

University of Dundee

## Machine learning and data mining frameworks for predicting drug response in cancer

Vougas, Konstantinos; Sakelaropoulos, Theodore; Kotsinas, Athanassios; Foukas, George-Romanos P.; Ntargaras, Andreas; Koinis, Filippos

*Published in:*  
Pharmacology & Therapeutics

*DOI:*  
[10.1016/j.pharmthera.2019.107395](https://doi.org/10.1016/j.pharmthera.2019.107395)

*Publication date:*  
2019

*Document Version*  
Peer reviewed version

[Link to publication in Discovery Research Portal](#)

### *Citation for published version (APA):*

Vougas, K., Sakelaropoulos, T., Kotsinas, A., Foukas, G-R. P., Ntargaras, A., Koinis, F., Polyzos, A., Myrianthopoulos, V., Zhou, H., Narang, S., Georgoulas, V., Alexopoulos, L., Aifantis, I., Townsend, P. A., Sfrikakis, P., Fitzgerald, R., Thanos, D., Bartek, J., Petty, R., ... Gorgoulis, V. G. (2019). Machine learning and data mining frameworks for predicting drug response in cancer: An overview and a novel *in silico* screening process based on association rule mining. *Pharmacology & Therapeutics*, 203, [107395].  
<https://doi.org/10.1016/j.pharmthera.2019.107395>

### General rights

Copyright and moral rights for the publications made accessible in Discovery Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from Discovery Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain.
- You may freely distribute the URL identifying the publication in the public portal.

### Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

## Accepted Manuscript

Machine learning and data mining frameworks for predicting drug response in cancer: an overview and a novel in silico screening process based on Association Rule Mining

Konstantinos Vougas, Theodore Sakelaropoulos, Athanassios Kotsinas, George-Romanos P. Foukas, Andreas Ntargaras, Filippos Koinis, Alexander Polyzos, Vassilis Myrianthopoulos, Hua Zhou, Sonali Narang, Vassilis Georgoulis, Leonidas Alexopoulos, Iannis Aifantis, Paul A. Townsend, Petros Sfikakis, Rebecca Fitzgerald, Dimitris Thanos, Jiri Bartek, Russell Petty, Aristotelis Tsirigos, Vassilis G. Gorgoulis



PII: S0163-7258(19)30138-X  
DOI: <https://doi.org/10.1016/j.pharmthera.2019.107395>  
Article Number: 107395  
Reference: JPT 107395  
To appear in: *Pharmacology and Therapeutics*

Please cite this article as: K. Vougas, T. Sakelaropoulos, A. Kotsinas, et al., Machine learning and data mining frameworks for predicting drug response in cancer: an overview and a novel in silico screening process based on Association Rule Mining, *Pharmacology and Therapeutics*, <https://doi.org/10.1016/j.pharmthera.2019.107395>

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

**Machine learning and data mining frameworks for predicting drug response in cancer: an overview and a novel *in silico* screening process based on Association Rule Mining**

Konstantinos Vougas<sup>1,2,18,\*</sup>, Theodore Sakelaropoulos<sup>3,4,\*</sup>, Athanassios Kotsinas<sup>2\*</sup>, George-Romanos P Foukas<sup>2</sup>, Andreas Ntargaras<sup>2</sup>, Filippos Koinis<sup>2</sup>, Alexander Polyzos<sup>5</sup>, Vassilis Myrianthopoulos<sup>2,6</sup>, Hua Zhou<sup>7</sup>, Sonali Narang<sup>3,4</sup>, Vassilis Georgoulas<sup>8</sup>, Leonidas Alexopoulos<sup>9</sup>, Iannis Aifantis<sup>3,4</sup>, Paul A Townsend<sup>10</sup>, Petros Sfikakis<sup>11,12</sup>, Rebecca Fitzgerald<sup>13</sup>, Dimitris Thanos<sup>1</sup>, Jiri Bartek<sup>14,15,16</sup>, Russell Petty<sup>17</sup>, Aristotelis Tsirigos<sup>3,4,7,18,#</sup> and Vassilis G. Gorgoulis<sup>1,2,10,12,18,#</sup>

1. Biomedical Research Foundation of the Academy of Athens, 4 Soranou Ephessiou Str., Athens, GR-11527, Greece
2. Molecular Carcinogenesis Group, Department of Histology and Embryology, School of Medicine, National and Kapodistrian University of Athens, 75 Mikras Asias Str, Athens, GR-11527, Greece
3. Department of Pathology, NYU School of Medicine, New York, NY 10016, USA
4. Laura and Isaac Perlmutter Cancer Center, NYU School of Medicine, New York, NY 10016, USA
5. Sanford I. Weill Department of Medicine, Sandra and Edward Meyer Cancer Center, Weill Cornell Medicine, New York, NY 10021 USA
6. Division of Pharmaceutical Chemistry, School of Pharmacy, National and Kapodistrian University of Athens, Athens, Greece

7. Applied Bioinformatics Laboratories, NYU School of Medicine, New York, NY 10016, USA
8. Laboratory of Tumour Cell Biology, School of Medicine, University of Crete, Heraklion, Crete, Greece
9. School of Mechanical Engineering, National Technical University of Athens, Zografou 15780, Greece
10. Division of Cancer Sciences, Faculty of Biology, Medicine and Health, Manchester Academic Health Science Centre, Manchester Cancer Research Centre, NIHR Manchester Biomedical Research Centre, University of Manchester, Manchester, M20 4GJ, UK
11. 1st Department of Propaedeutic Internal Medicine, Medical School, Laikon Hospital, National and Kapodistrian University of Athens, 75 Mikras Asias Str, Athens, GR-11527, Greece.
12. Center for New Biotechnologies and Precision Medicine, Medical School, National and Kapodistrian University of Athens, 75 Mikras Asias Str, Athens, GR-11527, Greece
13. Medical Research Council Cancer Unit, Hutchison/Medical Research Council Research Centre, University of Cambridge, Cambridge, UK
14. Genome Integrity Unit, Danish Cancer Society Research Centre, Strandboulevarden 49, Copenhagen DK-2100, Denmark
15. Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University, Hněvotínská, Olomouc 1333/5 779 00, Czech Republic
16. Science for Life Laboratory, Division of Translational Medicine and Chemical Biology, Department of Medical Biochemistry and Biophysics, Karolinska Institute, Stockholm SE-171 77, Sweden.
17. Division of Molecular and Clinical Medicine, Ninewells Hospital and School of Medicine,



University of Dundee, Dundee DD1 9SY, Scotland

18. Co-senior author

**Key words:** Drug Response Prediction; Precision Medicine; Data mining; Machine Learning; Association Rule Mining

\* These authors contributed equally to this work.

# To whom correspondence should be addressed:

Vassilis G. Gorgoulis (Lead Contact), E-mail: [vgorg@med.uoa.gr](mailto:vgorg@med.uoa.gr); Tel.: +30-210-7462352

Aristotelis Tsirigos, E-mail: [aristotelis.tsirigos@nyulangone.org](mailto:aristotelis.tsirigos@nyulangone.org); Tel.: +1 646 501 2693

**ABSTRACT**

A major challenge in cancer treatment is predicting the clinical response to anti-cancer drugs on a personalized basis. The success of such a task largely depends on the ability to develop computational resources that integrate big “omic” data into effective drug-response models. Machine learning is both an expanding and an evolving computational field that holds promise to cover such needs. Here we provide a focused overview of: 1) the various supervised and unsupervised algorithms used specifically in drug response prediction applications, 2) the strategies employed to develop these algorithms into applicable models, 3) data resources that are fed into these frameworks and 4) pitfalls and challenges to maximize model performance. In this context we also describe a novel *in silico* screening process, based on Association Rule Mining, for identifying genes as candidate drivers of drug response and compare it with relevant data mining frameworks, for which we generated a web application freely available at: <https://compbio.nyumc.org/drugs/>. This pipeline explores with high efficiency large sample-spaces, while is able to detect low frequency events and evaluate statistical significance even in the multidimensional space, presenting the results in the form of easily interpretable rules. We conclude with future prospects and challenges of applying machine learning based drug response prediction in precision medicine.

**Abbreviations list*****Molecular terms***

ARHGDIB: Rho GDP dissociation inhibitor beta

BCL2: BCL2 Apoptosis Regulator

BRCA1: Breast cancer type 1 Susceptibility Protein

BRAF: B-Raf proto-oncogene, Serine/Threonine Kinase

CCND3: Cyclin D3

CD151: Tetraspanin-24

CDC6: Cell cycle division 6

CDKN2A: Cyclin dependent kinase inhibitor 2A

CTCF: 11-zinc finger protein or CCCTC-binding factor

DDR: DNA damage response

EGFR: Epidermal growth factor receptor

EMT: Epithelial to mesenchymal transition

ERK: Extracellular regulated kinase

FLT3: Fms related tyrosine kinase 3

GHRH: Growth hormone-releasing hormone

GMIP: GEM interacting protein

ID1: Inhibitor of DNA binding 1

KRAS: Kirsten rat sarcoma proto-oncogene

LYL1: Lymphoblastic leukemia associated hematopoiesis regulator 1

MAGI3: Membrane-Associated Guanylate Kinase 3

MAPK: Mitogen-activated protein kinase

MAP2K3: Mitogen-activated protein kinase kinase 3

MDM2: Mouse double minute 2

MDR1: Multidrug resistance 1

MEK: Mitogen-activated protein kinase kinase

MLL2: KMT2D - Histone-Lysine N-Methyltransferase MLL2

mTOR: mechanistic target of rapamycin kinase

MYC: MYC Proto-Oncogene, BHLH Transcription Factor

NPTN: Neuropilin

NQO1: NAD(P)H dehydrogenase 1

NSCLC: Non-small cell lung cancer

PARP: Poly(ADP-ribose) polymerase

PDIA3: ERp57/PDIA3: Protein disulfide isomerase family

PI3K: Phosphoinositide 3-kinase

PIK3CA: Phosphatidylinositol-4,5-bisphosphate 3-kinase

POF1B: Premature Ovarian Failure Protein 1B

PTEN: Phosphatase and tensin homolog

REV7: MAD2L2 - Mitotic Arrest Deficient 2 Like 2

SAMSN1: SAM domain, SH3 domain and nuclear localization signals 1

SCLC: Small cell lung cancer

SHLD1-3: Shieldin complex subunit 1-3

SMAD3: Mothers Against Decapentaplegic Homolog 3

TKI: Tyrosine Kinase Inhibitor

TP53: Tumor Protein p53

ZCCHC7: Zinc finger CCHC-type containing 7

ZNF22: Zinc finger protein 22

***Statistical, machine learning and cell lines databases terms***

ACC: Accuracy

ANOVA: Analysis of variance

ARM: Association Rule Mining

AUC: Area under the ROC curve

BATTLE: Biomarker-integrated Approaches of Targeted Therapy for Lung Cancer Elimination

BEMKL: Bayesian efficient multiple kernel learning

CCLE: Cancer Cell Line Encyclopedia

CCLP: Cosmic Cell Line Project

CNV: Copy Number Variations

CTRP: Cancer Therapeutic Response Portal

cwKBMF: component-wise Kernelized Bayesian matrix factorization

DLNN: Deep Learning Neural Networks

DREAM: Dialogue on Reverse Engineering Assessment and Methods

FDR: False Discovery Rate

FN: False Negative

FNR: False Negative Rate

FOR: False Omission Rate

FP: False Positive

FPR: False positive rate

GBMS: Gradient Boosting machines

GDSC: Genomics of Drug Sensitivity in Cancer

KF-CV: k-fold cross-validations

KNN: K-nearest neighbors

LOBICO: Logic Optimization for Binary Input to Continuous Output

MCDA: Multi-criteria decision analysis

MKL: Multiple Kernel Learning

Mut: Mutations

NCI-60: National Cancer Institute drug screening panel

NPV: Negative Predictive Value

PCA: Principal Component Analysis

PPV: Positive Predictive Value

SNE: Stochastic Neighbor Embedding

RMSE: Root Mean Square Error

STREAM: Scalable-Time Ridge Estimator by Averaging of Models

SVM: Support Vector Machines

TCGA: The Cancer Genome Atlas

TCPA: The Cancer Proteome Atlas

TN: True Negative

TP: True Positive

TNR: True Negative Rate

TPR: True Positive Rate

## Table of Contents

1. Introduction: The urge for “big data” analyzers in precision medicine	11
2. The tools it takes to grasp valuable clinical information	12
2a. The overall <i>in silico</i> strategy	12
2b. Data resources and categories of input data	13
2c. Computational techniques and selection of prediction models	15
2d. Testing the prediction models	18
2e. Clinical applications and challenges to be met	19
3. Deep Learning neural networks (DLNN): an emerging “key player”	22
4. A novel <i>in silico</i> screening process based on <i>Association Rule Mining</i> (ARM study)	24
4a. Rule verification based on prior knowledge	28
4b. Rule Experimental Validation	32
5. Comparison of ARM study with other frameworks	38
6. Perspectives and future challenges	40
Acknowledgments	43
References	44



## 1. Introduction: The urge for “big data” analyzers in precision medicine

Predicting the clinical response to therapeutic agents is a major challenge in cancer treatment. Traditional features such as, histopathological characteristics of tumors, although always useful, have reached their limit and are unable to solely guide “precise” therapeutic solutions. The advent of multiple high-throughput platforms producing “omics” data has provided to the biomedical community, over the last decade, a huge molecular repository of data (**big data**) – <sup>for terms in bold see Table 1 for machine learning terminology</sup> that is continuously expanding and promises to pave the way to precision medicine approaches. Such information merged with detailed clinical records, including response to therapy, will enable scientists to dissect the molecular events that are known to drive carcinogenesis and alter major downstream processes, such as gene expression (Halazonetis et al., 2008; Negrini et al., 2010; Galanos et al., 2018; Alexandrov et al., 2013; Zhang et al., 2016). Ultimately, molecular disease signatures are anticipated to be delivered and matched with the most effective therapeutic interventions.

