFR was found in a significantly deleted region, confirming the results observed by Shah et al. (2012) [18].
Figure 5 and Appendix E highlight the same five loci, respectively for their high p-value, and for their high G-score, a score calculated by Gistic and based on the amplification or deletion peak and the frequency of alterations across the samples. These 5 peaks include two loci that were identified significantly amplified, and three others that were found significantly deleted. All of these regions were previously identified as altered in other studies concerning breast cancer. 1q21.3 was found significantly amplified in 10 to 30% of breast primary tumours, and in more than 70% of recurrent cases. Indeed, 1q21.3 would be amplified in tumour infiltrating cells (TICs), which are characteristic of tumour heterogeneity and enhance resistance to treatments and relapse. 1q21.3 amplification was shown responsible for the enhanced phosphorylation of IRAK1, a protein involved in tumour recurrence [103]. 8q24.21, another locus significantly amplified in many samples, is known to contain mostly long non-coding RNAs (lnRNAs) and miRNAs, and only a few protein-coding genes. Some lnRNAs contained in this locus (PCAT1, CCAT1, CCAT2) are known to be involved in breast cancer progression which suggests a correlation between the amplification of this locus and their abnormal activity in this type of cancer [104]. The most important protein-coding gene it contains is the oncogene MYC, which is thought to be a key target gene in breast cancer, and especially in TNBC because of its over-expression in many cases [105, 106]. We found that 10q23.3 was significantly deleted across the tumour samples. This locus codes for PTEN, a tumour suppressor known to be mutated in many cancers. Deletion of this locus has been observed in several cancer types, and loss of heterozygosity of the 10p23 region was observed in 40% of invasive carcinomas, suggesting that it plays a role in tumour progression [107, 108]. Similarly, the significant deletion of the 8p23.2 locus in the TNBC samples agree with the literature, since this locus was found deleted in 55% of primary breast tumours. This locus particularly codes for the gene CSMD1, a known tumour suppressor [109]. Lastly, the 19p13.3 locus was found significantly deleted as well, agreeing with previous results where this locus was found with allele loss in breast cancers, and especially in hereditary breast cancers. It is coding for genes that are thought to be tumor suppressor [110, 111]. These results agree with the scientific literature and confirm that the analysis of SCNAs gives important insights into the mutational profile of TNBC.

3.1.3 Frequently Mutated Genes

The third omics analysis conducted in this study highlighted the most frequently mutated genes in the TNBC samples. 157 TNBC samples with Simple Nucleotide Variation (SNV) data were retrieved from the TCGA-BRCA repository. They were all female samples. A MAF file was prepared using Maftools, and used for further analysis on MutSigCV. The algorithm of MutSigCV takes into account the mutational heterogeneity in cancers and returns the most frequently mutated genes compared to the gene-specific background mutation rate. Surprisingly, only three genes were identified as significantly mutated across all the patients' samples, namely TP53, PIK3CA, and PTEN, with a FDR-corrected p-value lower than 0.05. They respectively had a mutation frequency among all 157 samples of 87.3%, 15.9%, and 7.6%. **repository** in the *MutSigCV_results* folder. The previous study using the CASCADE 3.0 colorectal cancer model have found between 6 and 55 significant results using MutSigCV [61], however, using the same tool, Jiang et al. (2016) only found TP53 as recurrent somatic mutation in TNBC [71]. Furthermore, it was already known that TP53, PIK3CA, and PTEN were frequently mutated in TNBC [18, 112]. Shah et al. (2012) additionally reported USH2A and MYO3D among the most frequently mutated genes, with an even higher frequency than PTEN. Divergence in the methods for the detection of frequently mutated genes could explain these differences.

3.1.4 Differentially Methylated Regions

The last omics data analysis performed was a Differentially Methylated Regions (DMR) analysis, which highlighted the most significantly hypo-methylated or hyper-methylated regions of the genome of the TNBC patients compared to the group of normal samples. Using the function *TCGAanalyze_DMC*, we identified 255 significantly differentially methylated CpG islands, including 100 hyper-methylated and 155 hypo-methylated regions. The probes' names were converted to Gene Symbols corresponding to the gene encoded in these regions, resulting in a total of 88 hyper-methylated genes and 115 hypo-methylated genes. The detailed results of this analysis are available in the Github **repository** in the *DMR_results.csv* file. These regions are represented as a volcano plot in Figure 6.

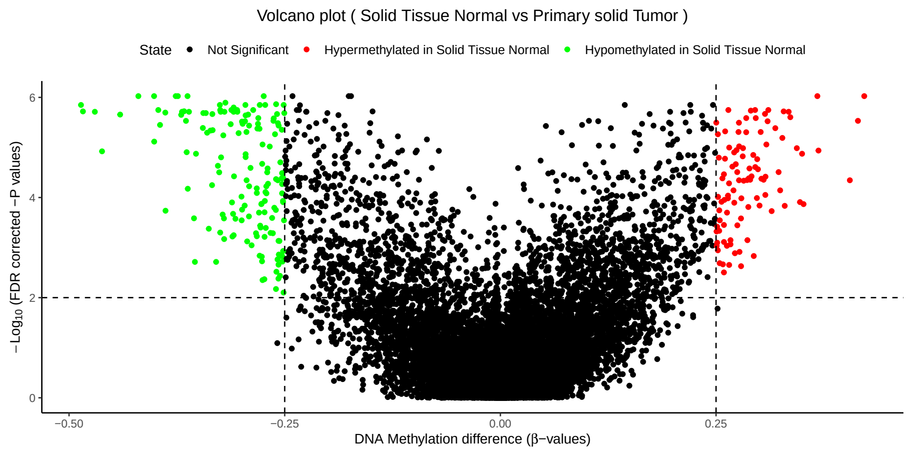


Figure 6: Volcano plot of the Differentially Methylated Regions (DMR). The dashed lines represent the threshold of beta-value and the logarithmic value of the FDR-corrected p-value threshold used to identify the DMR.

Basal-like breast cancers, which are a majority of TNBC, are known to have low levels of methylation, compared to the other breast cancer subtypes [7, 113]. However, we would have expected an hypermethylation of BRCA1 promoter region since it has been observed in several TNBC cases, leading to an epigenetic silencing of the BRCA1 gene [71]. It was not found significantly differentially methylated in our study.

3.2 Composition of the Gene Lists Used for the Enrichment Analysis

The omics analyses previously described gave rise to lists of aberrant genes in TNBC. The number of genes resulting from each of these analyses is summarized in Table 4. Different gene lists were used to perform pathway Enrichment Analyses (EA). As described in 2.5, we performed a separate EA on each of the gene lists made of the results from the DEA, the SCNA analysis, and the DMR analysis. They respectively contained 2260, 2062, and 203 gene names. The list of the DEGs and the differentially methylated genes contained 2426 genes, and 37 genes were found in common between the two separate lists. The list composed of the genes from the DEA, the DMR analysis, and the SCNA analysis contained 4304 distinct genes, and 184 were found in common between each separate lists. The last list, which additionally included the RMGs contained 4307 genes. The different lists of genes used for the EA can be found in **Github** in the *gene_lists_EA_* folder.

Table 4: Number of genes identified in each of the omics analysis performed. These genes were later used for pathway enrichment.

Omics Analysis	Number of Genes
Differentially Expressed Genes	2260
Amplified or Deleted Genes	2062
Differentially Methylated Genes	203
Recurrently Mutated Genes	3

As it was done in Wang et Guda (2005), we could have built gene lists based on the number of omics analyses in which the genes were highlighted [114]. For example, we could have selected the genes that were differentially expressed and altered in at least one of the other types of omics. However, doing so would have restricted the list to 210 genes, and the limited mapping of the entities by Reactome and KEGG databases would have restricted this number even more. Furthermore, Tsirvouli et al. (2020) used an approach where all the genes that were found in either of the omics analyses were used for enrichment against a pathway database, so we decided to apply the same method in this project [61].

3.3 Identification of Over-Represented Pathways

The Enrichment Analysis performed on the gene lists described above provided a large number of results with an FDR≤0.05 which were thus considered significantly over-represented and needed to be filtered. Indeed, Reactome provided much more results than KEGG in most of the analyses, as shown in Table 5. These results needed to be manually checked in a pre-selection step, in order to remove those which were not related to cancer. The final selection of the pathways composing the model was done by combining the knowledge resulting both from the EA and from the literature mining. These two processes were done simultaneously, and the reasons why some pathways found in the EA were discarded or selected will mostly be given in the literature mining part (3.4).

Table 5: Number of significantly over-represented pathways identified in each of the gene lists (FDR≤0.05).

Gene list	Reactome	KEGG
DEA	173	29
SCNA	6	32
DMR	2	1
DEA ∪ DMR	165	30
DEA ∪ SCNA ∪ DMR	189	52
DEA ∪ SCNA ∪ DMR ∪ RMG	207	51

The enrichment analysis of the differentially methylated genes did not give a lot of significant results (FDR≤0.05) compared to the other gene lists (”Neuroactive ligand-receptor interaction” against KEGG, ”Class A/1 (Rhodopsin-like receptors)” and ”Peptide ligand-binding receptors” against Reactome).

This is a surprising result since several studies agree on the fact that TNBC cases show extensive hypo-methylation, which is directly linked to the high activity of some genes involved in the immune system and revealing the presence of tumour infiltrating lymphocytes for example [7, 115]. Another study showed that differentially methylated genes in TNBC that were negatively correlated to the expression of differentially expressed genes were enriched in several important cancer pathways such as angiogenesis, extracellular matrix organization, inflammatory response and cell adhesion. However, the identification of the DMR was done using different statistical methods than us, and they also used another pathway EA tool [116].
These results suggest that there might be important differences in the way differentially methylated genes are identified, which would lead to important variations in the results of following analyses.

3.3.1 Results of the Enrichment Analysis against KEGG

KEGG provided a reasonable amount of results compared to Reactome, but a pre-selection step was still required to remove the pathways that were not related to cancer. As we wanted to build a model constrained in size, we also needed to identify the pathways which contained the most important genes. Pathways such as ”Alcoholism”, or ”Systemic lupus erythematosus” had low q-value but were considered irrelevant for the modelling part as they were not signalling processes involved in cancer. Most of the pathways present in CASCADE 2.0 were over-represented in at least one of the gene lists (data not shown).
Table 6 only displays the pathways that were found enriched in our analysis and that were selected during the literature curation as new modules. The extensive results are not shown for practical reasons, but Appendix F shows plots of the 30 over-represented pathways with the lowest q-value for each EA against KEGG. The DEGs were enriched in genes involved in cAMP signalling and in cell cycle, two pathways already present in CASCADE 2.0.
The list resulting from the SCNA was enriched in MAPK cascade, p53 signalling, and PI3K/Akt signalling genes, three pathways present in CASCADE 2.0. The VEGFR and the cPLA2 signalling cascades were also shown to be over-represented in this gene set.
The EA indicated that the differentially methylated genes were enriched in ”neuroactive ligand-binding interaction” genes, which was not selected for the modelling part since it included many receptors and their ligands. These individual receptors are not interacting with each other, and the pathway itself is not directly known as a main actor of cancer, even though it is not excluded that it could be involved in some cancers. However, for practical reasons, it was not selected for the following steps of the modelling. To see if more results could be obtained from the DMR, an EA was performed on both DEA and DMR results. It was enriched in cAMP signalling and in cell cycle, two pathways that were also over-represented in the DEGs gene set.
We then combined the genes resulting from the DEA, the DMR analysis, and the SCNA analysis, and performed an EA on the 4,307 genes of this list. This was repeated on the same list to which we added the three recurrently mutated genes found by the MutSigCV algorithm, TP53, PTEN, and PIK3CA. The results were the same: cAMP signalling, cell cycle, MAPK cascade, p53 and PI3K/Akt signalling were over-represented and present in CASCADE 2.0. No new pathways were highlighted in these two EA.

Table 6: Pathways identified by the Enrichment Analyses of the different gene lists against KEGG

Gene list	cAMP	Cell Cycle	cPLA2	MAPK	PI3K/Akt	p53	VEGFR
DEA	🟦	🟦					
SCNA			🟦	🟦	🟦	🟦	🟦
DMR							
DEA+DMR	🟦	🟦					
DEA+SCNA +DMR	🟦	🟦		🟦	🟦	🟦	
DEA+SCNA + DMR+RMG	🟦	🟦		🟦	🟦	🟦	

Overall, the EA against KEGG mostly highlighted that the lists were enriched in genes involved in cAMP, MAPK, PI3K/Akt, p53, VEGFR, and cPLA2 signalling, and cell cycle. Most of these pathways are already present in CASCADE 2.0 and will be kept in the model. cPLA2 and VEGFR signalling were only over-represented in the SCNAs gene list, but will be studied later in the literature review.

These findings show that the individual omics alterations of TNBC cases, when taken together, bring up the same functional alterations than what is represented in CASCADE 2.0. It is also bringing up new pathways that can be further studied to extend CASCADE 2.0 in a TNBC-specific model.

3.3.2 Results of the Enrichment Analysis against Reactome

The same analyses were conducted against Reactome, and resulted in much more significantly over-represented pathways than with KEGG. Indeed, the pathway names in Reactome often refer to sub-processes of larger pathways, and many results were redundant. Several pathways were considered relevant in the cancer context but will be referred as only one because they are overall forming one more general pathway. For example, ”DNA Double Strand Break Response” and ”DNA Double Strand Break Repair” were named ”DNA repair”.

Similarly, Table 7 only displays the pathways that were found enriched in our analysis and that were selected during the literature curation as new modules. The extensive results are not shown for practical reasons, but Appendix G shows plots of the 30 over-represented pathways with the lowest q-value for each EA against Reactome. The DEGs were enriched in several relevant pathways, including in cell cycle, matrix remodelling, and Rho GTPases signalling which are present in CASCADE 2.0, but also in Notch, FGFR and BRCA signalling and DNA repair.

Surprisingly, the EA of the SCNA genes against Reactome did not provide a lot of significant results compared to the other gene lists. This list was enriched in pathways related to glycerophospholipids biosynthesis. The genes involved in this process are part of the arachidonic acid-derived signalling cascade, such as cPLA2.

Similarly to the EA against KEGG, the list of differentially methylated genes was enriched in two pathways involving G protein-coupled receptors which were not considered in the following steps. The list combining the DEGs and the differentially methylated genes was enriched in the same pathways than the DEGs alone, along with PI3K/Akt signalling genes.

The EA on all the aberrant genes except the RMGs showed that they were enriched in cell cycle, matrix remodelling, PI3K/Akt, and Rho GTPases signalling, which are already present in CASCADE. However, BRCA signalling and DNA repair, Notch, cPLA2 and FGFR signalling were also over-represented.

Performing an EA on all the aberrant genes (including the RMGs) gave additional results: apoptosis, cell adhesion, p53 and RTKs signalling were over-represented too.

Overall, similarly to the EA against KEGG, most of the pathways found over-represented are already included in CASCADE 2.0: apoptosis, cell adhesion, cell cycle, matrix remodelling, PI3K/Akt, p53, Rho GTPases, and RTKs signalling. This suggests that EA is a powerful tool to leverage knowledge at a functional level in cancer models. However, some pathways were highlighted by Reactome and not by KEGG: BRCA signalling and DNA repair, cPLA2, FGFR, and Notch signalling could, among others, be considered as new modules for the TNBC model. A combination of several pathway enrichment tools could enable us to complete the results and thus, reveal a wider range of abnormally functioning pathways in diseases.

Table 7: Pathways identified by the Enrichment Analyses of the different gene lists against Reactome

Gene list	Apoptosis	BRCA /DNA repair	Cell adhesion	Cell Cycle	cPLA2	FGFR	Matrix remodelling	Notch	PI3K / Akt	p53	Rho GTPases	RTKs
DEA		🟦		🟦		🟦					🟦	
SCNA					🟦							
DMR												
DEA+ DMR		🟦		🟦		🟦			🟦		🟦	
DEA+ SCNA+ DMR		🟦		🟦	🟦	🟦	🟦	🟦	🟦		🟦	
DEA+ SCNA+ DMR+ RMG	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦

3.4 Identification of Pathways in the Literature

Scientific literature related to breast cancer and especially to TNBC helped identify a number of pathways known to be abnormally regulated in the context of TNBC.
The literature curation confirmed that all of the pathways already present in CASCADE 2.0 are also known to be involved in TNBC mechanisms, and often targeted by therapies.

For instance, cell cycle genes are often found affected in TNBC or in Basal-like breast cancer, and have an impact not only on cell cycle, but also on the cytoskeleton with the organization of the centrosomes, of the chromosomes, and of the mitotic spindles [117]. Notably, various kinases have been found to be implied in the misregulation of the cell cycle, and have an impact on tumour growth and metastasis through the control of the cytoskeleton and the Epithelial to Mesenchymal Transition (EMT) [118, 119, 120, 121]. Cell cycle genes are also closely related to apoptosis, a pathway that is largely represented in CASCADE 2.0 [122, 123]. Genes such as MDM2, RB1, MYC, EP300, and cyclins were clearly identified as key genes in the misregulation of the cell cycle in TNBC, and were already present in CASCADE 2.0 [124, 125, 126, 127]. Examples of cell cycle genes potentially targeted by therapies are PLK1, the Aurora kinases A and B, and the cAMP-response element-binding binding protein (CREBBP) along with Beta-catenin and FOXM1 [128, 120, 123]. Overall, these elements concerning affected cell cycle genes, and their effects on tumour growth, EMT and proliferation, as well as apoptosis, confirmed that the cell cycle, matrix remodelling, and apoptosis pathways are central to tumour progression of TNBC, and that some genes are specific to TNBC cases and would need to be added to the model.

Other pathways present in CASCADE 2.0 such as PI3K/Akt-mTOR, MAPK cascade, WNT, and TGFβ signalling were often mentioned in the literature concerning TNBC. These pathways were not extensively investigated - and except for a few new genes, not extended - since CASCADE 2.0 was considered as a generic cancer model, with proven efficiency to describe cancer models, and that we decided that we would not remove any of its components. However, it is important to mention that several models of TNBC took these pathways into account, as in Tognetti et al. (2021) or in Narrandes et al. (2018) [117, 50]. Therefore, these studies highlight that the PI3K/Akt-mTOR signalling elements are among the most central ones and that the incidence of mutations of genes such as PIK3CA and PTEN (member of the PI3K/Akt signalling) is higher in TNBC than in other breast cancers, making it key therapeutic targets for the future of TNBC, as already suggested in other studies [129, 130]. The MAPK signalling cascade was found hyperactivated in a subset of TNBC cases [131] and genes of the MAPK cascade have been found hyperactivated in TNBC compared to other breast cancer subtypes [80]. The WNT pathway and its receptors, the TGFβ, NF-κB, and Jak/STAT signalling pathways are also misregulated in TNBC, and constitute potential targets for treatment [117, 129, 131].

FGFR Signalling: One of the pathways frequently mentioned in the literature is the Fibroblast Growth Factor Receptors (FGFR) signalling cascade. FGFR is a family of four Receptor Tyrosine Kinases (RTKs), named FGFR1, FGFR2, FGFR3, and FGFR4. RTKs signalling is a pathway that is partly represented in CASCADE2.0, but no FGFR is included. It has been shown that a large number of alterations and mutations in these receptors are responsible for the abnormal regulation of the signalling cascade triggered by the RTKs, and participate to carcinogenesis [132]. Specifically, the binding of FGFR's ligands triggers a panel of signalling cascade involved in developmental processes as well as in physiological processes in adults. Furthermore, FGFR alterations have been observed in different cancers, and this family of receptors is strongly considered for therapy [133]. In TNBC specifically, FGFR2 was shown to be amplified in 4% of the cases while it was never amplified in other breast cancer subtypes [134]. In the case of FGFR1, it is amplified in 10% of all breast cancer cases, and especially in luminal breast cancer [135]. However, it is amplified in the TNBC cell line CAL120 too [136]. Some TNBC cell lines also have high expression of FGFR1, which was shown to be linked to the Overall Survival (OS) of TNBC patients, cases with lower expression having better prognosis than the others [137]. In another study, FGFR3 was shown to be highly expressed and phosphorylated in cell lines harboring a FGFR3-TACC3 gene fusion, and could therefore be targeted for therapy in these cases [138]. Furthermore, TNBC cell lines showed a high sensitivity to FGFR inhibitors [136], which makes it a key therapeutic target for TNBC treatments.

cPLA2 Signalling: Eicosanoids are a family of lipid Signalling molecules principally derived from the metabolism of arachidonic acid (AA) [139]. Phospholipase A2s (PLA2s), including calcium-dependent PLA2 (cPLA2), are responsible for the release of AA into the cytoplasm, which can trigger the synthesis of different types of eicosanoids: prostaglandins and thromboxanes (called the prostanoids) are synthesized through the cyclo-oxygenase (COX1 and COX2) pathway, hydroxyeicosatetraenoic acids (HETEs) and leukotrienes through the lipoxygenase (LOX5, LOX12, and LOX15) pathways, and epoxygenated fatty acids (EETs) through the cytochrome P450 pathway [140, 141, 142]. COX1 and COX2 are responsible for the synthesis of prostanoids such as PGE2 or TXA2. Indeed, cPLA2 helps release AA from its membrane, which is then being metabolized into prostanoids. The production of these molecules is followed by their release from the cell, where they can act on biological processes by binding to their associated receptors such as EP1 to EP4 and TP for PGE2 and TXA2, respectively. Specifically, cAMP and Ca2+ production is enhanced under the activation of these receptors. On the other hand, lipoxygenases are responsible for the production of HETEs and leukotrienes through the same kind of mechanisms. The main receptors of HETEs and leukotrienes are BLT1 and GPR31. The most well characterised LOX pathway is the one triggered by LOX5, but LOX12 and LOX15 have been shown to have an important role in cancer development as well. [142]. In breast cancer especially, it has been showed that the overexpression of COX2 is responsible for tumour development and metastasis. PGE2 is reported to bind particularly to EP2, EP3, and EP4, which would lead to an over-production of cAMP and activation of SRC and EGFR, which activate Signalling molecules of the MAPK, PI3K/Akt, and ERK cascades, pathways promoting cell survival, proliferation, and migration [143]. Basal-like and TNBC cases show a tendency to overexpress cPLA2, and this overexpression would lead to a low rate of relapse-free survival [144]. Also, in TNBC, the activation of BLT1, the receptor of LTB4, would enhance cell migration, and LOX inhibition would block it [145]. Overall, the cPLA2 signalling pathway represent a potential therapeutic target, both at the level of cPLA2, AA-derived molecules, or their receptors, and could be considered for addition in CASCADE2.0.

EGFR Signalling The Epidermal Growth Factor Receptor (EGFR) is a member of the RTKs, and more precisely, a member of the subfamily of c-erbB along with HER2. Its signalling is triggered by the binding of one of its numerous ligands to its extracellular binding sites [146]. EGFR, also referred to as HER or ERBB1, is able to form heterodimers with HER2 (ERBB2) and ERBB3, which results in a wider range of possibilities for the regulation of their downstream targets. The main downstream targets of EGFR activation are PLCγ and the MAPK family, PI3K, and Signal Transducer and Activator of Transcription (STAT) genes [146, 147]. EGFR signalling is known to be altered in many cancers, promoting cellular growth and proliferation, preventing apoptosis, and also for being an important prognostic factor in different types of cancers [148]. It has been shown that EGFR is overexpressed in TNBC cases compared to non-TNBC, but that the rate of EGFR-overexpressing cases varies widely between the studies [15]. Gene copy number of the EGFR would also be altered in TNBC cases, being high in 33% of them, and significantly correlated with a poor disease-free survival [149]. Targeting EGFR in TNBC therapy is a promising strategy which is still under investigation. Several FDA-approved-drugs such as Erlotinib and Lapatinib inhibit EGFR [129].

BRCA Signalling/DNA Repair 5 to 10% of breast cancer cases are due to inherited factors, among which 52% would result from a mutation in BRCA1, and 35% from a mutation in BRCA2 [150]. Both genes play a role in DNA damage response and transcription of genes involved in DNA repair, cell cycle, and apoptosis. More precisely, BRCA1 and BRCA2 are thought to form a complex with Rad51 to regulate DNA Double-Strand Break (DSB) repair and Homologous Recombination (HR). BRCA1 is also involved in cell cycle checkpoints and transcription mechanisms as it interacts with p53 and other cell cycle regulators such as p21. The G2/M transition of the cell cycle would be regulated by BRCA1 as well. Indeed, the phosphorylation of BRCA1 by ATM and the activation of checkpoint kinases such as CHK1, are thought to be essential steps to trigger cell cycle checkpoints and detect DNA damages [122]. In TNBC, the prevalence of BRCA1 mutations was shown to be higher than in non-TNBC, with approximately 70% of breast cancers with BRCA1 mutations showing a triple-negative profile [15, 129]. The mutations in BRCAs lead to deficiency in HR, which can be used for a treatment based on synthetic lethality: Polyadenosine diphosphate-ribose polymerase (PARP) is a protein that recognizes Single-Strand Breaks (SSB) of DNA, and triggers its repair by Base Excision Repair (BER). The inhibition of this protein blocks the repair of SSB, leading to more abnormalities in DNA, which, combined to the HR deficiency due to BRCA mutations, leads to cell death [151]. Currently, Olaparib and Talazoparib are PARP inhibitors used to treat TNBC [152].

VEGFR Signalling The Vascular Endothelial Growth Factor (VEGF) Receptor 2 (VEGFR2), or Kinase-Insert-Domain-Containing Receptor (KDR), is a member of the RTK family, and the main receptor for the ligands VEGF. VEGFR2, referred to as VEGFR in this report, is activated through binding of its ligands, and has fundamental vascular functions. It is responsible for vasculogenesis, angiogenesis, and differentiation of endothelial cell progenitors, but it also has a role in major physiological functions such as survival, proliferation, and differentiation. Therefore, it is implied in many diseases, and especially cancers, as its abnormal regulation can impact all of these processes [153]. Breast cancers progression would be partly due to VEGF expression, with an Overall Survival (OS) inversely correlated to VEGF expression, and a lower response to treatment for the cases with high expression. Especially, it has been found that the levels of VEGF is higher in TNBC and basal-like breast cancer than in the non-basal-like and non-TNBC, and that the microvascular density, thought to be a prognostic factor in breast cancer, is higher as well [154]. Targeting of VEGFR signalling is indeed investigated for TNBC treatment with the use of bevacizumab as a neoadjuvant, for example [155]. However, more investigations need to be done, but the VEGFR pathway remains a promising target.

Notch, Hippo, and Hedgehog Signalling The Notch, Hippo, and Hedgehog signalling pathways are not present in CASCADE 2.0, but they were added to CASCADE 3.0 as they were identified as misregulated pathways in colorectal cancer, and in cancer in general as well. These three pathways were documented in several papers concerning TNBC. The Notch pathway is a signalling cascade resulting from the potential binding of five ligands, DLL1, DLL3, DLL4 and Jagged1-2 to the Notch receptors. Notch1-Notch4 are single-pass transmembrane receptors, which, when activated, trigger the transcription of their target genes [1]. Notch signalling has a central role in cell fate decision and cell death, in tissue development and cell proliferation, and therefore, is involved in the development of cancers when it is misregulated. Notch signalling is involved in the maintenance of mammary cancer stem cells and has been found hyperactive in breast cancers in general, but Notch1/3/4 receptors have been found specifically overexpressed in TNBC. Mutations in Notch1 and Notch3 would lead to an hyperactivation of Notch signalling and its downstream targets such as NF-κB or Akt-mTOR. Notch3 would also be amplified in TNBC compare to other breast cancer subtypes [156, 157]. Overall, Notch signalling seems to play a key role in TNBC tumorigenesis, is hyperactivated because of alterations, and its downstream targets are key signalling components in cancer.

Similarly, the Hippo pathway has been found hyperactivated in TNBC [131, 158, 159]. The Hippo signalling is a kinase cascade leading to the inactivation of YAP/TAZ, cofactors involved in the transcription of genes in control of cell growth, cell proliferation, and cell death [160]. Alterations in the behaviour of Hippo genes and their upstream regulators would provoke a high signalling activity, thus promoting tumour growth and progression of TNBC.

Lastly, the Hedgehog signalling pathway has been shown to be involved in breast cancer, and some of its members are specifically altered in TNBC. It is a signalling cascade triggered by the binding of three ligands (Sonic, Desert, and Indian Hedgehog, respectively referred to as SHH, DHH, IHH) to the receptors PTCH1 and SMO, which subsequently activate the glioma-associated oncogenes GLI which regulate the transcription of cancer promoter genes. It is especially involved in embryogenesis, stem cell renewal and tissue repair and promotes metastasis in the context of cancers [129, 161]. Studies have found that the expression of SHH, PTCH and SMO were prognostic factors of TNBC, and that, overall, the Hedgehog signalling pathway was upregulated in TNBC more than in other breast cancer subtypes [162, 163].

3.5 Composition of the Prior Knowledge Network - the TNBC Model

The goal was to extend CASCADE 2.0 with new modules, to build a gene regulatory network representing the signalling processes involved in TNBC. The literature curation highlighted the importance of the pathways from CASCADE 2.0. Many of these pathways were also over-represented in the lists of altered genes from 3.3. For instance, the cell cycle, the MAPK cascade, the PI3K/Akt, and the Rho GTPases signalling were both over-represented, and present in CASCADE 2.0. All the pathways and nodes from CASCADE 2.0 were kept for the modelling, and the same topology was respected, except for a few exceptions. Indeed, CASCADE 2.0 included some genes annotated with "_g", and their protein product (e.g, PTEN_g and PTEN). The genes were often either only regulating the activity of their gene product, or only an intermediate between two nodes (e.g, their activity was regulated by one other node, and they were regulating the activity of their gene product). The impact of such nodes on the dynamic of a boolean network is low and we decided to remove them.
Overall, the middle-out modelling approach revealed eight new pathways that were selected and added to the model. The DNA repair and Notch pathways were highlighted both in the EA and in the literature curation. Furthermore, they are present in CASCADE 3.0 as colorectal cancer-specific pathways [61]. Besides this, Hippo and Hedgehog signalling, two pathways from CASCADE 3.0 were documented as important altered processes in TNBC. As a consequence, they were slightly modified in order to be more specific to TNBC, and were added to the PKN. Here, the literature was used to find the most important genes composing these pathways in the context of TNBC, and SIGNOR was retrieved to build the topology of the modules.
Four other new modules, namely cPLA2, EGFR, FGFR, and VEGFR were composed using mainly SIGNOR and the literature about TNBC.
Furthermore, the progesterone and estrogen receptors with their ligands, and the human epidermal growth factor HER2 were added to a module named "Receptors". As we tried to constrain the size of the model, some interactions of the new modules that could be simplified without impacting the general dynamic of the network were modified, the nodes with only one incoming and one outgoing edges were removed, and if a node was not reported in the literature as being altered in most cancer types or in TNBC, and was not a central node of the module, it was generally removed.