The only efficient means to exploit the multi-dimensionality of the generated large data sets and to achieve the goal of predicting drug responses are computational technologies (**Figure 1**). Presently, *in silico* tools have propelled a widespread effort to effectively translate the growing wealth of high-throughput profiling data into clinically meaningful, personalised treatment strategies required by precision medicine (van't Veer and Bernards, 2008; Ali and Aittokallio, 2019; Azuaje F, 2017). However, the computational prediction of drug responses in cancer involves significant research challenges and questions including: **i)** which data set should be selected, **ii)** which computational setting is suitable for application, **iii)** are the

produced models valid to all types of cancer or a specific one, and **iv)** how is the efficacy of the models evaluated and validated. Herein, we address these critical questions by: **i)** presenting and discussing current trends in **machine learning** and **data mining** methodologies related to drug response (**Table 1, Figures 1, 2**) and **ii)** suggesting means to increase their competence. Finally, we present a novel *in silico* screening process that is based on an **unsupervised** data mining method called ***Association Rule Mining***—for terms in bold and italics see Table 2 for description of computational algorithms (**ARM**)\*-see Abbreviations table that is capable of generating simple rules linking a specific gene(s) status with drug response.

## 2. The tools it takes to grasp valuable clinical information

**2a. The overall *in silico* strategy:** The standard scheme to develop a computational model for predicting biological outcomes includes three key steps (**Figures 1-3**): **i)** opting the data set, **ii)** selecting the **algorithm** and **training** it to develop a prediction model, and **iii)** **testing** it in unseen data sets (**Figure 3, and for terminology see Table 1**). In the *first step* the desired data set(s) is selected and pre-processed. The latter includes **feature selection**, **normalization**, when more than one data set is combined, and filtering of noise or irrelevant information. Choosing the proper features is a pivotal stage for algorithms to be effective in **classification**, **regression** and **pattern recognition** (see paragraph 2b). The *second step* involves the **training** phase that aims in building the fittest model for drug response prediction. There is a wide range of computational approaches that are used to process the data sets (**Figures 1, 2**). A list and a brief description of the most commonly applied ones is presented in **Table 2**. Basically, they are divided into **supervised** and **unsupervised** learning techniques (**Figure 1, Table 2**). Although

the former methods are the most widely used, it is notable that the latter ones can provide the ground for generating prediction models, as they carry out fundamental tasks, such as **clustering** and **sample stratification** data, *prior* to the implementation of supervised learning (Zhao et al., 2014; Azuaje F, 2017; Byers et al., 2013; Moghaddas Gholami et al., 2013), as well as provide critical insights and knowledge extraction. Actually, unsupervised clustering represented the basis for traditional analytical strategies trying to identify efficient treatments in distinct patient sub-clusters (Hoadley et al., 2014, Campbell et al., 2017), or alternatively starting from treatment response clustering and then moving into the molecular context that could explain drug behaviour (Pemovska et al. 2013; Tyner et al. 2013; Frismantas et al. 2017; Andersson et al. 2018). The *third step*, also termed **independent evaluation**, is the decisive one as it will **test** if the candidate model, after training, can accurately predict response on unseen settings either experimental ones such as, cell lines, xenografts or animal models, or preferably in clinical samples.

In the following subsections each step will be discussed in more detail, including comparison of methodologies, pointing out potential weak spots that need to be improved in the future to maximize the predictive power of these **artificial intelligence**-based frameworks (Figures 1, 2).

**2b. Data resources and categories of input data:** A proficient prediction model largely depends on the “quantity and quality” of the input data. With the term “quality” we refer mainly to normalization and the source of the data. Normalization is an essential process when different data sets are merged ensuring that bias during the analysis is avoided, and includes operations such as, **matching**, **batch effect removal** and **data imputation** (when data are missing in one or more of the data sets) (Hastie et al., 2001). Ideally, to develop promising drug prediction

models the origin of the data should be clinical derived material and to a large extent success has been hindered by the lack of such reliable sources. Nevertheless, despite the fact that individual cancer cell lines do not reflect the complexity of clinical cancer tissues with fidelity (**Weinstein, 2012**), when compiled in large panels, it appears that they are able to recapitulate the genomic diversity of human cancers (**Iorio et al., 2016**). These panels can be readily utilised as platforms upon which expert systems for the prediction of pharmacological response may be developed. Currently the most significant resources of input data for drug response studies are publically available cell line repositories that include dose response data for a large number of compounds. Particularly, the Cancer Cell Line Encyclopedia (CCLE)\*, the Genomics of Drug Sensitivity in Cancer (GDSC)\* project and the National Cancer Institute drug screening panel (NCI-60)\* are the most widely used panels as they offer: **i**) baseline data (i.e molecular features from untreated samples) containing mutation, gene copy number, gene expression, and in the case of NCI-60 protein data information, and **ii**) various measurements of drug sensitivity in a large number of compounds (**Table 3**). Notably, NCI-60 has information for more than 1500 compounds, but in only 59 cell lines from 9 tissues, which makes CCLE and GDSC much more popular as they have data for more that 1000 cell lines derived from 15 and 36 cancer types, respectively (**Table 3**). Finally another unique resource that needs to be mentioned is the AstraZeneca-Sanger DREAM\* challenge drug-synergy dataset that contains 910 pairwise combinations of 118 drugs tested on 85 cell lines whose ‘omic’ profiling is available through GDSC (**Table 3**).

Another important issue in developing an efficient prediction model is the type of data used (**feature selection**). In general, in most models the input information consists mainly of single-nucleotide mutations, copy number variations, gene expression and of course the performance

to the therapeutic agent(s)/scheme (Jang et al., 2014; Costello et al., 2014; Daemen et al., 2013; Geeleher et al., 2014). Comparative analyses until now have demonstrated that in most cases gene expression determines the most powerful predictive features. On the other hand, integrated approaches, combining various “omic” modalities only marginally affect drug response (Jang et al., 2014; Costello et al., 2014). However, there are exceptions in this general tendency suggesting that more studies are required including combination of genomic, transcriptomic, epigenomic and proteomic profiles as data types (Moghaddas Gholami et al., 2013; Corte’s-Ciriano et al., 2016; Mendenet et al., 2013; Fey et al., 2015; Zhang et al., 2015; Niepel et al., 2013). Recently, simulations of signalling pathway activity has become the focus of investigation in prognostic models providing promising results (Fey et al., 2015); thus exemplifying that apart from developing novel computational methodologies, blending of different data types could help overcome study constraints.

**2c. Computational techniques and selection of prediction models:** The machine learning algorithms used to building drug response prediction models are mainly based on supervised learning techniques, although, as mentioned above, in many cases unsupervised methods provide the basis for the former (Moghaddas Gholami et al., 2013; Byers et al., 2013; Nicolau et al., 2011) (Table 2, Figures 1, 2). The methods presented in Table 2 can be broadly grouped in supervised and unsupervised learning methods. *Linear, Ridge, Lasso* and *Elastic Net regression* are examples of linear supervised learning, while kernel-based *support vector machines, decision-trees/random-forests* and artificial *neural networks* (shallow and deep) are examples of *non-linear supervised learning* (Table 2). *Principle Components Analysis (PCA)\** and *t-SNE\** (Table 2) are characteristic examples of linear and non-linear

*dimensionality reduction* techniques, respectively, which fall under *unsupervised learning*, along with clustering methodologies such as *k-means*, *hierarchical* and *k-nearest-neighbor clustering* (Table 2).

Although, all methods have pros and cons (Table 2) and no single approach can consistently surpass others on different settings, it appears that regression models tend to perform better when applied in diverse data sets (Stransky et al., 2015; Jang et al., 2014). Nonetheless, the ascertainment that no “true winner” exists has led to the development of different model building strategies. Ensembling different techniques and learning frameworks have emerged as a promising approach (a process termed *ensemble learning* – see Table 2). A characteristic example is the DREAM7 Challenge setup which utilized the Bayesian efficient multiple kernel learning (BEMKL)\* method that leveraged four machine-learning principles: i) **kernelized regression**, ii) **multi-view learning**, iii) **multi-task learning**, and iv) **Bayesian inference** (Costello et al., 2014). Particularly, *Kernel regression* gave mainly the advantage to capture non-linear relationships between the selected features and drug response, *multiview learning* integrated heterogeneous input data (views), even various representations of the same data set, into a single model, *multitask learning* shared information across all drugs implying simultaneous modelling, and finally *Bayesian inference* handled uncertainty from small sample size. Overall, BEMKL demonstrated improved predictive performance as depicted by the significant increase of signal-to-noise ratio (Costello et al., 2014). A variation of BEMKL is component-wise MKL (cwKBMF)\* which has the ability to identify groups of output variables and apply MKL providing supplementary information regarding the biological and structural characteristics of the drugs. In this manner it further refines the use of prior knowledge for various subsets, such as pathway information, thus enabling one the link the

target to the drug's mechanism of action (Ali and Aittokallio, 2019, Ammad-ud-din et al., 2016). The STREAM\* algorithm that combines Bayesian inference with Ridge regression is another paradigm of integrated approach trained and tested on public data (Neto et al., 2014) whereas, improved prediction was reported when *Elastic net* was combined with *Principle Component Analysis* (Park et al., 2014). **Network-based data representations** is a noteworthy method in which similarity networks among cell lines and between drugs are built independently, based on their expression and structural correlations, respectively. Subsequently, the two networks are integrated by linking the components of the first (cell lines) with the corresponding items (drugs) of the second producing a **weighted model** that reported drug response predictions (Fey et al., 2015; Zhang et al., 2015; Wang et al., 2014). It is apparent that the list of methodologies will grow as long as the “philosopher’s stone” of machine learning has yet to be invented. It is possible that the key to this challenge lies in artificial neural networks (**Table 2**) as discussed in **section 3**.

Once the desired computational algorithm is selected, it must be trained to the input data (**Figure 3**). During training, fine tuning of the **algorithm parameters** will lead to the model with optimal performance. The most widely used method in order to optimize the model parameters, without over-fitting, is **k-fold cross-validation (KF-CV)\*** (Stone M., 1974). According to KF-CV the data item set is divided in k subsets and the k-1 ones are used for training, while the model is evaluated in the k<sup>th</sup> item set. The process is iterated until all subsets are trained. Subsequently, the trained model is evaluated applying various **metrics of performance**, depending on the type modelling (regression vs classification) (**Table 4**). Recently, the power for drug response prediction was

shown to be further boosted by a process termed **transfer learning (TL)**. TL is a way of incorporating supporting information among different cell lines. In principal, training data include expression profiles and drug responses of tissue-specific cell material (cell-lines/samples) as well as material of related origin (tissue-type), while only expression status is required for the testing samples (**Turki et al., 2018**). For confusion avoidance it must be noted that the term Transfer Learning is also used in machine learning with Artificial Neural Networks where the model weights trained in one subdomain are transferred to another. This procedure has been shown to reduce training time and increase predictive accuracy (**Weiss et al., 2016**).

**2d. Testing the prediction models:** The ultimate goal of training is to build a model that **fits** to data beyond the ones utilized for developing the model (**Figure 3**). The best way to test the latter it is to implement it to blind data sets (**Figure 3**), preferentially clinical panels as the final objective of the whole workflow is to deliver tools that could help towards identifying tailored therapies for individual cancer patients (precision and personalized oncology) (see following **sub-section 2e**). In case a fully trained model fails to **generalise** then we are dealing with **overfitting** of the model (**Dietterich T., 1995**). Overfitting corresponds to an analysis that is adapted too close or exactly to a specific data set (the training data set) and falls short to predict additional data reliably, a.k.a fails to **generalise**. On the other end, there is **underfitting** when an *in silico* pipeline is unable to capture the underlying structure of a particular data set. In machine learning these conditions are termed **overtraining** and **undertraining**, respectively (**Dietterich T., 1995**). Especially, overfitting represents a crucial topic in the machine learning community and a number of factors appear to be responsible,



with the amount and diversity (number of features  $\gg$  samples) of the training data being the most important. The discrepancy in model performance in **testing** vs **training** steps is mathematically reflected by **cost functions** (Mehta et al., 2019) and the aim is to minimize as much as possible the cost effect. This is achieved by a process called **regularization** that intends to reduce the **variance** by increasing the **bias** in a step-wise manner. In simple words, **regularization**, which among others can be achieved with L1 (*Lasso*) or L2 (*Ridge*) penalisation, in combination with KF-CV schemes optimizes the parameters of the model and delivers the best model for eventual clinical validation.

**2e. Clinical applications and challenges to be met:** At the clinical level, research has been hampered mostly by the lack of large clinical cohorts that include both detailed “omic” data, especially genomic and transcriptomic profiles and responses to therapeutic agents. Most of the training-testing scenarios, as mentioned, are based on publicly available cell-line data resources (Table 5). Although the worth of cancer cell lines in everyday cancer research cannot be questioned, particularly in data mining procedures, as they offer a rapidly available set to screen (see section 4), their ability to develop drug prediction models for direct clinical use poses certain challenges (Caponigro and Sellers, 2011; Ross and Wilson, 2011). The most important one is that cancers are heterogeneous in nature and molecular matching with a cell line is not feasible, leading to leak of information during the *in silico* analysis (Hanahan and Weinberg, 2000; Hanahan and Weinberg, 2011; Turajlic et al., 2019). It has been suggested and shown that this hurdle can be circumvented by acquiring fresh patient material, keeping it under short-term culture; thus capturing better tumor heterogeneity

and the genomic/transcriptomic profile of the primary tumor site (Tentler et al., 2012; Day et al, 2015). Another important issue is that cell lines lack the influence of the tumor microenvironment (Hanahan and Weinberg, 2000; Hanahan and Weinberg, 2011). The tumor-microenvironment interplay determines not only cancer development but in certain ways also response to treatment (Wu and Dai, 2017).