Initially, CASCADE 2.0 contained 11 modules, 144 nodes and 367 edges. The resulting PKN, which we named the TNBC model, contains 20 modules, 221 nodes and 716 edges. The names of all the modules of the PKN, the number of nodes it contains, in which analyses they were highlighted, and the sources used to build them are summarized in Table 8.

The gene regulatory network was built and visualised on GINsim. We attributed a color to each module and the nodes were colored according to the module to which they belong. It was then exported to a format that could be read by Cytoscape, because it provided more layouts than GINsim and enabled us to compute some topological properties of the network. The general topological properties of the network are summarized in Table 9. As a comparison, CASCADE 2.0 has a lower average degree (5.0) and clustering coefficient (0.043), and slightly higher diameter (13) and average path length (5.6).

We searched how many altered genes from the omics data analyses resulted in our network. In total, without taking into account the genes with synonym names, 52 genes resulting from the combined omics analyses ended up in the TNBC model. The name of these genes and their topological properties are displayed in Table 10, and they are sorted by decreasing degree and decreasing betweenness centrality. The average degree of these nodes is 7.1, which is higher than the average degree of the network. Most of the nodes were already represented in CASCADE 2.0, but new and central nodes are present as well, such as the estrogen and progesterone receptors ESR1 and PGR, the epidermal growth factor receptor EGFR, the fibroblast growth factor receptor FGFR, the vascular endothelial growth factor family of genes VEGF_f, BRCA2, or the PARP family of genes.

Name	Module	Degree	Betweenness Centrality
TP53	Cell cycle	22	5.603
MYC	Cell cycle	18	5.9975

Continuation of Table 10

Name	Module	Degree	Betweenness Centrality
ESR1	receptor	15	9.9636
PIK3CA	PI3K/Akt	13	15.475
SMAD2	TGFB	13	1.6258
DVL f	WNT	12	3.8308
CSNK1D_E	HIPPO	12	0.9252
LRP_f	WNT	11	3.5625
IKBKB	NFKB	11	3.5042
EGFR	EGFR	11	0.3241
PTEN	PI3K/Akt	10	4.3569
BIRC_f	Apoptosis	10	3.9562
PRKCD	MAPK cascade	10	3.3665
CDC25A	Cell cycle	10	2.6611
PLK1	Cell cycle	10	1.6808
CHEK1	DNA repair/BRCA	9	1.8349
RHOA	Rho GTPases	8	2.1093
FOS	MAPK cascade	8	1.9776
CASP3	Apoptosis	7	9.2462
RB1	Cell cycle	7	5.4097
CASP9	Apoptosis	7	3.666
S6K_f	MAPK cascade	7	0.5639
SKI	TGFB	7	0.5074
cPLA2a	cPLA2	6	4.8241
DUSP1	MAPK cascade	6	4.0068
VEGF_f	VEGFR	6	2.2758
GRB2	MAPK cascade	6	1.3551
STK_f	HIPPO	6	0.9367
RAS_f	MAPK cascade	5	1.514
CASP8	Apoptosis	5	1.4795
MMP_f	MAPK cascade	5	1.2981
EGR1	MAPK cascade	5	1.2314
TGFBR2	TGFB	5	0.7239
AURKA	Cell cycle	5	0.518
CCNE1	Cell cycle	5	0.2689
AURKB	Cell cycle	5	0.1902
VEGFR2	VEGFR	4	2.6992
FZD_f	WNT	4	1.2347
CDKN2A	Cell cycle	4	0.9764
SRF	Rho GTPases	4	0.7944
E2F1	Cell cycle	4	0.137
CCNB1	Cell cycle	3	1.1614
FGFR	FGFR	3	0.9446
RAD51	DNA repair/BRCA	3	0.9097
PGR	receptor	3	0.8637
BRCA2	DNA repair/BRCA	3	0.4864
cAMP	PI3K/Akt	3	0.3402
PARP_f	Cell cycle	3	0.2159
CREBBP	Cell cycle	3	0.2077
MAP3K4	MAPK cascade	2	0.3094
DLL1_3	NOTCH	2	0.0253
FGF_f	FGFR	1	0

As seen in Table 10, the results of the omics analyses highlighted 52 nodes that were eventually represented in different pathways of the TNBC models. These nodes have an average degree higher than the average degree of the network, suggesting that omics data analyses are able to unveil nodes

Table 8: Modules present in the TNBC model. Their size is the number of nodes that were attributed to the pathway. The analysis describes in which of the enrichment analysis (EA) or literature curation steps (or both) the pathway was highlighted. The source is the main source(s) used to build the pathway at a component's scale.

Module	Size	Analysis	Source
Apoptosis	15	EA ; Literature	CASCADE 2/3
Cell cycle	19	EA ; Literature	CASCADE 2/3
cPLA2 signalling	17	EA ; Literature	Literature
DNA repair/BRCA signalling	12	EA ; Literature	CASCADE 3 ; Literature
EGFR signalling	3	Literature	Literature ; SIGNOR
FGFR signalling	4	EA ; Literature	Literature
HIPPO signalling	3	Literature	CASCADE 3
Hedgehog signalling	4	Literature	CASCADE 3
Jak/STAT signalling	8	None	CASCADE 2
MAPK signalling	43	EA ; Literature	CASCADE 2/3
mTOR signalling	5	Literature	CASCADE 2
NF-κB signalling	8	Literature	CASCADE 2/3
Notch signalling	9	EA ; Literature	CASCADE 3 ; Literature ; SIGNOR
PI3K/Akt signalling	8	EA ; Literature	CASCADE 2 ; Literature
Rho GTPases signalling	11	EA ; Literature	CASCADE 2
RTKs signalling	8	EA ; Literature	CASCADE 2
TGFβ signalling	16	EA ; Literature	CASCADE 2
VEGFR signalling	5	EA ; Literature	Literature, SIGNOR
WNT signalling	16	EA ; Literature	CASCADE 2

Table 9: Topological properties of the PKN

Nodes	Edges	Average degree	Diameter	Charact. path length	Clustering coefficient
221	716	6.3	12	4.9	0.076

that are somehow important in the dynamic of the network.

An exhaustive list of the nodes of the network, the pathway they belong to, along with topological parameters is given on the **Github** repository in the *node_measures_PKN.csv* file. A visualisation of the PKN is displayed in Figure 7. It is represented as a directed network, and the edges can either be activating or inhibitory. The different modules can be distinguished by their color.

As seen in Section 3.3 and 3.4, the results from the EA revealed pathways that were often known in the literature as being altered in TNBC. This suggests that an EA is a reliable statistical tool to unveil the most important pathways underlying a disease. Thus, if further extension steps of the model should be done, it would be interesting to investigate the results of the EA that we did not consider in this work. Several pathways were recurrent in the significant results of the EA, such as the the Calcium, IL17, or PPAR signalling pathways, and could probably bring new insights into the dynamic of the model. Additionally, we omitted the modelling of the androgen receptor (AR) which has been revealed by various studies as a key over-expressed receptor in TNBC cases of the LAR subtype [15, 19, 20]. However, addition of new pathways would end up in an even larger network, and reduction or refinement steps could be done in order to keep the network in a reasonable size. For example, GINsim allows the reduction of Boolean models by removing iteratively the nodes that the user wants to possibly remove. From there, the user can compare the expected dynamic of the system with the dynamic of the reduced model, and conclude on the usefulness of the removed node(s) [52, 164].

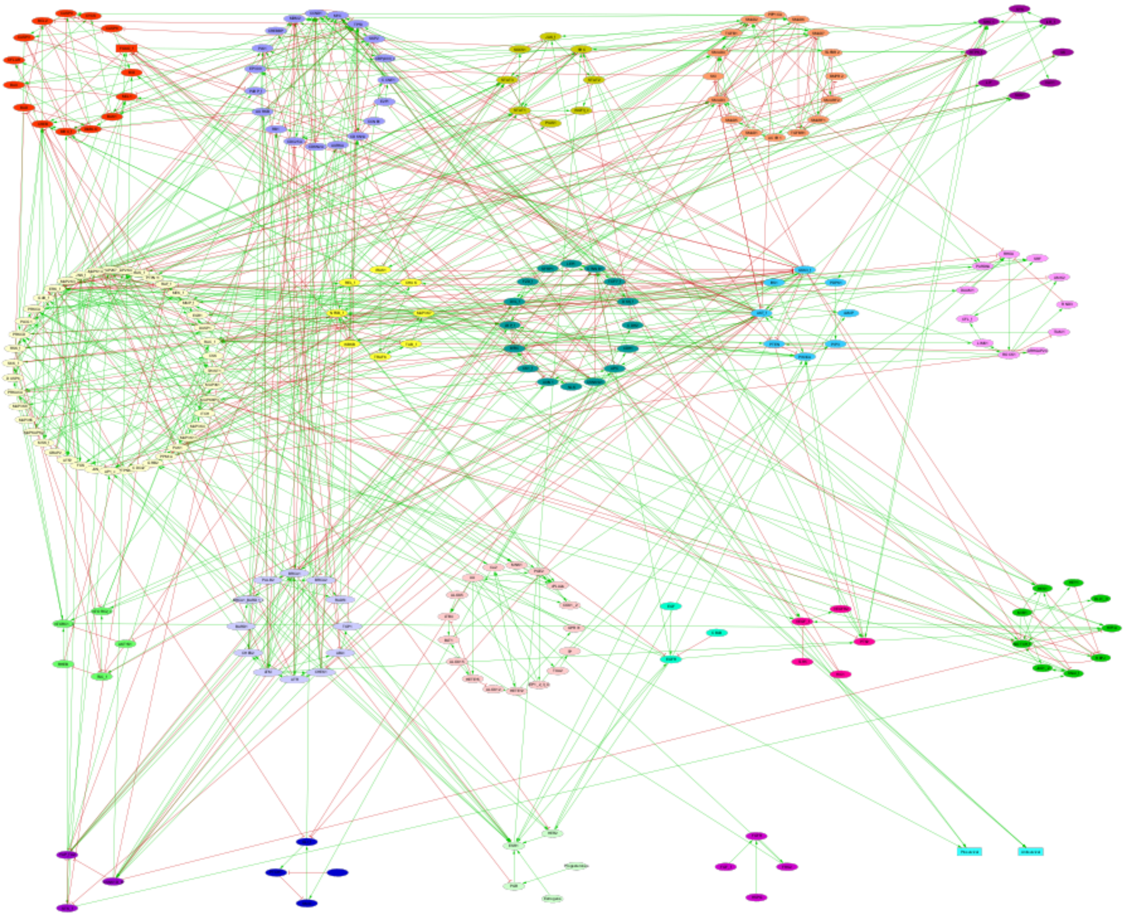


Figure 7: Representation of the Prior Knowledge Network (PKN) - the TNBC model. The edges are directed and colored in green if activating, in red if inhibitory. The ten modules on top of the figure with the light green module on the left of the third row are the ones from CASCADE 2.0, possibly completed with new nodes. The nine new modules correspond to the following colors: DNA repair/BRCA signalling is in light purple, cPLA2 signalling is in light pink, EGFR signalling is in light blue/green, VEGFR signalling is in fuschia, Notch signalling is in green, Hippo signalling is in dark purple, Hedgehog signalling is in marine blue, the receptors are in light green, and FGFR signalling is in purple. The two light blue rectangular nodes are the phenotype nodes *Pro-survival* and *Anti-survival*.

3.6 Conversion of the Prior Knowledge Network into a Boolean Network

3.6.1 Training of the Models on Data from TNBC Cell Lines

The PKN was transformed into a Boolean network using Gitsbe, which in practice produces several different Boolean networks that could fit the reality. The four cell lines used to train the model are expected to produce different dynamic behaviours since their RNA expression data differ from each other. As detailed in Section 2.9.1, the RNA expression data of the four cell lines were scaled to values between 0 and 1 to be adapted to Gistic. A heatmap of the scaled RNA expression data of 184 genes represented in the model is displayed in Figure 8. Hierarchical clustering has been applied to the cell lines, and we can see that MDA-MB-231 and MDA-MB-453 are clustering together on one side, and BT549 and HS-578T on the other side. Clusters of genes also have been formed, and we clearly see three groups of genes that have a high, medium, or low expression. However, differences in the expression of some genes between the four cell lines remain.

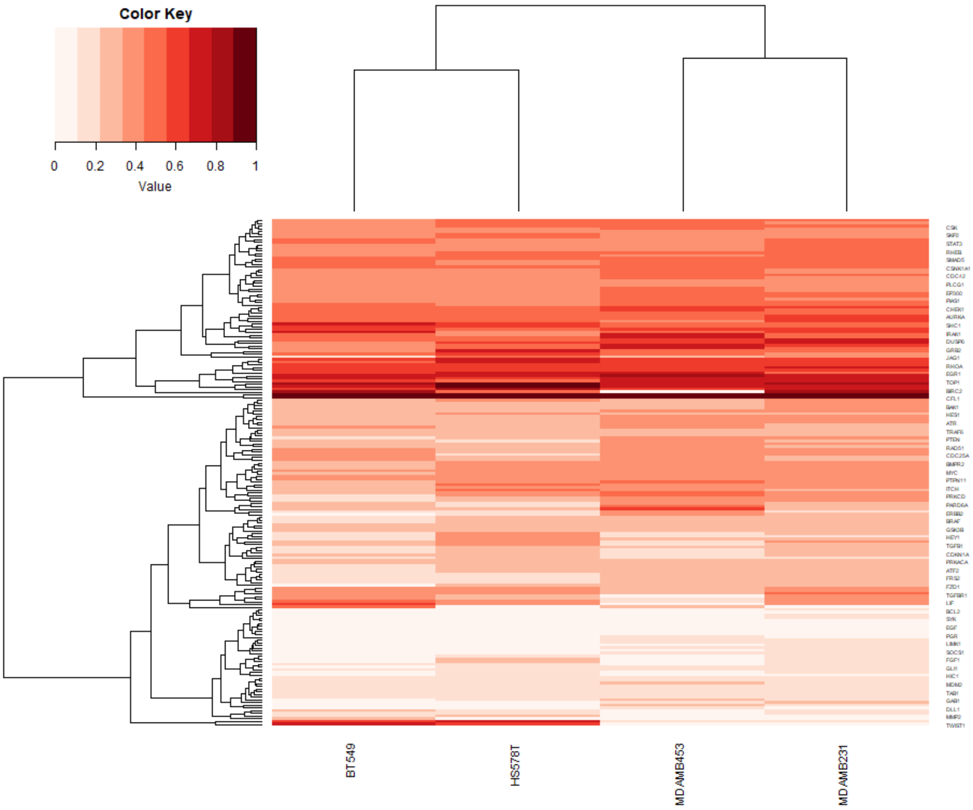


Figure 8: Heatmap of the RNA expression of 184 genes of the network in the four cell lines from CCLE.

In order to be aware of which genes should play a role in the differentiation of the dynamics of the models trained on the four cell lines, we investigated which genes had the most variable expression across them. The full results can be found in the **Github** repository in the *variance_RNAexpr_training.csv* file, and the top ten genes with the highest variance are listed below:

- TWIST1: is a gene attributed to the MAPK signalling cascade and is positively regulated by ERK_f and JNK_f in the model. It was added by extension to the model, during the literature curation phase.
- DKK1: is present in the model under the gene family DKK_f.
- CDH2: is a newly added gene from the WNT module.
- TP53: was already present in CASCADE 2.0, and was found to be the most frequently mutated genes in the TNBC omics data analyses.
- CDKN2A: was added by extension of the cell cycle module.
- JAG1: is present in the NOTCH pathway, and it was merged with JAG2 as they were contributing to the same dynamics. It was present in CASCADE 3.0.
- LIMK2: is a node from CASCADE 2.0, belonging to the Rho GTPases pathway.
- FOS: is involved in MAPK signalling node from CASCADE 2.0.
- STAT1: is present in CASCADE 2.0 along with its family members STAT2 and STAT3. They are part of the Jak/STAT signalling module.
- PTPN6: is part of the MAPK module from CASCADE 2.0.

We also investigated the state of ESR1, PGR, and HER2 (ERBB2) in the RNA expression data. The scaled expression value of these nodes in the four cell lines are listed in Table 11. However, if we expect a low protein expression of these receptors, we do not necessarily expect their mRNA levels to be low.

The scaled expression level of ESR1 is lower than 0.2 in all the cell lines, and the scaled expression level of PGR is even lower than 0.1 in all the cell lines. If we consider that a low level of expression is below 0.5, then these genes are weakly expressed in these four TNBC cell lines. The ERBB2 receptor also has low levels of scaled expression in all the cell lines, but we noticed that its expression in MDA-MB-453 was more than twice higher than in the other cell lines (0.414).

As for the frequently mutated genes, TP53 is the least expressed in MDA-MB-453 (0.086), and the most highly expressed in BT549 (0.590), while PTEN and PIK3CA both have levels of expression ranging from approximately 0.200 to 0.350 in the four cell lines.

Table 11: Expression status of relevant nodes in the CCLE data of the four cell lines used to train the models. The expression ranges from 0 to 1 because of the transformation applied and described in 2.9.1.

Gene name	**MDA-MB-231**	**MDA-MB-453**	**HS-578T**	**BT549**
ESR1	0.109	0.087	0.048	0.050
PGR	0.065	0.098	0.027	0.013
ERBB2	0.197	0.414	0.201	0.155
TP53	0.429	0.086	0.435	0.590
PTEN	0.352	0.340	0.273	0.205
PIK3CA	0.249	0.287	0.310	0.205

The mRNA levels of the three receptors remain low in the four cell lines ($\leq$0.5). This is correct to train the models, since we expect them to be inactive in the steady-states. We will discuss later of the resulting nodes' states obtained after training of the models (see section 3.8.4).

TP53 is reported as mutated in all the cell lines except MDA-MB-453 in Neve et al. (2006) and most mutations affecting TP53 are missense mutations where there is a loss of function of the protein [165, 166]. Indeed, using RNA expression data of TP53 could be erroneous in this case, since the functionality of the protein as a tumour suppressor might be more important for the final dynamic of the model, and it is not reflected in the training data.

PTEN is also a tumour suppressor with loss of function mutations in many cancers, and especially in TNBC. It is also thought that epigenetic silencing of the promoter regions of PTEN could lead to loss of activity of the PTEN protein, and thus, to lower its tumour suppressor activity [167]. PTEN would be mutated in MDA-MB-453 and in BT549 [20].

PIK3CA is the catalytic subunit p110α of PI3K, and is an oncogene upstream of the PI3K/Akt signalling pathway. The mutations affecting PIK3CA are activating mutations, and lead to an over-activation of the PI3K/Akt pathway in TNBC [168].

These observations both on PTEN and PIK3CA show that their protein activity in the four cell lines studied could be respectively lower and higher than what we observe for their RNA expression. We will discuss later about their local state in the trained models, and verify if they match the biological reality.

To improve this bias, we could try to train the models by forcing the state of these proteins to 0 or 1, according to their expected state. Other differences in the mutational profile of the cell lines could be introduced in the model, for instance, MDA-MB-231 has mutations in KRAS, BRAF, and CDKN2A which could be reported in the simulations, and BT549 has mutations in RB1 [20].

3.6.2 Training of the Models on TCGA Data

The hierarchical clustering of the 173 TCGA-BRCA TNBC samples with the four cell lines' data from the CCLE was performed based on the 1000 most variable genes across all the cell lines, and we chose to form two clusters. The four CCLE cell lines were clustered together in a group with 88 TCGA samples. The code to obtain those results can be found on the **Github** repository in the *clustering_TCGA_CCLE.R* file. From this clustering, we note that the CCLE cell lines cluster together before clustering with any other TCGA sample, meaning that the closest samples to any of these cell lines are the other CCLE cell lines. These findings will be discussed below. The 88 selected TCGA samples were then used to train the models in another set of simulations. Similarly to the CCLE cell lines, we plotted a heatmap of the RNA expression values from the TCGA samples for 184 genes of our model, which is displayed in Figure 9. This time, the training data were discretised, therefore we can only see the active or inactive status of the genes in the samples. However, we can clearly see two clusters of genes that are either active or inactive in all of the samples. Other genes have a more mitigated activation status across all the samples, sometimes being active, and sometimes inactive, showing the heterogeneity of TNBC cases.

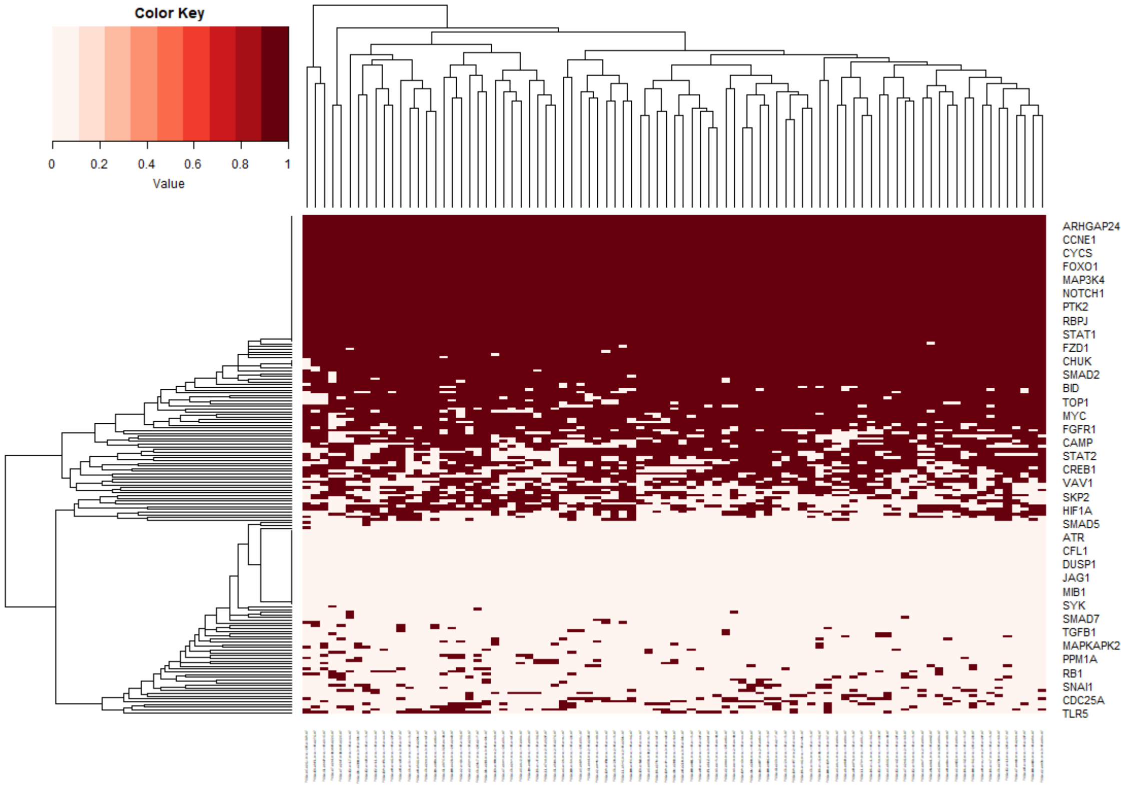


Figure 9: Heatmap of the RNA expression of 184 genes of the network in 88 TCGA samples which clustered with the four CCLE cell lines.

The discretisation of the data was motivated by previous findings in the Druglogics group: it has been observed that using boolean states to train a model in the pipeline enabled the generation of models with better fitting scores, and with better prediction performances [61].

3.7 Drug Synergy Predictions

The pipeline was run with many different configuration and input parameters. As highlighted by the results of the predictions on CASCADE 2.0 [58], no more than 150 simulations need to be run to obtain good performance results.

3.7.1 Tested Drugs and Their Targets

The panel of drugs tested with Drabme was specific to each cell line. The drugs involved in synergistic combinations were retrieved from Synergx, DrugComb, and DrugCombDB, and their targets and effect on their targets were retrieved from DrugBank, as explained in section 2.9.3. Appendix D summarizes the tested drugs, their effect, and their targets for each of the cell lines. Several drugs target several nodes of the network, which was possible to specify in the input files

of Drabme. For example, Tamoxifen inhibits ESR1, PRKCA and PRKCD, which were then forced to 0 for the predictions. Also, several drugs have the same effect on the same targets in the context of the TNBC model (e.g. Nilotinib and Imatinib, or Fulvestrant and Tamoxifen).
MDA-MB-453 is not included and was not considered for further analyses, since we could not find any synergistic drug combinations targeting nodes of the model for this cell line. In the remaining parts of this section, we will not talk about MDA-MB-453 anymore, and we will refer to MDA-MB-231, HS-578T and BT549 as the three CCLE cell lines.
In total, 253 combinations were tested for MDA-MB-231, and among them 52 were observed synergistic based on the BLISS score. For HS-578T, 210 combinations were tested among which 102 were observed synergistic based on the BLISS score. And finally, 231 combinations were tested for BT549, and 76 were observed synergistic based on the BLISS score. The lists of observed synergistic combinations (based on the BLISS score) for each of the cell line can be found on the **Github** repository: *observed_synergies* → *observed_synergies_bliss_name.tab*.

In general, the cell lines showed different synergies (see drugpanel in Appendix D), but one part of those synergies was common to all. We looked for the combinations that were observed synergistic across all the cell lines. We found that 18 synergy combinations were common to all of them, and 11 drugs were needed to test all of these combinations. The 18 combinations common to the four CCLE cell lines are displayed in Table 12

Table 12: Synergistic drug combinations observed for the three CCLE cell lines

Drug A	**Drug B**	**Targets**
Paclitaxel	Lapatinib	BCL2 ; EGFR ; ERBB2
Celecoxib	Paclitaxel	PDPK1 ; BCL2
Dasatinib	Fulvestrant	ABL1 ; SRC ; ESR1
Vismodegib	Vandetanib	SMO ; EGFR
Vismodegib	Celecoxib	SMO ; PDPK1
Ruxolitinib	Paclitaxel	JAK_f ; BCL2
Axitinib	Paclitaxel	VEGFR2 ; BCL2
Vismodegib	Paclitaxel	SMO ; BCL2
Vandetanib	Paclitaxel	EGFR ; BCL2
Vemurafenib	Paclitaxel	RAF_f ; BCL2
Vismodegib	Ruxolitinib	SMO ; JAK_f
Vandetanib	Ruxolitinib	EGFR ; JAK_f
Vemurafenib	Dasatinib	RAF_f ; ABL1 ; SRC
Vandetanib	Dasatinib	EGFR ; ABL1 ; SRC
Ruxolitinib	Dasatinib	JAK_f ; ABL1 ; SRC
Vismodegib	Fulvestrant	SMO ; ESR1
Vemurafenib	Nilotinib	RAF_f ; ABL1
Vandetanib	Nilotinib	EGFR ; ABL1

3.7.2 Relevance of Using TCGA Samples for Drug Synergy Predictions

The clustering of samples based on their gene expression profile is commonly used to find subgroups of patients with the same expression profiles, and for which we can find specific biomarkers that could be treatment targets [101, 169]. Indeed, the clustering of TCGA samples and CCLE cell lines revealed that certain cell lines had a more similar expression profile to certain groups of TCGA patients. From this observation, we hypothesized that this subset of TCGA patients would likely respond to drug perturbations in the same way as the four CCLE cell lines did. Thus, to run predictions on the models calibrated on the expression data of this subset of TCGA patients, we reduced the drug panel to the drugs present in Table 12. Reducing the drug panel to drugs that were synergistic in all the CCLE cell lines reduces the risk of testing drug combinations that are efficient only in one specific cell line, and maximizes our chances that the TCGA patients respond to it.
However, it is important to note that the use of cell lines to build cancer models is controversial. It is true that it was found that breast cancer cell lines generally represent quite well the tran-

scriptomic profile of breast cancer patients (e.g., TCGA-BRCA samples). But it was also found that some important differences remain [92, 170]. Indeed, since cell lines are cultured *in vitro*, they do not account for all the events happening *in vivo* and the environment of the tumour cells. On the other hand, TCGA samples are composed of tumour cells, but also of the cells in their environment such as immune cells, fibroblasts or epithelial cells, and the ratio of tumour cells over the other types of cells is called the ”tumour purity” [171]. This particularity makes TCGA data analyses account for the genomic features of not only the tumour cells, but also of the tumour micro-environment, contrary to cell lines which are cultivated without this environment. Vincent et al. (2015) noted that this could lead to a forced epithelial-to-mesenchymal transition in the culture of the cell lines, thus losing the epithelial features of the cells. They also found that 99% of the genes that were usually up-regulated in cancer stromal cells (compared to malignant cells) were significantly down-regulated in the breast cancer cell lines from CCLE, proving the importance of the tumour micro-environment in the expression of some groups of genes and on the biological processes they lead. This could explain the findings mentioned in section 3.6.2: there are fundamental differences between the expression of subsets of genes in *in vitro* cultured cell lines and patients samples, thus CCLE cell lines cluster together before clustering with any other TCGA sample. Additionally, Jiang et al. (2016) found that the mutational and the protein expression profiles of most breast cancer cell lines correlated poorly with the TCGA-BRCA data [92].

These results are contrasted by the fact that Vincent et al. (2015) also found that the basal/ER- breast cancer cell lines were more representative of the TCGA-BRCA basal/ER- tumours than the luminal/ER+ cell lines are of the luminal/ER+ TCGA-BRCA data. This is interesting for us since, according to the PAM50 classification, MDA-MB-231, HS-578T and BT549 are classified as basal-like, while MDA-MB-453 is classified as luminal B [92, 170]. Thus, MDA-MB-231, HS-578T and BT549 would be more likely to represent the TCGA samples (for which the majority is basal-like and ER-).

These findings provide reasonable evidence to assume that building logical models calibrated on TCGA samples could provide good fitness scores, since TCGA patients' samples are more representative of TNBC. However, predicting synergies on a subset of TCGA samples that clustered well with the four CCLE cell lines could lead to wrong results because of the assumptions that TCGA data will respond in the same way as CCLE cell lines.

3.7.3 Specificity of the Drug Panels

The drugs selected for the *in silico* predictions were observed synergistic in the CCLE cell lines. However, we have used drugs such as Sorafenib to target RAF_f, VEGFR2, and FGFR. Sorafenib is known to be an inhibitor of the RAF/MEK/ERK pathway, or the MAPK/ERK pathway [172], but it has been shown to be targeting other molecules such as VEGFR2/3 or the platelet-derived growth factor β [173]. We searched the International Centre for Kinase profiling database which provides information on the effects of some kinase inhibitors [174]. We observe that, administered at a fixed concentration, Sorafenib will inhibit the activity of many molecules to a certain extent, the most efficiently inhibited being ERK8 (93% of activity inhibited). It is also inhibiting the activity of other molecules such as Src (56% of activity inhibited), or JNK2 (35% of activity inhibited), which is not represented in the predictions. Indeed, these inconsistencies rely on the fact that the model is limited in size, contrary to *in vivo* or *in vitro* experiments where the cells and most of their components are present and affected by the drugs.