As a result of these constraints there is a pressing need to evaluate the *in silico* technology that is constantly developed in “real patients”. In this vein, there are a number of studies that have implemented this approach testing machine learning models in patient-derived data from clinical trials or other patient cohorts. The most prominent ones are subsequently presented and discussed (see also Table 5). Geeleher et al. (Geeleher et al., 2014) trained models (*Ridge Regression* – Table 2) on gene expression data and drug responses from the Cancer Genome Project that is a subset of the GDSC, and tested them independently on publicly available data (TCGA\*) (Table 3) from clinical trials in myeloma and non-small-cell lung cancers (NSCLC)\*. Another group following a breast cancer cell-line based training approach but applying other algorithms (*Support Vector Machine* and *Random Forest* – Table 2) tested the built model in independent patient-derived data from TCGA (Daemen et al., 2013). On both occasions, the cell-line trained models predicted the therapeutic response, including relapse-free survival. The Biomarker-integrated Approaches of Targeted Therapy for Lung Cancer Elimination (BATTLE)\* study represents an important patient data source to evaluate (Kim et al., 2011) and discover consequential links between molecular markers and drug response. Based on this data resource, Byers and colleagues applying *hierarchical clustering* and *principal component analysis* (Table 2) identified a 76-gene expression signature that could distinguish non-small cell lung cancer samples with and

without EMT (epithelial to mesenchymal transition)\* features, demonstrating resistance of the former to EGFR\* inhibitors and how to overcome it (Byers et al., 2013). Likewise, using the BATTLE trial study Blumenschein et al, developed a gene expression signature of sorafenib efficacy (Blumenschein et al., 2013). Implementing an *elastic net* model (Table 2) in B-cell lymphoma cell lines with available gene expression datasets, Falgreen and collaborators generated a resistance gene signature in diffuse large B-cell lymphoma patients treated with CHO (Cyclophosphamide, Doxorubicin and Vincristine) (Falgreen et al., 2015). In colorectal cancer, Guinney et al., showed the prospective clinical utility of modelling specific cancer phenotypes and molecular traits. Specifically, training an *elastic net* model (Table 2) on a large series of colorectal cancer tissues according to their K-ras phenotype they were able to predict resistance to cetuximab, an anti-EGFR antibody used in K-ras wild-type colorectal cancer patients (Guinney et al., 2014). Using an **iterative rule-based approach** Chen et al., (Chen et al., 2015) revealed in ovarian cancer a 61-transcript expression signature for predicting patient's response (poor vs good survival groups) to platinum-taxane chemotherapy. Notably, when the expression signature was combined with BRCA1/2\* mutation status, a traditional prognostic marker for ovarian and breast cancer, patient stratification was further improved. The latter signifies the importance of combining molecular features, in certain cases.

Overall, the encouraging results of these studies render essential: **i)** the formulation of large patient-derived data-bases that will include apart from traditional clinical information detailed molecular high-throughput profiles and **ii)** a “methodological road map” that will guide the scientific community (basic researchers, bioinformaticians and clinician) in selecting the “best tool” for the “right question”.

### 3. Deep Learning neural networks (DLNN): an emerging “key player”

A new promising player with increased performance in the “arena” of machine learning is *neural networks* (Table 2). Particularly, its advanced form, Deep Learning neural networks (DLNN)\*, have the ability to “understand” complexity and multidimensionality, while have been effectively applied in various fields (e.g. image analysis, text mining, etc.) with increased classification accuracy compared to classical computational methods (Figure 4a) (Schmidhuber, 2015). DLNN is based on the modelling of high-level neural networks in flexible, multilayer systems of connected and interacting neurons, which perform numerous data abstractions and transformations (LeCun et al., 2015) (Figure 4b).

The basic unit in the model (Figure 4c) is the neuron, a biologically inspired model of the human neuron. In humans, the varying strengths of the neurons’ output signals travel along the synaptic junctions and are then aggregated as input for a connected neuron’s activation. In the DLNN models, the weighted combination ( $\alpha = \sum_{i=1}^n w_i x_i + b$ ) of input signals is aggregated, and then an output signal  $f(\alpha)$  transmitted by the connected neuron. The function  $f$  represents the nonlinear activation function used throughout the network and the bias  $b$  represents the neuron’s activation threshold. Multi-layer, feed-forward neural networks consist of many layers of interconnected neuron units (Figure 4b-c), starting with an input layer to match the feature space, followed by multiple (hidden) layers of nonlinearity, and ending with a linear classification layer to match the output space. The inputs and outputs of the model’s units follow the basic logic of the single neuron described above. Bias units are included in each non-output layer of the network. The weights linking neurons and biases with other neurons fully determine the output of the entire network. Learning occurs when these weights are adapted to minimize the error on

the labelled training data. More specifically, for each training example  $j$ , the objective is to minimize the loss function,  $L(W, B | j)$ . After the completion of the Test-set prediction, the classification performance is measured by calculating the Area Under the Curve (AUC)\* of the ROC-curve, Youden's Index, Sensitivity, Specificity, Accuracy (ACC)\*, Positive and Negative Predictive Values (PPV and NPV)\* and False Positive Rate (FPR)\* of the prediction (**Table 4**).

In a recent surge of interest, DLNN has been effectively applied to extract features from various large and complex data sets, including predicting drug-target interactions (**Wang et al., 2014b**), drug toxicity in the liver (**Xu et al., 2015**), pharmacological properties of drugs (**Aliper et al., 2016**) and automated diagnosis of histopathology slides (**Coudray et al., 2018**), among others. Altogether, studies using the DLNN architecture demonstrate its suitability for the analysis of complex biological data, as it can automatically construct complex features and allows for **multi-task learning** (**Bengio et al., 2013**). One of the main shortcomings of DLNNs apart from the computationally intensive long training times required, is their tendency to overfit due to the huge number of available model weights through fully connecting multiple hidden layers. This problem was however effectively addressed by a regularisation technique called dropout (**Hinton et al., 2012**). Dropout reduces overfitting by omitting a random percentage of the feature detectors on each training round, thus allowing the successful generalisation of the DLNN.

To the best of our knowledge and at the time of preparation of this review, there is only one report applying Deep Learning for response to therapy in clinical settings, namely Chiu and collaborators (**Chiu et al., 2019**) who applied deep learning models to predict drug response in 9059 tumors of 33 cancer types from TCGA. The authors identified as effective, drugs that are known to be potent in specific cancers, such as EGFR inhibitors in non-small cell lung cancer, as

well as novel drugs for a specific type of cancer, such as vinorelbine for TTN-mutated tumors. Notably, the authors of the aforementioned study used a type of DLNN called *autoencoder*. Autoencoders are unsupervised DLNNs that are trained to reconstruct their input (**Table 2**) (**Hinton et al., 1994**). In order for the networks to do so, they learn the most meaningful structures and relationships among the input features by compressing the information through a bottleneck hidden layer in the middle of the hidden layer stack that forces the network to discard all unnecessary information. These kind of networks have a wide range of applications, namely:

- (i) Dimensionality reduction, where the neurons of the bottleneck layers are used as non-linear multi-dimensional principal components (**Taghanaki et al., 2017**),
- (ii) Compression, where the structure learned is stored as the compressed version of the original information (**Tan and Eswaran, 2011**),
- (iii) Missing value imputation, where the intricate relationships of the input features that were learned were used to impute missing values (**Talwar et al., 2018**),
- (iv) Denoising, where the structure learned was used to reconstruct the input without the noise which was discarded during the learning process (**Creswell and Bharath, 2019**).

Interestingly, Rampásek and collaborators demonstrated the use of deep autoencoders to integrate drug response information along with gene expression perturbation for building more effective predictive models of drug response in cell lines (**Rampásek et al., 2019**).

#### **4. A novel *in silico* screening process based on Association Rule Mining (ARM study)**

Given their molecular profiling data, both large cell-line panels (CCLE and GDSC) have been utilized in attempts to identify biomarkers for predicting drug response of specific cancer cell-lines (**Barretina et al., 2012; Garnett et al., 2012**). Previous efforts to define biomarkers of

drug response primarily employ **general linear models**, penalized linear modelling techniques, to identify cooperative interactions among multiple genes and transcripts across the genome and define response signatures for each drug (**Forbes et al., 2015**). While efficient, these algorithms suffer certain limitations since when used for feature selection, as described in previous studies (**Barretina et al., 2012; Garnett et al., 2012**), the derived results are simple associations between a single gene and drug response. If, however, one wishes to explore the relevance of a more complex feature-space relationship (two or three-way interactions among simple features in all possible combinations) to the drug response, the process is convoluted. This is primarily due to the fact that these algorithms fall-short in automatically evaluating all possible combinations including multi-way interactions of a large number of features against a response variable without further implementation. Furthermore, multi-feature models generated by such algorithms are difficult to interpret in terms of biological relevance. When utilised as a classifier to predict whether a sample will be resistant or sensitive to a drug, given its molecular profile, the **general linear** algorithms do not perform optimally. This is due to the fact that at the core of these algorithms lays linear regression, as opposed to non-linear classifiers, such as **Random-Forests** and **Kernel-based** models. The later have been shown to outperform the **general linear** algorithms in the task of actually predicting drug response, as demonstrated in a recent proof of concept study on a panel of 53 breast cancer cell lines evaluated for pharmacological response against 28 anti-cancer drugs (**Iorio et al., 2016**).

A promising methodology used by large businesses that overcomes the primary limitations of the **general linear models** for feature selection, yet capable of analysing enormous volume of transaction data to discover all possible associations between the data features is the **Association Rule Mining** (ARM) (**Table 2, Figure 5**). Previous studies moved along the same lines to

produce easily interpretable logical rules out of similar pharmacogenomic datasets (**Iorio et al., 2016; Masica and Karchin, 2013**). Within this context we developed a resource of rules linking candidate genes as cancer drivers to drug response using this *in silico* methodology. The reason is that association rule mining provides an efficient big-data ready framework that is able to evaluate a huge sample space of associations among features including multi-way interactions with more than 30 different objective measures (**Tan et al., 2004**). Additionally, the output of the algorithm comes in the form of easily interpretable rules, making knowledge extraction and meta-analysis a more straightforward process.

First, a comprehensive dataset was constructed using the GDSC and Cosmic Cell line project (CCLP) databases (**Figure 5a**). This task was achieved by merging data from the CCLP and GDSC. GDSC was used (**Garnett et al., 2012**) as a drug response data source for 251 therapeutic compounds, which provided IC<sub>50</sub> values for each compound, as well as information on tissue origin. Information on total gene mRNA expression, number of DNA copies and mutational status was obtained from the Cosmic Cell line project (CCLP) (**Forbes et al., 2015**). CCLP was preferred over CCLE as a data source since it provides profiles on 1,074 cancer cell lines and is not limited to the mutational status of only 1,600 genes, as is the case with CCLE. GDSC contains dose response data for the 1,001 CCLP cell lines only and therefore only those were used in our analysis. Although NCI-60 contains the largest number of therapeutic compounds tested for pharmacologic activity, it was excluded as a data source, as the number of cell lines presented is very small compared to the other resources used. A summary of the compiled pharmacogenomics dataset is presented in **Supplementary Figure 1**.

Applying the Apriori algorithm (**Agrawal et al., 1993**) significant associations from all of the possible combinations of the features from the main dataset (tissue of origin, gene expression,



mutation status, CNV plus drug response) were extracted, in order to generate a large rule-set, containing all *tissue-to-gene*, *tissue-to-drug*, *gene-to-gene*, *gene-to-drug* and *drug-to-drug* associations. The main bottleneck in the application of association rule mining is the computationally intensive requirements. While this will likely improve as computing power increases, due to hardware limitations in the currently presented resource we maintained only the *tissue-to-drug*, *gene-to-drug* and *drug-to-drug* associations for the present study. Gene-to-gene associations, which constitute an enormous RAM intensive rule-set, were discarded. Details and metrics of the Apriori algorithm can be found in **Figures 5 and 6**. The basic interest metrics, available by the *arules* R package, and utilised were *support*, *confidence* and *lift*. *Support* is the frequency of the rule occurrence in the total dataset, while *confidence* is the frequency of rule occurrence in the cases of the dataset fulfilling the left hand side of the rule and *lift* is the factor by which, the co-occurrence of A and B exceeds the expected probability of A and B co-occurring, had they been independent. Relationships between confidence and support metrics (for top 10,000 one-way and 100,000 two-way rules) are visualized in the scatterplots in **Supplementary Figure 2**. To select significant non-random rules by controlling the false positive rate (FPR)\*, a randomization approach was applied based on running the Apriori algorithm on a permuted version of the initially employed dataset (see “**Association Rule Mining: Apriori Algorithm / Dynamic Thresholding**” in Supportive material section). At 5% FPR, 1,326,251 1-way rules were identified: 2,124 of them where tissue to drug, 989,163 gene-expression to drug, 110,442 gene-CNV\* to drug and 224,522 gene-mutation to drug (**Supplementary File 1, “one\_way\_rule\_count”**). All identified rules are available online via an interactive Rshiny application: <https://compbio.nyumc.org/drugs/> (**Supplementary File 2**). Representative outputs from the web application, confirming prior-knowledge, are presented in

**Figure 7 and Supplementary Figure 3.** The user can search for a tissue, gene or drug of interest, filter using different metrics and visualize the results and download the data. The biological relevance of the rules generated was examined both computationally (based on prior knowledge) and experimentally, as demonstrated in the following sections.