3.8 Optimization of Gitsbe and Drabme Results

This section provides the results of several runs of the DrugLogics pipeline on the TNBC model [91]. The results of the pipeline can be influenced by some user-defined parameters. We did a series of optimization steps to select which parameters were the best for the TNBC model. The first subsections detail the results obtained with the variation of one main parameter of the pipeline, such as the number of simulations, the number of calibrated nodes, or the origin of the calibration data (CCLE or TCGA). Then, we summarize the results of the best predictions for each cell line and the parameters that we used to obtain them. Lastly, we compare the predictions on the TNBC model with the predictions run on CASCADE 2.0.

The results of this section were obtained with the introduction of 40 topology mutations and no balance mutations, as it was shown for CASCADE 2.0 that a combination of both types of mutations led to less performing predictions. Finally, the BLISS score was used as the main synergy metric. Other results obtained with different types of mutations and another synergy metric will be mentioned later in this report and are presented in the Appendices.
We computed the accuracy of the predictions for each run, and the best runs are the ones which obtained the highest ROC-AUC. As detailed in Section 2.9.4, the ROC curve plots the True Positive Rate (TPR) as a function of the False Positive Rate (FPR). A ROC-AUC of 1 shows that the model is able to perfectly categorize the combinations as synergistic or not. An AUC of 0.5 shows that the model provides random predictions, and is not able to differentiate between synergistic or non-synergistic combinations. If the AUC is lower than 0.5, the model is classifying some of the combinations in the opposite category. Indeed, we expect our predictions to lead to ROC-AUC higher than 0.5.

3.8.1 Evolution of the Performance with an Increasing Number of Simulations

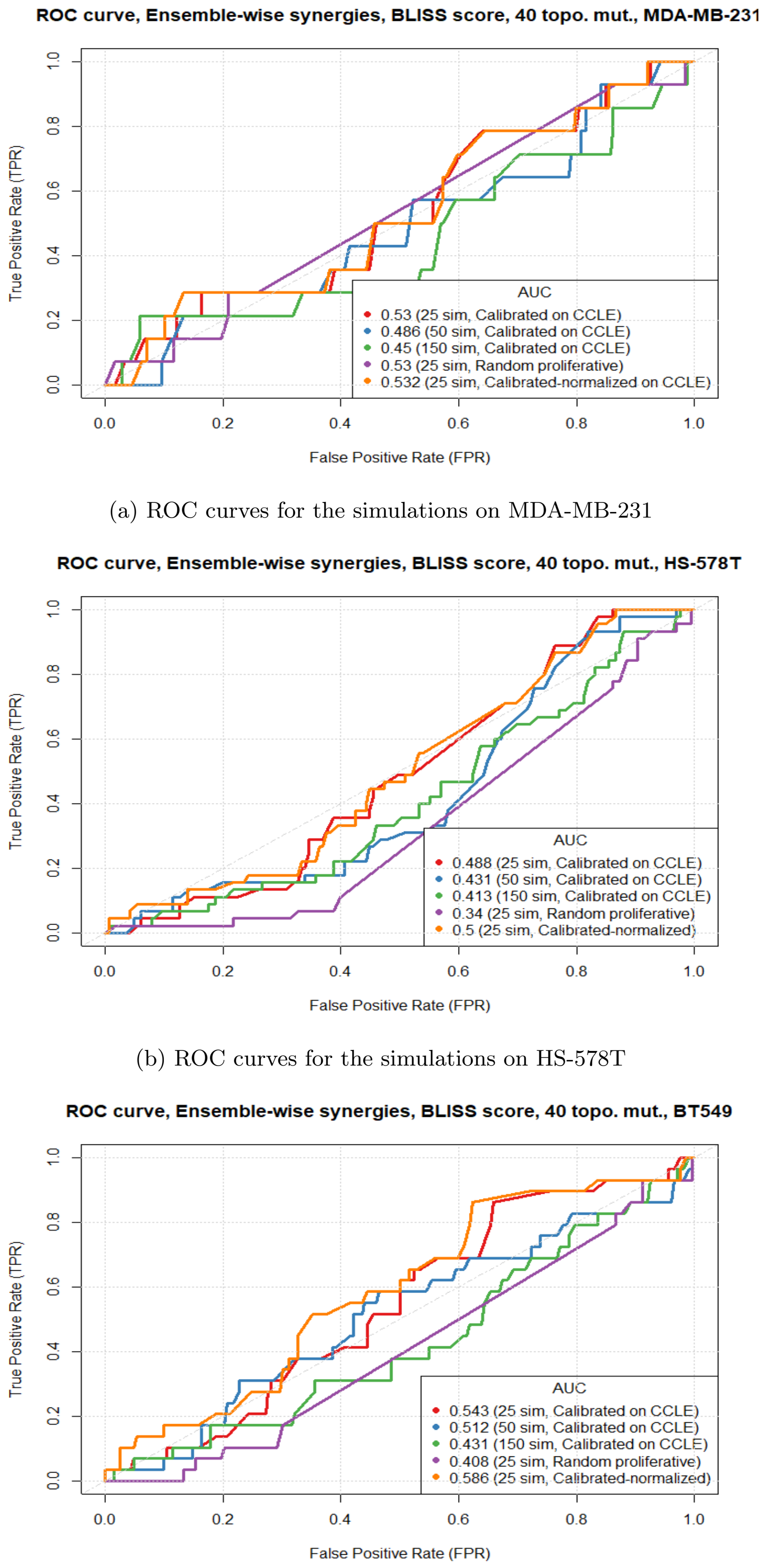
Observations from the DrugLogics group on the simulations ran on CASCADE 1.0 and CASCADE 2.0 led to think that performing no more than 150 simulations with Gitsbe and using 450 models for the synergy predictions was leading to the best performances. For example, CASCADE 2.0 trained on the AGS cell line data provided a ROC-AUC of 0.721 with 50 simulations, 0.712 with 100 simulations, 0.731 with 150 simulations, and it dropped at an AUC of 0.715 for 200 simulations [58].
In this section, we aim to explore if there would generally be a preferred number of simulations for the TNBC model as well, and we compare the results of predictions on the three CCLE cell lines ran in the exact same conditions, except the number of simulations which varies between 25, 50, and 150.
When running Gitsbe, the user can choose the number of models that Gitsbe will save after every simulation. The models saved are the ones with the best fitness score of the simulation, and will be used as a starting point in the next simulations to build the new models. This is the principle of a genetic algorithm which mimics natural evolution processes. Thus, by increasing the number of simulations, we allow Gitsbe to apply more variations to the models to make it evolve and to improve it further. We expect that a number of simulations close to 150 would provide the best results, as observed for CASCADE 2.0.
Here, the calibration was done on 184 nodes of the model.
We also compare these results to the results of predictions on models trained with unperturbed conditions, also called random proliferative training. As this training mode is supposed to reproduce the dynamic of cancers, which is "if we do not perturb the system, it keeps proliferating", we wanted to see if it could produce better results when we normalize the calibrated results with the random results as it sometimes did with CASCADE 2.0 [58, 91]. For each cell line, we decided to run random proliferative-trained predictions by using the number of simulations that gave the best results for the corresponding cell line, and normalize the best results with random proliferative predictions.

Figure 10a shows the predictions performances of the TNBC model on the MDA-MB-231 cell line data. We observe that the highest AUC was obtained for 25 simulations. The random proliferative training gave the same performances than calibrated models, and the calibrated-normalized models did not significantly improve the predictions, but did not worsen it either. In the case of MDA-MB-231, increasing the number of simulations did not improve the results, but it produced lower AUC-ROC. Only the runs of 25 simulations allowed an AUC over 0.5, and thus, performed better than random predictions.

For the case of HS-578T, using 25 simulations and therefore 75 models for the predictions was also the best option, and increasing the number of simulations (and selected models) lowered the prediction performances. However, the AUC for this cell line were below 0.5, as Figure 10b shows. Here, contrary to the predictions on MDA-MB-231 data, the calibrated-normalized predictions provided a slightly higher AUC-ROC (0.5) than for the models calibrated on HS-578T CCLE data.

Lastly, Figure 10c shows the results of the same investigation for the performances of BT549 models. As it was observed before, performing 25 simulations and using the 75 best models to predict drug synergies was the most efficient option (AUC-ROC of approximately 0.54), and the normalization slightly improved these results with a ROC-AUC of 0.586. Overall, this cell line obtained better results than the previous ones, even though we also observe that the AUC of the ROC curves decreases with an increasing number of simulations, and that using 150 simulations gave an AUC lower than 0.5. The random proliferative-trained models had the worst AUC, but helped improving the predictions when used for normalization.

Overall, we observe that models using HS-578T have poorer prediction performances than MDA-MB-231 and BT549 under these conditions. Surprisingly, we did not observe an increase in the performances of the models to predict drug synergies when the number of simulations increased, as it was observed for CASCADE 2.0. The best predictions are thus obtained by running 25 simulations, and the calibrated and calibrated-normalized gave approximately the same performances. However, the performances remain low compared to the predictions of Niederdofer et al. (2020), and we decided to make other parameters vary to optimize the performances. When running further analyses, we prioritized predictions using 25 simulations.



(a) ROC curves for the simulations on MDA-MB-231

(b) ROC curves for the simulations on HS-578T

(c) ROC curves for the simulations on BT549

Figure 10: Evolution of the prediction performances in terms of ROC-AUC for an increasing number of simulations using the three CCLE cell lines calibration and synergy data

3.8.2 Evolution of the Performances and Calibration on Best Determinative Power Nodes

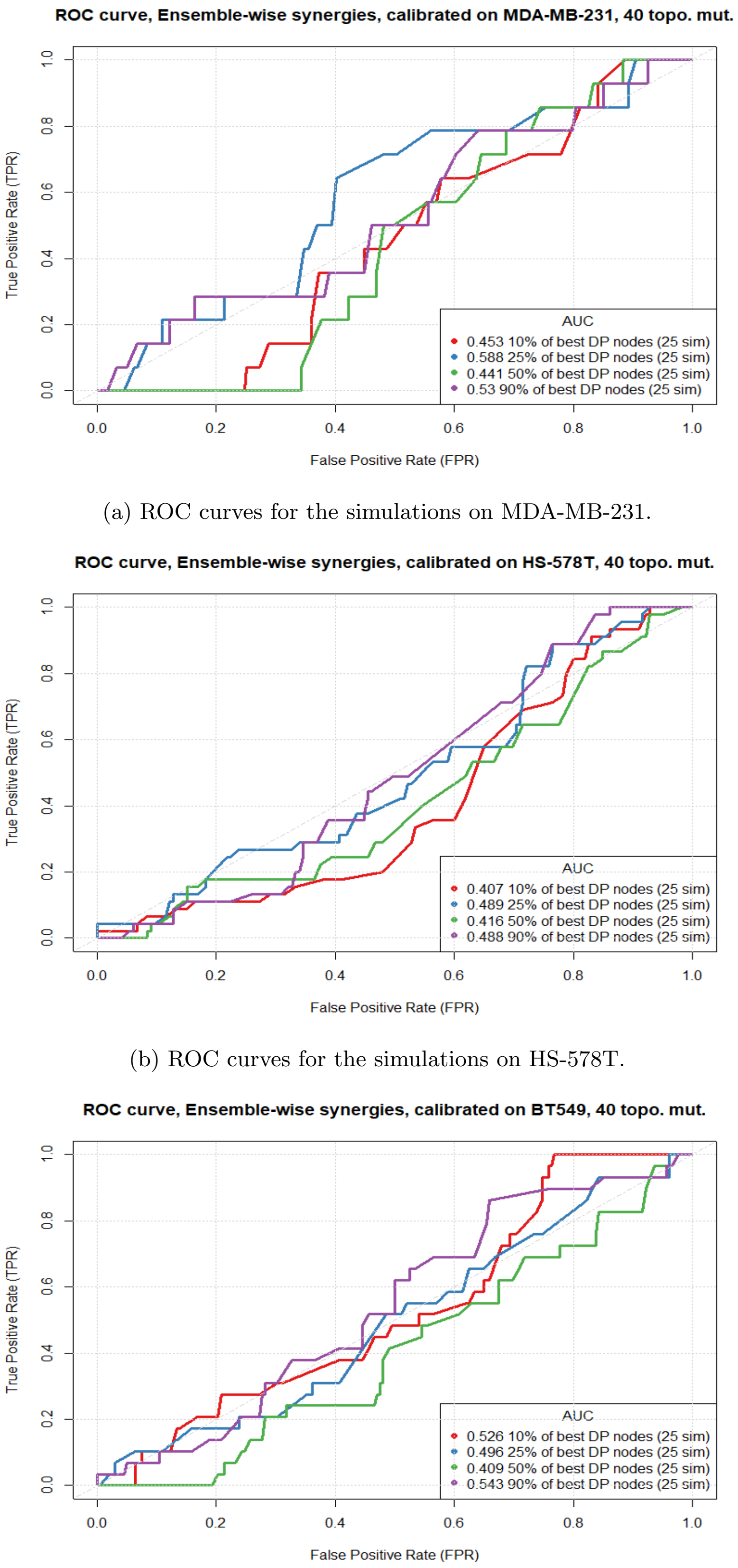
One important parameter on which we could play is the number, and the nature of the nodes that are calibrated during the generation of the Boolean models. As explained in section 2.9.1, we tried to calibrate less nodes, by choosing the most influential ones. We used the Determinative Power (DP) of the nodes, and calibrated only the 10%, 25%, and 50% for which we had RNA expression data and which had the best DP. We also plotted the ROC curve corresponding to the predictions on the models calibrated on 184 nodes, which represents 90% of the total number of nodes. Figure 11 gathers the results of simulations performed on the different cell lines. As mentioned earlier, we are now running 25 simulations, therefore the predictions are based on the 75 best models extracted from the 25 simulations.

The results corresponding to MDA-MB-231 are displayed in Figure 11a. The highest AUC (approximately 0.59) was obtained by calibrating the 25% of nodes with the best DP.

In agreement with what we observed in section 3.8.1, the results on HS-578T were still very low, as shown in Figure 11b. The AUC of two ROC were comparable (approximately 0.49) and were obtained by calibration of 25% of the nodes with the best DP and 90% of the nodes without distinction based on their DP.

Concerning BT549, the best results were obtained with 90% and 10% of calibration (approximately 0.54 and 0.53 respectively).

Overall, we do not observe a drop in the AUC value when we calibrate more than 50% nodes, as it was expected. We generally observe good results with 90% of calibrated nodes, and 25% calibration gave good results for MDA-MB-231 and HS-578T.



(a) ROC curves for the simulations on MDA-MB-231.

(b) ROC curves for the simulations on HS-578T.

(c) ROC curves for the simulations on BT549.

Figure 11: Graphs of the evolution of the performance of the synergy predictions in terms of ROC-AUC with an increasing number of calibrated nodes. Each graph shows the ROC curves for the drug synergy predictions on models calibrated with the 10%, 25%, 50%, and 90% of nodes having the best Determinative Power

As seen in Figure 11, the calibration of less, but more determinant nodes can improve the predictions in some cases. In other cases, such as for BT549 models, it did not improve it. However, the AUC only slightly vary when the number of calibrated nodes vary, and it would be necessary to run more predictions to deduce the effects of calibration on each of these cell lines.

3.8.3 Comparison of Prediction Performances With TCGA Calibration Data

Additionally, the TNBC model was trained on TCGA-BRCA data. This set of TCGA samples was part of the samples used in the omics analysis (2.4.1). These samples were considered relevant calibration data to enable drug synergy predictions that were comparable to what we did with the CCLE cell line data. Indeed, this set of samples clustered well with the four TNBC cell lines, as discussed in section 3.7.2.

The subset of TCGA samples was normalised using counts-per-million, the expression values were then scaled between 0 and 1 (0 for the gene with the lowest expression, and 1 for the gene with the higher expression), and these scaled-expression values were discretised: value lower or equal to 0.5 were set to 0, and values higher than 0.5 were set to 1. The same general parameters were used to generate the models and the predictions (40 topology mutations, Bliss synergy score), but the panel of drugs tested only contained 11 drugs, all involved in combinations observed across the four CCLE cell lines. Figure 12 shows the evolution of the prediction performances with the number of simulations. In agreement with what we observed for the models trained on CCLE data, the best AUC-ROC is obtained by running 25 simulations. However, we observe here that the scores are lower than 0.5. The normalization of the calibrated predictions provided lower performances in terms of ROC-AUC (0.393), while the random proliferative models outperformed the others (0.484). As a comparison, the best predictive results obtained in these conditions using BT549 CCLE data and drug synergies gave an AUC of 0.54 and 0.59 for calibrated-normalized predictions (fig. 14).

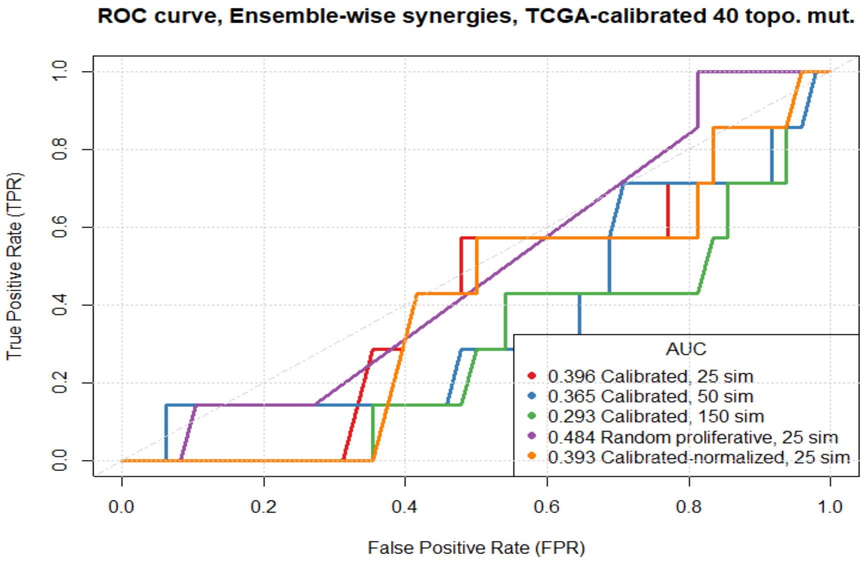


Figure 12: Evolution of the prediction performances in terms of ROC-AUC of models trained on TCGA-BRCA data with an increasing number of simulations.

Similarly to the previous work on the CCLE cell lines' models, we trained fewer nodes of the TNBC model on the TCGA RNA expression data. In five different runs, we chose the 10%, 25%, 50% nodes with the highest DP, another one where we trained all the nodes for which we have TCGA RNA expression data (which represent 90% of the network), and finally, we trained the network on random proliferative data. The ROC curves of these simulations are displayed in Figure 13.

As previously observed, the best scores were obtained with 10 or 25% of the best DP-nodes calibration. Surprisingly, random proliferative training again provided the best score out of all these attempts, whereas the calibrated-normalized predictions gave poor performances (0.393 of ROC-AUC).

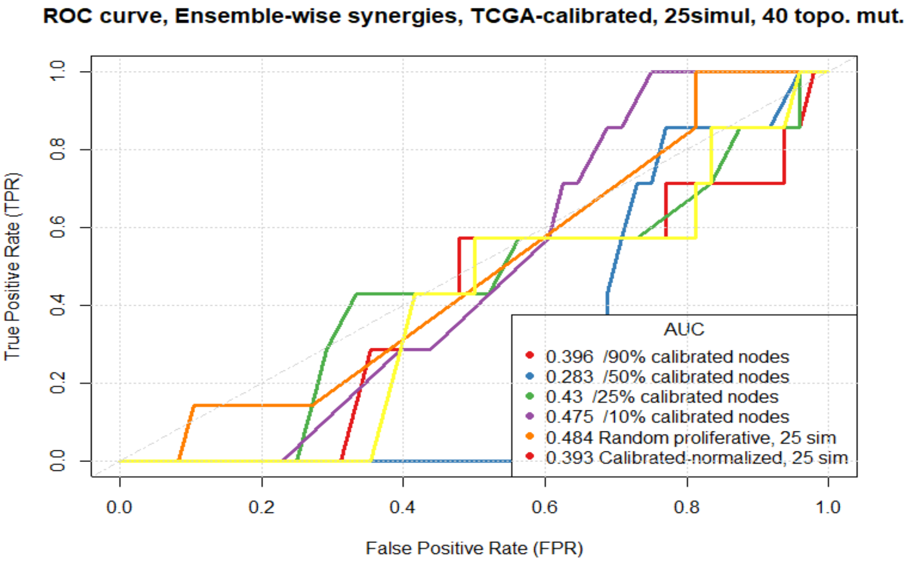


Figure 13: Evolution of the prediction performances of models trained on TCGA-BRCA data with an increasing number of calibrated nodes.

As seen in Figure 13, the performances obtained in this round of simulations are lower than the performances of models calibrated with data from the TNBC cell lines. The reason for this is likely to be related to the fact that we trained the model on TCGA data, and predicted combinations observed on CCLE data. Even if the TCGA samples are more representative of TNBC and its micro-environment than the CCLE cell lines (see section 3.7.2), it is very likely that the drug panel used for the predictions is coupled to the cell lines, and thus, that the synergies observed are specific to the cell lines. The low number of common drug synergies observed between the CCLE cell lines also shows that the cell lines themselves do not respond in the same way in many cases. If basal/ER- cell lines are more representative of the TCGA-BRCA samples, it is probably not enough to conclude that the same drug synergies will be observed across all of them [170].
Vincent et al. (2015) observed that the average gene expression of all the CCLE breast cancer cell lines has a higher correlation to the gene expression of each individual breast cancer subtypes from TCGA [170]. It could be interesting to average the expression of the CCLE cell lines that we used, and use it instead of TCGA data to predict the 18 common drug synergies observed across these cell lines.

3.8.4 Best Predictive Results

In this section, we will highlight the best predictions that we obtained for each cell line, using the basic parameters stated in section 3.7. Even though some parameters remained unchanged, we tried to find the best parameterization for each cell line, meaning that the best results for one cell line were probably obtained with slightly different parameters than for another one. For the best performing runs of the pipeline, we looked at the average fitness score of the best models selected for synergy predictions:

- MDA-MB-231 data-calibrated best models have an average fitness score of 0.59
- HS-578T data-calibrated best models have an average fitness score of 0.59

- BT549 data-calibrated best models have an average fitness score of 0.57.

Interestingly, the average fitness score does not vary significantly between the cell lines. Across the simulations run by the Druglogics group, it has been observed that the fitness score of the models is improved by the discretisation of the training data. Indeed, the fitness score is calculated based on the difference between the state of a node in its steady state (active or inactive, translated by 1 or 0), and its state in the training file. If the states of the nodes in the training file are continuous values ranging from 0 to 1, it is very likely that there will be a high difference (on a scale from 0 to 1) with the steady states of the nodes.
We performed new predictions using the same general parameters as in most of the previous predictions (25 simulations, 40 topology mutations, calibration of 184 nodes), except that we discretised the training data to two different states: 0 if the continuous values were lower than 0.5 or 1 if not. Then, we compared the fitness scores of the different predictions by tracing the plot of the evolution of the fitness score across the generations (see Appendix I). We observe no significant improvement in the fitness score of the models when using discretised data, suggesting that the fitness score remains low for other possible reasons such as that the topology of the network does not translate well the dynamics of TNBC, or that the training data are not adapted. This will be further discussed later.

We investigated the steady states of some very important nodes of the network: the breast cancer-specific receptors ESR1, PGR, and ERBB2 are expected to be inactive in TNBC, and the three most frequently mutated genes TP53, PTEN, and PIK3CA are expected to be active for the latter, and inactive for the two others. Their activity is important in the dynamic of the network, and their local states reflect the ability of the network to represent the signalling events of TNBC. Table 13 summarizes the ratio of activation of these nodes across all the best models for each cell line.
The receptors are expected to be inactive in TNBC. Here, we observe that they are mostly active, except for PGR which tends to be less active (active in less than 50% of the best models) in HS-578T and BT549 models. ESR1 and ERBB2 are also less active in BT549 models (active in 40% and 49% of the models, respectively). TP53 and PTEN tend to be inactive in all cell line-models, but PIK3CA is activated in most of them. Overall, compared to the other cell lines, the activity of these nodes remains lower in BT549 models.

Table 13: Rate of activity of the breast cancer-specific receptors, the frequently mutated genes, and the phenotypic nodes in the best models selected for the predictions

Node	**Activity in MDA-MB-231 models (%)**	**Activity in HS-578T models (%)**	**Activity in BT549 models (%)**
ESR1	96	92	40
PGR	79	43	24
ERBB2	67	64	49
TP53	19	40	33
PIK3CA	93	99	79
PTEN	8	4	19

The main characteristic of TNBC is the absence of the ER (ESR1), PR (PGR), and HER2 (ERBB2) receptors when detected by IHC, that is, the absence of the protein. As seen in section 3.6.1, the training data were able to represent this since the mRNA expression level of the corresponding genes were low. However, it seems like it was not translated through the dynamic of the models, as the steady-states of the receptors are active in a majority of them, except in BT549 models. It is surprising since the expression of these receptors were very close both in BT549 and in HS-578T, especially for ESR1 (0.050 and 0.048, respectively), and that it is eventually active in 92% and 40% of the HS-578T and BT549 models, respectively.
This could be explained by differences in the mRNA expression of other nodes of the network, which influence the expression of ESR1. To counteract this effect, we tried simulations in which the three receptors' states were forced to 0. The ROC curve of the predictions is displayed in Appendix J. The AUC obtained is 0.46, indeed it has no predictive benefits to force the receptors' states to 0.

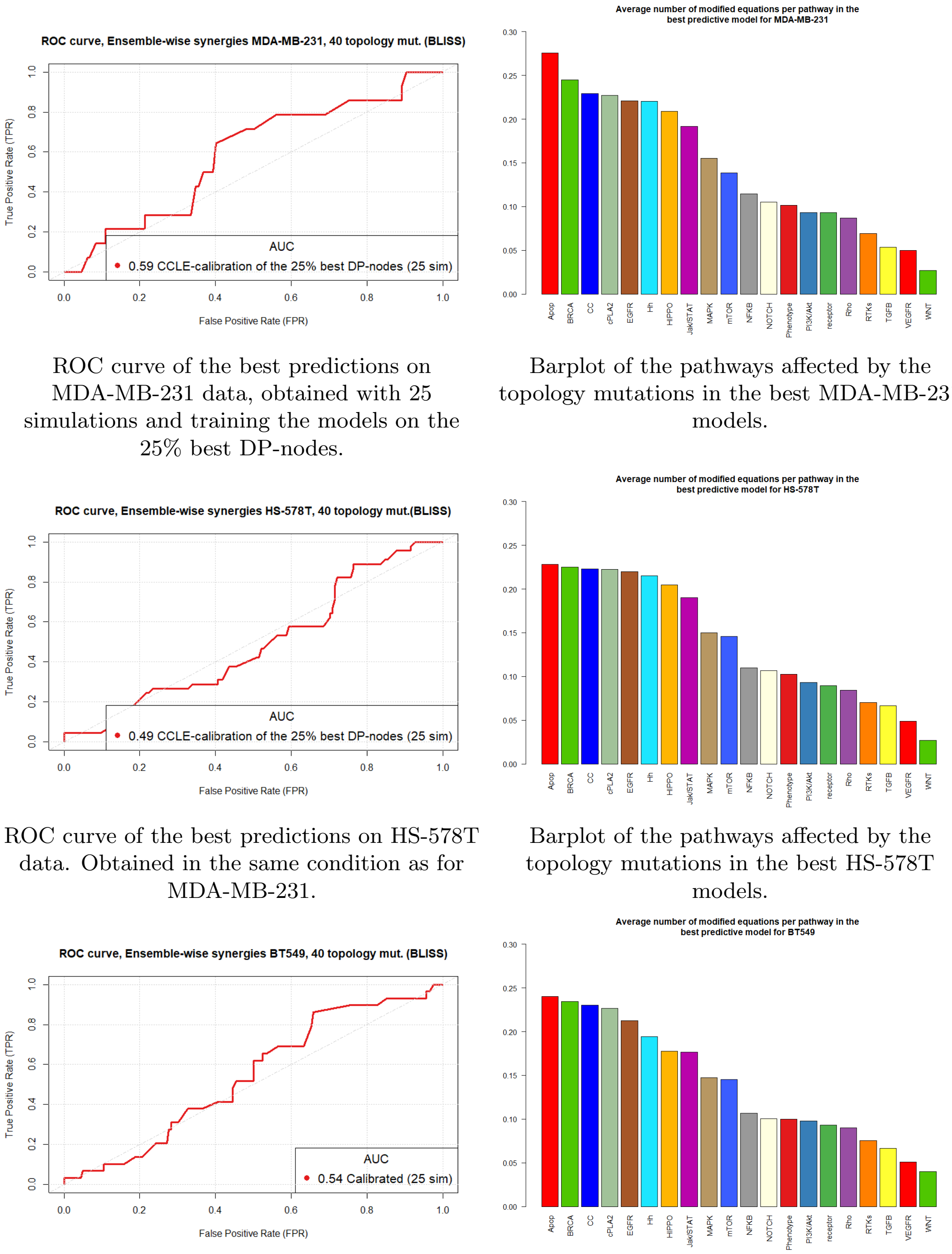
As for the frequently mutated genes, we previously mentioned that TP53 and PTEN were tumour suppressor genes that usually had loss of function mutations in TNBC, whereas PIK3CA had an increased activity. The states of these nodes in most of the best models are in agreement with these characteristics.

The fitness scores obtained for these models and the steady states of ESR1, PGR, and ERBB2 are somewhat disappointing results. Further investigation should be done on the relevance of the calibration data used in this study.
CASCADE 2.0 was trained on protein activity data, unlike the TNBC model which we trained on mRNA expression data. Indeed, Niederdofer et al. (2020) used both protein activity data generated by a pathway activity inference algorithm, and protein activity observations extracted from the literature. Combining two different types of calibration data, they observed that it reduced the chances to use incorrect protein activity data, and that it was beneficial to the models' performances.
These results suggest that the use of mRNA calibration data could induce erroneous dynamics in the system, and that the use of proteomics data would be strongly preferred to infer the expected signalling processes of TNBC.

We computed the ROC curve for each of these predictions (fig. 14). Also, for each cell line, we analysed the effect of the topology mutations included during the simulations. We looked at the modules that were the most affected by these mutations. A barplot representing the average number of mutations per module across the best models selected from the predictions is displayed along with the ROC curves in Figure 14.