#### ***4a. Rule verification based on prior knowledge***

To explore the potential biological relevance of our statistically significant association rules, we examined whether: (1) known predictors of drug response are present in our rule set, and, (2) drugs and their targets are present together in sensitivity-associated rules if the target(s) are mutated and/or over-expressed.

##### ***MAPK and PI3K signalling pathway***

Initially, we followed an unbiased approach, where we performed k-means clustering (see **“Association Rule Mining – Apriori Algorithm” in Supportive material section**) of the 1000 rules with the largest support (k=50) for drug sensitivity associated with: (a) the ERK/MAPK<sup>\*</sup> signalling, and, (b) the PI3K<sup>\*</sup> signalling (**Supplementary File 1: “1-way rules” and Supplementary File 3: “Drugs”**). First, the clustering of the top rules associated with ERK/MAPK signalling revealed that mutated BRAF<sup>\*</sup> (known to be essential to ERK/MAPK signalling (**McCain, 2013**)) was present among the top 50 cluster centres (**Figure 8a**). Additionally, this clustering revealed that the melanoma cell lines are expected to be highly sensitive to BRAF and MEK<sup>\*</sup> inhibitors, a prediction that can be verified in the literature with studies showing that combined BRAF and MEK inhibition is one of the most effective treatments for melanomas (**Figure 8a**) (**Long et al., 2014**). The *half maximal inhibitory*

*concentration* (IC<sub>50</sub>) values of the drugs included in this group indicate increased sensitivity for melanoma cell lines and for cell lines carrying mutated BRAF as compared to the total dataset (p-value < 0.05) (**Figure 8b**). Second, the clustering of the top rules associated with PI3K signalling revealed the presence of mutated PTEN among the top 50 cluster centres (**Figure 9a**). PTEN\* is a direct PIK3CA\* suppressor (**Carracedo and Pandolfi, 2008**) that is frequently mutated in cancer with loss-of-function mutations (**Rodriguez-Escudero et al., 2011**), which in turn leads to increased PIK3CA activity. Notably, mutated PIK3CA was also present in the mutated-PTEN cluster (**Figure 9b, right panel**). Given that both, PTEN and PIK3CA, belong to the same pathway, the fact that the onco-suppressor (PTEN) is deactivated at the same time that the oncogene (PIK3CA) is further activated by hot-spot gain-of-function mutations can be conceptualized as a variation of the Knudson double-hit hypothesis (**Knudson, 1971**). IC<sub>50</sub> heatmaps (**Figure 9c, right panel**) indicate that cell lines with PIK3CA mutations are significantly more responsive (p-value < 0.01) to inhibitors targeting the PI3K pathway compared to cell lines with wild-type PIK3CA, which seem to be resistant to the same inhibitors. These observations confirm that clustering of significant rules can provide relevant insights regarding the molecules that are related to responsiveness to certain classes of drugs.

#### *Multiple drug response, p53 and PARPi resistance*

To further validate our models, we also looked for specific genes known to be implicated in drug resistance and/or sensitivity. We observed that the *ABCB1* gene that encodes the Multidrug-Resistance-1 (MDR1)\* protein, was found in our rule set to be linked with resistance to multiple drugs when it is over-expressed (55 out of 57 drugs), while when suppressed it is linked with sensitivity (7 out of 9 drugs) (**Supplementary File 1: “1-way rules”**). In addition, our rules

indicate that EGFR over-expression and suppression are significantly associated with Lapatinib sensitivity and resistance, respectively, which is in agreement with previous findings demonstrating that EGFR expression can efficiently affect response to this tyrosine kinase inhibitor (TKI)\* (**Rusnak et al., 2007**) (**Supplementary File 1: “1-way rules”**). Moreover, we observed that known predictors of drug response are highly ranked in our rule set. For example, suppressed NAD(P)H dehydrogenase 1 (NQO1)\* and over-expressed MDM2, a p53 inhibitor, which are known predictors of sensitivity for the drugs 17-AAG (Tanespimycin) and Nutlin-3, respectively (**Kelland et al., 1999; Muller et al., 2007**), are present in our rule-set with lift values 4 and 4.06, respectively, which are in the top 25% quantile of lift values in our list of significant 1-way rules (**Supplementary File 1: “1-way rules”**). Of note, three recent reports demonstrating that inactivation of genes encoding subunits of the shieldin complex (REV7, SHLD1-3)\* cause resistance to poly(ADP-ribose) polymerase inhibition (PARPi)\* in BRCA1-deficient cells and tumours (**Mirman et al., 2018; Noordermeer et al., 2018; Dev et al., 2018**), were also confirmed by the Apriori data mining process (**Supplementary File 1: “1-way rules”; Figure 7 and Supplementary Figure 3**). In addition, and within the same context, we identified in the literature a list of 96 genes whose status was experimentally linked with PARPi (**Figure 10; Supplementary File 1**). We queried our database to identify rules associating these 96 genes with all PARP inhibitors enlisted. We found a total of 166 rules describing associations of 71/96 (74%) genes with PARP inhibitors. Specifically, we spotted 24 rules with gene mutations, 13 rules with gene copy-number variations (CNVs) and 129 rules with gene-expression (**Supplementary File 1, “PARPi”**). To exclude the possibility that the observed matches were due to chance alone, we performed a Monte-Carlo simulation taking into account all relevant parameters (see **Supplementary Materials, section 2.5**). We demonstrated (**Figure 10**) that the

number of the reported matches could not have been observed randomly ( $p$ -value = 0.008766261), highlighting the effectiveness of the data mining process applied.

#### *Drug response in small-cell lung cancer*

The following two examples indicate how the association rules, when allowing for interactions (2-way), can be used to gain further insight in the molecular mechanisms of drug resistance in Small-Cell Lung Cancer (SCLC)\* and identify potential points of intervention.

The 1-way rules indicate a large pattern of multi-drug resistance (93 drugs) involving SCLC (**Supplementary File 1: “1-way rules”**). SCLC accounts for approximately 15% of all lung cancer cases (**Planchard and Le Pechoux, 2011**). It is considered one of the most aggressive malignancies mainly due to the rapid development of multi-drug resistance (**Yeh et al., 2005**), which is in agreement with our finding. The 2-way rules (**Supplementary File 1: “2-way rules”**), indicate that the Growth hormone-releasing hormone (GHRH)\* over-expression greatly increases the lift-value (hence statistical significance) to 39 of the above drugs, suggesting it may be involved in multi-drug resistance mechanisms. It is known that inhibition of GHRH activity using antagonists yields high anti-tumour activity by impeding cell proliferation (**Kiaris et al., 2000; Popovics et al., 2017**). Furthermore, GHRH activity has been linked to drug-resistance in triple negative breast cancer (**Perez et al., 2014**). Herein, by including interactions in association rule mining we were able to infer that GHRH antagonists could be potentially used in combination with specific chemotherapeutic agents for the effective treatment of SCLC. This is further supported by the fact that in preclinical models monotherapy with novel GHRH antagonists resulted in significant suppression of SCLC and NSCLC tumor growth (**Wang et al., 2018**).

In a separate example, with the 1-way rules (**Supplementary File 1: “1-way rules”**), we observed statistically significant resistance to Obatoclax-Mesylate, a BCL<sup>\*</sup>-family inhibitor, with a lift-value of 2.47 in 22 out of 66 SCLC cell lines (33.3%). With the 2-way rules (**Supplementary File 1: “2-way rules”**), we noted that SMAD3<sup>\*</sup> down-regulation greatly increases the lift-value to 4.77, since resistance to Obatoclax-Mesylate is observed in 9 out of 14 SCLC cell lines under-expressing SMAD3 (64.3%). SMAD3 is known to promote apoptosis through transcriptional inhibition of BCL-2 (Yang et al., 2006). SCLC cell lines under-expressing SMAD3 clearly possess increased levels of BCL-2, which correlates well with the phenotype of resistance to a BCL-2 inhibitor, such as Obatoclax-Mesylate. In this example, association rule mining precisely elucidated a specific mechanism of resistance of SCLC tumors to BCL-family inhibitors, by highlighting a unique molecule that presents high mechanistic relevance to BCL-inhibition.

#### **4b. Rule Experimental Validation**

##### *Drug-specific target selection and experimental validation*

The generated 1-way rule-set consists of 1,326,251 statistically significant rules (**Supplementary File 1: “1-way rules”**) as selected by the Dynamic Thresholding procedure. In order to ascertain that our rule-set consists of meaningful rules in an unbiased and systematic way, we devised a systematic 4-step rule-based gene-selection algorithm (**Supplementary Figure 4b1; “Validation procedure” in Supportive material section**) to identify novel therapeutic targets and then we proceeded with their experimental validation. Particularly this algorithm associates gene expression with drug resistance patterns across a big number of diverse drugs and is designed to narrow down the long list of more than 16,000 genes to one with

only few selected candidates, the silencing of which should increase the efficacy of a specifically applied treatment. Using this algorithm 128 rules corresponding to 128 genes per drug were identified on average, summing to a total of 30,639 rules (**Supplementary File 3: “si\_t\_resistance\_genes\_all\_drugs”**). We applied the algorithm on all available drugs (**Supplementary File 3: “t\_resistance\_genes\_all\_drugs”**), but in order to provide a practical application we focused on the efficacy enhancement of Doxorubicin (**Supplementary File 3: “DoxoTargetsSelectionResGenes”**). The experimental validation of the algorithm was designed to monitor whether Doxorubicin treatment in combination with the silencing of each identified target resulted in a synergistic increase in efficacy across four cancer cell lines, namely A549 (lung carcinoma), NCI-H1299 (lung carcinoma derived from metastatic site), MCF7 (breast adenocarcinoma derived from metastatic site) and Saos-2 (osteosarcoma). Our algorithm selected 72 out of 16445 total genes available from our initial dataset (**Supplementary File 3: “DoxoTargetsSelectionResGenes”**). We randomly chose five targets from the list, for experimental validation, namely *MAGI3*<sup>\*</sup>, *POF1B*<sup>\*</sup>, *PDIA3*<sup>\*</sup>, *CD151*<sup>\*</sup> and *NPTN*<sup>\*</sup>, none of which are specifically connected with Doxorubicin efficacy in the biomedical literature (**Supplementary File 3: “DoxoTargetsSelectionResGenes”**). As predicted by our algorithm, in all cases siRNA treatment led to a significant sensitization of the examined cells to Doxorubicin (**Supplementary Figure 4a, 4b1, 4b2i-ii, 4c1, 4d2; Supplementary File 3: “Doxorubicin\_IC50”; Supplementary Materials**). Decreased soft agar colony formation further supported these findings (**Supplementary Figure 4e1**). Potential mechanistic insights underlying these results are proposed in **Table 6**. As a negative control, we reversed the algorithm for all drugs to select genes that upon silencing should decrease efficacy of Doxorubicin (**Supplementary File 3: “si\_t\_sensitivityGenes\_all\_durgs”**). We randomly chose

again 5 targets from the list for experimental validation namely *TP53*<sup>\*</sup>, *CTCF*<sup>\*</sup>, *CCND3*<sup>\*</sup>, *ARHBD1B*<sup>\*</sup> and *ZCCHC7*<sup>\*</sup> (**Supplementary File 3: “DoxoTargetsSelectionSensGenes”**). In accordance to our predictions, siRNA treatments led to a significant increase in resistance to Doxorubicin (**Supplementary Figure 4a, 4b1, 4b3i-ii, 4c2, 4d3, 4e2; Supplementary File 3: “Doxorubicin\_IC50”; Supplementary Materials**). Presumable underlying mechanisms of increased resistance are proposed in **Table 7**.

#### *ID1 as a biomarker of response to PI3K-targeted therapies*

After demonstrating that rule-clustering delivers relevant results, we present an example of how the rules can be used to gain novel insights on biomarker discovery for drug response. The PI3K signalling pathway rule clustering, links the suppression of the *ID1*<sup>\*</sup> gene to sensitivity to 10 out of 16 drugs targeting the PI3K pathway with high lift and support values (**Figure 9a**). Inhibitor of DNA binding 1 (*ID1*) is a transcription regulator, widely reported as linked to tumour metastasis when over-expressed (**Eisfeld et al., 2017; Jin et al., 2016**) and known to activate the PI3K pathway (**Li et al., 2012**), while inhibition of *ID1* expression suppresses cancer invasion and progression (**Murase et al., 2016; Tominaga et al., 2016**). IC<sub>50</sub> heatmaps (**Figure 9b,c; left panel**) indicate that cell lines under-expressing *ID1* are significantly more responsive to inhibitors targeting the PI3K pathway compared to cell lines over-expressing *ID1* ( $p < 0.01$ ). These results imply that apart from being used as a therapeutic target *per se*, *ID1* could be utilised as a predictive biomarker for response to PI3K-targeted therapies, as its expression seems to distinguish sensitive from resistant cell lines more efficiently than the actual *PIK3CA* mutation status (**Figure 9b,c; right panel; Figure 11a**). Within this context, we recently demonstrated that chronic expression of the tumor-suppressor *p21*<sup>WAF/Cip1</sup>, in a *p53*-deficient



environment, exhibited an oncogenic behaviour, by “escaping” from the antitumor barrier of senescence and generating aggressive and chemo-resistant clones (**Figure 11b**) (**Galanos et al., 2016**). In line with the above observations, ID1 was found up-regulated in these cells (**Galanos et al., 2016**). To experimentally validate the *in silico* prediction, we interrogated the sensitivity of the p21<sup>WAF/Cip1</sup> “escaped” clones for two PI3K inhibitors, namely CAL-101 and ZSTK474 from our panel (**Figure 9a; Figure 11a**), before and after *ID1* silencing. As shown in **Figure 11c (left panel)**, the “escaped” p21<sup>WAF/Cip1</sup> cells showed IC<sub>50</sub> values of 0.141  $\mu$ M and 1.26  $\mu$ M for CAL-101 and ZSTK474, respectively. Concurrent silencing of *ID1* with administration of each inhibitor significantly reduced the corresponding IC<sub>50</sub> values and decreased colony formation (**Figure 11c right panel**), suggesting that inhibition of ID1 confers to PI3K chemo- sensitivity in accordance with the *in silico* model (**Figure 9; Figure 11a**).