The ROC curve and especially the AUC-ROC show if the model is able to correctly predict synergies and not wrongly predict synergistic combinations that are in reality, not synergistic. As we can see in Figure 14, the AUC-ROC is higher than 0.5 for MDA-MB-231 (0.59) and for BT549 (0.54), but slightly lower for HS-578T (0.49). This is what we generally observed in all the other predictions that we ran: HS-578T is less performing than the two other cell lines, and the predictions are generally worse than what could be expected by chance.
The best predictions for MDA-MB-231 and HS-578T were obtained in the same conditions, which were the calibration of only 25% of the network's nodes on the respective cell line's CCLE RNA expression data. These nodes were selected based on their Determinative Power (DP) as detailed in section 2.9.1. As for BT549, the best predictions were obtained by calibration of 184 nodes of the model on BT549 CCLE RNA expression data. This is the only cell line for which we observed that normalizing the calibrated predictions improved the predictions performances. The 184 nodes are the genes of the model for which we had expression data, they were not selected based on any network's property. All of these predictions involved 25 simulations, and 75 best selected models.
Besides the ROC curves, the barplots reveal the tendency of a module of being modified or even dismissed during the introduction of the mutations. We accounted for the size of each module by dividing the total amount of modified equations by the size of the module. Here, we observe that the Apoptosis module is the most modified one, closely followed by the DNA repair/BRCA, Cell Cycle, cPLA2, EGFR, and Hedgehog modules. Overall, the modules are affected in a similar fashion for the three cell lines. Overall, the least modified modules are TGFβ, VEGFR and WNT, with less than 5% of modifications. Interestingly, TGFβ was the most dismissed module in CASCADE 2.0, along with MAPK.

If the predictions made with the MDA-MB-231 and BT549 models are slightly positive, they remain generally low. HS-578T provided overall bad results. There does not seem to be reasonable biological evidence to this, and HS-578T might require other parameters to perform better. Other simulations' results can be found in Appendix K, in which we introduced the use of balance mutations, alone or in combination with topology mutations too. Balance mutations can change the equation of the nodes by changing the impact of their regulators on them (e.g., it can change an "AND" relationship to an "OR"). We also ran simulations using the HSA score as a drug synergy metric. The best AUC that we could obtain for HS-578T was by using a combination of three balance mutations with 40 topology mutations, and the HSA score (the observed synergy files were modified to account for the observed synergies based on HSA score, and can be found on the **Github** repository: *observed_synergies* → *observed_synergies_hsa_name.tab*). The best AUC obtained is 0.61, which is significantly higher than what we previously obtained.



ROC curve of the best predictions on MDA-MB-231 data, obtained with 25 simulations and training the models on the 25% best DP-nodes.

Barplot of the pathways affected by the topology mutations in the best MDA-MB-231 models.

ROC curve of the best predictions on HS-578T data. Obtained in the same condition as for MDA-MB-231.

Barplot of the pathways affected by the topology mutations in the best HS-578T models.

ROC curve of the best predictions on BT549 data. It was obtained by running 25 simulations on 184 calibrated nodes

Barplot of the pathways affected by the topology mutations in the best BT549 models.

Figure 14: Best predictive results for each of the CCLE cell line. On the left, the ROC curve of the prediction. On the right, the barplot of the pathways affected by the topology mutations in the best models. The y-axis is the average rate of mutations that were applied to a module across the best models.

We also improved the results of MDA-MB-231 and BT549 models by using a combination of three balance mutations and topology mutations (40 and 20, respectively), and by computing the HSA score. The ROC curves are displayed in Appendix M and Appendix L, respectively. For MDA-MB-231, the AUC was 0.62, and for BT-549, the best AUC was 0.57. Other predictions on BT549 outperformed the previous results as well, by using varying number of simulations and mutations. These results are close from the ones that we previously obtained, but they still outperformed it.

More investigations could be done using the HSA score and introducing balance mutations.

3.8.5 Prediction Performances with CASCADE 2.0

As the prediction performances of the the TNBC model remain poor, we tried to run predictions on CASCADE 2.0. Figure 15 shows the ROC curves and their respective AUC for predictions on CASCADE 2.0 where 116 nodes were calibrated on the respective cell line data. These predictions were run on the 75 best models generated across 25 simulations, using the same parameters as we used for the TNBC model (40 topology mutations and using the BLISS score as synergy metric). For the three cell lines, the predictions obtained an AUC-ROC≥0.5, and this score was higher for HS-578T and BT549 (0.52 and 0.59, respectively) than the best predictions run on the TNBC model using similar parameters (0.49 and 0.54, respectively). However, the AUC remain close between CASCADE 2.0 and the TNBC model predictions.

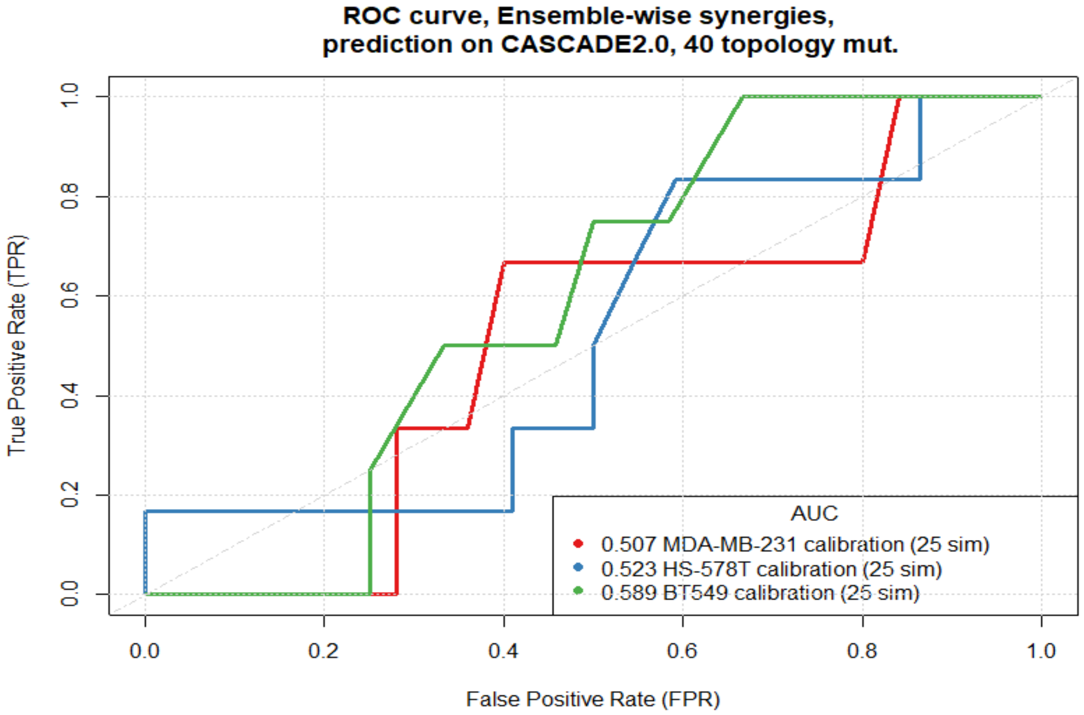


Figure 15: Performance of the predictions on CASCADE 2.0 using CCLE data calibration.

We also computed the average fitness score of the best models for each cell line predictions run on CASCADE 2.0. MDA-MB-231 and HS-578T both obtained a score of 0.58, and BT549 obtained a score of 0.59, which is similar to what we obtained for the predictions on the TNBC model. As a comparison, the fitness scores obtained on CASCADE 2.0 using the AGS cell line as calibration could reach more than 0.8.
The prediction performances of CASCADE 2.0 are poor compared to what we expected given the good results obtained on four cell lines in Niederdofer et al. (2020). However, a notable difference in our predictions is the type of calibration data used. As mentioned in Section 3.8.4, CASCADE 2.0 performed well when calibrated on protein activity data, whereas we calibrated it on mRNA expression data from the three CCLE cell lines.

3.8.6 Predictive Performances of the TNBC Model

In Section 3.7, we first investigated the performances of the TNBC model by predicting drug synergies based on three CCLE cell lines calibration data. With a varying number of simulations and calibrated nodes, we observed that the models surprisingly tend to perform better when trained over 25 simulations, while we expected better performances with more simulations (from 50 to 150). The number of calibrated nodes, contrary to what we expected from previous work in our group, would not play a determinant role even though improvements were observed when calibrating 25% or less of the best DP nodes. A reason for this could be that the determinative power of the nodes is calculated based on the logical equations ruling the network, and that we used the default network (where each equation is in the form of equation 1) as a reference for the DP. The inconsistency of this measure is that the topology of the network has been changed through the simulations by introducing mutations. Indeed, the DP of the nodes could have been calculated again for each network.

As we tried to improve the predictions by using TCGA calibration data, which are supposedly more representative of TNBC tumours, we observed a decrease in the AUC of the ROC curves of the different rounds of predictions. Indeed, the prediction of synergies usually observed in TNBC cell lines are probably not appropriate considering the conditions in which the cell lines are cultured. The CCLE cell lines most likely do not account for the tumour micro-environment as the TCGA samples do, and lead to an epithelial-to-mesenchymal transition. This could be a major difference since most of the TNBC tumours are basal-like and luminal-like, two subtypes exerting phenotypes of epithelial cells [8, 92, 170]. A strong hypothesis from Vincent et al. (2015) would be to use the average value of the mRNA expression of the CCLE cell lines as training data, since they could represent better the heterogeneity of the disease.

Overall, even for the best predictive models, we observed that their fitness score remain low compared to what has been observed for CASCADE 2.0 and the AGS cell line, and that the states of ESR1, PR, and HER2 were mostly active, which is not representative of their state in TNBC. However, forcing the receptors to their inactive state did not improve the predictions, as shown in Appendix J. Discretisation of the training data as it was done for the predictions of CASCADE 2.0 on the AGS cell line did not significantly improve the fitness of the models or the predictions either as shown in AppendixI. As we observed through the barplots of topology mutations, some pathways are preferentially modified. Among them, the DNA repair/BRCA pathway, the cPLA2, the EGFR, or the Hedgehog signalling cascades were added to CASCADE 2.0, suggesting that their topology does not account for the dynamic of the events in TNBC. The apoptosis and cell cycle pathways were also among the most frequently modified ones, whereas they were almost completely made of nodes and interactions already present in CASCADE 2.0, which could mirror inconsistencies in these modules regarding the events in TNBC, and could explain why CASCADE 2.0 did not perform well on the TNBC cell lines as seen in Section 3.8.5.

Furthermore, we observed that HS-578T models usually have lower performances than the other CCLE cell lines. Since HS-578T is classified as basal-like in the PAM50 subtypes, and as mesenchymal-like in Lehmann et al. (2016) classification, exactly like MDA-MB-231 and BT549, there is no clear reason that leads us to think that the topology of the network poorly represents it compared to the others.

Also, the RNA expression data used for the training of the models do not account for other types of genetic events such as epigenetic silencing or other post-transcriptional modifications which can widely vary between the cell lines, and may indicate underlying cellular phenomenon specific to HS-578T cell line [20, 175].

4 Conclusion

The aim of this project was to investigate whether we could develop a logical model that allows accurate representation and system behaviour predictions of the signalling events of triple negative breast cancer (TNBC).

The TNBC model was developed in a pathway-centric manner using modelling methods that have a proven scientific value. Integration of multi-omics data into disease models is widely used today and allows the identification of individual aberrations, but most importantly of more complex processes involving different layers of the genome. Indeed, tumours are caused by the emergence of somatic mutations, which we studied through two types of data: single nucleotide variation and somatic copy number variation of TNBC patients. Both give precious, but limited information on the resulting dynamics leading the development steps of a cancer, from the emergence of the first tumour cells to the formation of metastases. Systems biology relies on the overlay and linkage of many components of a system to unveil underlying processes specific to the disease context.

The middle-out modelling approach deployed to build the TNBC model enabled the definition of modules representing sets of genes, proteins, and complexes that are highly related to specific signalling pathways. In addition to the mutation data, we integrated gene expression and methylation data of TNBC patients to conduct a top-down analysis. We leveraged alterations from the individual omics types first, and assessed the pathways related to these alterations in an over-representation analysis. A literature curation was conducted in parallel in order to verify the findings of the top-down analysis, and to proceed to the bottom-up approach. Indeed, the bottom-up step consisted in the addition of missing components of the pathway-modules based on previous literature findings focused on TNBC.

The resulting model is classified as a CASCADE 2.0-family model, extended for TNBC. It is composed of 20 pathway-modules, 221 nodes and 716 edges. As we tried to limit the size of the network, the nodes and edges were kept only if they were considered central in the dynamics of the pathways. A good proportion of these nodes (52/221, approximately 23%) were found altered in the omics data, suggesting that key components of the signalling events of TNBC can be leveraged through multi-omics analyses. The pathway Enrichment Analysis (EA) also gave reliable results since most of the pathways found in CASCADE 2.0 were over-represented in the altered genes of the TNBC samples too. Other cancer-related pathways that were not present in CASCADE 2.0 but over-represented in the EA often proved to be specifically misregulated in TNBC during the literature curation phase, showing the reliability of EA for unveiling new key pathways to extend existing models.

Using the DrugLogics pipeline [91], the Prior Knowledge Network (PKN) was converted into logical models to account for the dynamic ruling the activity of its nodes, and was used in *in silico* drug synergy predictions with the ambition of being able to correctly predict true synergies, and even new ones.

The TNBC model was transformed into several cell line-specific logical models using mRNA expression data of three TNBC cell lines. The generation of Boolean models with the DrugLogics pipeline was done by calibration of the states of the PKN on the mRNA expression data. This iterative process produced the best possible fitting models, which can be assessed by computing the fitness score of the resulting Boolean models. The fitness score of the best predictive models produced throughout this project remained low for all the cell lines compared to the observations of Niederdofer et al. (2020) concerning CASCADE 2.0 [58]. We tried to improve this fitness score by discretising the calibration data as it was suggested in previous studies of the group, in vain. We also observed that key nodes such as the estrogen receptor (ESR1), progesterone receptor (PGR), and the human epidermal growth factor receptor 2 (ERBB2) do not respect the characteristics of TNBC cases, because their stable states show them active in most models, while they should be inactive. It reflects that the dynamic of the model cannot be further improved by the genetic algorithm to match the biological reality.

We explain it through several possible weaknesses of the project, the first one being that the topology of the network might cause a blockage in the network, making it impossible to better fit the calibration data. However, the introduction of topology mutations during the mutation phase of Gitsbe should have solved this issue. A second hypothesis is that the mRNA expression data of CCLE cell lines are not representative of real TNBC cases. This could have been handled by the use of calibration data retrieved from TCGA, but the difficulty with those data layed on the fact that the drug synergies listed by drug synergy databases were specific to the TNBC cell lines, and not to TNBC patients in general. A last hypothesis to improve the dynamic of the Boolean models is that protein activity data would be strongly preferred over mRNA expression data to train the TNBC model. Indeed, mRNA play an important role in transcriptional processes, but what primes in the dynamic of the cells is the activity of the proteins, which is regulated by many post-transcriptional processes, not reflected in the mRNA expression. Thus, the dynamic of the models could certainly be improved by directly using a combination of inferred or literature-based protein activity data as it was done for CASCADE 2.0 [58].

Despite our efforts to find the best possible parameters to generate the cell line-specific models and predict drug synergies, the prediction performances of the TNBC model generally remained disappointing. The HS-578T cell line typically showed the poorest performances in terms of Area Under the Receiver Operating Characteristic Curve (ROC-AUC). The HS-578T models most often misclassified more combinations than it successfully did. On the other hand, MDA-MB-231 and BT549 provided positive results under certain conditions, but no biological reason was found to explain that they outperformed HS-578T models. However, we do not exclude the hypothesis that complex transcriptional or post-transcriptional processes could happen in HS-578T, whereas they are not reported in the mRNA data used for the calibration of the models.

Areas of improvement for the models probably lie in the refinement of the pathway-modules. Typically, the combinations that were misclassified could be investigated, as it was done in Niederdofer et al. (2020), and the functional processes affected by these drugs should be carefully revised in order to improve the topology of the network. To further improve the topology of the network, the pathways mostly affected by the introduction of topology mutations should also be revised. Indeed, the topology mutations were applied by the algorithm in order to improve the fitness score of the model, which could reveal weaknesses in the network building process. Additionally, efforts should be made in the calibration of the network with protein activity data.

From these three axes of improvements, the Boolean models should better represent the biological reality and dynamics of TNBC, and more investigation could be done on what parameters used by the DrugLogics pipeline provide the best results. We found that 25 simulations generally optimised the prediction performances, and that calibrating a fraction of the nodes with the best determinative power did not have the expected effect on the ROC-AUC, which is not in agreement with the previous findings of our group. Additionally, the normalization of calibrated predictions with the scores obtained by the random proliferative models did not significantly improve the predictive power of the models, as it was expected. Other efforts in the optimisation of the pipeline could be to investigate the use of the HSA score instead of BLISS, since we saw on a few results that it could improve our predictions.

Bibliography

[1] Bruce Alberts. *Molecular biology of the cell.* en. Sixth edition. New York, NY: Garland Science, Taylor and Francis Group, 2015.

[2] Douglas Hanahan and Robert A. Weinberg. 'Hallmarks of Cancer: The Next Generation'. en. In: *Cell* 144.5 (Mar. 2011), pp. 646–674. ISSN: 00928674. DOI: 10.1016/j.cell.2011.02.013. URL: https://linkinghub.elsevier.com/retrieve/pii/S0092867411001279 (visited on 30th Apr. 2022).

[3] Douglas Hanahan and Robert A Weinberg. 'The Hallmarks of Cancer'. en. In: *Cell* 100.1 (Jan. 2000), pp. 57–70. ISSN: 00928674. DOI: 10.1016/S0092-8674(00)81683-9. URL: https://linkinghub.elsevier.com/retrieve/pii/S0092867400816839 (visited on 17th Mar. 2022).

[4] Hyuna Sung et al. 'Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries'. en. In: *CA: A Cancer Journal for Clinicians* 71.3 (May 2021), pp. 209–249. ISSN: 0007-9235, 1542-4863. DOI: 10.3322/caac.21660. URL: https://onlinelibrary.wiley.com/doi/10.3322/caac.21660 (visited on 30th Apr. 2022).

[5] Lindsey A. Torre et al. 'Global cancer statistics, 2012: Global Cancer Statistics, 2012'. en. In: *CA: A Cancer Journal for Clinicians* 65.2 (Mar. 2015), pp. 87–108. ISSN: 00079235. DOI: 10.3322/caac.21262. URL: http://doi.wiley.com/10.3322/caac.21262 (visited on 30th Apr. 2022).

[6] Lindsey A. Torre et al. 'Global Cancer in Women: Burden and Trends'. en. In: *Cancer Epidemiology Biomarkers & Prevention* 26.4 (Apr. 2017), pp. 444–457. ISSN: 1055-9965, 1538-7755. DOI: 10.1158/1055-9965.EPI-16-0858. URL: http://cebp.aacrjournals.org/lookup/doi/10.1158/1055-9965.EPI-16-0858 (visited on 30th Apr. 2022).

[7] The Cancer Genome Atlas Network. 'Comprehensive molecular portraits of human breast tumours'. en. In: *Nature* 490.7418 (Oct. 2012), pp. 61–70. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature11412. URL: http://www.nature.com/articles/nature11412 (visited on 13th Apr. 2021).

[8] Charles M. Perou et al. 'Molecular portraits of human breast tumours'. en. In: *Nature* 406.6797 (Aug. 2000), pp. 747–752. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/35021093. URL: http://www.nature.com/articles/35021093 (visited on 6th Apr. 2021).

[9] Pankaj Kumar and Rupali Aggarwal. 'An overview of triple-negative breast cancer'. en. In: *Archives of Gynecology and Obstetrics* 293.2 (Feb. 2016), pp. 247–269. ISSN: 0932-0067, 1432-0711. DOI: 10.1007/s00404-015-3859-y. URL: http://link.springer.com/10.1007/s00404-015-3859-y (visited on 2nd Nov. 2021).

[10] Aamir Ahmad, ed. *Breast Cancer Metastasis and Drug Resistance: Challenges and Progress.* en. Vol. 1152. Advances in Experimental Medicine and Biology. Cham: Springer International Publishing, 2019. DOI: 10.1007/978-3-030-20301-6. URL: http://link.springer.com/10.1007/978-3-030-20301-6 (visited on 1st May 2022).

[11] Charles M. Perou. 'Molecular Stratification of Triple-Negative Breast Cancers'. en. In: *The Oncologist* 16.S1 (Jan. 2011), pp. 61–70. ISSN: 1083-7159, 1549-490X. DOI: 10.1634/theoncologist.2011-S1-61. URL: https://academic.oup.com/oncolo/article/16/S1/61/6401790 (visited on 1st May 2022).

[12] Xiaoxian Li et al. 'Triple-negative breast cancer has worse overall survival and cause-specific survival than non-triple-negative breast cancer'. en. In: *Breast Cancer Research and Treatment* 161.2 (Jan. 2017), pp. 279–287. ISSN: 0167-6806, 1573-7217. DOI: 10.1007/s10549-016-4059-6. URL: http://link.springer.com/10.1007/s10549-016-4059-6 (visited on 1st May 2022).

[13] Elena Vagia, Devalingam Mahalingam and Massimo Cristofanilli. 'The Landscape of Targeted Therapies in TNBC'. en. In: *Cancers* 12.4 (Apr. 2020), p. 916. ISSN: 2072-6694. DOI: 10.3390/cancers12040916. URL: https://www.mdpi.com/2072-6694/12/4/916 (visited on 6th Apr. 2021).

[14] Ana M. Gonzalez-Angulo et al. 'Incidence and Outcome of *BRCA* Mutations in Unselected Patients with Triple Receptor-Negative Breast Cancer'. en. In: *Clinical Cancer Research* 17.5 (Mar. 2011), pp. 1082–1089. ISSN: 1078-0432, 1557-3265. DOI: 10.1158/1078-0432.CCR-10-2560. URL: http://clincancerres.aacrjournals.org/lookup/doi/10.1158/1078-0432.CCR-10-2560 (visited on 1st May 2022).

[15] Li Zhang et al. 'Androgen Receptor, EGFR, and BRCA1 as Biomarkers in Triple-Negative Breast Cancer: A Meta-Analysis'. en. In: *BioMed Research International* 2015 (2015), pp. 1–12. ISSN: 2314-6133, 2314-6141. DOI: 10.1155/2015/357485. URL: http://www.hindawi.com/journals/bmri/2015/357485/ (visited on 2nd Nov. 2021).

[16] Priyanka Sharma et al. 'Germline BRCA mutation evaluation in a prospective triple-negative breast cancer registry: implications for hereditary breast and/or ovarian cancer syndrome testing'. en. In: *Breast Cancer Research and Treatment* 145.3 (June 2014), pp. 707–714. ISSN: 0167-6806, 1573-7217. DOI: 10.1007/s10549-014-2980-0. URL: http://link.springer.com/10.1007/s10549-014-2980-0 (visited on 1st May 2022).

[17] Hermela Shimelis et al. 'Triple-Negative Breast Cancer Risk Genes Identified by Multigene Hereditary Cancer Panel Testing'. en. In: *JNCI: Journal of the National Cancer Institute* 110.8 (Aug. 2018), pp. 855–862. ISSN: 0027-8874, 1460-2105. DOI: 10.1093/jnci/djy106. URL: https://academic.oup.com/jnci/article/110/8/855/5062996 (visited on 1st Dec. 2021).

[18] Sohrab P. Shah et al. 'The clonal and mutational evolution spectrum of primary triple-negative breast cancers'. en. In: *Nature* 486.7403 (June 2012), pp. 395–399. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature10933. URL: http://www.nature.com/articles/nature10933 (visited on 1st May 2022).

[19] Giampaolo Bianchini et al. 'Triple-negative breast cancer: challenges and opportunities of a heterogeneous disease'. en. In: *Nature Reviews Clinical Oncology* 13.11 (Nov. 2016), pp. 674–690. ISSN: 1759-4774, 1759-4782. DOI: 10.1038/nrclinonc.2016.66. URL: http://www.nature.com/articles/nrclinonc.2016.66 (visited on 28th Oct. 2021).

[20] Brian D. Lehmann et al. 'Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies'. en. In: *Journal of Clinical Investigation* 121.7 (July 2011), pp. 2750–2767. ISSN: 0021-9738. DOI: 10.1172/JCI45014. URL: http://www.jci.org/articles/view/45014 (visited on 25th Jan. 2021).

[21] Brian D. Lehmann et al. 'Refinement of Triple-Negative Breast Cancer Molecular Subtypes: Implications for Neoadjuvant Chemotherapy Selection'. en. In: *PLOS ONE* 11.6 (June 2016). Ed. by Anna Sapino, e0157368. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0157368. URL: https://dx.plos.org/10.1371/journal.pone.0157368 (visited on 16th Feb. 2021).

[22] Kwang-Ai Won and Charles Spruck. 'Triple-negative breast cancer therapy: Current and future perspectives (Review)'. en. In: *International Journal of Oncology* 57.6 (Oct. 2020), pp. 1245–1261. ISSN: 1019-6439, 1791-2423. DOI: 10.3892/ijo.2020.5135. URL: http://www.spandidos-publications.com/10.3892/ijo.2020.5135 (visited on 2nd May 2022).

[23] Peter Schmid et al. 'Atezolizumab and Nab-Paclitaxel in Advanced Triple-Negative Breast Cancer'. en. In: *New England Journal of Medicine* 379.22 (Nov. 2018), pp. 2108–2121. ISSN: 0028-4793, 1533-4406. DOI: 10.1056/NEJMoa1809615. URL: http://www.nejm.org/doi/10.1056/NEJMoa1809615 (visited on 2nd May 2022).

[24] Elan Gorshein et al. 'Durable Response to PD1 Inhibitor Pembrolizumab in a Metastatic, Metaplastic Breast Cancer'. en. In: *Case Reports in Oncology* 14.2 (June 2021), pp. 931–937. ISSN: 1662-6575. DOI: 10.1159/000515510. URL: https://www.karger.com/Article/FullText/515510 (visited on 2nd May 2022).

[25] Mark Robson et al. 'Olaparib for Metastatic Breast Cancer in Patients with a Germline *BRCA* Mutation'. en. In: *New England Journal of Medicine* 377.6 (Aug. 2017), pp. 523–533. ISSN: 0028-4793, 1533-4406. DOI: 10.1056/NEJMoa1706450. URL: http://www.nejm.org/doi/10.1056/NEJMoa1706450 (visited on 2nd May 2022).

[26] Jennifer K. Litton et al. 'Talazoparib in Patients with Advanced Breast Cancer and a Germline *BRCA* Mutation'. en. In: *New England Journal of Medicine* 379.8 (Aug. 2018), pp. 753–763. ISSN: 0028-4793, 1533-4406. DOI: 10.1056/NEJMoa1802905. URL: http://www.nejm.org/doi/10.1056/NEJMoa1802905 (visited on 2nd May 2022).

[27] J Craig Venter et al. 'The Sequence of the Human Genome'. en. In: *THE HUMAN GENOME* 291 (2001), p. 50.

[28] Chandra Shekhar Pareek, Rafal Smoczynski and Andrzej Tretyn. 'Sequencing technologies and genome sequencing'. en. In: *Journal of Applied Genetics* 52.4 (Nov. 2011), pp. 413–435. ISSN: 1234-1983, 2190-3883. DOI: 10.1007/s13353-011-0057-x. URL: http://link.springer.com/10.1007/s13353-011-0057-x (visited on 2nd May 2022).

[29] Jason A. Reuter, Damek V. Spacek and Michael P. Snyder. 'High-Throughput Sequencing Technologies'. en. In: *Molecular Cell* 58.4 (May 2015), pp. 586–597. ISSN: 10972765. DOI: 10.1016/j.molcel.2015.05.004. URL: https://linkinghub.elsevier.com/retrieve/pii/S1097276515003408 (visited on 2nd May 2022).

[30] Yehudit Hasin, Marcus Seldin and Aldons Lusis. 'Multi-omics approaches to disease'. en. In: *Genome Biology* 18.1 (Dec. 2017), p. 83. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1215-1. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1215-1 (visited on 2nd May 2022).

[31] Shancheng Ren et al. 'Integration of Metabolomics and Transcriptomics Reveals Major Metabolic Pathways and Potential Biomarker Involved in Prostate Cancer'. en. In: *Molecular & Cellular Proteomics* 15.1 (Jan. 2016), pp. 154–163. ISSN: 15359476. DOI: 10.1074/mcp.M115.052381. URL: https://linkinghub.elsevier.com/retrieve/pii/S1535947620337142 (visited on 2nd May 2022).

[32] the NCI CPTAC et al. 'Proteogenomic characterization of human colon and rectal cancer'. en. In: *Nature* 513.7518 (Sept. 2014), pp. 382–387. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature13438. URL: http://www.nature.com/articles/nature13438 (visited on 2nd May 2022).

[33] Matteo Bersanelli et al. 'Methods for the integration of multi-omics data: mathematical aspects'. en. In: *BMC Bioinformatics* 17.S2 (Dec. 2016), S15. ISSN: 1471-2105. DOI: 10.1186/s12859-015-0857-9. URL: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0857-9 (visited on 19th May 2021).

[34] Sajib Chakraborty et al. 'Onco-Multi-OMICS Approach: A New Frontier in Cancer Research'. en. In: *BioMed Research International* 2018 (Oct. 2018), pp. 1–14. ISSN: 2314-6133, 2314-6141. DOI: 10.1155/2018/9836256. URL: https://www.hindawi.com/journals/bmri/2018/9836256/ (visited on 2nd May 2022).

[35] Jingwen Yan et al. 'Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data'. en. In: *Briefings in Bioinformatics* (June 2017). ISSN: 1467-5463, 1477-4054. DOI: 10.1093/bib/bbx066. URL: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbx066 (visited on 2nd May 2022).

[36] NCI CPTAC et al. 'Proteogenomics connects somatic mutations to signalling in breast cancer'. en. In: *Nature* 534.7605 (June 2016), pp. 55–62. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature18003. URL: http://www.nature.com/articles/nature18003 (visited on 2nd May 2022).

[37] Manel Esteller. 'CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future'. en. In: *Oncogene* 21.35 (Aug. 2002), pp. 5427–5440. ISSN: 0950-9232, 1476-5594. DOI: 10.1038/sj.onc.1205600. URL: http://www.nature.com/articles/1205600 (visited on 15th Mar. 2022).

[38] M. Esteller. 'Promoter Hypermethylation and BRCA1 Inactivation in Sporadic Breast and Ovarian Tumors'. en. In: *Journal of the National Cancer Institute* 92.7 (Apr. 2000), pp. 564–569. ISSN: 14602105. DOI: 10.1093/jnci/92.7.564. URL: https://academic.oup.com/jnci/article-lookup/doi/10.1093/jnci/92.7.564 (visited on 2nd May 2022).

[39] Hans V Westerhoff and Bernhard O Palsson. 'The evolution of molecular biology into systems biology'. en. In: *Nature Biotechnology* 22.10 (Oct. 2004), pp. 1249–1252. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt1020. URL: http://www.nature.com/articles/nbt1020 (visited on 3rd May 2022).

[40] Anthony Trewavas. 'A Brief History of Systems Biology: *"Every object that biology studies is a system of systems." Francois Jacob (1974).*' en. In: *The Plant Cell* 18.10 (Oct. 2006), pp. 2420–2430. ISSN: 1040-4651, 1532-298X. DOI: 10.1105/tpc.106.042267. URL: https://academic.oup.com/plcell/article/18/10/2420-2430/6115339 (visited on 3rd May 2022).

[41] Leroy Hood et al. 'Systems Biology and New Technologies Enable Predictive and Preventative Medicine'. en. In: *Science* 306.5696 (Oct. 2004), pp. 640–643. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1104635. URL: https://www.science.org/doi/10.1126/science.1104635 (visited on 3rd May 2022).

[42] Hiroaki Kitano. 'Systems Biology: A Brief Overview'. en. In: *Science* 295.5560 (Mar. 2002), pp. 1662–1664. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1069492. URL: https://www.science.org/doi/10.1126/science.1069492 (visited on 3rd May 2022).

[43] Trey Ideker, Timothy Galitski and Leroy Hood. 'A new approach to decoding life: Systems Biology'. en. In: *Annual Review of Genomics and Human Genetics* 2.1 (Sept. 2001), pp. 343–372. ISSN: 1527-8204, 1545-293X. DOI: 10.1146/annurev.genom.2.1.343. URL: https://www.annualreviews.org/doi/10.1146/annurev.genom.2.1.343 (visited on 3rd May 2022).

[44] Charles Auffray and Leroy Hood. 'Editorial: Systems biology and personalized medicine - the future is now'. en. In: *Biotechnology Journal* 7.8 (Aug. 2012), pp. 938–939. ISSN: 18606768. DOI: 10.1002/biot.201200242. URL: https://onlinelibrary.wiley.com/doi/10.1002/biot.201200242 (visited on 3rd May 2022).

[45] Q. Tian, N. D. Price and L. Hood. 'Systems cancer medicine: towards realization of predictive, preventive, personalized and participatory (P4) medicine: Key Symposium: systems cancer medicine'. en. In: *Journal of Internal Medicine* 271.2 (Feb. 2012), pp. 111–121. ISSN: 09546820. DOI: 10.1111/j.1365-2796.2011.02498.x. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2796.2011.02498.x (visited on 3rd May 2022).

[46] Albert-László Barabási and Márton Pósfai. *Network science*. Cambridge: Cambridge University Press, 2016. URL: http://barabasi.com/networksciencebook/.

[47] Albert-László Barabási and Zoltán N. Oltvai. 'Network biology: understanding the cell's functional organization'. en. In: *Nature Reviews Genetics* 5.2 (Feb. 2004), pp. 101–113. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/nrg1272. URL: http://www.nature.com/articles/nrg1272 (visited on 23rd Apr. 2022).

[48] B Dutta et al. 'A network-based, integrative study to identify core biological pathways that drive breast cancer clinical subtypes'. en. In: *British Journal of Cancer* 106.6 (Mar. 2012), pp. 1107–1116. ISSN: 0007-0920, 1532-1827. DOI: 10.1038/bjc.2011.584. URL: http://www.nature.com/articles/bjc2011584 (visited on 14th Mar. 2021).

[49] T. Ideker et al. 'Discovering regulatory and signalling circuits in molecular interaction networks'. en. In: *Bioinformatics* 18.Suppl 1 (July 2002), S233–S240. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/18.suppl_1.S233. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/18.suppl_1.S233 (visited on 3rd May 2022).

[50] Marco Tognetti et al. 'Deciphering the signaling network of breast cancer improves drug sensitivity prediction'. en. In: *Cell Systems* 12.5 (May 2021), 401–418.e12. ISSN: 24054712. DOI: 10.1016/j.cels.2021.04.002. URL: https://linkinghub.elsevier.com/retrieve/pii/S2405471221001113 (visited on 23rd May 2021).

[51] Felix M Weidner et al. 'Capturing dynamic relevance in Boolean networks using graph theoretical measures'. en. In: *Bioinformatics* 37.20 (Oct. 2021). Ed. by Janet Kelso, pp. 3530–3537. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btab277. URL: https://academic.oup.com/bioinformatics/article/37/20/3530/6275260 (visited on 8th Apr. 2022).

[52] Wassim Abou-Jaoudé et al. 'Logical Modeling and Dynamical Analysis of Cellular Networks'. eng. In: *Frontiers in Genetics* 7 (2016), p. 94. ISSN: 1664-8021. DOI: 10.3389/fgene.2016.00094.

[53] René Thomas. 'Boolean Formalization of Genetic Control Circuits'. en. In: (1973), p. 23.

[54] Sui Huang, Ingemar Ernberg and Stuart Kauffman. 'Cancer attractors: A systems view of tumors from a gene network dynamics and developmental perspective'. en. In: *Seminars in Cell & Developmental Biology* 20.7 (Sept. 2009), pp. 869–876. ISSN: 10849521. DOI: 10.1016/j.semcdb.2009.07.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S1084952109001499 (visited on 12th May 2021).

[55] Aurélien Naldi et al. 'Diversity and Plasticity of Th Cell Types Predicted from Regulatory Network Modelling'. en. In: *PLoS Computational Biology* 6.9 (Sept. 2010). Ed. by Richard Bonneau, e1000912. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1000912. URL: https://dx.plos.org/10.1371/journal.pcbi.1000912 (visited on 12th May 2021).

[56] Mihaela T. Matache and Valentin Matache. 'Logical Reduction of Biological Networks to Their Most Determinative Components'. en. In: *Bulletin of Mathematical Biology* 78.7 (July 2016), pp. 1520–1545. ISSN: 0092-8240, 1522-9602. DOI: 10.1007/s11538-016-0193-x. URL: http://link.springer.com/10.1007/s11538-016-0193-x (visited on 7th May 2022).

[57] Bhanwar Lal Puniya et al. 'Systems Perturbation Analysis of a Large-Scale Signal Transduction Model Reveals Potentially Influential Candidates for Cancer Therapeutics'. en. In: *Frontiers in Bioengineering and Biotechnology* 4 (Feb. 2016). ISSN: 2296-4185. DOI: 10.3389/fbioe.2016.00010. URL: http://journal.frontiersin.org/Article/10.3389/fbioe.2016.00010/abstract (visited on 12th May 2021).

[58] Barbara Niederdorfer et al. 'Strategies to Enhance Logic Modeling-Based Cell Line-Specific Drug Synergy Prediction'. en. In: *Frontiers in Physiology* 11 (July 2020), p. 862. ISSN: 1664-042X. DOI: 10.3389/fphys.2020.00862. URL: https://www.frontiersin.org/article/10.3389/fphys.2020.00862/full (visited on 25th Jan. 2021).

[59] Francesca Vitali et al. 'A Network-Based Data Integration Approach to Support Drug Repurposing and Multi-Target Therapies in Triple Negative Breast Cancer'. en. In: *PLOS ONE* 11.9 (Sept. 2016). Ed. by Jianhua Ruan, e0162407. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0162407. URL: https://dx.plos.org/10.1371/journal.pone.0162407 (visited on 19th Apr. 2021).

[60] Åsmund Flobak et al. 'Discovery of Drug Synergies in Gastric Cancer Cells Predicted by Logical Modeling'. eng. In: *PLoS computational biology* 11.8 (Aug. 2015), e1004426. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004426.

[61] Eirini Tsirvouli et al. 'A Middle-Out Modeling Strategy to Extend a Colon Cancer Logical Model Improves Drug Synergy Predictions in Epithelial-Derived Cancer Cell Lines'. eng. In: *Frontiers in Molecular Biosciences* 7 (2020), p. 502573. ISSN: 2296-889X. DOI: 10.3389/fmolb.2020.502573.

[62] Mark A. Jensen et al. 'The NCI Genomic Data Commons as an engine for precision medicine'. en. In: *Blood* 130.4 (July 2017). Publisher: The American Society of Hematology, p. 453. DOI: 10.1182/blood-2017-03-735654. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5533202/ (visited on 21st Mar. 2022).

[63] The Cancer Genome Atlas Research Network et al. 'The Cancer Genome Atlas Pan-Cancer analysis project'. en. In: *Nature Genetics* 45.10 (Oct. 2013), pp. 1113–1120. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng.2764. URL: http://www.nature.com/articles/ng.2764 (visited on 21st Mar. 2022).

[64] *R: The R Project for Statistical Computing.* URL: https://www.r-project.org/ (visited on 28th Apr. 2022).

[65] *RStudio — Open source & professional software for data science teams - RStudio.* URL: https://www.rstudio.com/ (visited on 29th Apr. 2022).

[66] *Bioconductor - Home.* URL: https://bioconductor.org/ (visited on 29th Apr. 2022).

[67] Gordon Guyatt, Harry Shannon and Stephen Walter. 'Hypothesis testing'. en. In: *CAN MED ASSOC J* (1995), p. 6.

[68] Chittaranjan Andrade. 'The P Value and Statistical Significance: Misunderstandings, Explanations, Challenges, and Alternatives'. en. In: *Indian Journal of Psychological Medicine* 41.3 (May 2019), pp. 210–215. ISSN: 0253-7176, 0975-1564. DOI: 10.4103/IJPSYM.IJPSYM_193_19. URL: http://journals.sagepub.com/doi/10.4103/IJPSYM.IJPSYM_193_19 (visited on 10th May 2022).

[69] Yoav Benjamini. 'Discovering the false discovery rate: False Discovery Rate'. en. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.4 (Aug. 2010), pp. 405–416. ISSN: 13697412. DOI: 10.1111/j.1467-9868.2010.00746.x. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2010.00746.x (visited on 8th May 2022).

[70] Mark E. Glickman, Sowmya R. Rao and Mark R. Schultz. 'False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies'. en. In: *Journal of Clinical Epidemiology* 67.8 (Aug. 2014), pp. 850–857. ISSN: 08954356. DOI: 10.1016/j.jclinepi.2014.03.012. URL: https://linkinghub.elsevier.com/retrieve/pii/S0895435614001127 (visited on 10th May 2022).

[71] Tingting Jiang et al. 'Predictors of Chemosensitivity in Triple Negative Breast Cancer: An Integrated Genomic Analysis'. en. In: *PLOS Medicine* 13.12 (Dec. 2016). Ed. by Marc Ladanyi, e1002193. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1002193. URL: https://dx.plos.org/10.1371/journal.pmed.1002193 (visited on 3rd Mar. 2021).

[72] Antonio Colaprico et al. 'TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data'. en. In: *Nucleic Acids Research* 44.8 (May 2016), e71–e71. ISSN: 1362-4962, 0305-1048. DOI: 10.1093/nar/gkv1507. URL: https://academic.oup.com/nar/article/44/8/e71/2465925 (visited on 9th Mar. 2021).

[73] Davide Risso et al. 'GC-Content Normalization for RNA-Seq Data'. en. In: *BMC Bioinformatics* 12.1 (Dec. 2011), p. 480. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-480. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-480 (visited on 3rd Mar. 2022).

[74] Rameen Beroukhim et al. 'Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma'. en. In: *Proceedings of the National Academy of Sciences* 104.50 (Dec. 2007), pp. 20007–20012. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0710052104. URL: https://pnas.org/doi/full/10.1073/pnas.0710052104 (visited on 17th Mar. 2022).

[75] Craig H Mermel et al. 'GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers'. en. In: *Genome Biology* 12.4 (Apr. 2011), R41. ISSN: 1474-760X. DOI: 10.1186/gb-2011-12-4-r41. URL: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2011-12-4-r41 (visited on 2nd Feb. 2021).

[76] Michael S. Lawrence et al. 'Mutational heterogeneity in cancer and the search for new cancer-associated genes'. en. In: *Nature* 499.7457 (July 2013), pp. 214–218. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature12213. URL: http://www.nature.com/articles/nature12213 (visited on 11th Mar. 2021).

[77] Martin Widschwendter and Peter A Jones. 'DNA methylation and breast carcinogenesis'. en. In: *Oncogene* 21.35 (Aug. 2002), pp. 5462–5482. ISSN: 0950-9232, 1476-5594. DOI: 10.1038/sj.onc.1205606. URL: https://www.nature.com/articles/1205606 (visited on 29th Mar. 2022).

[78] Peter W. Laird. 'The power and the promise of DNA methylation markers'. en. In: *Nature Reviews Cancer* 3.4 (Apr. 2003), pp. 253–266. ISSN: 1474-175X, 1474-1768. DOI: 10.1038/nrc1045. URL: https://www.nature.com/articles/nrc1045 (visited on 11th May 2022).

[79] Peter W. Laird. 'Principles and challenges of genome-wide DNA methylation analysis'. en. In: *Nature Reviews Genetics* 11.3 (Mar. 2010), pp. 191–203. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/nrg2732. URL: http://www.nature.com/articles/nrg2732 (visited on 15th Mar. 2022).

[80] Xiaosheng Wang and Chittibabu Guda. 'Integrative exploration of genomic profiles for triple negative breast cancer identifies potential drug targets'. en. In: *Medicine* 95.30 (July 2016), e4321. ISSN: 0025-7974. DOI: 10.1097/MD.0000000000004321. URL: https://journals.lww.com/00005792-201607260-00046 (visited on 22nd Mar. 2021).

[81] Anita Grigoriadis et al. 'Molecular characterisation of cell line models for triple-negative breast cancers'. en. In: *BMC Genomics* 13.1 (2012), p. 619. ISSN: 1471-2164. DOI: 10.1186/1471-2164-13-619. URL: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-13-619 (visited on 14th Mar. 2021).

[82] Pan Du et al. 'Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis'. en. In: *BMC Bioinformatics* 11.1 (Dec. 2010), p. 587. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-587. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-587 (visited on 15th Mar. 2022).

[83] Bijay Jassal et al. 'The reactome pathway knowledgebase'. en. In: *Nucleic Acids Research* (Nov. 2019), gkz1031. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkz1031. URL: https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkz1031/5613674 (visited on 11th May 2022).

[84] Minoru Kanehisa et al. 'KEGG as a reference resource for gene and protein annotation'. en. In: *Nucleic Acids Research* 44.D1 (Jan. 2016), pp. D457–D462. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkv1070. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1070 (visited on 11th May 2022).

[85] Guangchuang Yu and Qing-Yu He. 'ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization'. eng. In: *Molecular bioSystems* 12.2 (Feb. 2016), pp. 477–479. ISSN: 1742-2051. DOI: 10.1039/c5mb00663e.

[86] Tianzhi Wu et al. 'clusterProfiler 4.0: A universal enrichment tool for interpreting omics data'. en. In: *The Innovation* 2.3 (Aug. 2021), p. 100141. ISSN: 26666758. DOI: 10.1016/j.xinn.2021.100141. URL: https://linkinghub.elsevier.com/retrieve/pii/S2666675821000667 (visited on 1st Apr. 2022).

[87] *PubMed*. URL: https://pubmed.ncbi.nlm.nih.gov/ (visited on 23rd Apr. 2022).

[88] Aurélien Naldi et al. 'Logical Modeling and Analysis of Cellular Regulatory Networks With GINsim 3.0'. eng. In: *Frontiers in Physiology* 9 (2018), p. 646. ISSN: 1664-042X. DOI: 10.3389/fphys.2018.00646.

[89] Luana Licata et al. 'SIGNOR 2.0, the SIGnaling Network Open Resource 2.0: 2019 update'. en. In: *Nucleic Acids Research* (Oct. 2019), gkz949. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkz949. URL: https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkz949/5608992 (visited on 23rd Apr. 2022).

[90] Paul Shannon et al. 'Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks'. en. In: *Genome Research* 13.11 (Nov. 2003), pp. 2498–2504. ISSN: 1088-9051. DOI: 10.1101/gr.1239303. URL: http://genome.cshlp.org/lookup/doi/10.1101/gr.1239303 (visited on 23rd Apr. 2022).

[91] Åsmund Flobak et al. *Logical modeling: Combining manual curation and automated parameterization to predict drug synergies*. en. preprint. Systems Biology, July 2021. DOI: 10.1101/2021.06.28.450165. URL: http://biorxiv.org/lookup/doi/10.1101/2021.06.28.450165 (visited on 13th May 2022).

[92] Guanglong Jiang et al. 'Comprehensive comparison of molecular portraits between cell lines and tumors in breast cancer'. en. In: *BMC Genomics* 17.S7 (Aug. 2016), p. 525. ISSN: 1471-2164. DOI: 10.1186/s12864-016-2911-z. URL: http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-2911-z (visited on 5th May 2022).

[93] Reinhard Heckel, Steffen Schober and Martin Bossert. 'Harmonic analysis of Boolean networks: determinative power and perturbations'. en. In: *EURASIP Journal on Bioinformatics and Systems Biology* 2013.1 (Dec. 2013), p. 6. ISSN: 1687-4153. DOI: 10.1186/1687-4153-2013-6. URL: https://bsb-eurasipjournals.springeropen.com/articles/10.1186/1687-4153-2013-6 (visited on 7th May 2022).

[94] Heewon Seo et al. 'SYNERGxDB: an integrative pharmacogenomic portal to identify synergistic drug combinations for precision oncology'. en. In: *Nucleic Acids Research* 48.W1 (July 2020), W494–W501. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkaa421. URL: https://academic.oup.com/nar/article/48/W1/W494/5842189 (visited on 15th May 2022).

[95] Bulat Zagidullin et al. 'DrugComb: an integrative cancer drug combination data portal'. en. In: *Nucleic Acids Research* 47.W1 (July 2019), W43–W51. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkz337. URL: https://academic.oup.com/nar/article/47/W1/W43/5486743 (visited on 7th Apr. 2022).

[96] Hui Liu et al. 'DrugCombDB: a comprehensive database of drug combinations toward the discovery of combinatorial therapy'. en. In: *Nucleic Acids Research* (Oct. 2019), gkz1007. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkz1007. URL: https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkz1007/5609522 (visited on 15th May 2022).

[97] David S Wishart et al. 'DrugBank 5.0: a major update to the DrugBank database for 2018'. en. In: *Nucleic Acids Research* 46.D1 (Jan. 2018), pp. D1074–D1082. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkx1037. URL: http://academic.oup.com/nar/article/46/D1/D1074/4602867 (visited on 7th Apr. 2022).

[98] Wei Zhang et al. 'Effector CD4+ T Cell Expression Signatures and Immune-Mediated Disease Associated Genes'. en. In: *PLoS ONE* 7.6 (June 2012). Ed. by R. Lee Mosley, e38510. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0038510. URL: https://dx.plos.org/10.1371/journal.pone.0038510 (visited on 13th May 2022).

[99] F. Finotello and B. Di Camillo. 'Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis'. en. In: *Briefings in Functional Genomics* 14.2 (Mar. 2015), pp. 130–142. ISSN: 2041-2649, 2041-2657. DOI: 10.1093/bfgp/elu035. URL: https://academic.oup.com/bfg/article-lookup/doi/10.1093/bfgp/elu035 (visited on 3rd May 2022).

[100] Cornelia Braicu et al. 'Aberrant miRNAs expressed in HER-2 negative breast cancers patient'. en. In: *Journal of Experimental & Clinical Cancer Research* 37.1 (Dec. 2018), p. 257. ISSN: 1756-9966. DOI: 10.1186/s13046-018-0920-2. URL: https://jeccr.biomedcentral.com/articles/10.1186/s13046-018-0920-2 (visited on 25th Nov. 2021).

[101] Luciano Cascione et al. 'Integrated MicroRNA and mRNA Signatures Associated with Survival in Triple Negative Breast Cancer'. en. In: *PLoS ONE* 8.2 (Feb. 2013). Ed. by Robert Lafrenie, e55910. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0055910. URL: https://dx.plos.org/10.1371/journal.pone.0055910 (visited on 3rd May 2022).

[102] George A. Calin and Carlo M. Croce. 'MicroRNA signatures in human cancers'. en. In: *Nature Reviews Cancer* 6.11 (Nov. 2006), pp. 857–866. ISSN: 1474-175X, 1474-1768. DOI: 10.1038/nrc1997. URL: http://www.nature.com/articles/nrc1997 (visited on 3rd May 2022).

[103] Jian Yuan Goh et al. 'Chromosome 1q21.3 amplification is a trackable biomarker and actionable target for breast cancer recurrence'. en. In: *Nature Medicine* 23.11 (Nov. 2017), pp. 1319–1330. ISSN: 1078-8956, 1546-170X. DOI: 10.1038/nm.4405. URL: http://www.nature.com/articles/nm.4405 (visited on 13th May 2022).

[104] Claire Wilson and Aditi Kanhere. '8q24.21 Locus: A Paradigm to Link Non-Coding RNAs, Genome Polymorphisms and Cancer'. en. In: *International Journal of Molecular Sciences* 22.3 (Jan. 2021), p. 1094. ISSN: 1422-0067. DOI: 10.3390/ijms22031094. URL: https://www.mdpi.com/1422-0067/22/3/1094 (visited on 13th May 2022).

[105] Yassi Fallah et al. 'MYC-Driven Pathways in Breast Cancer Subtypes'. en. In: *Biomolecules* 7.4 (July 2017), p. 53. ISSN: 2218-273X. DOI: 10.3390/biom7030053. URL: http://www.mdpi.com/2218-273X/7/3/53 (visited on 6th Apr. 2021).

[106] Dai Horiuchi et al. 'MYC pathway activation in triple-negative breast cancer is synthetic lethal with CDK inhibition'. en. In: *Journal of Experimental Medicine* 209.4 (Apr. 2012), pp. 679–696. ISSN: 1540-9538, 0022-1007. DOI: 10.1084/jem.20111512. URL: https://rupress.org/jem/article/209/4/679/41340/MYC-pathway-activation-in-triplenegative-breast (visited on 6th Apr. 2021).

[107] C B Knobbe et al. 'The roles of PTEN in development, physiology and tumorigenesis in mouse models: a tissue-by-tissue survey'. en. In: *Oncogene* 27.41 (Sept. 2008), pp. 5398–5415. ISSN: 0950-9232, 1476-5594. DOI: 10.1038/onc.2008.238. URL: https://www.nature.com/articles/onc2008238 (visited on 14th May 2022).

[108] Shikha Bose et al. 'Allelic loss of chromosome 10q23 is associated with tumor progression in breast carcinomas'. en. In: *Oncogene* 17.1 (July 1998), pp. 123–127. ISSN: 0950-9232, 1476-5594. DOI: 10.1038/sj.onc.1201940. URL: https://www.nature.com/articles/1201940 (visited on 13th May 2022).

[109] Changqing Ma et al. 'Characterization CSMD1 in a large set of primary lung, head and neck, breast and skin cancer tissues'. en. In: *Cancer Biology & Therapy* 8.10 (May 2009), pp. 907–916. ISSN: 1538-4047, 1555-8576. DOI: 10.4161/cbt.8.10.8132. URL: http://www.tandfonline.com/doi/abs/10.4161/cbt.8.10.8132 (visited on 14th May 2022).

[110] Annika Bergman et al. ‘Genome-wide linkage scan for breast cancer susceptibility loci in Swedish hereditary non-BRCA1/2 families: Suggestive linkage to 10q23.32-q25.3’. In: *Genes, Chromosomes and Cancer* 46.3 (2007), pp. 302–309. DOI: https://doi.org/10.1002/gcc.20405. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/gcc.20405. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/gcc.20405.

[111] Tseng-Long Yang et al. ‘High-resolution 19p13.2–13.3 allelotyping of breast carcinomas demonstrates frequent loss of heterozygosity’. In: *Genes, Chromosomes and Cancer* 41.3 (2004), pp. 250–256. DOI: https://doi.org/10.1002/gcc.20080. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/gcc.20080. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/gcc.20080.

[112] Robert T. Lawrence et al. ‘The Proteomic Landscape of Triple-Negative Breast Cancer’. en. In: *Cell Reports* 11.4 (Apr. 2015), pp. 630–644. ISSN: 22111247. DOI: 10.1016/j.celrep.2015.03.050. URL: https://linkinghub.elsevier.com/retrieve/pii/S2211124715003411 (visited on 3rd Feb. 2022).

[113] Katarzyna Szarc vel Szic et al. ‘Epigenetic silencing of triple negative breast cancer hallmarks by Withaferin A’. en. In: *Oncotarget* 8.25 (June 2017), pp. 40434–40453. ISSN: 1949-2553. DOI: 10.18632/oncotarget.17107. URL: https://www.oncotarget.com/lookup/doi/10.18632/oncotarget.17107 (visited on 14th Mar. 2021).

[114] ZhaoYi Wang et al. ‘Identification, cloning, and expression of human estrogen receptor-36, a novel variant of human estrogen receptor-66’. en. In: *Biochemical and Biophysical Research Communications* 336.4 (Nov. 2005), pp. 1023–1027. ISSN: 0006291X. DOI: 10.1016/j.bbrc.2005.08.226. URL: https://linkinghub.elsevier.com/retrieve/pii/S0006291X05019510 (visited on 2nd Nov. 2021).

[115] Sarah Dedeurwaerder et al. ‘DNA methylation profiling reveals a predominant immune component in breast cancers’. en. In: *EMBO Molecular Medicine* 3.12 (Dec. 2011), pp. 726–741. ISSN: 1757-4676, 1757-4684. DOI: 10.1002/emmm.201100801. URL: https://onlinelibrary.wiley.com/doi/10.1002/emmm.201100801 (visited on 13th May 2022).

[116] Yinqi Gao et al. ‘Identification of a DNA Methylation-Based Prognostic Signature for Patients with Triple-Negative Breast Cancer’. en. In: *Medical Science Monitor* 27 (Mar. 2021). ISSN: 1643-3750. DOI: 10.12659/MSM.930025. URL: https://www.medscimonit.com/abstract/index/idArt/930025 (visited on 13th May 2022).

[117] Shavira Narrandes et al. ‘The exploration of contrasting pathways in Triple Negative Breast Cancer (TNBC)’. en. In: *BMC Cancer* 18.1 (Dec. 2018), p. 22. ISSN: 1471-2407. DOI: 10.1186/s12885-017-3939-4. URL: https://bmccancer.biomedcentral.com/articles/10.1186/s12885-017-3939-4 (visited on 21st Oct. 2021).

[118] Stephanie Colón-Marrero et al. ‘Mitotic kinases as drivers of the epithelial-to-mesenchymal transition and as therapeutic targets against breast cancers’. en. In: *Experimental Biology and Medicine* 246.9 (May 2021), pp. 1036–1044. ISSN: 1535-3702, 1535-3699. DOI: 10.1177/1535370221991094. URL: http://journals.sagepub.com/doi/10.1177/1535370221991094 (visited on 22nd Jan. 2022).

[119] P Cappello et al. ‘Role of Nek2 on centrosome duplication and aneuploidy in breast cancer cells’. en. In: *Oncogene* 33.18 (May 2014), pp. 2375–2384. ISSN: 0950-9232, 1476-5594. DOI: 10.1038/onc.2013.183. URL: http://www.nature.com/articles/onc2013183 (visited on 21st Oct. 2021).

[120] Xueqian Gong et al. ‘Aurora A Kinase Inhibition Is Synthetic Lethal with Loss of the *RB1* Tumor Suppressor Gene’. en. In: *Cancer Discovery* 9.2 (Feb. 2019), pp. 248–263. ISSN: 2159-8274, 2159-8290. DOI: 10.1158/2159-8290.CD-18-0469. URL: http://cancerdiscovery.aacrjournals.org/lookup/doi/10.1158/2159-8290.CD-18-0469 (visited on 22nd Jan. 2022).

[121] Daniel G. Hayward and Andrew M. Fry. ‘Nek2 kinase in chromosome instability and cancer’. en. In: *Cancer Letters* 237.2 (June 2006), pp. 155–166. ISSN: 03043835. DOI: 10.1016/j.canlet.2005.06.017. URL: https://linkinghub.elsevier.com/retrieve/pii/S0304383505005616 (visited on 21st Oct. 2021).

[122] Kiyotsugu Yoshida and Yoshio Miki. 'Role of BRCA1 and BRCA2 as regulators of DNA repair, transcription, and cell cycle in response to DNA damage'. en. In: *Cancer Science* 95.11 (Nov. 2004), pp. 866–871. ISSN: 1347-9032, 1349-7006. DOI: 10.1111/j.1349-7006.2004.tb02195.x. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1349-7006.2004.tb02195.x (visited on 3rd Nov. 2021).

[123] Alexander Ring et al. 'CBP/-Catenin/FOXM1 Is a Novel Therapeutic Target in Triple Negative Breast Cancer'. en. In: *Cancers* 10.12 (Dec. 2018), p. 525. ISSN: 2072-6694. DOI: 10.3390/cancers10120525. URL: http://www.mdpi.com/2072-6694/10/12/525 (visited on 26th Oct. 2021).

[124] Chong Gao et al. 'Context-dependent roles of MDMX (MDM4) and MDM2 in breast cancer proliferation and circulating tumor cells'. en. In: *Breast Cancer Research* 21.1 (Dec. 2019), p. 5. ISSN: 1465-542X. DOI: 10.1186/s13058-018-1094-8. URL: https://breast-cancer-research.biomedcentral.com/articles/10.1186/s13058-018-1094-8 (visited on 26th Oct. 2021).

[125] Erik S Knudsen et al. 'RB loss contributes to aggressive tumor phenotypes in MYC-driven triple negative breast cancer'. en. In: *Cell Cycle* 14.1 (Jan. 2015), pp. 109–122. ISSN: 1538-4101, 1551-4005. DOI: 10.4161/15384101.2014.967118. URL: http://www.tandfonline.com/doi/full/10.4161/15384101.2014.967118 (visited on 8th Apr. 2021).

[126] Alexander Ring, Pushpinder Kaur and Julie E. Lang. 'EP300 knockdown reduces cancer stem cell phenotype, tumor growth and metastasis in triple negative breast cancer'. en. In: *BMC Cancer* 20.1 (Dec. 2020), p. 1076. ISSN: 1471-2407. DOI: 10.1186/s12885-020-07573-y. URL: https://bmccancer.biomedcentral.com/articles/10.1186/s12885-020-07573-y (visited on 26th Oct. 2021).