Moreover, in the ID1 rule-cluster, over-expression of 4 other genes was found to be highly related with sensitivity to PI3K-pathway inhibitors, namely *ZNF22*<sup>\*</sup>, *GMIP*<sup>\*</sup>, *LYL1*<sup>\*</sup> and *SAMSNI*<sup>\*</sup> (**Figure 9b, left panel**). Interestingly, LYL1 (Lymphoblastic Leukemia Associated Hematopoiesis Regulator 1) is known to be implicated in the development of leukemia (**Meng et al., 2005**) and lymphoma (**Zhong et al., 2007**), both representing promising target groups for anti-PI3K/mTOR<sup>\*</sup> agents (**Bertacchini et al., 2015; Blachly and Baiocchi, 2014**). SAMSNI (SAM Domain, SH3 Domain And Nuclear Localization Signals 1) is an intriguing case since it appears to act as a tumour suppressor in certain malignancies such as multiple myeloma (**Noll et al., 2014**), gastric cancer (**Kanda et al., 2016**), lung cancer (**Yamada et al., 2008**) and hepatocellular carcinoma (**Sueoka et al., 2015**), whereas its over-expression has been associated with poor survival in glioblastoma multiforme (**Yan et al., 2013**), a malignancy where drug resistance represents a major challenge (**Haar et al., 2012**). Its detection in the rule-set concurs

with recent developments suggesting that targeting the PI3K pathway could be a potential therapeutic option to overcome drug resistance in glioblastoma multiforme (**Sami and Karsy, 2013**).

*CDC6 overexpression as an indicator of resistance to MAPK pathway inhibitors*

Among the results extracted from the Apriori data mining process we noticed three rules that drew our attention as they were related with the role of deregulated replication licensing in cancer, one of the main research fields of our group (**Karakaidos et al, 2004; Bartkova et al., 2006; Lontos et al., 2007; Sideridou et al., 2011; Petrakis et al., 2016; Galanos et al., 2016**). They linked CDC6 (Cell division cycle 6)\* overexpression (termed oncogenic CDC6) with resistance to MAPK (Mitogen-Activated Protein Kinase) inhibition (**Figure 12a**). In most cases, this type of resistance is associated with mutations that either render the MAPK pathway insensitive to treatment or reactivate alternative components of the signaling route bypassing the inhibitory block (**Logue and Morrison, 2012; Pritchard and Hayward, 2013; Varmus et al., 2016**) (**Figure 12b**). We and others have shown that CDC6 is deregulated in many types of cancer from their earliest stages and is an indicator of poor prognosis (**Karakaidos et al, 2004; Bartkova et al., 2006; Lontos et al., 2007; Sideridou et al., 2011; Petrakis et al., 2016; Galanos et al., 2016**) (**Supplementary Figure 5a**). According to the oncogene-induced DNA damage model for cancer development (**Halazonetis et al., 2008**), oncogenic CDC6 fuels genomic instability by causing replication stress and DNA damage (**Lontos et al., 2007; Gorgoulis et al., 2018; Petrakis et al., 2016; Sideridou et al., 2011; Galanos et al., 2016; Galanos et al., 2018; Komseli et al., 2018**). As DNA damage accumulates the DDR (DNA Damage Response)\* and the error-free repair pathways are overwhelmed leading, due to selective

pressure, to inactivation or exhaustion of vital DDR/R (DDR and Repair) components. Consequently, there is a shift to error-prone repair that leads to escape from the anti-tumor barriers of senescence and apoptosis, by generating a landscape of mutations that promote cancer development (**Halazonetis et al., 2008; Galanos et al., 2016; Galanos et al., 2018; Gorgoulis et al., 2018**). As CDC6 functions downstream of the RAS-RAF-MEK1/2-ERK1/2 pathway (**Lunn et al., 2010; Liu et al., 2010; Steckel et al., 2012; Di Micco et al., 2006; Sideridou et al., 2011; Hills & Diffley 2014; Petrakis et al., 2016**) and mutational activation of the MAPK signalling is a prominent feature of many cancer types (**Fang and Richardson, 2005; Dhillon et al., 2007; Kim and Choi, 2010; Logue and Morrison, 2012; Pritchard and Hayward, 2013**), we postulated that the aforementioned rules (**Figure 12a**) possibly reflect one way of how oncogenic CDC6 promotes cancer development. In particular, tumors with high levels of CDC6 would at some point select to rewire cellular signalling to another pathway, parallel to MAPK cascade that does not comprise RAF and MEK1/2\*, thus rendering these tumors unresponsive to MEK1/2 inhibitors, such as Trametinib or RDEA119. In other words, it is most unlikely that a RAF or MEK1/2 inhibitor would be effective when a downstream effector of this pathway is overexpressed and active. From cancer biology perspective activation of a parallel pathway would exert an additive tumor promoting effect phenocopying the activation of the RAS-RAF-MEK1/2-ERK1/2 pathway, as suggested in colon cancer (**Hanahan and Weinberg, 2010**).

To test this hypothesis we employed a CDC6-inducible normal cellular model that recapitulates in relatively short period all stages of cancer development (**Komseli et al., 2018**) (**Figure 12c**). Briefly, and in accordance to our model (**Halazonetis et al., 2008**), chronic CDC6 expression triggered the anti-tumor barrier of senescence (precancerous stage) that was eventually overridden leading to the emergence of aggressive clones (cancerous stage) (**Figure**

**12c) (Komseli et al., 2018).** We performed three biological replicates of this cancer evolution experiment and examined by WGS (whole genome sequencing) means the genetic alterations acquired. Interestingly, and in accordance to our assumption, among the alterations found, all three clones harbored an R55T amino-acid substitution located in codon 55 of MAP2K3 (Mitogen-Activated Protein Kinase Kinase 3) (**Figure 12d**). This is a key component of the stress/cytokines-induced p38 MAPK pathway located upstream of its end-effector, the p38 kinase (**Cuadrado and Nebreda, 2010**). It acts in parallel with the RAS-RAF-MEK1/2-ERK1/2 signaling route and has a significant role in cell proliferation and malignant transformation (**Cuadrado and Nebreda, 2010; Baldari et al., 2015**) (**Figure 12b**). This mutation has been also reported in colorectal cancer (<https://hive.biochemistry.gwu.edu/biomuta/proteinview/P46734>) and potentially affects the structure and function of MAP2K3 (see details in **Supplementary Figure 5b**). Of note, as we previously showed the activated p38 pathway promotes colon cancer progression (**Gupta et al., 2014**). A strong indication that this mutation is associated with activation of the MAPK p38 pathway is the increased phosphorylation levels of its downstream effector p38 in the escaped-from-senescence aggressive clones (**Figure 12e**). Within the same line and in support to the rules, the escaped-from-senescence cells harboring high levels of CDC6 were significantly more resistant to the MEK1/2 inhibitor PD98059 than the non-induced (OFF) cells with very low CDC6 levels (**Figure 12f; Supplementary Figure 5c**).

## 5. Comparison of ARM study with other frameworks

We compared our rules with the respective ones identified in various databases, namely GDSC (Genomics of Drug Sensitivity in Cancer) (**Iorio et al., 2016**), CCLE (Cancer Cell Line

Encyclopedia) (**Barretina et al., 2012**) and CTRP (Cancer Therapeutic Response Portal) (**Seashore-Ludlow et al., 2015**) (see **Supplementary Materials, Supplementary File 1**).

### ***GDSC-Genomics of Drug Sensitivity in Cancer***

ANOVA: i) *Mutations*: **Iorio et al., (Iorio et al., 2016)** identified 268 one-way mutated gene-to-drug relationships, of which 82 were matched with our one-way rules (overlap 34.75%). Interestingly, for genes bearing clinical relevance such as *BRAF*, *EGFR*, *PTEN*, *TP53*, *FLT3*<sup>\*</sup>, *KRAS* and *PIK3CA*, the overlap of our one-way rules with Iorio et al., was: 92.31%, 60.00%, 100.00%, 30.77%, 33.33%, 83.33% and 100.00%, respectively (**Supplementary File 1 "Mut Clinically Relevant Iorio V", Figure 13**) (**Sethi et al., 2013**). ii) *Copy number variations*: They (**Iorio et al., 2016**) identified 10,201 gain/losses related to drug responses, of which 827 were also present in our rules (overlap 8.11%). iii) *Gene expression*: 5361 drug response interaction were identified, 1089 of which were also identified by our pipeline (overlap 20.3%).

LOBICO: Regarding the comparison of our rules with the multiple relationship models generated by **Iorio et al.**, through LOBICO (**Iorio et al., 2016**), we identified 114 out of a total of 1112 LOBICO models that could be compared with our one-way rules, of which 38 were present in our rule-set (overlap 33.33%), and 2 rules that could be compared with our two-way rules, namely “CDKN2A-loss AND MYC-gain => Etoposide-Sensitivity” and “CDKN2A-loss AND MLL2-mutation=>SB52334-Sensitivity”. Although loss of CDKN2A is connected with Etoposide and SB52334 Sensitivity in our two-way rule-set (**Supplementary File 1 “two-way rules”**), MYC-gain and MLL2-mutation were not identified. It must be noted that the 1112 LOBICO models contain multiple genes combined together through the logic operators AND, OR and NOT which are then connected to a specific drug response. As a result, this scheme produces rules that cannot be directly compared to our rule-set. Therefore, no statistical

conclusion may be drawn due to the low number of compatible rules extracted.

### ***CCLE-Cancer Cell Line Encyclopedia***

Data were drawn from CCLE as follows (Stransky et al., 2015). i) *Mutations*: 421 mutation-drug response interactions were identified in the CCLE data-set, with 14 being in common with our rules (overlap 3.3%). ii) *Copy number variations*: From the 103 identified copy number variation-drug response interactions, 4 were also found in our rules (overlap 3.9%). iii) *Gene expression*: Finally 7382 gene-expression to drug response interaction were identified, 1000 of which were in common with the current study (overlap 13.55%) (Supplementary File 1).

### ***CTRP-Cancer Therapeutic Response Portal\****

Seashore-Ludlow et al., was utilized as the CTRP data-source (Seashore-Ludlow et al., 2015). The particular analysis was performed at a level connecting gene mutations to drug-cluster wide response, the common element of the cluster being the molecular target. An in-house R-script (see Supplementary Materials section 2.4) was utilised to subset the CTRP dataset to our collection of drugs and identify relevant rules from our dataset. From the 10829 gene mutation to drug cluster response interactions, 1811 were represented in our rules (overlap 16.72%).

## **6. Perspectives and future challenges**

Hitherto, the degrees of overlap that the various *in silico* settings demonstrate (Figure 13), suggest the necessity of applying multiple analytical techniques to maximize information retrieval. Moreover, although all *in silico* pipelines suffer to certain extent from false positive and negative outcomes it is possible that several, at first glance, contradictory results could simply reflect a U-shaped curve drug response or behaviour (Figure 14). In other words, deviation from optimal activity, either too little or too much has the same impact. A characteristic example is

mTOR1, where both, low and high activity, lead to insulin resistance (**Laplane and Sabatini, 2012**). The complexity of biological processes is indeed evident in everyday clinical practice. For example, not all patients with EGFR mutations respond to treatment with EGFR TKIs (Tyrosine Kinase Inhibitors) (**Zhong et al., 2017**). On the other hand, a subset of patients with wild type EGFR also responds to EGFR TKIs (**Ulivi et al., 2015; Xu et al., 2016; Koinis et al., 2018**). Likewise, vemurafenib-resistant melanomas that depend on the drug to proliferate can become re-sensitized following a "drug holiday period" (**Das-Thakur et al., 2013, Schreuer et al., 2017**).

Among the other methods described the screening pipeline based on ARM's could be effectively applied in the future in Biomarker-Guided Adaptive Clinical Trial Designs (**Antoniou et al., 2016**). Patient's molecular profile can be obtained and compared against the extracted from ARM's gene-drug response rules. These results can form the basis to design appropriate sophisticated target gene interventions. Initially they could be tested on patient-derived primary 2D and 3D cancer cell cultures (**Das et al., 2015**) and/or on xenograft models (**Siolas and Hannon, 2013**). The most effective schemes could be applied in clinical trials, constant monitoring for administration of personalised dosing and use of circulating tumour cell assays and ctDNA for early detection of the emergence of resistance (**Palmirotta et al., 2018**). Moreover, the pharmacogenetic databases could be further expanded by increasing the number of cancer cell lines, including patient-derived cell lines, as well as by increasing the number of therapeutic genes analysed by the system. Additionally, integration of other layers of "omics" information, including meta-genomics, proteomics, phospho-proteomics, interactomics and metabolomics will further enhance the applicability of this method, eventually increasing the power of the presented *in silico* process. Last but not least, the algorithm may be implemented in

a wider expert decision support system (artificially intelligence based) to assist oncologists in predicting drug response and selecting the best drug candidates for precision based therapy.



**ACKNOWLEDGMENTS**

We would like to thank Drs K Evangelou, T Karamitros, A Papanastasiou and P Vasileiou for their valuable help. Financial support was from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grants agreement No. 722729 (SYNTRAIN); the Welfare Foundation for Social & Cultural Sciences (KIKPE), Greece; Pentagon Biotechnology Ltd, UK; DeepMed IO Ltd, UK and NKUA-SARG grants No 70/3/9816, 70/3/12128. Dr. Tsirigos and the NYU Applied Bioinformatics Laboratories (ABL) are partially supported by the Cancer Center Support Grant P30CA016087 at the Laura and Isaac Perlmutter Cancer Center (A.T.).

**Conflict**

None of the authors have any competing interests.

## References

- Aas, T., Børresen, A.L., Geisler, S., Smith-Sørensen, B., Johnsen, H., Varhaug, J.E., et al. (1996) Specific P53 mutations are associated with de novo resistance to doxorubicin in breast cancer patients. *Nat Med* 2: 811-4.
- Abrams, S.L., Steelman, L.S., Shelton, J.G., Wong, E.W., Chappell, W.H., Bäsecke, J., et al (2010) The Raf/MEK/ERK pathway can govern drug resistance, apoptosis and sensitivity to targeted therapy. *Cell Cycle* 9: 1781-91.
- Agrawal, R., Imielinski, T., and Swami, A. (1993) Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data* (Washington, D.C., USA: ACM), pp. 207-216.
- Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., et al. (2013) Signatures of mutational processes in human cancer. *Nature* 500: 415-421.
- Ali, M., and Aittokallio, T. (2019) Machine learning and feature selection for drug response prediction in precision oncology applications. *Biophys Rev.* 11: 31-39.
- Ammad-ud-din, M., Khan, S.A., Malani, D., Murumägi, A., Kallioniemi, O., Aittokallio, T., et al. (2016) Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. *Bioinformatics* 32: i455–i463

Aliper, A., Plis, S., Artemov, A., Ulloa, A., Mamoshina, P., and Zhavoronkov, A. (2016). Deep Learning Applications for Predicting Pharmacological Properties of Drugs and Drug Repurposing Using Transcriptomic Data. *Molecular Pharmaceutics* 13, 2524-2530.

Andersson, E.I., Pützer, S., Yadav, B., et al. (2018) Discovery of novel drug sensitivities in T-PLL by high-throughput ex vivo drug testing and mutation profiling. *Leukemia* 32: 774–787.

Antoniou, M., Jorgensen, A.L., and Kolamunnage-Dona, R. (2016) Biomarker-Guided Adaptive Trial Designs in Phase II and Phase III: A Methodological Review. *PloS One* 11: e0149803.

Azuaje F. (2017) Computational models for predicting drug responses in cancer research. *Brief Bioinform.* 18: 820-829.

Baldari, S., Ubertini, V., Garufi, A., D'Orazi, G., and Bossi, G. (2015) Targeting MKK3 as a novel anticancer strategy: molecular mechanisms and therapeutical implications. *Cell Death Dis* 6: e1621.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., et al. (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483: 603-607.

Bartkova, J., Rezaei, N., Lontos, M., Karakaidos, P., Kletsas, D., Issaeva, N., et al. (2006) Oncogene-induced senescence is part of the tumorigenesis barrier imposed by DNA damage

checkpoints. *Nature* 444: 633-7.

Bartlett, D.W., and Davis, M.E. (2006) Insights into the kinetics of siRNA-mediated gene silencing from live-cell and live-animal bioluminescent imaging. *Nucleic Acids Res* 34: 322-33.

Beesley, P.W., Herrera-Molina, R., Smalla, K.H., and Seidenbecher, C. (2014) The Neuroplastin adhesion molecules: key regulators of neuronal plasticity and synaptic function. *J Neurochem* 131: 268-83.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *IEEE transactions on pattern analysis and machine intelligence*.

Bertacchini, J., Heidari, N., Mediani, L., Capitani, S., Shahjahani, M., Ahmadzadeh, A., et al. (2015) Targeting PI3K/AKT/mTOR network for treatment of leukemia. *Cellular and Molecular Life Sciences* 72: 2337-2347.

Blachly, J.S., and Baiocchi, R.A. (2014) Targeting PI3-kinase (PI3K), AKT and mTOR axis in lymphoma. *Br J Haematology* 167: 19-32.

Boissel, N., Cayuela, J.M., Preudhomme, C., Thomas, X., Grardel, N., Fund, X., et al. (2002) Prognostic significance of FLT3 internal tandem repeat in patients with de novo acute myeloid leukemia treated with reinforced courses of chemotherapy. *Leukemia* 16: 1699-704.

Breiman, L. (2001) Random Forests. *Machine Learning*. 45: 5–32.

Breiman, L., Friedman, J., Stone, C.J., and Olshen, R.A. (1984) *Classification and Regression Trees*. Chapman and Hall/CRC ISBN 9780412048418 - CAT# C4841

Brookshear J.G. (2008) *Computer Science: An Overview*. Addison-Wesley Publishing Company  
USA ISBN: 9780321524034

Byers, L.A., Diao, .L, Wang, J., et al. (2013) An epithelial-mesenchymal transition gene signature predicts resistance to EGFR and PI3K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance. *Clin Cancer Res* 19: 279–90.

Byrne, H.J., Bonnier, F., and Farhane, Z. (2017) Doxorubicin kinetics and effects on lung cancer cell lines using in-vitro Raman microspectroscopy: binding signatures, drug resistance and DNA repair. *J Biophotonics* 10: 1333-1346.

Byron, S.A., Loch, D.C., Pollock, P.M. (2012) Fibroblast growth factor receptor inhibition synergizes with Paclitaxel and Doxorubicin in endometrial cancer cells. *Int J Gynecol Cancer* 22: 1517-26.

Campbell, B.B., Light, N., Fabrizio, D., et al. (2017) Comprehensive analysis of hypermutation in human cancer. *Cell* 171: 1042–1056

Canela, A., Maman, Y., Jung, S., Wong, N., Callen, E., Day, A., et al. (2017) Genome Organization Drives Chromosome Fragility. *Cell* 170: 507-521.

Caponigro, G., and Sellers, W.R. (2011) Advances in the preclinical testing of cancer therapeutic hypotheses. *Nat Rev Drug Discov* 10: 179–87.

Carracedo, A., and Pandolfi, P.P. (2008) The PTEN-PI3K pathway: of feedbacks and cross-talks. *Oncogene* 27: 5527-5541.

Chang, C.I., Xu, B.E., Akella, R., Cobb, M.H., and Goldsmith, E.J. (2002) Crystal structures of MAP kinase p38 complexed to the docking sites on its nuclear substrate MEF2A and activator MKK3b. *Mol Cell* 9: 1241-9.

Chen, MA. (2009) Co-delivery of doxorubicin and Bcl-2 siRNA by mesoporous silica nanoparticles enhances the efficacy of chemotherapy in multidrug resistant cancer cells. *Small* 5: 2673–2677.

Chen, P., Huhtinen, K., Kaipio, K., et al. (2015) Identification of prognostic groups in high-grade serous ovarian cancer treated with platinum-taxane chemotherapy. *Cancer Res* 75: 2987–98.

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., et al. (2016) Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* 32: 1220-1222.

Chen, Y., Bathula, S.R., Li, J., and Huang, L. (2010) Multifunctional nanoparticles delivering small interfering RNA and doxorubicin overcome drug resistance in cancer. *J Biol Chem* 285: 22639-50.

Chen, Y., Wu, J.J., and Huang, L. (2010b) Nanoparticles targeted with NGR motif deliver c-myc siRNA and doxorubicin for anticancer therapy. *Mol Ther* 18: 828-34.

Chiu, Y.C., Chen, H.H., Zhang, T2, Zhang S., Gorthi, A., Wang, L.J., et al. (2019) Predicting drug response of tumors from integrated genomic profiles by deep neural networks. *BMC Med Genomics*. 12(Suppl 1): 18.

Cortes, C., and Vapnik, V.N. (1995) Support-vector networks. *Machine Learning*. 20: 273–297.

Corté's-Ciriano, I., van Westen, G.J., Bouvier, G., et al. (2016) Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel. *Bioinformatics* 32: 85–95.

Costello, J.C., Heiser, L.M., Georgii, E., et al. (2014) A community effort to assess and improve drug sensitivity prediction algorithms. *Nat Biotechnol* 32: 1202–12.

Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., and Tsirigos, A. (2018). Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med* 24, 1559-1567.

Crespi, A., Bertoni, A., Ferrari, I., Padovano, V., Della Mina, P., Berti, E., et al. (2015) POF1B localizes to desmosomes and regulates cell adhesion in human intestinal and keratinocyte cell lines. *J Invest Dermatol* 135: 192-201.

Creswell, A., and Bharath A.A. (2019) Denoising Adversarial Autoencoders. *IEEE Trans Neural Netw Learn Syst.* 30: 968-984.

Cuadrado, A., and Nebreda, A.R. (2010) Mechanisms and functions of p38 MAPK signalling. *Biochem J* 429: 403-17.

Daemen, A., Griffith, O.L., Heiser, L.M., et al. (2013) Modeling precision treatment of breast cancer. *Genome Biol* 14: R110.

Das Thakur, M., Salangsang, F., Landman, A.S., Sellers, W.R., Pryer, N.K., Levesque, M.P., et al. (2013) Modelling vemurafenib resistance in melanoma reveals a strategy to forestall drug resistance. *Nature* 494: 251-5.

Das, V., Bruzzese, F., Konecny, P., Iannelli, F., Budillon, A., and Hajdich, M. (2015) Pathophysiologically relevant in vitro tumor models for drug screening. *Drug Discovery Today* 20: 848-855.

Day, C.P., Merlino, G., Van Dyke, T. (2015) Preclinical mouse cancer models: a maze of



opportunities and challenges. *Cell* 163: 39–53.

Dev, H., Chiang, T.W., Lescale, C., de Krijger, I., Martin, A.G., Pilger, D., et al. (2018) Shieldin complex promotes DNA end-joining and counters homologous recombination in BRCA1-null cells. *Nat Cell Biol* 20: 954-965.

Dhillon, A.S., Hagan, S., Rath, O., and Kolch, W. (2007) MAP kinase signalling pathways in cancer. *Oncogene* 26: 3279-90.

Dietterich, T. (1995) Overfitting and undercomputing in machine learning. *ACM Computing Surveys (CSUR)*. 27: 326-327.

Di Micco, R., Fumagalli, M., Cicalese, A., Piccinin, S., Gasparini, P., Luise, C., et al. (2006) Oncogene-induced senescence is a DNA damage response triggered by DNA hyper-replication. *Nature* 444: 638-42.

Eisfeld, A.K., Kohlschmidt, J., Mrozek, K., Volinia, S., Blachly, J.S., Nicolet, D., et al. (2017) Mutational Landscape and Gene Expression Patterns in Adult Acute Myeloid Leukemias with Monosomy 7 as a Sole Abnormality. *Cancer Research* 77: 207-218.

Eliades, T., Pratsinis, H., Athanasiou, A.E., Eliades, G., and Kletsas, D. (2009) Cytotoxicity and estrogenicity of Invisalign appliances. *American Journal of Orthodontics and Dentofacial Orthopedics* 2009; 136: 100-103.

Enslen, H., Branchio, D.M., and Davis, R.J. (2000) Molecular determinants that mediate selective activation of p38 MAP kinase isoforms. *EMBO J* 19: 1301-11.

Evangelou, K., Lougiakis, N., Rizou, S.V., Kotsinas, A., Kletsas, D., Muñoz-Espín, D., et al. (2017) Robust, universal biomarker assay to detect senescent cells in biological specimens. *Aging Cell* 16: 192-197.

Falgreen, S., Dybkær, K., Young, K.H., et al. (2015) Predicting response to multidrug regimens in cancer patients using cell line experiments and regularised regression models. *BMC Cancer* 15: 235.

Fang, J.Y., and Richardson, B.C. (2005) The MAPK signalling pathways and colorectal cancer. *Lancet Oncol* 6: 322-7.

Fey, D., Halasz, M., Dreidax, D., et al. (2015) Signaling pathway models as biomarkers: patient-specific simulations of JNK activity predict the survival of neuroblastoma patients. *Sci Signal* 8: ra130.

Fiedler, W., Kayser, S., Kebenko, M., Janning, M., Krauter, J., Schittenhelm, M., et al (2015) A phase I/II study of sunitinib and intensive chemotherapy in patients over 60 years of age with acute myeloid leukaemia and activating FLT3 mutations. *Br J Haematol* 169: 694-700.

Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., et al. (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research* 43: D805-811.

Friedman, J.H. (2001) Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* 29: 1189-1232.

Frismantas, V., Dobay, M.P., Rinaldi, A., et al. (2017) Ex vivo drug response profiling detects recurrent sensitivity patterns in drug-resistant acute lymphoblastic leukemia. *Blood* 129: e26–e37

Galanos P, Pappas G, Polyzos A, Kotsinas A, Svolaki I, Giakoumakis NN, et al. (2018) Mutational signatures reveal the role of RAD52 in p53-independent p21-driven genomic instability. *Genome Biology* 19: 37.

Galanos, P., Vougas, K., Walter, D., Polyzos, A., Maya-Mendoza, A., Haagensen, E.J., et al. (2016) Chronic p53-independent p21 expression causes genomic instability by deregulating replication licensing. *Nature Cell Biology* 18: 777-789.

Garnett, M.J., Edelman, E.J., Heidorn, S.J., Greenman, C.D., Dastur, A., Lau, K.W., et al. (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* 483: 570-575.

Geeleher, P., Cox, N.J., and Huang, S.R. (2014) Clinical drug response can be predicted using

baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biol* 15: R47.

Gel, B., and Serra, E. (2017) “karyoploteR: an R / Bioconductor package to plot customizable genomes displaying arbitrary data.” *Bioinformatics* 33, 3088-3090.

Gillet, J.P., Calcagno, A.M., Varma, S., et al. (2011) Redefining the relevance of established cancer cell lines to the study of mechanisms of clinical anti-cancer drug resistance. *Proc Natl Acad Sci USA* 108: 18708–13.