[127] Zi-Ming Zhao et al. 'CCNE1 amplification is associated with poor prognosis in patients with triple negative breast cancer'. en. In: *BMC Cancer* 19.1 (Dec. 2019), p. 96. ISSN: 1471-2407. DOI: 10.1186/s12885-019-5290-4. URL: https://bmccancer.biomedcentral.com/articles/10.1186/s12885-019-5290-4 (visited on 22nd Jan. 2022).

[128] Iris Alejandra García et al. 'Therapeutic opportunities for PLK1 inhibitors: Spotlight on BRCA1-deficiency and triple negative breast cancers'. en. In: *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 821 (May 2020), p. 111693. ISSN: 00275107. DOI: 10.1016/j.mrfmmm.2020.111693. URL: https://linkinghub.elsevier.com/retrieve/pii/S0027510719301265 (visited on 27th Oct. 2021).

[129] Mauricio A. Medina et al. 'Triple-Negative Breast Cancer: A Review of Conventional and Advanced Therapeutic Strategies'. en. In: *International Journal of Environmental Research and Public Health* 17.6 (Mar. 2020), p. 2078. ISSN: 1660-4601. DOI: 10.3390/ijerph17062078. URL: https://www.mdpi.com/1660-4601/17/6/2078 (visited on 6th Sept. 2021).

[130] Mohammad A. Khan et al. 'PI3K/AKT/mTOR pathway inhibitors in triple-negative breast cancer: a review on drug discovery and future challenges'. en. In: *Drug Discovery Today* 24.11 (Nov. 2019), pp. 2181–2191. ISSN: 13596446. DOI: 10.1016/j.drudis.2019.09.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S1359644619303459 (visited on 5th Nov. 2021).

[131] Yin He et al. 'Classification of triple-negative breast cancers based on Immunogenomic profiling'. en. In: *Journal of Experimental & Clinical Cancer Research* 37.1 (Dec. 2018), p. 327. ISSN: 1756-9966. DOI: 10.1186/s13046-018-1002-1. URL: https://jeccr.biomedcentral.com/articles/10.1186/s13046-018-1002-1 (visited on 18th Jan. 2022).

[132] Mark A. Lemmon and Joseph Schlessinger. 'Cell Signaling by Receptor Tyrosine Kinases'. en. In: *Cell* 141.7 (June 2010), pp. 1117–1134. ISSN: 00928674. DOI: 10.1016/j.cell.2010.06.011. URL: https://linkinghub.elsevier.com/retrieve/pii/S0092867410006653 (visited on 15th Apr. 2022).

[133] Ellen Margrethe Haugsten et al. 'Roles of Fibroblast Growth Factor Receptors in Carcinogenesis'. en. In: *Molecular Cancer Research* 8.11 (Nov. 2010), pp. 1439–1452. ISSN: 1541-7786, 1557-3125. DOI: 10.1158/1541-7786.MCR-10-0168. URL: http://mcr.aacrjournals.org/lookup/doi/10.1158/1541-7786.MCR-10-0168 (visited on 15th Apr. 2022).

[134] N Turner et al. 'Integrative molecular profiling of triple negative breast cancers identifies amplicon drivers and potential therapeutic targets'. en. In: *Oncogene* 29.14 (Apr. 2010), pp. 2013–2023. ISSN: 0950-9232, 1476-5594. DOI: 10.1038/onc.2009.489. URL: https://www.nature.com/articles/onc2009489 (visited on 15th Apr. 2022).

[135] Nicholas Turner et al. '*FGFR1* Amplification Drives Endocrine Therapy Resistance and Is a Therapeutic Target in Breast Cancer'. en. In: *Cancer Research* 70.5 (Mar. 2010), pp. 2085–2094. ISSN: 0008-5472, 1538-7445. DOI: 10.1158/0008-5472.CAN-09-3746. URL: http://cancerres.aacrjournals.org/lookup/doi/10.1158/0008-5472.CAN-09-3746 (visited on 18th Oct. 2021).

[136] Rachel Sharpe et al. 'FGFR Signaling Promotes the Growth of Triple-Negative and Basal-Like Breast Cancer Cell Lines Both *In Vitro* and *In Vivo*'. en. In: *Clinical Cancer Research* 17.16 (Aug. 2011), pp. 5275–5286. ISSN: 1078-0432, 1557-3265. DOI: 10.1158/1078-0432.CCR-10-2727. URL: http://clincancerres.aacrjournals.org/lookup/doi/10.1158/1078-0432.CCR-10-2727 (visited on 18th Oct. 2021).

[137] Chee Leong Cheng et al. 'Expression of FGFR1 is an independent prognostic factor in triple-negative breast cancer'. en. In: *Breast Cancer Research and Treatment* 151.1 (May 2015), pp. 99–111. ISSN: 0167-6806, 1573-7217. DOI: 10.1007/s10549-015-3371-x. URL: http://link.springer.com/10.1007/s10549-015-3371-x (visited on 18th Oct. 2021).

[138] Nicole J. Chew et al. 'FGFR3 signaling and function in triple negative breast cancer'. en. In: *Cell Communication and Signaling* 18.1 (Dec. 2020), p. 13. ISSN: 1478-811X. DOI: 10.1186/s12964-019-0486-4. URL: https://biosignaling.biomedcentral.com/articles/10.1186/s12964-019-0486-4 (visited on 16th Sept. 2021).

[139] Amber M. Johnson, Emily K. Kleczko and Raphael A. Nemenoff. 'Eicosanoids in Cancer: New Roles in Immunoregulation'. en. In: *Frontiers in Pharmacology* 11 (Nov. 2020), p. 595498. ISSN: 1663-9812. DOI: 10.3389/fphar.2020.595498. URL: https://www.frontiersin.org/articles/10.3389/fphar.2020.595498/full (visited on 11th Oct. 2021).

[140] W L Smith. 'The eicosanoids and their biochemical mechanisms of action'. en. In: *Biochemical Journal* 259.2 (Apr. 1989), pp. 315–324. ISSN: 0264-6021, 1470-8728. DOI: 10.1042/bj2590315. URL: https://portlandpress.com/biochemj/article/259/2/315/24577/The-eicosanoids-and-their-biochemical-mechanisms (visited on 15th Apr. 2022).

[141] Francesco Caiazza, Brian J. Harvey and Warren Thomas. 'Cytosolic Phospholipase A2 Activation Correlates with HER2 Overexpression and Mediates Estrogen-Dependent Breast Cancer Cell Growth'. en. In: *Molecular Endocrinology* 24.5 (May 2010), pp. 953–968. ISSN: 0888-8809, 1944-9917. DOI: 10.1210/me.2009-0293. URL: https://academic.oup.com/mend/article/24/5/953/2706140 (visited on 21st Apr. 2021).

[142] Renata Nascimento Gomes, Souza Felipe da Costa and Alison Colquhoun. 'Eicosanoids and cancer'. eng. In: *Clinics (Sao Paulo, Brazil)* 73.suppl 1 (Aug. 2018), e530s. ISSN: 1980-5322. DOI: 10.6061/clinics/2018/e530s.

[143] Mousumi Majumder et al. 'EP4 as a Therapeutic Target for Aggressive Human Breast Cancer'. en. In: *International Journal of Molecular Sciences* 19.4 (Mar. 2018), p. 1019. ISSN: 1422-0067. DOI: 10.3390/ijms19041019. URL: http://www.mdpi.com/1422-0067/19/4/1019 (visited on 7th Oct. 2021).

[144] Michael G. Chiorazzo et al. 'Detection and Differentiation of Breast Cancer Sub-Types using a cPLA2 Activatable Fluorophore'. eng. In: *Scientific Reports* 9.1 (Apr. 2019), p. 6122. ISSN: 2045-2322. DOI: 10.1038/s41598-019-41626-y.

[145] Atasi De Chatterjee et al. 'Arachidonic Acid Induces the Migration of MDA-MB-231 Cells by Activating Raft-associated Leukotriene B4 Receptors'. en. In: *Clinical Cancer Drugs* 5.1 (Nov. 2018), pp. 28–41. ISSN: 2212697X. DOI: 10.2174/2212697X05666180418145601. URL: http://www.eurekaselect.com/161381/article (visited on 11th Oct. 2021).

[146] Suzanne A. Eccles. 'The epidermal growth factor receptor/Erb-B/HER family in normal and malignant breast biology'. en. In: *The International Journal of Developmental Biology* 55.7-8-9 (2011), pp. 685–696. ISSN: 0214-6282. DOI: 10.1387/ijdb.113396se. URL: http://www.intjdevbiol.com/paper.php?doi=113396se (visited on 18th Apr. 2022).

[147] Alan Wells. *EGF receptor*. eng. 1999. (Visited on 2nd Nov. 2021).

[148] R.I Nicholson, J.M.W Gee and M.E Harper. 'EGFR and cancer prognosis'. en. In: *European Journal of Cancer* 37 (Sept. 2001), pp. 9–15. ISSN: 09598049. DOI: 10.1016/S0959-8049(01)00231-3. URL: https://linkinghub.elsevier.com/retrieve/pii/S0959804901002313 (visited on 18th Apr. 2022).

[149] Heae Surng Park et al. 'High EGFR gene copy number predicts poor outcome in triple-negative breast cancer'. en. In: *Modern Pathology* 27.9 (Sept. 2014), pp. 1212–1222. ISSN: 0893-3952, 1530-0285. DOI: 10.1038/modpathol.2013.251. URL: http://www.nature.com/articles/modpathol2013251 (visited on 2nd Nov. 2021).

[150] D. Ford et al. 'Genetic Heterogeneity and Penetrance Analysis of the BRCA1 and BRCA2 Genes in Breast Cancer Families'. en. In: *The American Journal of Human Genetics* 62.3 (Mar. 1998), pp. 676–689. ISSN: 00029297. DOI: 10.1086/301749. URL: https://linkinghub.elsevier.com/retrieve/pii/S0002929707638488 (visited on 18th Apr. 2022).

[151] Jill J. J. Geenen et al. 'PARP Inhibitors in the Treatment of Triple-Negative Breast Cancer'. en. In: *Clinical Pharmacokinetics* 57.4 (Apr. 2018), pp. 427–437. ISSN: 0312-5963, 1179-1926. DOI: 10.1007/s40262-017-0587-4. URL: http://link.springer.com/10.1007/s40262-017-0587-4 (visited on 1st Dec. 2021).

[152] Xupeng Bai et al. 'Triple-negative breast cancer therapeutic resistance: Where is the Achilles' heel?' en. In: *Cancer Letters* 497 (Jan. 2021), pp. 100–111. ISSN: 03043835. DOI: 10.1016/j.canlet.2020.10.016. URL: https://linkinghub.elsevier.com/retrieve/pii/S0304383520305425 (visited on 26th Oct. 2021).

[153] I. Zachary. 'VEGF signalling: integration and multi-tasking in endothelial cell biology'. en. In: *Biochemical Society Transactions* 31.6 (Dec. 2003), pp. 1171–1177. ISSN: 0300-5127, 1470-8752. DOI: 10.1042/bst0311171. URL: https://portlandpress.com/biochemsoctrans/article/31/6/1171/64507/VEGF-signalling-integration-and-multitasking-in (visited on 1st Nov. 2021).

[154] Domenico Ribatti et al. 'Angiogenesis and Antiangiogenesis in Triple-Negative Breast cancer'. en. In: *Translational Oncology* 9.5 (Oct. 2016), pp. 453–457. ISSN: 19365233. DOI: 10.1016/j.tranon.2016.07.002. URL: https://linkinghub.elsevier.com/retrieve/pii/S1936523316300717 (visited on 19th Apr. 2022).

[155] Gagan K. Gupta et al. 'Perspectives on Triple-Negative Breast Cancer: Current Treatment Strategies, Unmet Needs, and Potential Targets for Future Therapies'. en. In: *Cancers* 12.9 (Aug. 2020), p. 2392. ISSN: 2072-6694. DOI: 10.3390/cancers12092392. URL: https://www.mdpi.com/2072-6694/12/9/2392 (visited on 1st Nov. 2021).

[156] M. V. Giuli et al. 'Notch Signaling Activation as a Hallmark for Triple-Negative Breast Cancer Subtype'. en. In: *Journal of Oncology* 2019 (July 2019), pp. 1–15. ISSN: 1687-8450, 1687-8469. DOI: 10.1155/2019/8707053. URL: https://www.hindawi.com/journals/jo/2019/8707053/ (visited on 17th Nov. 2021).

[157] Fokhrul Hossain et al. 'Notch Signaling Regulates Mitochondrial Metabolism and NF-B Activity in Triple-Negative Breast Cancer Cells via IKK-Dependent Non-canonical Pathways'. en. In: *Frontiers in Oncology* 8 (Dec. 2018), p. 575. ISSN: 2234-943X. DOI: 10.3389/fonc.2018.00575. URL: https://www.frontiersin.org/article/10.3389/fonc.2018.00575/full (visited on 17th Nov. 2021).

[158] Jianmin Zhang et al. 'Genetic variations in the Hippo signaling pathway and breast cancer risk in African American women in the AMBER Consortium'. en. In: *Carcinogenesis* 37.10 (Oct. 2016), pp. 951–956. ISSN: 0143-3334, 1460-2180. DOI: 10.1093/carcin/bgw077. URL: https://academic.oup.com/carcin/article-lookup/doi/10.1093/carcin/bgw077 (visited on 21st Apr. 2022).

[159] Damiano Cosimo Rigiracciolo et al. 'IGF-1/IGF-1R/FAK/YAP Transduction Signaling Prompts Growth Effects in Triple-Negative Breast Cancer (TNBC) Cells'. en. In: *Cells* 9.4 (Apr. 2020), p. 1010. ISSN: 2073-4409. DOI: 10.3390/cells9041010. URL: https://www.mdpi.com/2073-4409/9/4/1010 (visited on 24th Jan. 2022).

[160] Zhonghao Wang et al. 'Regulation of Hippo signaling and triple negative breast cancer progression by an ubiquitin ligase RNF187'. en. In: *Oncogenesis* 9.3 (Mar. 2020), p. 36. ISSN: 2157-9024. DOI: 10.1038/s41389-020-0220-5. URL: http://www.nature.com/articles/s41389-020-0220-5 (visited on 24th Jan. 2022).

[161] Rosamaria Lappano, Yves Jacquot and Marcello Maggiolini. 'GPCR Modulation in Breast Cancer'. en. In: *International Journal of Molecular Sciences* 19.12 (Dec. 2018), p. 3840. ISSN: 1422-0067. DOI: 10.3390/ijms19123840. URL: http://www.mdpi.com/1422-0067/19/12/3840 (visited on 16th Sept. 2021).

[162] Araceli García-Martínez et al. 'Hedgehog gene expression patterns among intrinsic subtypes of breast cancer: Prognostic relevance'. en. In: *Pathology - Research and Practice* 223 (July 2021), p. 153478. ISSN: 03440338. DOI: 10.1016/j.prp.2021.153478. URL: https://linkinghub.elsevier.com/retrieve/pii/S0344033821001394 (visited on 24th Jan. 2022).

[163] Syeda Kiran Riaz et al. 'Involvement of hedgehog pathway in early onset, aggressive molecular subtypes and metastatic potential of breast cancer'. en. In: *Cell Communication and Signaling* 16.1 (Dec. 2018), p. 3. ISSN: 1478-811X. DOI: 10.1186/s12964-017-0213-y. URL: https://biosignaling.biomedcentral.com/articles/10.1186/s12964-017-0213-y (visited on 24th Jan. 2022).

[164] Luca Grieco et al. 'Integrative Modelling of the Influence of MAPK Network on Cancer Cell Fate Decision'. en. In: *PLoS Computational Biology* 9.10 (Oct. 2013). Ed. by Satoru Miyano, e1003286. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003286. URL: https://dx.plos.org/10.1371/journal.pcbi.1003286 (visited on 12th May 2021).

[165] Richard M. Neve et al. 'A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes'. en. In: *Cancer Cell* 10.6 (Dec. 2006), pp. 515–527. ISSN: 15356108. DOI: 10.1016/j.ccr.2006.10.008. URL: https://linkinghub.elsevier.com/retrieve/pii/S153561080600314X (visited on 4th May 2022).

[166] Laxmi Silwal-Pandit, Anita Langerød and Anne-Lise Børresen-Dale. 'TP53 Mutations in Breast and Ovarian Cancer'. en. In: *Cold Spring Harbor Perspectives in Medicine* 7.1 (Jan. 2017), a026252. ISSN: 2157-1422. DOI: 10.1101/cshperspect.a026252. URL: http://perspectivesinmedicine.cshlp.org/lookup/doi/10.1101/cshperspect.a026252 (visited on 4th May 2022).

[167] Virginia Álvarez-Garcia et al. 'Mechanisms of PTEN loss in cancer: It's all about diversity'. en. In: *Seminars in Cancer Biology* 59 (Dec. 2019), pp. 66–79. ISSN: 1044579X. DOI: 10.1016/j.semcancer.2019.02.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S1044579X18300592 (visited on 4th May 2022).

[168] J. Pascual and N.C. Turner. 'Targeting the PI3-kinase pathway in triple-negative breast cancer'. en. In: *Annals of Oncology* 30.7 (July 2019), pp. 1051–1060. ISSN: 09237534. DOI: 10.1093/annonc/mdz133. URL: https://linkinghub.elsevier.com/retrieve/pii/S0923753419312396 (visited on 4th May 2022).

[169] Ronglai Shen, Adam B. Olshen and Marc Ladanyi. 'Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis'. en. In: *Bioinformatics* 25.22 (Nov. 2009), pp. 2906–2912. ISSN: 1460-2059, 1367-4803. DOI: 10.1093/bioinformatics/btp543. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp543 (visited on 18th May 2021).

[170] Krista Marie Vincent, Scott D. Findlay and Lynne Marie Postovit. 'Assessing breast cancer cell lines as tumour models by comparison of mRNA expression profiles'. en. In: *Breast Cancer Research* 17.1 (Dec. 2015), p. 114. ISSN: 1465-542X. DOI: 10.1186/s13058-015-0613-0. URL: http://breast-cancer-research.biomedcentral.com/articles/10.1186/s13058-015-0613-0 (visited on 5th May 2022).

[171] Dvir Aran, Marina Sirota and Atul J. Butte. 'Systematic pan-cancer analysis of tumour purity'. en. In: *Nature Communications* 6.1 (Dec. 2015), p. 8971. ISSN: 2041-1723. DOI: 10.1038/ncomms9971. URL: http://www.nature.com/articles/ncomms9971 (visited on 5th May 2022).

[172] J F Lyons et al. 'Discovery of a novel Raf kinase inhibitor.' en. In: *Endocrine-related cancer* (Sept. 2001), pp. 219–225. ISSN: 1351-0088. DOI: 10.1677/erc.0.0080219. URL: https://erc.bioscientifica.com/view/journals/erc/8/3/11566613.xml (visited on 5th May 2022).

[173] Scott M. Wilhelm et al. 'BAY 43-9006 Exhibits Broad Spectrum Oral Antitumor Activity and Targets the RAF/MEK/ERK Pathway and Receptor Tyrosine Kinases Involved in Tumor Progression and Angiogenesis'. en. In: *Cancer Research* 64.19 (Oct. 2004), pp. 7099–7109. ISSN: 0008-5472, 1538-7445. DOI: 10.1158/0008-5472.CAN-04-1443. URL: http://cancerres.aacrjournals.org/lookup/doi/10.1158/0008-5472.CAN-04-1443 (visited on 5th May 2022).

[174] *Home — International Centre for Kinase Profiling*. URL: https://www.kinase-screen.mrc.ac.uk/ (visited on 6th May 2022).

[175] Milene Pereira Moreira et al. ‘Phenotypic, structural, and ultrastructural analysis of triple-negative breast cancer cell lines and breast cancer stem cell subpopulation’. en. In: *European Biophysics Journal* 48.7 (Oct. 2019), pp. 673–684. ISSN: 0175-7571, 1432-1017. DOI: 10.1007/s00249-019-01393-0. URL: http://link.springer.com/10.1007/s00249-019-01393-0 (visited on 7th May 2022).

Appendix

A Versions of the Bioinformatics Tools Used

Versions of the bioinformatics tools used in the project

Tool	Version
R	4.1.3
RStudio	1.4.1103
Bioconductor	3.13
Gistic	2.0
MutSigCV	1.3.5
Reactome	79 ; 80
KEGG	99.1 ; 100-102.0
GINsim	3.0
Cytoscape	3.9.0
SIGNOR	2.0
DrugBank	5.0
DrugCombDB	1.0
Synergx	1.0
DrugComb	1.5

B TCGA Barcodes of the TNBC Patients

173 TNBC patient's barcodes from TCGA-BRCA project available on the GDC repository

TCGA-A1-A0SK	TCGA-AN-A0FJ	TCGA-BH-A0B9	TCGA-E2-A1LH
TCGA-A1-A0SO	TCGA-AN-A0FL	TCGA-BH-A0BG	TCGA-E2-A1LI
TCGA-A1-A0SP	TCGA-AN-A0FX	TCGA-BH-A0BL	TCGA-E2-A1LK
TCGA-A2-A04P	TCGA-AN-A0G0	TCGA-BH-A0E0	TCGA-E2-A1LL
TCGA-A2-A04T	TCGA-AN-A0XU	TCGA-BH-A0E6	TCGA-E2-A1LS
TCGA-A2-A04U	TCGA-AO-A03U	TCGA-BH-A0WA	TCGA-E2-A573
TCGA-A2-A0CM	TCGA-AO-A0J2	TCGA-BH-A18G	TCGA-E2-A574
TCGA-A2-A0D0	TCGA-AO-A0J4	TCGA-BH-A18Q	TCGA-E9-A1N8
TCGA-A2-A0D2	TCGA-AO-A0J6	TCGA-BH-A18T	TCGA-E9-A1ND
TCGA-A2-A0EQ	TCGA-AO-A124	TCGA-BH-A18V	TCGA-E9-A1RH
TCGA-A2-A0SX	TCGA-AO-A128	TCGA-BH-A1EW	TCGA-E9-A22G
TCGA-A2-A0T0	TCGA-AO-A129	TCGA-BH-A1F6	TCGA-E9-A244
TCGA-A2-A0T2	TCGA-AO-A12F	TCGA-BH-A1FC	TCGA-E9-A5FL
TCGA-A2-A0YM	TCGA-D8-A1JG	TCGA-BH-A42U	TCGA-EW-A1OV
TCGA-A2-A1G6	TCGA-AO-A1KR	TCGA-C8-A12K	TCGA-EW-A1OW
TCGA-A2-A3XS	TCGA-AQ-A04J	TCGA-C8-A12V	TCGA-EW-A1P1
TCGA-A2-A3XT	TCGA-AQ-A54N	TCGA-C8-A131	TCGA-EW-A1P4
TCGA-A2-A3XU	TCGA-AR-A0TP	TCGA-C8-A134	TCGA-EW-A1P7
TCGA-A2-A3XX	TCGA-AR-A0TS	TCGA-C8-A1HJ	TCGA-EW-A1P8
TCGA-A2-A3XY	TCGA-AR-A0U4	TCGA-C8-A26X	TCGA-EW-A1PB
TCGA-A7-A0CE	TCGA-AR-A1AH	TCGA-C8-A26Y	TCGA-EW-A1PH
TCGA-A7-A0DA	TCGA-AR-A1AI	TCGA-C8-A27B	TCGA-GI-A2C9
TCGA-A7-A26F	TCGA-AR-A1AJ	TCGA-C8-A3M7	TCGA-GM-A2DB
TCGA-A7-A26G	TCGA-AR-A1AO	TCGA-D8-A13Z	TCGA-GM-A2DD
TCGA-A7-A26I	TCGA-AR-A1AQ	TCGA-D8-A142	TCGA-GM-A2DF
TCGA-A7-A4SD	TCGA-AR-A1AR	TCGA-D8-A143	TCGA-GM-A2DH
TCGA-A7-A4SE	TCGA-AR-A1AY	TCGA-D8-A147	TCGA-GM-A2DI
TCGA-A7-A5ZV	TCGA-AR-A256	TCGA-D8-A1JF	TCGA-GM-A3XL

TCGA-A7-A6VV	TCGA-AR-A2LH	TCGA-D8-A1JL	TCGA-HN-A2NL
TCGA-A7-A6VW	TCGA-AR-A2LR	TCGA-D8-A1XK	TCGA-LL-A5YO
TCGA-A7-A6VY	TCGA-AR-A5QQ	TCGA-D8-A1XQ	TCGA-LL-A73Y
TCGA-A8-A07C	TCGA-B6-A0I2	TCGA-D8-A27F	TCGA-LL-A740
TCGA-A8-A07O	TCGA-B6-A0I6	TCGA-D8-A27H	TCGA-OL-A5D6
TCGA-A8-A07R	TCGA-B6-A0IE	TCGA-D8-A27M	TCGA-OL-A5D7
TCGA-A8-A08R	TCGA-B6-A0IQ	TCGA-E2-A14N	TCGA-OL-A5RW
TCGA-AC-A2BK	TCGA-B6-A0RE	TCGA-E2-A14R	TCGA-OL-A66I
TCGA-AC-A2QH	TCGA-B6-A0RS	TCGA-E2-A150	TCGA-OL-A66P
TCGA-AC-A2QJ	TCGA-B6-A0RT	TCGA-E2-A158	TCGA-OL-A6VO
TCGA-AC-A6IW	TCGA-B6-A0RU	TCGA-E2-A159	TCGA-OL-A97C
TCGA-AC-A8OQ	TCGA-B6-A0WX	TCGA-E2-A1AZ	TCGA-PL-A8LV
TCGA-AN-A04D	TCGA-B6-A402	TCGA-E2-A1B6	TCGA-PL-A8LZ
TCGA-AN-A0AL	TCGA-BH-A0AV	TCGA-E2-A1L7	TCGA-S3-AA10
TCGA-AN-A0AR	TCGA-BH-A0B3	TCGA-E2-A1LG	TCGA-S3-AA15
TCGA-AN-A0AT			

C Model Outputs

Model outputs used to compute the global output response of the models in unperturbed conditions, or after perturbation.

Name	**Weight**
CASP8	-1
CASP9	-1
FOXO_f	-1
RSK_f	1
CCND1	1
MYC	1

D Drug Panels of the CCLE Cell Lines

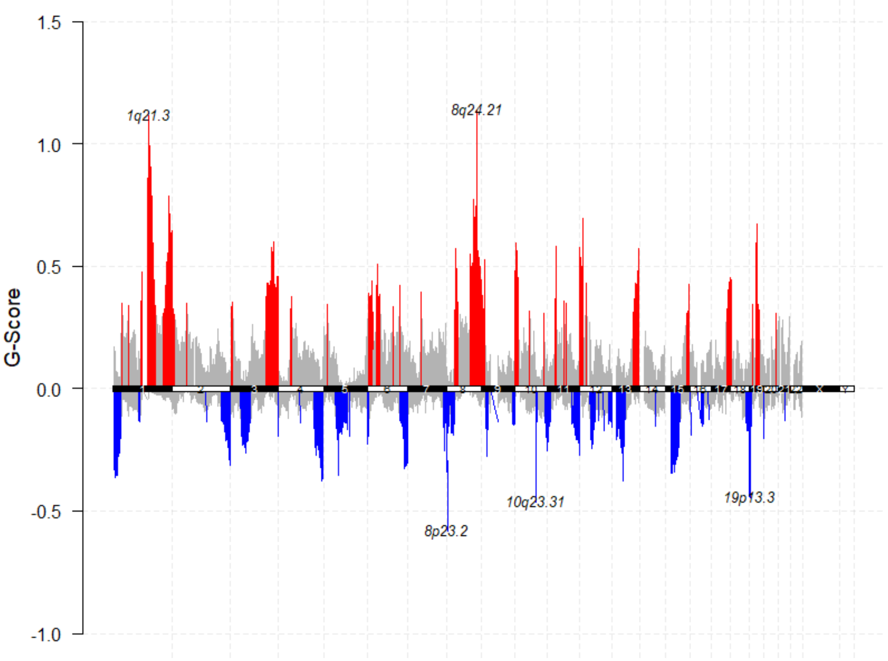
Drugs tested for each cell line, their activation or inhibitory effect, and their targets.