Gorgoulis, V.G., Pefani, D.E., Pateras, I.S., and Trougakos, I.P. (2018) Integrating the DNA damage and protein stress responses during cancer development and treatment. *J Pathol* 246: 12-40.

Guinney, J., Ferte', C., Dry, J., et al. (2014) Modeling RAS phenotype in colorectal cancer uncovers novel molecular traits of RAS dependency and improves prediction of response to targeted agents in patients. *Clin Cancer Res* 20: 265–72.

Gupta, J., del Barco Barrantes, I., Igea, A., Sakellariou, S., Pateras, I.S., Gorgoulis, V.G., et al. (2014) Dual function of p38 $\alpha$  MAPK in colon cancer: suppression of colitis-associated tumor initiation but requirement for cancer cell survival. *Cancer Cell* 25: 484-500.

Haar, C.P., Hebbar, P., Wallace, G.C.t., Das, A., Vandergrift, W.A., 3rd, Smith, J.A., et al. (2012) Drug resistance in glioblastoma: a mini review. *Neurochemical Research* 37: 1192-1200.

Haeuw, J.F., Goetsch, L., Bailly, C., Corvaia, N. (2011) Tetraspanin CD151 as a target for antibody-based cancer immunotherapy. *Biochem Soc Trans* 39: 553-8.

Hahsler, M., Grün, B., and Hornik, K. (2005) arules - A Computational Environment for Mining Association Rules and Frequent Item Sets. *Journal of Statistical Software* 14: 1 - 25.

Halazonetis, T.D., Gorgoulis, V.G., and Bartek, J. (2008) An oncogene-induced DNA damage model for cancer development. *Science* 319: 1352-1355.

Hanahan, D., and Weinberg, R.A. (2000) The hallmarks of cancer. *Cell* 100: 57-70.

Hanahan, D., and Weinberg, R.A. (2011) Hallmarks of cancer: the next generation. *Cell* 144: 646-74.

Hastie, T., Tibshirani, R., and Friedman, J. (2001) *The Elements of Statistical Learning*. Springer Series in Statistics Springer New York Inc., New York, NY, USA.

Henderson, D.J., and Parmeter, C.F. (2015) *Applied Nonparametric Econometrics* by Daniel J. Henderson. Cambridge Core. doi: 10.1017/CBO9780511845765

Hills, S.A., and Diffley, J.F. (2014) DNA replication and oncogene-induced replicative stress. *Curr Biol* 24: R435-44.

Hinton, G.E. (2007) Learning multiple layers of representation. Trends in Cognitive Sciences. 11: 428–434.

Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R. (2012) Improving neural networks by preventing co-adaptation of feature detectors. arXiv:1207.0580

Hinton, G. E., and Zemel, R. S. (1994) Autoencoders, minimum description length and Helmholtz free energy. In Advances in neural information processing systems pp. 3-10.

Hoadley, K.A., Yau, C., Wolf, D.M., et al. (2014) Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell 158: 929–944

Holland, P., and Cooper, J. (1999) Protein modification: Docking sites for kinases. Current Biology 9: R329-R331.

Hui, Z, and Hastie, T. (2005) Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society, Series B: 301–320.

Hussmann, M., Janke, K., Kranz, P., Neumann, F., Mersch, E., Baumann, M., et al E. (2015) Depletion of the thiol oxidoreductase ERp57 in tumor cells inhibits proliferation and increases sensitivity to ionizing radiation and chemotherapeutics. Oncotarget 6: 39247-61.

Iorio, F., Knijnenburg, T.A., Vis, D.J., Bignell, G.R., Menden, M.P., Schubert, M., et al. (2016) A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* 166: 740-754.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013) *An Introduction to Statistical Learning with Applications in R*. Springer. ISBN 978-1-4614-7138-7

Jang, I.S., Neto, E.C., Guinney, J., et al. (2014) Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. *Pac Symp Biocomput* 63–74.

Jin, X., Jeon, H.M., Jin, X., Kim, E.J., Yin, J., Jeon, H.Y., et al. (2016) The ID1-CULLIN3 Axis Regulates Intracellular SHH and WNT Signaling in Glioblastoma Stem Cells. *Cell Reports* 16: 1629-1641.

Kanda, M., Shimizu, D., Sueoka, S., Nomoto, S., Oya, H., Takami, H., et al. (2016) Prognostic relevance of SAMS1 expression in gastric cancer. *Oncology Letters* 12: 4708-4716.

Karakaidos, P., Taraviras, S., Vassiliou, L.V., Zacharatos, P., Kastrinakis, N.G., Kougiou, D., et al. (2004) Overexpression of the replication licensing regulators hCdt1 and hCdc6 characterizes a subset of non-small-cell lung carcinomas: synergistic effect with mutant p53 on tumor growth and chromosomal instability--evidence of E2F-1 transcriptional control over hCdt1. *Am J Pathol* 165: 1351-65.

Kasthuber, E.R., Lowe, S.W. (2017) Putting p53 in context. *Cell* 170: 1062-1078.

Kelland, L.R., Sharp, S.Y., Rogers, P.M., Myers, T.G., and Workman, P. (1999) DT-Diaphorase expression and tumor cell sensitivity to 17-allylamino, 17-demethoxygeldanamycin, an inhibitor of heat shock protein 90. *Journal of the National Cancer Institute* 91: 1940-1949.

Kiaris, H., Schally, A.V., and Varga, J.L. (2000) Suppression of tumor growth by growth hormone-releasing hormone antagonist JV-1-36 does not involve the inhibition of autocrine production of insulin-like growth factor II in H-69 small cell lung carcinoma. *Cancer Letters* 161: 149-155.

Kim, E.K., and Choi, E.J. (2010) Pathological roles of MAPK signaling pathways in human diseases. *Biochim Biophys Acta* 1802: 396-405.

Kim, E.S., Herbst, R.S., Wistuba, I.I., Lee, J.J., Blumenschein, G.R. Jr, Tsao. A., et al. (2011) The BATTLE trial: personalizing therapy for lung cancer. *Cancer Discov* 1: 44-53.

Kleppmann M. (2017) *Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems*, ISBN-13: 978-1449373320

Knudson, A.G., Jr. (1971) Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A* 68: 820-823.

Koinis, F., Voutsina, A., Kalikaki, A., Koutsopoulos, A., Lagoudaki, E., Tsakalaki, E., et al. (2018)



Long-term clinical benefit from salvage EGFR tyrosine kinase inhibitors in advanced non-small-cell lung cancer patients with EGFR wild-type tumors. *Clin Transl Oncol* 20: 140-149.

Komseli, E.S., Pateras, I.S., Krejsgaard, T., Stawiski, K., Rizou, S.V., Polyzos, A., et al. (2018) A prototypical non-malignant epithelial model to study genome dynamics and concurrently monitor micro-RNAs and proteins in situ during oncogene-induced senescence. *BMC Genomics* 19; 37.

Kragelj, J., Palencia, A., Nanao, M.H., Maurin, D., Bouvignies, G., Blackledge, M., et al. (2015) Structure and dynamics of the MKK7-JNK signaling complex. *Proc Natl Acad Sci U S A* 112: 3409-14.

Lacombe, A., Lee, H., Zahed, L., Choucair, M., Muller, J.M., Nelson, S.F., et al. (2006) Disruption of POF1B binding to nonmuscle actin filaments is associated with premature ovarian failure. *Am J Hum Genet* 79: 113-9.

Lapante, M., and Sabatini, D.M. (2012) mTOR Signaling in Growth Control and Disease *Cell* 149: 274-293.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436-444.

Lee, S.J., Park, J.W., Kang, B.S., Lee, D.S., Lee, H.S., Choi, S., et al. (2017) Chronophin activation is necessary in Doxorubicin-induced actin cytoskeleton alteration. *BMB Rep* 50: 335-340.

Li, H., and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler Transform. *Bioinformatics* 26: 589-95.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009) 1000 Genome Project Data Processing Subgroup. 1000 genome project data processing subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics* 25: 2078–2079.

Li, W., Wang, H., Kuang, C.Y., Zhu, J.K., Yu, Y., Qin, Z.X., et al. (2012) An essential role for the Id1/PI3K/Akt/NFkB/survivin signalling pathway in promoting the proliferation of endothelial progenitor cells in vitro. *Molecular and Cellular Biochemistry* 363: 135-145.

Liang, J., Tong, P., Zhao, W., et al. (2014) The REST gene signature predicts drug sensitivity in neuroblastoma cell lines and is significantly associated with neuroblastoma tumor stage. *Int J Mol Sci* 15: 11220–33.

Libbrecht, M.W., and Noble, W.S. (2015) Machine learning applications in genetics and genomics. *Nature Reviews Genetics* 16: 321-332.

Liontos, M., Koutsami, M., Sideridou, M., Evangelou, K., Kletsas, D., Levy, B., et al. (2007) Deregulated overexpression of hCdt1 and hCdc6 promotes malignant behavior. *Cancer Res* 67: 10899-909.

Lior, R., and Maimon, O. (2005) "Clustering methods." Data mining and knowledge discovery handbook. Springer US, 321-352.

Liu, T., Wang, S., Wang, L., Wang, J., Li, Y. (2015) Targeting CD151 by lentivirus-mediated RNA interference inhibits luminal and basal-like breast cancer cell growth and invasion. *Mol Cell Biochem* 407: 111-21.

Liu, Y., Hock, J.M., Sullivan, C., Fang, G., Cox, A.J., Davis, K.T., et al. (2010) Activation of the p38 MAPK/Akt/ERK1/2 signal pathways is required for the protein stabilization of CDC6 and cyclin D1 in low-dose arsenite-induced cell proliferation. *J Cell Biochem* 111: 1546-55.

Logue, J.S., and Morrison, D.K. (2012) Complexity in the signaling network: insights from the use of targeted inhibitors in cancer therapy. *Genes Dev* 26: 641-50.

Long, G.V., Stroyakovskiy, D., Gogas, H., Levchenko, E., de Braud, F., Larkin, J., et al. (2014) Combined BRAF and MEK inhibition versus BRAF inhibition alone in melanoma. *The New England Journal of Medicine* 371: 1877-1888.

Lovitt, C.J., Shelper, T.B., Avery, V.M. (2018) Doxorubicin resistance in breast cancer cells is mediated by extracellular matrix proteins. *BMC Cancer* 18: 41.

Lu, J.H., Shi, Z.F., Xu, H. (2014) The mitochondrial cyclophilin D/p53 complexation mediates doxorubicin-induced non-apoptotic death of A549 lung cancer cells. *Mol Cell Biochem* 389: 17-

24.

Lunn, C.L., Chrivia, J.C., and Baldassare, J.J. (2010) Activation of Cdk2/Cyclin E complexes is dependent on the origin of replication licensing factor Cdc6 in mammalian cells. *Cell Cycle* 9: 4533-41.

Luo, J., Schumacher, M., Scherer, A., Sanoudou, D., Megherbi, D., Davison, T., et al. (2010) A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. *Pharmacogenomics J.* 10: 278–291.

Maron, M.E. (1961) Automatic Indexing: An Experimental Inquiry. *Journal of the ACM.* 8: 404–417.

Masica, D.L., and Karchin, R. (2013) Collections of simultaneously altered genes as biomarkers of cancer cell drug response. *Cancer research* 73: 1699-1708.

McCain, J. (2013) The MAPK (ERK) Pathway: Investigational Combinations for the Treatment Of BRAF-Mutated Metastatic Melanoma. *P & T: a peer-reviewed journal for formulary management*, 38: 96-108.

McCulloch, W., and Pitts, W. (1943) A Logical Calculus of Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics.* 5: 115–133.

Mehta, P., Bukov, M., Wang, C.H., Alexandre, G.R.D., Richardson, C., et al. (2019) A high-bias, low-variance introduction to Machine Learning for physicists. arXiv:1803.08823v3

Mendelsohn, A.R., Hamer, J.D., Wang, Z.B., Brent, R. (2002) Cyclin D3 activates Caspase 2, connecting cell proliferation with cell death. *Proc Natl Acad Sci USA* 99: 6871-6.

Menden, M.P., Iorio, F., Garnett, M., et al. (2013) Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS One* 8: e61318.

Menden, M.P., Wang, D., Mason, M.J., Szalai, B., Bulusu, K.C., Guan, Y., et al. (2019) Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nat Commun.* 10: 2674.

Meng, H., Liong, M., Xia, T., Li, Z., Ji, Z., Zink, J.I., et al. (2010) Engineered design of mesoporous silica nanoparticles to deliver doxorubicin and P-glycoprotein siRNA to overcome drug resistance in a cancer cell line. *ACS Nano* 4: 4539-50.

Meng, H., Mai, W.X., Zhang, H., Xue, M., Xia, T., Lin, S., et al. (2013) Codelivery of an optimal drug/siRNA combination using mesoporous silica nanoparticles to overcome drug resistance in breast cancer in vitro and in vivo. *ACS Nano* 7: 994-1005.

Meng, Y.S., Khoury, H., Dick, J.E., and Minden, M.D. (2005) Oncogenic potential of the transcription factor LYL1 in acute myeloblastic leukemia. *Leukemia* 19: 1941-1947.

Milligan, G.W., and Cooper, M. C. (1988) A study of standardization of variables in cluster analysis. *Journal of Classification* 5: 181–204.

Min, X., Akella, R., He, H., Humphreys, J.M., Tsutakawa, S.E., Lee, S.J., Tainer, J.A., et al. (2009) The structure of the MAP2K MEK6 reveals an autoinhibitory dimer. *Structure* 17: 96-104.

Mirman, Z., Lottersberger, F., Takai, H., Kibe, T., Gong, Y., Takai, K., et al. (2018) 53BP1-RIF1-shieldin counteracts DSB resection through CST- and Pol $\alpha$ -dependent fill-in. *Nature* 560: 112-116.