MDA-MB-231			**HS-578T**			**BT549**		
Drug	**Effect**	**Targets**	**Drug**	**Effect**	**Targets**	**Drug**	**Effect**	**Targets**
Nilotinib	inhibits	ABL1	Nilotinib	inhibits	ABL1	Nilotinib	inhibits	ABL1
Imatinib	inhibits	ABL1	Imatinib	inhibits	ABL1	Imatinib	inhibits	ABL1
Dasatinib	inhibits	ABL1 SRC	Dasatinib	inhibits	ABL1 SRC	Dasatinib	inhibits	ABL1 SRC
Paclitaxel	inhibits	BCL2	Paclitaxel	inhibits	BCL2	Paclitaxel	inhibits	BCL2
Lapatinib	inhibits	EGFR ERBB2	Lapatinib	inhibits	EGFR ERBB2	Lapatinib	inhibits	EGFR ERBB2
Erlotinib	inhibits	EGFR	Erlotinib	inhibits	EGFR	Erlotinib	inhibits	EGFR
Vandetanib	inhibits	EGFR	Vandetanib	inhibits	EGFR	Vandetanib	inhibits	EGFR
Tamoxifen	inhibits	ESR1 PRKCA PRKCD	Tamoxifen	inhibits	ESR1 PRKCA PRKCD	Tamoxifen	inhibits	ESR1 PRKCA PRKCD
Fulvestrant	inhibits	ESR1	Fulvestrant	inhibits	ESR1	Fulvestrant	inhibits	ESR1
Raloxifene	activates	ESR1	Raloxifene	activates	ESR1	Raloxifene	activates	ESR1
Pazopanib	inhibits	VEGFR2 FGF_f	Pazopanib	inhibits	VEGFR2 FGF_f	Pazopanib	inhibits	VEGFR2 FGF_f
Sorafenib	inhibits	VEGFR2 FGFR RAF_f	Sorafenib	inhibits	VEGFR2 FGFR RAF_f	Sorafenib	inhibits	VEGFR2 FGFR RAF_f
Thalidomide	inhibits	NFKB_f	Thalidomide	inhibits	NFKB_f	Thalidomide	inhibits	NFKB_f
Ruxolitinib	inhibits	JAK_f	Ruxolitinib	inhibits	JAK_f	Ruxolitinib	inhibits	JAK_f

Celecoxib	inhibits	PDPK1	Celecoxib	inhibits	PDPK1	Celecoxib	inhibits	PDPK1
Vemurafenib	inhibits	RAF_f	Vemurafenib	inhibits	RAF_f	Vemurafenib	inhibits	RAF_f
Vismodegib	inhibits	SMO	Vismodegib	inhibits	SMO	Vismodegib	inhibits	SMO
Sunitinib	inhibits	VEGFR2	Sunitinib	inhibits	VEGFR2	Sunitinib	inhibits	VEGFR2
Gefitinib	inhibits	EGFR	Gefitinib	inhibits	EGFR	Cabozantinib	inhibits	VEGFR2
Afatinib	inhibits	EGFR ; ERBB2	Topotecan	inhibits	TOP1	Topotecan	inhibits	TOP1
Axitinib	inhibits	VEGFR2	Axitinib	inhibits	VEGFR2	Axitinib	inhibits	VEGFR2
Sonidegib	inhibits	SMO				Sonidegib	inhibits	SMO
Alisertib	inhibits	AURKA						

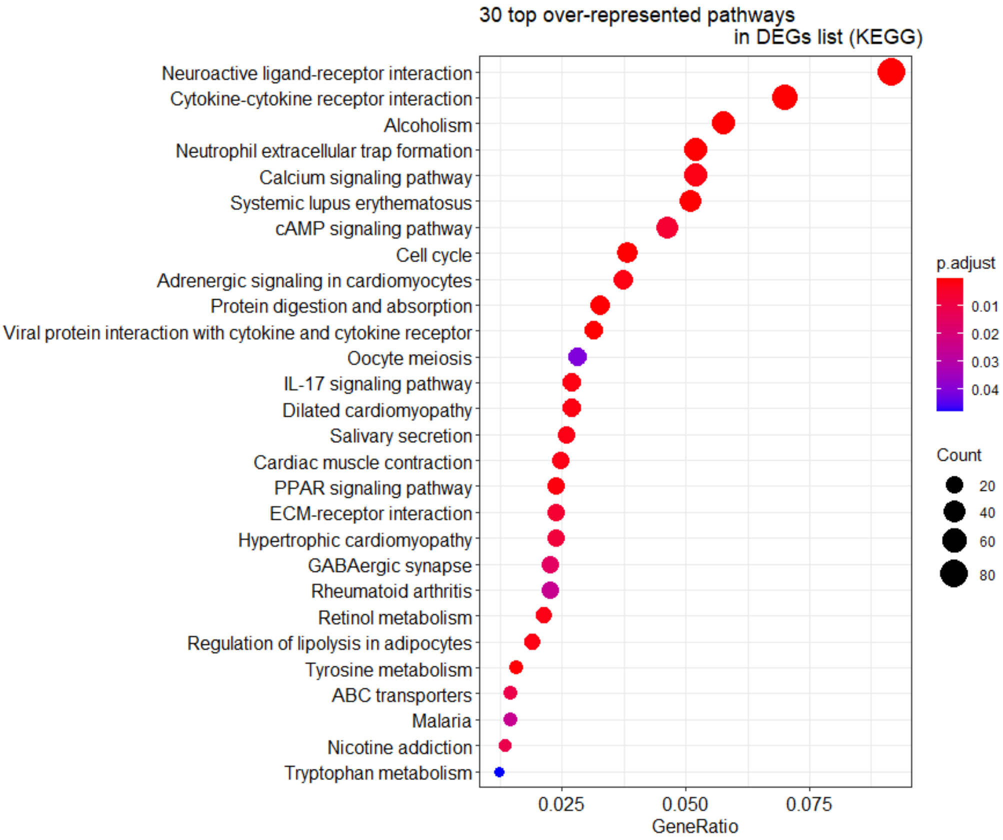
E Chromplot of the Somatic Copy Number Alterations

Chromplot of the significantly amplified and deleted regions of the genome. The y-axis represents the G-score associated to each peak. The G-score is calculated based on the amplitude of the aberration and its frequency across the samples.

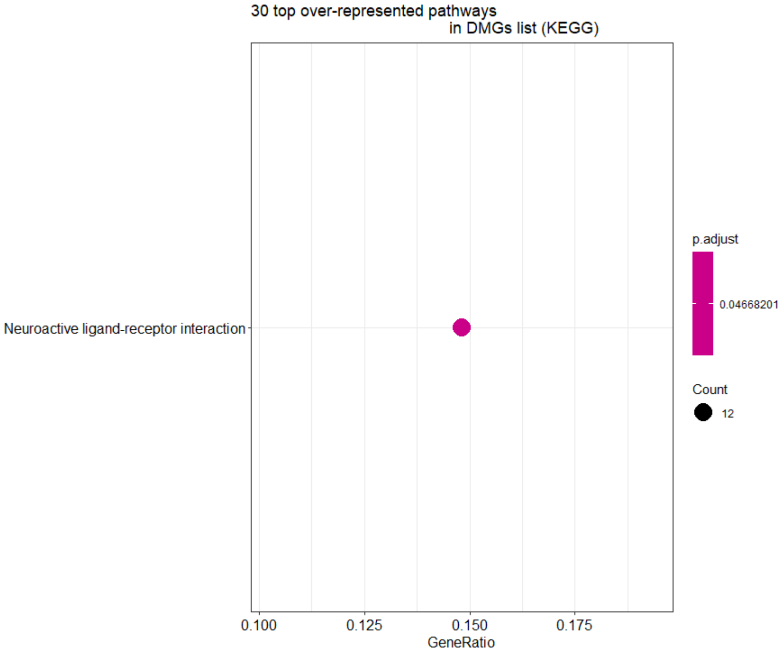


F Most Over-Represented Pathways Against KEGG

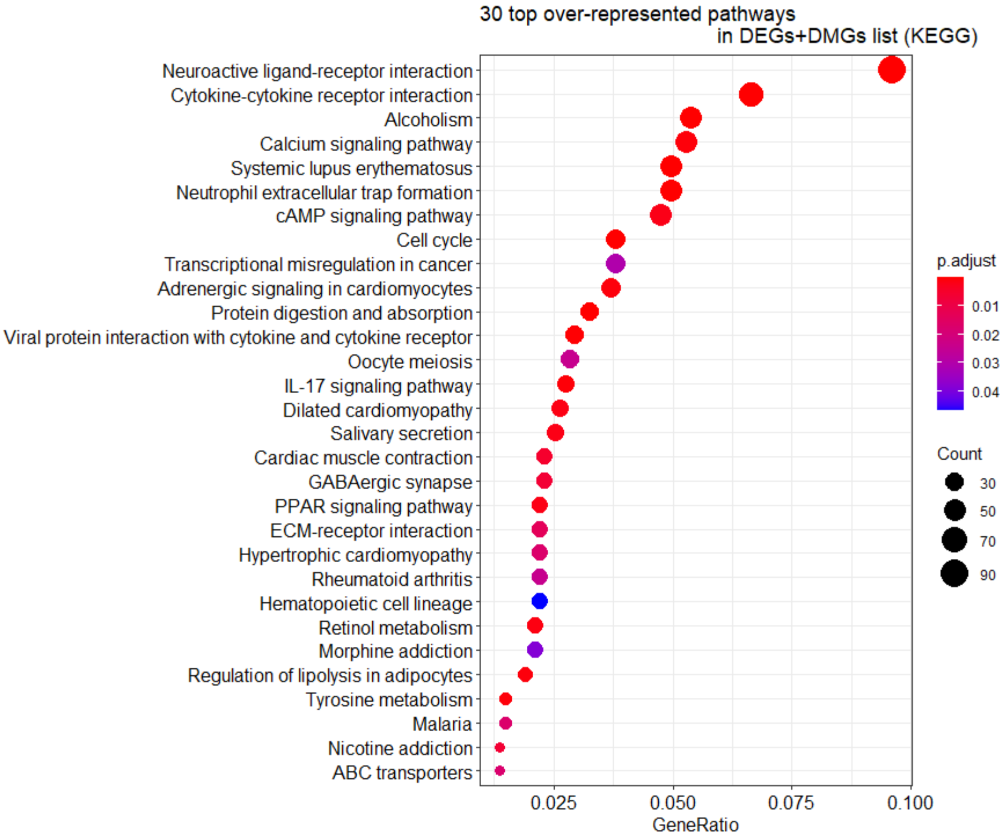
Dotplots of the 30 top over-represented pathways obtained by enrichment analysis against KEGG database. On the y-axis, the pathways are represented. The x-axis correspond to the ratio of genes that belong to the pathway of interest over the total number of genes in the corresponding gene list. The size of the dots represent the FDR-adjusted p-value.



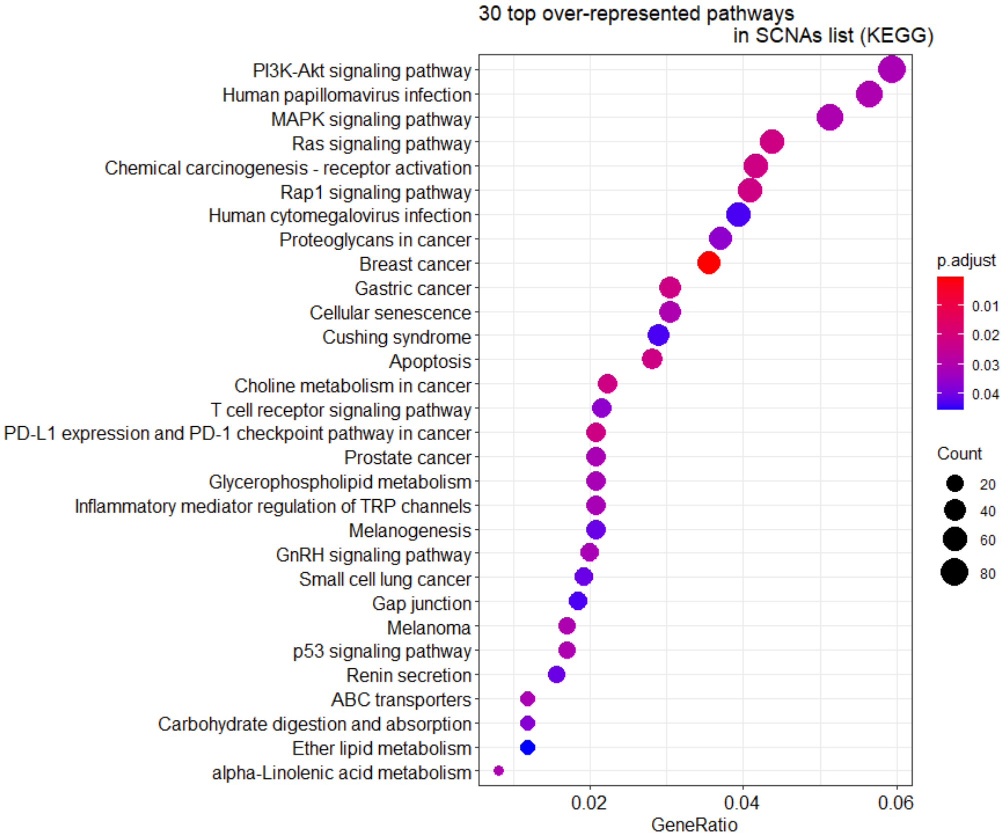
Results of the EA on DEGs against KEGG



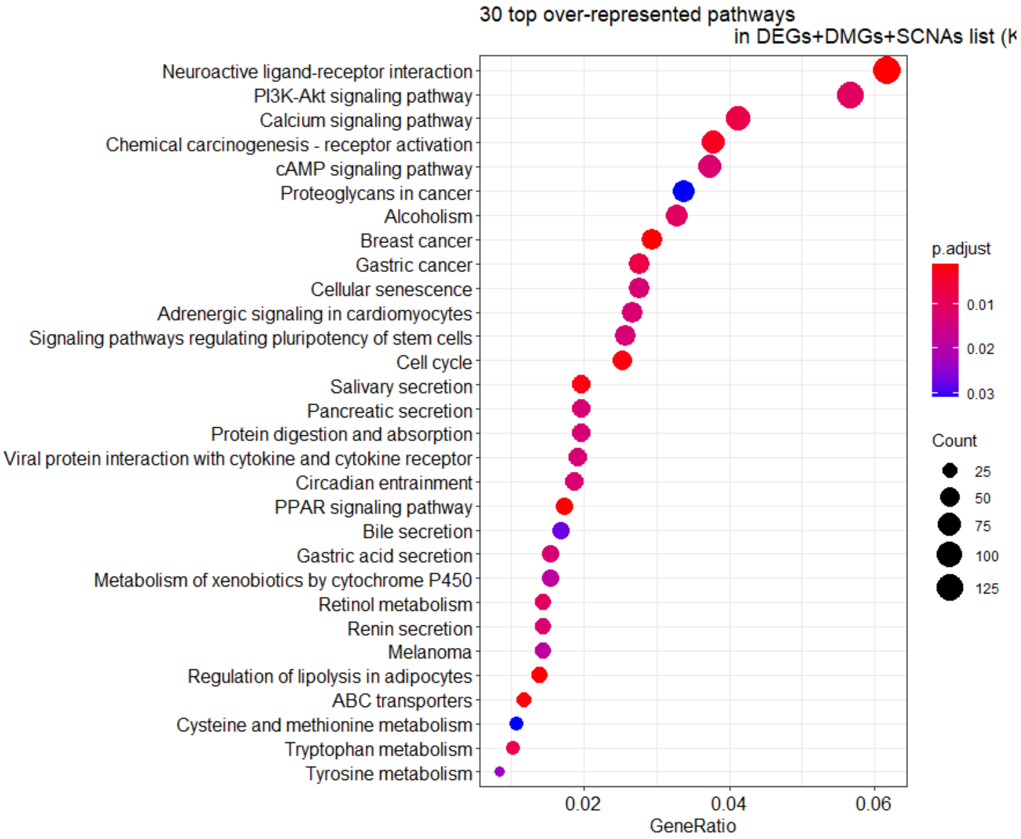
Results of the EA on DMGs against KEGG



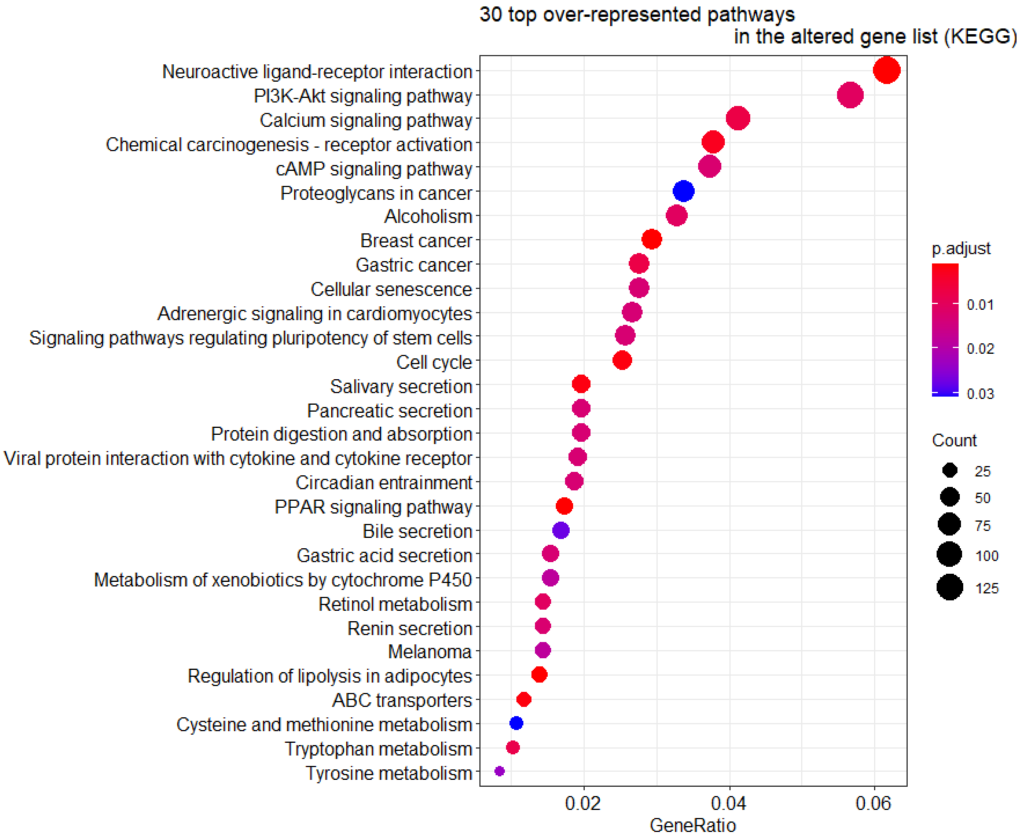
Results of the EA on DEGs and DMGs against KEGG



Results of the EA on SCNAs against KEGG



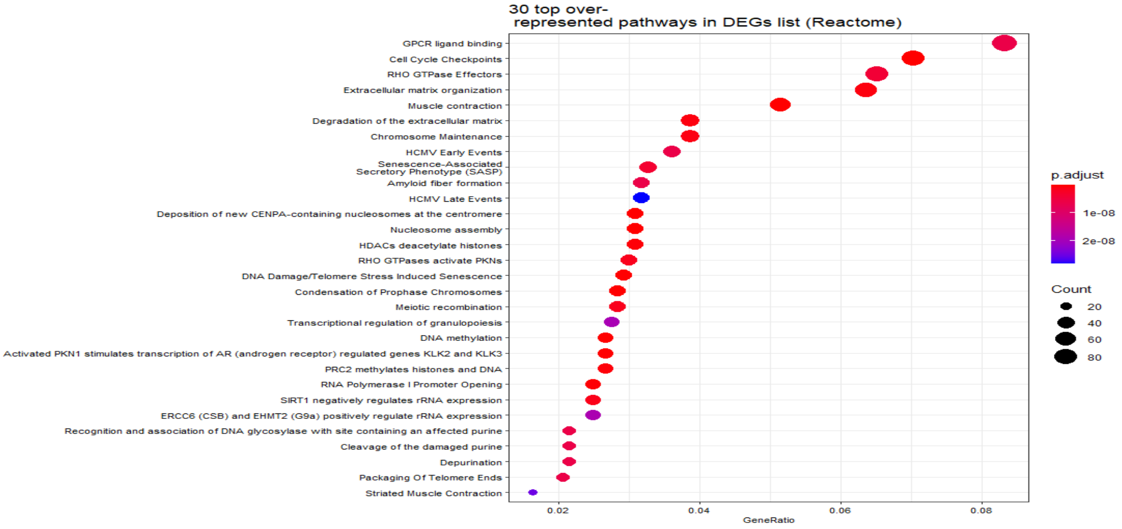
Results of the EA on DEGs and DMGs and SCNAs against KEGG



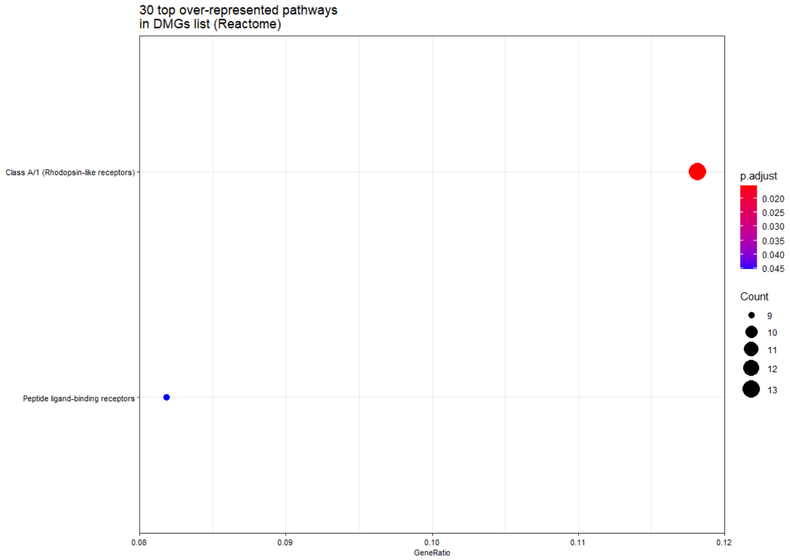
Results of the EA on all altered genes against KEGG

G Most Over-Represented Pathways Against Reactome

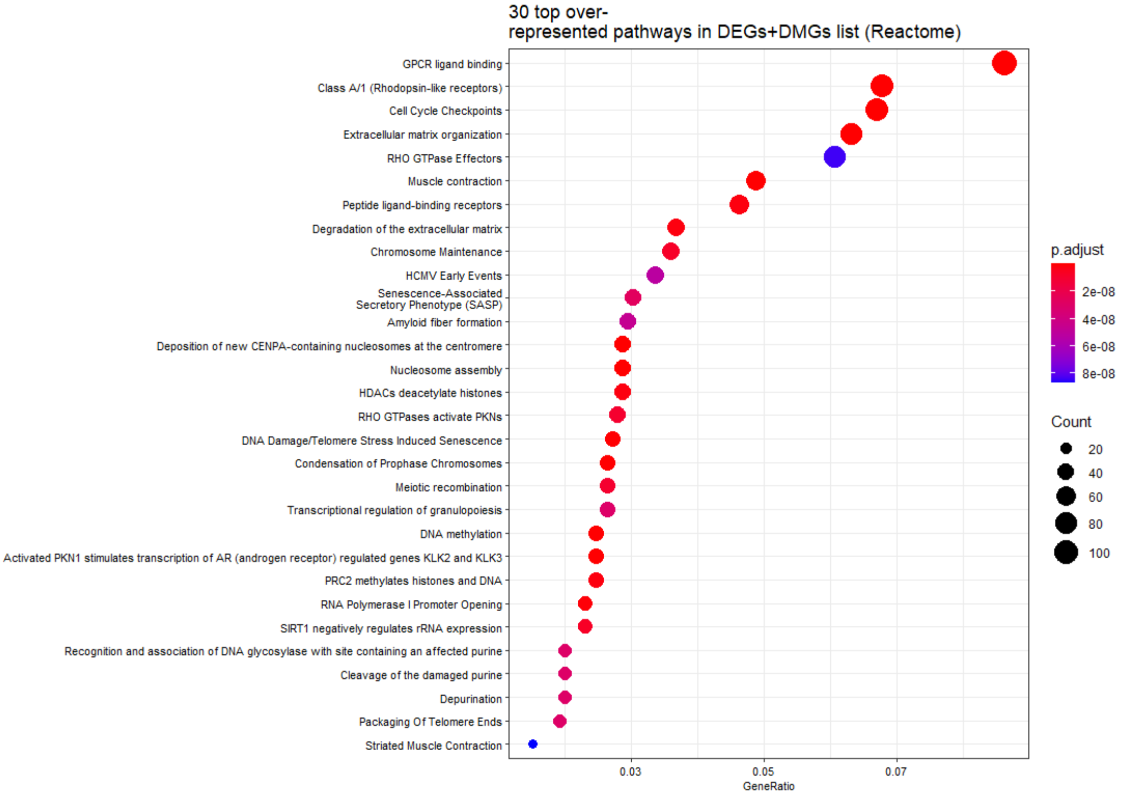
Dotplots of the 30 top over-represented pathways obtained by enrichment analysis against Reactome database. On the y-axis, the pathways are represented. The x-axis correspond to the ratio of genes that belong to the pathway of interest over the total number of genes in the corresponding gene list. The size of the dots represent the FDR-adjusted p-value.



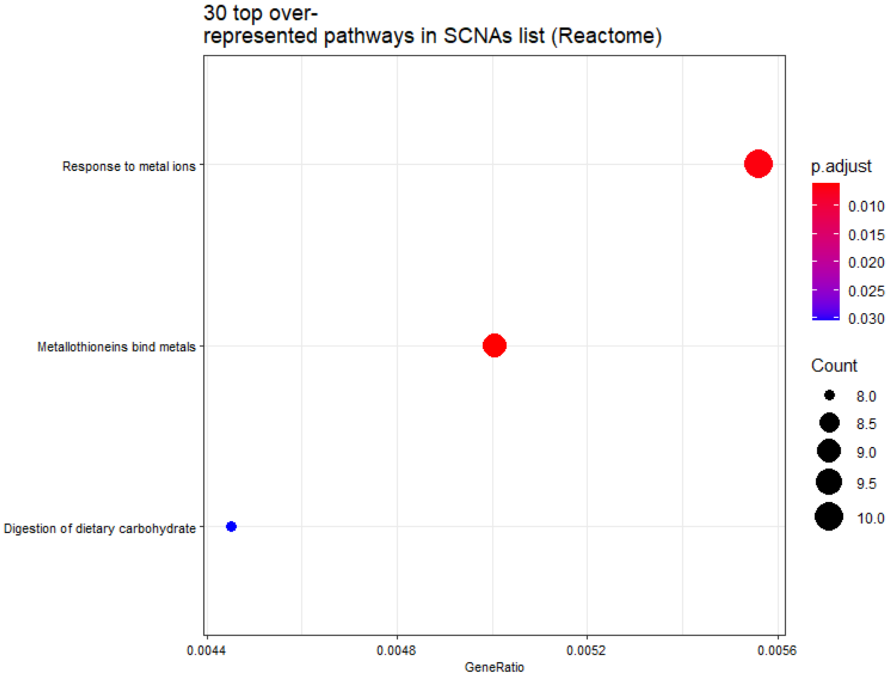
Results of the EA on DEGs against Reactome



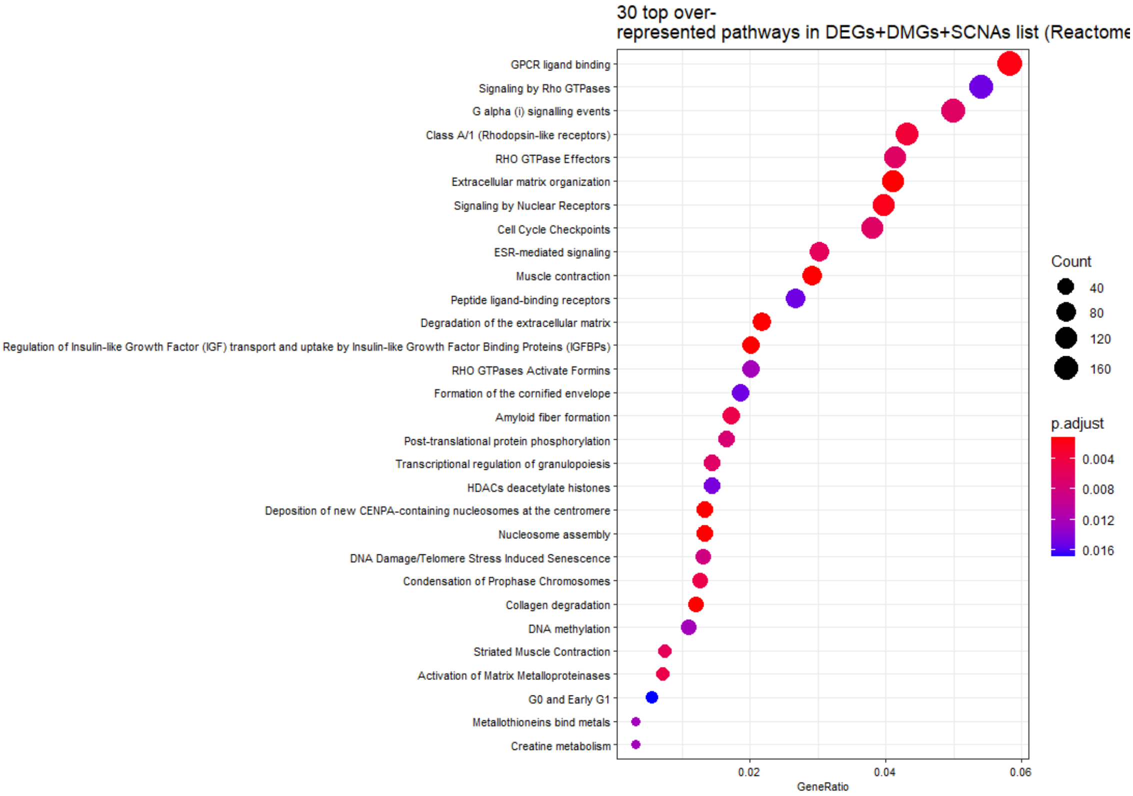
Results of the EA on DMGs against Reactome



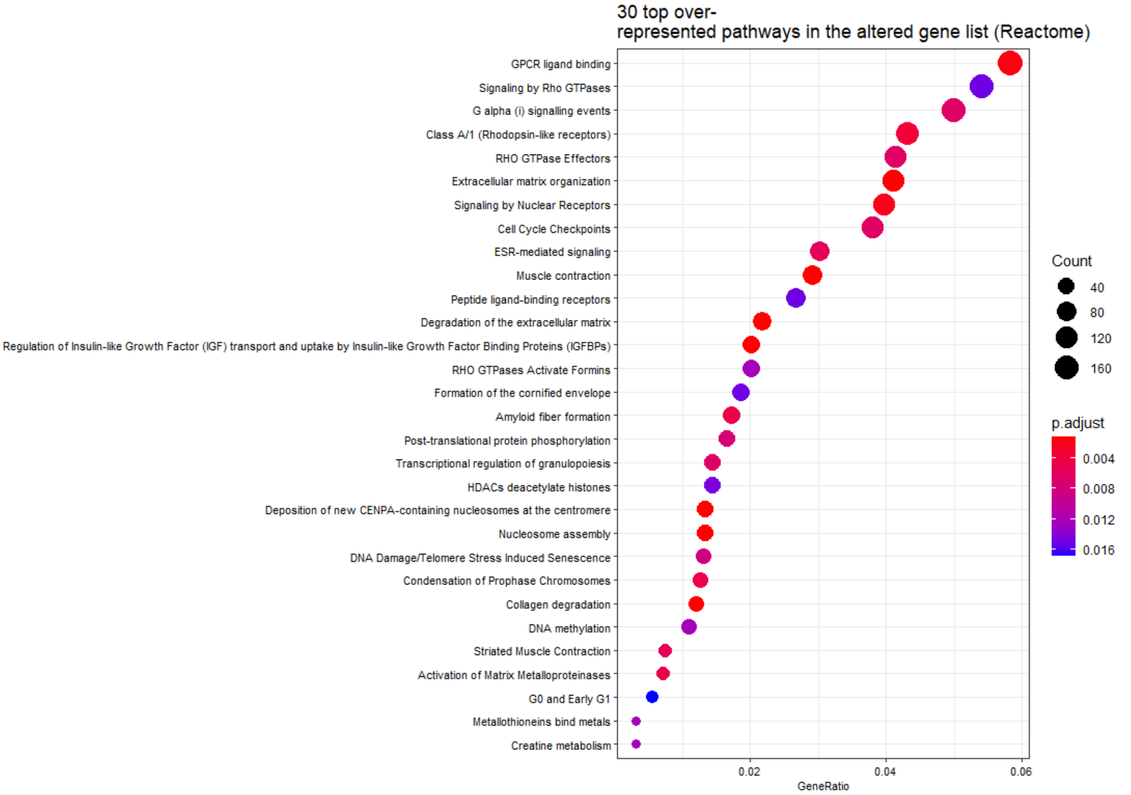
Results of the EA on DEGs and DMGs against Reactome



Results of the EA on SCNAs against Reactome



Results of the EA on DEGs and DMGs and SCNAs against Reactome



Results of the EA on all altered genes against Reactome

H Nodes of the TNBC Model and Their Topological Measures

Nodes of the PKN and their topological properties. Each node was classified in one pathway according to the pathway databases and literature used, but it can be ambiguous sometimes, and the composition of a pathway can vary from one study to another, or from one database to another.

name	Module	EdgeCount	Indegree	Outdegree	Betweenness Centrality	Omic node
BAD	Apoptosis	6	4	2	0.4537	NO
BAK1	Apoptosis	6	4	2	0.8863	NO
BAX	Apoptosis	5	3	2	0.8863	NO
BCL2	Apoptosis	7	4	3	0.6820	NO
BID	Apoptosis	6	3	3	0.3406	NO
BIRC_f	Apoptosis	10	7	3	3.9562	YES
CASP3	Apoptosis	7	3	4	9.2462	YES
CASP8	Apoptosis	5	2	3	1.4795	YES
CASP9	Apoptosis	7	5	2	3.6660	YES
CFLAR	Apoptosis	4	3	1	0.2597	NO
CREB	Apoptosis	14	10	4	2.0004	NO
CYCS	Apoptosis	4	3	1	1.1199	NO
DIABLO	Apoptosis	6	4	2	2.2287	NO
FOXO_f	Apoptosis	10	8	2	1.9905	NO
MCL1	Apoptosis	7	5	2	0.6004	NO
AURKA	Cell cycle	5	1	4	0.5180	YES
AURKB	Cell cycle	5	2	3	0.1902	YES
CBPp300_c	Cell cycle	6	3	3	1.5703	NO
CCNB1	Cell cycle	3	1	2	1.1614	YES
CCND1	Cell cycle	18	13	5	13.1642	NO
CCNE1	Cell cycle	5	4	1	0.2689	YES
CDC25A	Cell cycle	10	8	2	2.6611	YES
CDKN1A	Cell cycle	11	8	3	4.8084	NO
CDKN2A	Cell cycle	4	1	3	0.9764	YES
CREBBP	Cell cycle	3	1	2	0.2077	YES
E2F1	Cell cycle	4	2	2	0.1370	YES
EP300	Cell cycle	8	4	4	2.1570	NO
MDM2	Cell cycle	14	12	2	0.9194	NO
MYC	Cell cycle	18	10	8	5.9975	YES
PARP_f	Cell cycle	3	2	1	0.2159	YES
PLK1	Cell cycle	10	4	6	1.6808	YES
RB1	Cell cycle	7	5	2	5.4097	YES
SKP2	Cell cycle	5	3	2	2.3372	NO
TP53	Cell cycle	22	13	9	5.6030	YES
AA	cPLA2	7	1	6	4.4640	NO
ALOX12	cPLA2	2	1	1	0.0999	NO
ALOX15	cPLA2	1	0	1	0.0000	NO
ALOX5	cPLA2	1	0	1	0.0000	NO
BLT1	cPLA2	3	2	1	0.3719	NO
Ca2	cPLA2	7	3	4	0.5883	NO
COX1_2	cPLA2	3	2	1	0.3547	NO
cPLA2a	cPLA2	6	5	1	4.8241	YES
EP1_2_3_4	cPLA2	4	1	3	0.7313	NO
GPR31	cPLA2	3	1	2	0.3093	NO
HETE12	cPLA2	4	2	2	0.9497	NO
HETE15	cPLA2	5	2	3	0.8710	NO
LTB4	cPLA2	4	3	1	0.3854	NO
MNK1	cPLA2	3	2	1	0.3117	NO
PGE2	cPLA2	4	3	1	1.0957	NO
TP	cPLA2	2	1	1	0.1028	NO
TXA2	cPLA2	2	1	1	0.5178	NO

Continuation of Appendix B

name	Module	EdgeCount	Indegree	Outdegree	Betweenness Centrality	Omic node
ABL1	DNA repair/BRCA	6	2	4	5.6010	NO
ATM	DNA repair/BRCA	15	3	12	3.9430	NO
ATR	DNA repair/BRCA	10	2	8	3.2609	NO
BARD1	DNA repair/BRCA	1	0	1	0.0000	NO
BRCA1	DNA repair/BRCA	14	9	5	13.9395	NO
BRCA1_BARD1_c	DNA repair/BRCA	3	2	1	0.8848	NO
BRCA2	DNA repair/BRCA	3	2	1	0.4864	YES
CHEK1	DNA repair/BRCA	9	3	6	1.8349	YES
CHEK2	DNA repair/BRCA	7	3	4	0.1725	NO
PALB2	DNA repair/BRCA	3	0	3	0.0000	NO
RAD51	DNA repair/BRCA	3	2	1	0.9097	YES
TOP1	DNA repair/BRCA	2	1	1	0.1577	NO
CBLB	EGFR	1	0	1	0.0000	NO
EGF	EGFR	1	0	1	0.0000	NO
EGFR	EGFR	11	4	7	0.3241	YES
FGF_f	FGFR	1	0	1	0.0000	YES
FGFR	FGFR	3	2	1	0.9446	YES
FRS2	FGFR	3	1	2	0.9484	NO
HSPG	FGFR	1	0	1	0.0000	NO
HH	Hedgehog	1	0	1	0.0000	NO
GLI_f	Hedgehog	6	4	2	2.9109	NO
PTCH1	Hedgehog	3	2	1	0.8547	NO
SMO	Hedgehog	3	2	1	1.4884	NO
CSNK1D_E	HIPPO	12	2	10	0.9252	YES (only CSNK1D)
STK_f	HIPPO	6	3	3	0.9367	YES
YAP_TAZ	HIPPO	12	4	8	2.1273	NO
ISGF3_c	Jak/STAT	2	2	0	0.0000	NO
JAK_f	Jak/STAT	8	4	4	2.0768	NO
PIAS1	Jak/STAT	4	1	3	0.0499	NO
SOCS1	Jak/STAT	3	1	2	0.8726	NO
SRC	Jak/STAT	16	6	10	7.4399	NO
STAT1	Jak/STAT	9	6	3	2.3459	NO
STAT2	Jak/STAT	2	1	1	0.0268	NO
STAT3	Jak/STAT	15	10	5	4.4042	NO
AP1_c	MAPK cascade	9	7	2	3.1242	NO
ATF2	MAPK cascade	4	3	1	0.0644	NO
CDC42	MAPK cascade	3	2	1	1.0074	NO
CSK	MAPK cascade	2	1	1	0.0000	NO
DUSP1	MAPK cascade	6	4	2	4.0068	YES

Continuation of Appendix B

name	Module	EdgeCount	Indegree	Outdegree	Betweenness Centrality	Omic node
DUSP6	MAPK cascade	4	2	2	0.4026	NO
EGR1	MAPK cascade	5	3	2	1.2314	YES
ERK_f	MAPK cascade	34	5	29	12.5941	NO
FOS	MAPK cascade	8	5	3	1.9776	YES
GAB_f	MAPK cascade	6	4	2	1.9780	NO
GRAP2	MAPK cascade	3	1	2	0.2470	NO
GRB2	MAPK cascade	6	4	2	1.3551	YES
ITCH	MAPK cascade	5	1	4	0.4143	NO
JNK_f	MAPK cascade	15	4	11	3.1145	NO
JUN	MAPK cascade	4	2	2	0.1587	NO
MAP2K3	MAPK cascade	3	2	1	0.6549	NO
MAP2K4	MAPK cascade	7	5	2	1.2047	NO
MAP2K7	MAPK cascade	4	3	1	0.4174	NO
MAP3K1	MAPK cascade	3	1	2	0.1067	NO
MAP3K11	MAPK cascade	2	1	1	0.3094	NO
MAP3K4	MAPK cascade	2	1	1	0.3094	YES
MAP3K5	MAPK cascade	2	1	1	0.5935	NO
MAP3K8	MAPK cascade	2	1	1	0.1926	NO
MAPK14	MAPK cascade	20	3	17	5.1659	NO
MAPK8-IP3	MAPK cascade	4	1	3	0.7418	NO
MAPKA-PK2	MAPK cascade	6	1	5	0.6820	NO
MEK_f	MAPK cascade	6	4	2	1.2237	NO
MMP_f	MAPK cascade	5	4	1	1.2981	YES
MSK_f	MAPK cascade	6	2	4	0.2082	NO
PAK1	MAPK cascade	9	2	7	3.5622	NO
PLCG1	MAPK cascade	6	2	4	1.5686	NO
PPM1A	MAPK cascade	6	1	5	1.1006	NO
PRKACA	MAPK cascade	12	3	9	4.1990	NO
PRKCA	MAPK cascade	8	4	4	1.5084	NO

Continuation of Appendix B

name	Module	EdgeCount	Indegree	Outdegree	Betweenness Centrality	Omic node
PRKCD	MAPK cascade	10	5	5	3.3665	YES
PTPN11	MAPK cascade	9	3	6	2.8133	NO
PTPN6	MAPK cascade	3	2	1	0.5006	NO
RAC_f	MAPK cascade	11	6	5	5.2450	NO
RAF_f	MAPK cascade	9	5	4	1.6119	NO
RAS_f	MAPK cascade	5	3	2	1.5140	YES
RSK_f	MAPK cascade	11	2	9	0.4976	NO
S6K_f	MAPK cascade	7	2	5	0.5639	YES
TWIST1	MAPK cascade	7	2	5	0.8110	NO
AKT1S1	mTOR	2	1	1	0.0000	NO
mTORC1_c	mTOR	8	4	4	1.1242	NO
mTORC2_c	mTOR	6	3	3	1.3347	NO
RHEB	mTOR	3	1	2	0.2301	NO
TSC_f	mTOR	7	5	2	1.2786	NO
CHUK	NFKB	7	3	4	2.0440	NO
IKBKB	NFKB	11	4	7	3.5042	YES
IRAK1	NFKB	4	2	2	1.1561	NO
MAP3K7	NFKB	7	2	5	1.6131	NO
NFKB_f	NFKB	16	8	8	10.1381	NO
REL_f	NFKB	7	5	2	1.3836	NO
TAB_f	NFKB	3	2	1	0.3580	NO
TRAF6	NFKB	5	2	3	1.9332	NO
DLL1_3	NOTCH	2	1	1	0.0253	YES (DLL3)
HES1	NOTCH	8	4	4	1.1214	NO
HEY1	NOTCH	2	1	1	0.0000	NO
HIF1A	NOTCH	7	6	1	0.2690	NO
JAG1_2	NOTCH	3	2	1	0.0668	NO
MIB1	NOTCH	2	0	2	0.0000	NO
NOTCH_f	NOTCH	14	8	6	4.6014	NO
RBPJ	NOTCH	7	5	2	0.1920	NO
SNAI_f	NOTCH	8	4	4	1.0656	NO
Anti - survival	Phenotype	3	3	0	0.0000	NO
Prosurvival	Phenotype	3	3	0	0.0000	NO
AKT_f	PI3K/Akt	30	8	22	18.2698	NO
cAMP	PI3K/Akt	3	2	1	0.3402	YES
GSK3_f	PI3K/Akt	24	8	16	5.8081	NO
IRS1	PI3K/Akt	4	3	1	0.3821	NO
PDPK1	PI3K/Akt	7	1	6	1.3565	NO
PIK3CA	PI3K/Akt	13	7	6	15.4750	YES
PIP3	PI3K/Akt	4	3	1	0.4042	NO
PTEN	PI3K/Akt	10	5	5	4.3569	YES
ESR1	receptor	15	12	3	9.9636	YES
Estrogens	receptor	1	0	1	0.0000	NO
HER2	receptor	7	4	3	0.6545	NO
PGR	receptor	3	2	1	0.8637	YES
Progest -erones	receptor	1	0	1	0.0000	NO

Continuation of Appendix B

name	**Module**	**EdgeCount**	**Indegree**	**Outdegree**	**Betweenness Centrality**	**Omic node**
ARHGAP24	Rho GTPases	3	1	2	0.3594	NO
CFL_f	Rho GTPases	3	2	1	0.7211	NO
DAAM1	Rho GTPases	2	1	1	0.0638	NO
LIMK1	Rho GTPases	3	2	1	0.3371	NO
LIMK2	Rho GTPases	3	2	1	0.3182	NO
PARD6A	Rho GTPases	5	3	2	1.2072	NO
RHOA	Rho GTPases	8	7	1	2.1093	YES
RND3	Rho GTPases	2	1	1	0.0483	NO
ROCK1	Rho GTPases	9	2	7	4.9318	NO
SRF	Rho GTPases	4	3	1	0.7944	YES
TIAM1	Rho GTPases	2	1	1	0.2046	NO
ILK	RTKs	3	2	1	0.7722	NO
ILR_f	RTKs	6	2	4	2.2904	NO
LIF	RTKs	2	1	1	0.2220	NO
RTPK_f	RTKs	10	4	6	4.9895	NO
SHC1	RTKs	9	8	1	0.4128	NO
SOS1	RTKs	4	3	1	0.2438	NO
SYK	RTKs	4	2	2	1.4807	NO
VAV1	RTKs	2	1	1	0.4722	NO
ACVR1	TGFB	5	1	4	0.2134	NO
BMPR2	TGFB	3	2	1	0.5013	NO
PPP1CA	TGFB	9	2	7	4.0884	NO
SKI	TGFB	7	1	6	0.5074	YES
SMAD1	TGFB	10	8	2	1.4506	NO
SMAD2	TGFB	13	9	4	1.6258	YES
SMAD3	TGFB	17	12	5	2.2962	NO
SMAD4	TGFB	13	10	3	1.0097	NO
SMAD5	TGFB	4	3	1	0.0028	NO
SMAD6	TGFB	7	3	4	0.5498	NO
SMAD7	TGFB	14	8	6	6.0307	NO
SMURF1	TGFB	8	1	7	0.8142	NO
SMURF2	TGFB	7	1	6	0.7046	NO
TGFB1	TGFB	5	3	2	0.9564	NO
TGFBR1	TGFB	10	5	5	2.0618	NO
TGFBR2	TGFB	5	3	2	0.7239	YES
HIC1	VEGFR	2	0	2	0.0000	NO
PTK2	VEGFR	9	6	3	1.1879	NO
TLR5	VEGFR	1	0	1	0.0000	NO
VEGF_f	VEGFR	6	5	1	2.2758	YES
VEGFR2	VEGFR	4	1	3	2.6992	YES
APC	WNT	6	5	1	0.1993	NO
AXIN1	WNT	10	6	4	2.1368	NO
BTRC	WNT	6	4	2	0.8371	NO
CDH1	WNT	5	3	2	0.3800	NO
CDH2	WNT	3	2	1	0.3570	NO
CK1_f	WNT	3	1	2	0.1264	NO
CSNK1A1	WNT	9	1	8	1.4043	NO

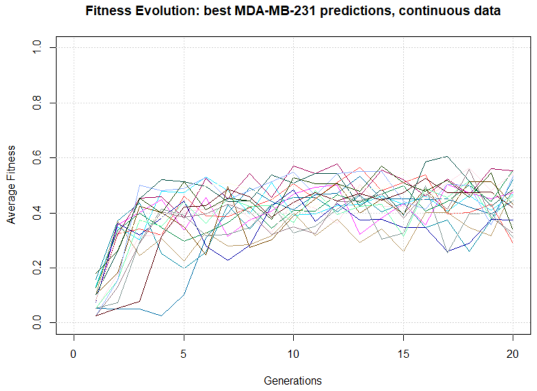
Continuation of Appendix B

name	**Module**	**EdgeCount**	**Indegree**	**Outdegree**	**Betweenness Centrality**	**Omic node**
CTNNB1	WNT	11	8	3	3.2706	NO
DKK_f	WNT	3	2	1	0.7962	NO
DVL_f	WNT	12	5	7	3.8308	YES (DVL1)
FZD_f	WNT	4	1	3	1.2347	YES
LEF1	WNT	4	2	2	0.4384	NO
LRP_f	WNT	11	7	4	3.5625	YES
NLK	WNT	3	1	2	0.0722	NO
SFRP1	WNT	2	1	1	1.1290	NO
TCF7_f	WNT	6	2	4	1.0908	NO

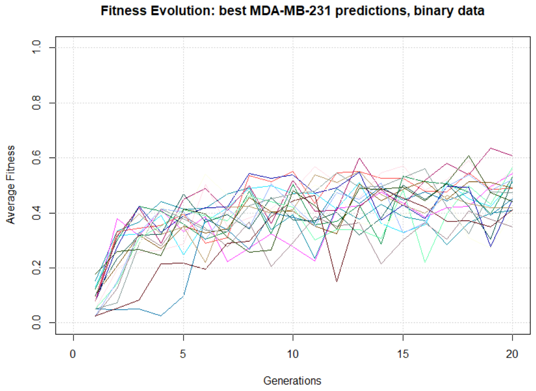
End of Appendix B

I Fitness Evolution of the Cell Line-Specific Models

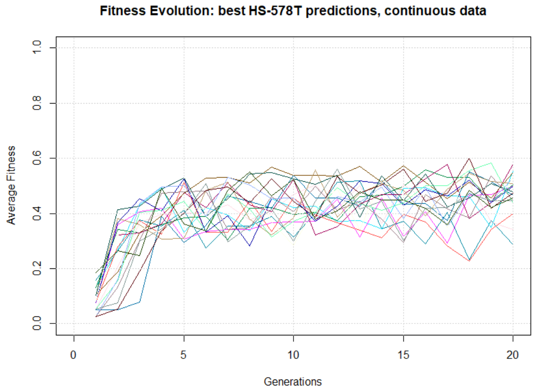
Comparison of the fitness evolution across the generations of the models trained on continuous (left) and binary data (right) of the CCLE cell lines



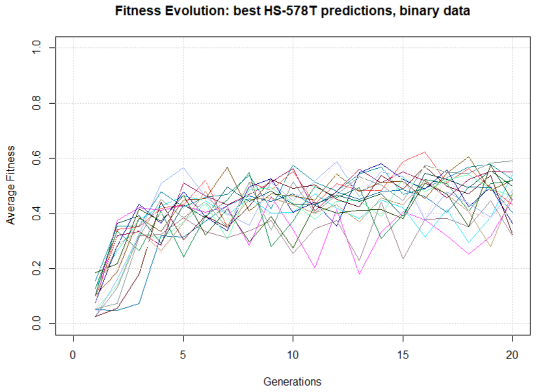
Fitness evolution of the models trained on continuous MDA-MB-231 data



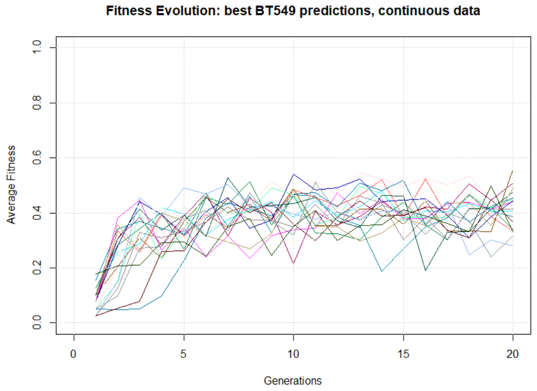
Fitness evolution of the models trained on binary MDA-MB-231 data



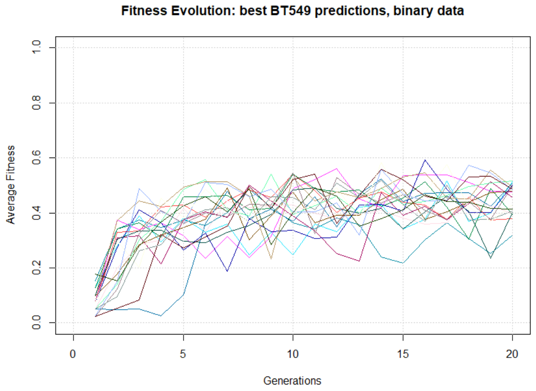
Fitness evolution of the models trained on continuous HS-578T data



Fitness evolution of the models trained on binary HS-578T data



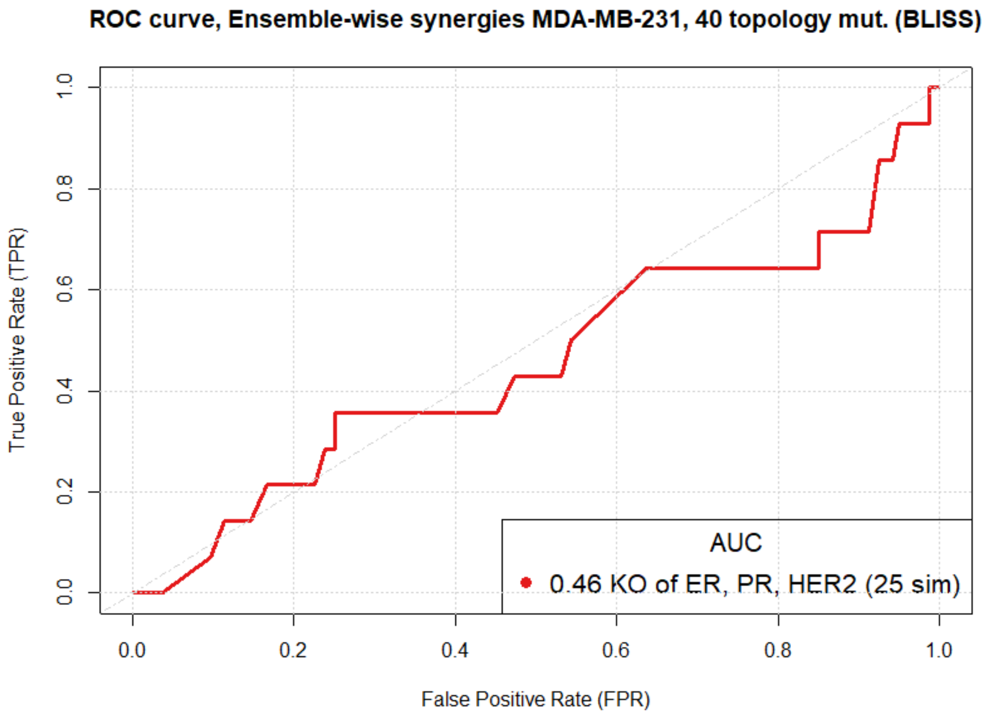
Fitness evolution of the models trained on continuous BT549 data



Fitness evolution of the models trained on binary BT549 data

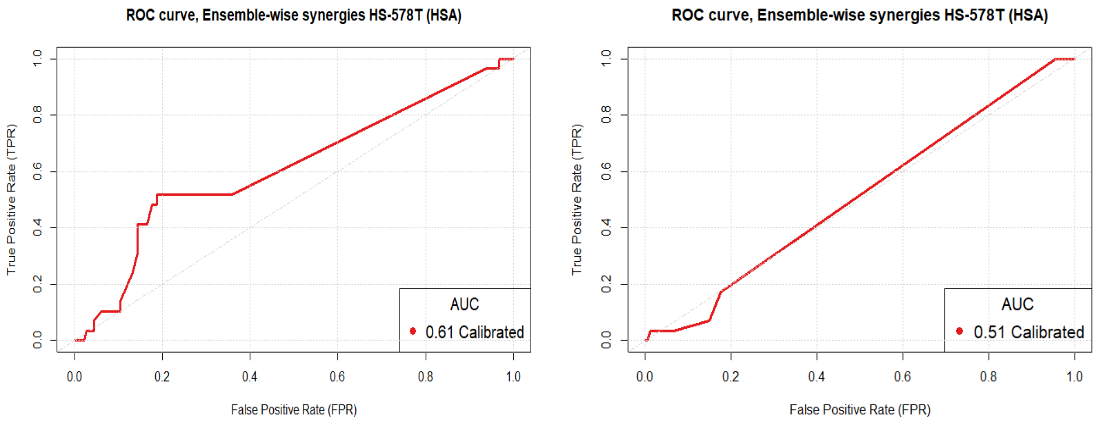
J Prediction Performances of MDA-MB-231 Models with the TNBC-Receptors Forced Inactive

ROC curve of the predictions run on MDA-MB-231-calibrated models, with the ER, PR and HER2 receptors set to 0.



K Prediction Performances of HS-578T Models Using the HSA Synergy Metric

ROC curves of the predictions run on HS-578T-calibrated models. The parameters were: 25 simulations, calibration of 184 nodes. The synergy score was calculated with the HSA metric. Different types of mutations were introduced between the two predictions.

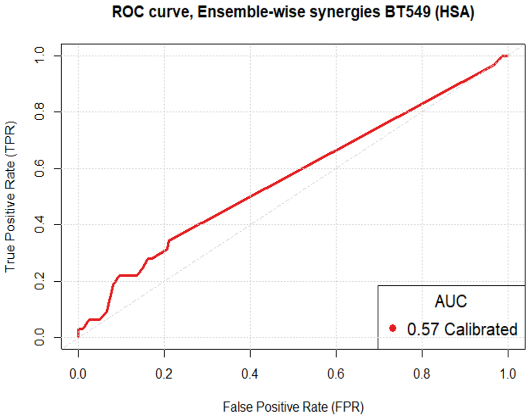


ROC curve obtained by combining 3 balance mutations to 40 topology mutations

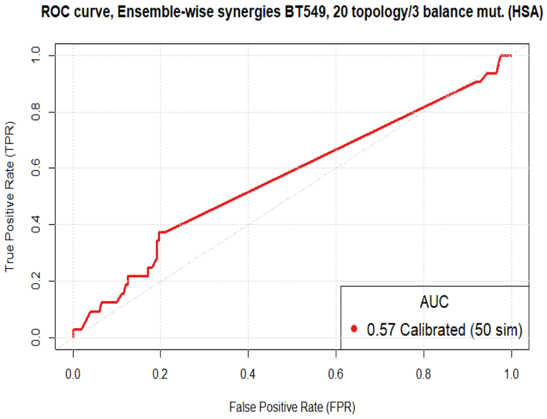
ROC curve obtained by introducing only 3 balance mutations

L Prediction Performances of BT549 Models Using the HSA Synergy Metric

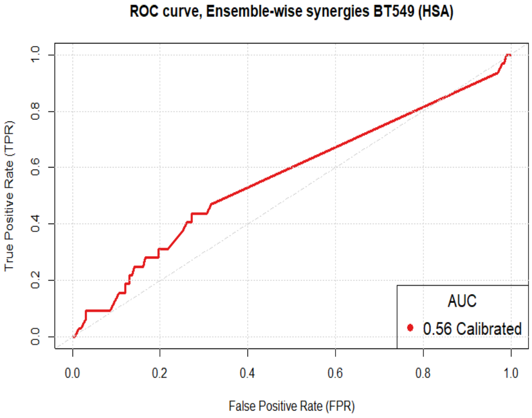
ROC curves of the predictions run on BT549-calibrated models. 184 nodes were calibrated. The synergy score was calculated with the HSA metric. Different number of simulations were run, and different types of mutations were introduced between the two predictions.



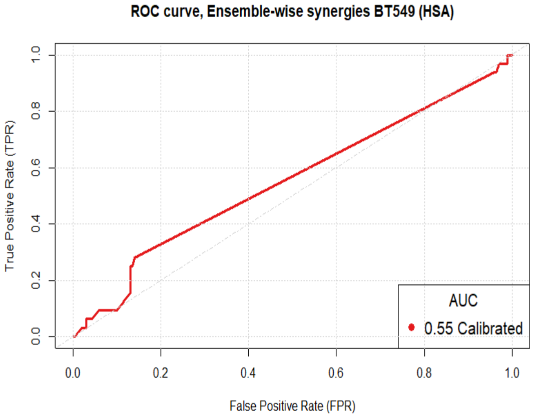
ROC curve obtained by combining 3 balance mutations to 20 topology mutations, in 25 simulations



ROC curve obtained by combining 3 balance mutations to 20 topology mutations, in 50 simulations



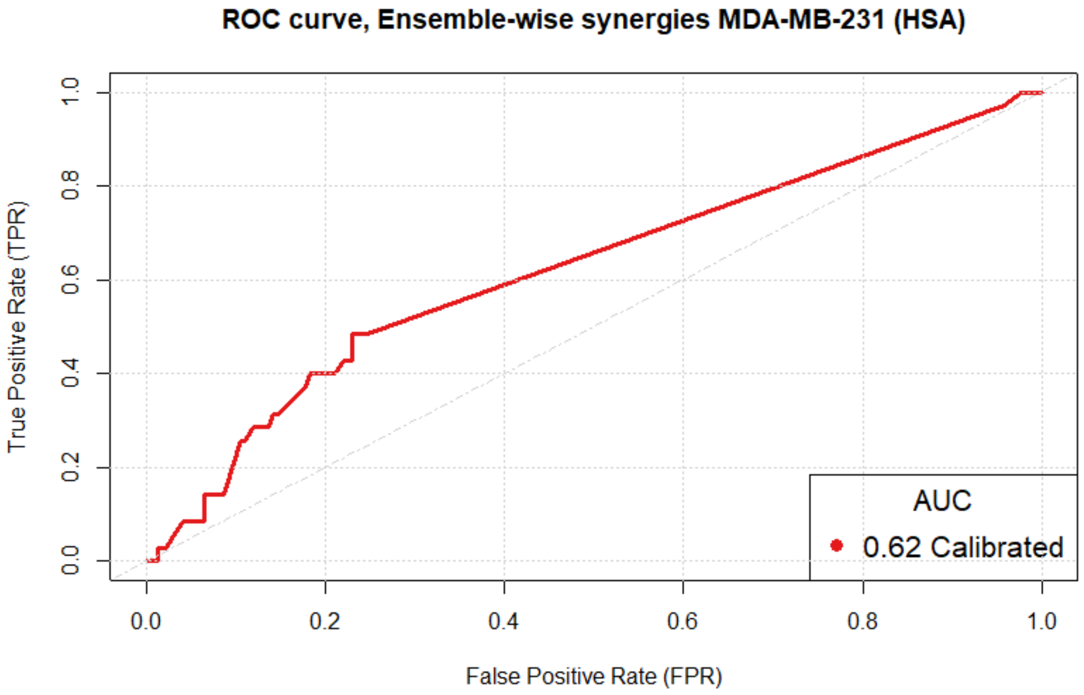
ROC curve obtained by using only 40 topology mutations, in 25 simulations



ROC curve obtained by combining 3 balance mutations to 40 topology mutations, in 25 simulations

M Prediction Performances of MDA-MB-231 Models Using the HSA Synergy Metric and a Combination of Topology and Balance Mutations.

ROC curve of the predictions run on MDA-MB-231-calibrated models. The parameters used were: 25 simulations, 3 balance mutations combined with 40 topology mutations. The synergy score was calculated with the HSA metric.


ROC curve, Ensemble-wise synergies MDA-MB-231 (HSA)
True Positive Rate (TPR)
False Positive Rate (FPR)
AUC
● 0.62 Calibrated