Moghaddas Gholami, A., Hahne, H., Wu, Z., et al. (2013) Global proteome analysis of the NCI-60 cell line panel. *Cell Rep* 4: 609–620.

Muller, C.R., Paulsen, E.B., Noordhuis, P., Pedeutour, F., Saeter, G., and Myklebost, O. (2007) Potential for treatment of liposarcomas with the MDM2 antagonist Nutlin-3A. *International Journal of Cancer* 121: 199-205.

Murase, R., Sumida, T., Kawamura, R., Onishi-Ishikawa, A., Hamakawa, H., McAllister, S.D., et al. (2016) Suppression of invasion and metastasis in aggressive salivary cancer cells through targeted inhibition of ID1 gene expression. *Cancer Letters* 377: 11-16.

Neal, B., Mittal, S., Baratin, A., Tania, V., Scicluna, M., Lacoste-Julien, S., and Mitliagkas, I.

(2018) A Modern Take on the Bias-Variance Tradeoff in Neural Networks. arXiv, 1810.08591.

Negrini, S., Gorgoulis, V.G., and Halazonetis, T.D. (2010) Genomic instability--an evolving hallmark of cancer. *Nature Reviews Molecular Cell Biology* 11: 220-228.

Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized Linear Models. *Journal of the Royal Statistical Society. Series A*, 135: 370-384.

Neto, E.C., Jang, I.S., Friend, S.H., et al. (2014) The STREAM algorithm: computationally efficient ridge-regression via Bayesian model averaging, and applications to pharmacogenomic prediction of cancer cell line sensitivity. *Pac Symp Biocomput* 27–38.

Nicolau, M., Levine, A.J., Carlsson G. (2011) Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proc Natl Acad Sci USA* 108: 7265–70.

Nidheesh, N., Abdul Nazeer, K.A., and Ameer, P.M. (2017) An enhanced deterministic K-Means clustering algorithm for cancer subtype prediction from gene expression data. *Comput Biol Med* 91: 213-221.

Niepel, M., Hafner, M., Pace, E.A., et al. (2013) Profiles of basal and stimulated receptor signaling networks predict drug response in breast cancer lines. *Sci Signal* 6: ra84.

Noll, J.E., Hewett, D.R., Williams, S.A., Vandyke, K., Kok, C., To, L.B., et al. (2014) SAMS1 is a tumor suppressor gene in multiple myeloma. *Neoplasia* 16: 572-585.

Noordermeer, S.M., Adam, S., Setiaputra, D., Barazas, M., Pettitt, S.J., Ling, A.K., et al. (2018) The shieldin complex mediates 53BP1-dependent DNA repair. *Nature* 560: 117-121.

Núñez-Enríquez, J.C., Bárcenas-López, D.A., Hidalgo-Miranda, A., Jiménez-Hernández, E., Bekker-Méndez, V.C., Flores-Lujano, J., et al. (2016) Gene Expression Profiling of Acute Lymphoblastic Leukemia in Children with Very Early Relapse. *Arch Med Res* 47: 644-655.

O'Connor, M.J. (2015) Targeting the DNA Damage Response in Cancer. *Mol Cell* 60: 547-60.

Padovano, V., Lucibello, I., Alari, V., Della Mina, P., Crespi, A., Ferrari, I., et al. (2011) The POF1B candidate gene for premature ovarian failure regulates epithelial polarity. *J Cell Sci* 124: 3356-68.

Palmirotta, R., Lovero, D., Cafforio, P., Felici, C., Mannavola, F., Pellè, E., et al. (2018) Liquid biopsy of cancer: a multimodal diagnostic tool in clinical oncology. *Ther Adv Med Oncol* 10: 1758835918794630.

Park, H., Shimamura, T. Miyano, S., et al. (2014) Robust prediction of anti-cancer drug sensitivity and sensitivity-specific bio- marker. *PLoS One* 9: e108990.



Patil, Y.B., Swaminathan, S.K., Sadhukha, T., Ma, L., and Panyam, J. (2010) The use of nanoparticle-mediated targeted gene silencing and drug delivery to overcome tumor drug resistance. *Biomaterials* 31; 358-65.

Pemovska, T., Johnson, E., Kontro, M., et al (2015) Axitinib effectively inhibits BCR-ABL1 (T315I) with a distinct binding conformation. *Nature* 519: 102–105

Pereira, E., Camacho-Vanegas, O., Anand, S., et al. (2015) Personalized circulating tumor DNA biomarkers dynamically predict treatment response and survival in gynecologic cancers. *PLoS One* 10: e0145754.

Perez, R., Schally, A.V., Popovics, P., Cai, R., Sha, W., Rincon, R., et al. (2014) Antagonistic analogs of growth hormone-releasing hormone increase the efficacy of treatment of triple negative breast cancer in nude mice with doxorubicin; A preclinical study. *Oncoscience* 1: 665-673.

Petrakis, T.G., Komseli, E.S., Papaioannou, M., Vougas, K., Polyzos, A., Myrianthopoulos, V., et al. (2016) Exploring and exploiting the systemic effects of deregulated replication licensing. *Semin Cancer Biol* 37-38: 3-15.

Pearson, K. (1901) "On Lines and Planes of Closest Fit to Systems of Points in Space". *Philosophical Magazine*. 2: 559–572.

Planchard, D., and Le Pechoux, C. (2011) Small cell lung cancer: new clinical recommendations and current status of biomarker assessment. *European Journal of Cancer* 47 Suppl 3: S272-283.

Popovics, P., Schally, A.V., Salgueiro, L., Kovacs, K., and Rick, F.G. (2017) Antagonists of growth hormone-releasing hormone inhibit proliferation induced by inflammation in prostatic epithelial cells. *Proc Natl Acad Sci U S A* 114: 1359-1364.

Porter, A.P., Papaioannou, A., Malliri, A. (2016) Deregulation of Rho GTPases in cancer. *Small GTPases* 7: 123-38.

Pritchard, A.L., and Hayward, N.K. (2013) Molecular pathways: mitogen-activated protein kinase pathway mutations and drug resistance. *Clin Cancer Res* 19: 2301-9.

Pritchard, J.R., Bruno, P.M., Hemann, M.T., Lauffenburger, D.A. (2013) Predicting cancer drug mechanisms of action using molecular network signatures. *Mol Biosyst* 9: 1604-19.

R Core Team R (2016): A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing).

Rangel, R., Guzman-Rojas, L., Kodama, T., Kodama, M., Newberg, J.Y., Copeland, N.G., et al. (2017) Identification of new tumor suppressor genes in triple-negative breast cancer. *Cancer Res pii: canres.0785.2017*.

Ramirez-Gonzalez, R.H., Leggett, R.M., Waite, D., Thanki, A., Drou, N., Caccamo, M., et al. (2013) StatsDB: platform-agnostic storage and understanding of next generation sequencing run metrics. *F1000Res*. 2: 248.

Rampášek, L., Hidru, D., Smirnov, P., Haibe-Kains, B., and Goldenberg, A. (2019) Improving drug response prediction via modeling of drug perturbation effects. *Bioinformatics* pii: btz158.

Rickardson, L., Fryknäs, M., Dhar, S., Lövborg, H., Gullbo, J., Rydåker, M., et al. (2005) Identification of molecular mechanisms for cellular drug resistance by combining drug activity and gene expression profiles. *Br J Cancer* 93: 483-92.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., et al. (2011) Integrative genomics viewer. *Nat Biotechnol* 29: 24-6.

Rodriguez-Escudero, I., Oliver, M.D., Andres-Pons, A., Molina, M., Cid, V.J., and Pulido, R. (2011) A comprehensive functional analysis of PTEN mutations: implications in tumor- and autism-related syndromes. *Human Molecular Genetics* 20: 4132-4142.

Roidl, A., Berger, H.J., Kumar, S., Bange, J., Knyazev, P., and Ullrich, A. (2009) Resistance to chemotherapy is associated with fibroblast growth factor receptor 4 up-regulation. *Clin Cancer Res* 15: 2058-66.

Ross, N.T., and Wilson, C.J. (2014) In vitro clinical trials: the future of cell-based profiling.

Front Pharmacol 5: 121.

Ruder, S. (2017) An Overview of Multi-Task Learning in Deep Neural Networks. arXiv, 1706.05098.

Rusnak, D.W., Alligood, K.J., Mullin, R.J., Spehar, G.M., Arenas-Elliott, C., Martin, A.M., et al. (2007) Assessment of epidermal growth factor receptor (EGFR, ErbB1) and HER2 (ErbB2) protein expression levels and response to lapatinib (Tykerb, GW572016) in an expanded panel of human normal and tumour cell lines. *Cell Proliferation* 40: 580-594.

Saad, M., Garbuzenko, O.B., and Minko T. (2008) Co-delivery of siRNA and an anticancer drug for treatment of multidrug-resistant cancer. *Nanomedicine* 3: 761-76.

Sahai, E., and Marshall, C.J. (2002) RHO-GTPases and cancer. *Nat Rev Cancer* 2: 133-42.

Sami, A., and Karsy, M. (2013) Targeting the PI3K/AKT/mTOR signaling pathway in glioblastoma: novel therapeutic agents and advances in understanding. *Tumour Biology* 34: 1991-2002.

Santana-Codina, N., Carretero, R., Sanz-Pamplona, R., Cabrera, T., Guney, E., Oliva, B., et al (2013) A transcriptome-proteome integrated network identifies endoplasmic reticulum thiol oxidoreductase (ERp57) as a hub that mediates bone metastasis. *Mol Cell Proteomics* 12: 2111-25.

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural networks: the official journal of the International Neural Network Society* 61, 85-117.

Schreuer, M., Jansen, Y., Planken, S., Chevolet, I., Seremet, T., Kruse, V., et al. (2017) Combination of dabrafenib plus trametinib for BRAF and MEK inhibitor pretreated patients with advanced BRAFV<sup>600</sup>-mutant melanoma: an open-label, single arm, dual-centre, phase 2 clinical trial. *Lancet Oncol* 18: 464-472.

Seashore-Ludlow, B., Rees, M.G., Cheah, J.H., Cokol, M., Price, E.V., Coletti, M.E., et al. (2015) Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer Discov* 5: 1210-23.

Sethi, S., Ali, S., Philip, P.A., and Sarkar, F.H. (2013) Clinical Advances in Molecular Biomarkers for Cancer Diagnosis and Therapy. *Int J Mol Sci* 14: 14771–14784.

Shoemaker, R.H. (2006) The NCI60 human tumour cell line anticancer drug screen. *Nature Reviews Cancer* 6: 813-823.

Sideridou, M., Zakopoulou, R., Evangelou, K., Lontos, M., Kotsinas, A., Rampakakis, E., et al. (2011) Cdc6 expression represses E-cadherin transcription and activates adjacent replication origins. *J Cell Biol* 195: 1123-40.

Siolas, D., and Hannon, G.J. (2013) Patient-derived tumor xenografts: transforming clinical samples into mouse models. *Cancer Research* 73: 5315-5319.

Steckel, M., Molina-Arcas, M., Weigelt, B., Marani, M., Warne, P.H., Kuznetsov, H., et al. (2012) Determination of synthetic lethal interactions in KRAS oncogene-dependent cancer cells reveals novel therapeutic targeting strategies. *Cell Res* 22: 1227-45.

Stone M. (1974) Cross-validatory choice and assessment of statistical predictions. *J. Royal Stat. Soc.* 36: 111–147.

Stransky, N., Ghandi, M., Kryukov, G.V., Garraway, L.A., Lehár, J., Liu, M., et al. (Cancer Cell Line Encyclopedia Consortium; Genomics of Drug Sensitivity in Cancer Consortium.) (2015) Pharmacogenomic agreement between two cancer cell line data sets. *Nature* 528: 84-7.

Su, B.B., Zhou, S.W., Gan, C.B., and Zhang, X.N. (2016) MiR-330-5p regulates tyrosinase and PDIA3 expression and suppresses cell proliferation and invasion in cutaneous malignant melanoma. *J Surg Res* 203: 434-40.

Sueoka, S., Kanda, M., Sugimoto, H., Shimizu, D., Nomoto, S., Oya, H., et al. Suppression of (2015) SAMS1 Expression is Associated with the Malignant Phenotype of Hepatocellular Carcinoma. *Annals of Surgical Oncology* 22: S1453-1460.

Sun, T.M., Du, J.Z., Yao, Y.D., Mao, C.Q., Dou, S., Huang, S.Y., et al. (2011) Simultaneous

delivery of siRNA and paclitaxel via a "two-in-one" micelleplex promotes synergistic tumor suppression. *ACS Nano* 5: 1483-94.

Sun, Y., Xia, P., Zhang, H., Liu, B., and Shi, Y. (2015) P53 is required for Doxorubicin-induced apoptosis via the TGF-beta signaling pathway in osteosarcoma-derived cells. *Am J Cancer Res* 6: 114-25.

Taghanaki, S.A., Kawahara, J., Miles, B., and Hamarneh, G. (2017) Pareto-optimal multi-objective dimensionality reduction deep auto-encoder for mammography classification. *Comput Methods Programs Biomed.* 145: 85-93.

Talwar, D., Mongia, A., Sengupta, D., and Majumdar, A. (2018) AutoImpute: Autoencoder based imputation of single-cell RNA-seq data. *Sci Rep.* 8(1): 16329.

Tan, C.C., and Eswaran, C. (2011) Using autoencoders for mammogram compression. *J Med Syst.* 35: 49-58.

Tan, P.N., Kumar, V., and Srivastava, J. (2004) Selecting the right objective measure for association analysis. *Information Systems* 29: 293-313.

Tentler, J.J., Tan, A.C., Weekes, C.D., et al. (2012) Patient-derived tumour xenografts as models for oncology drug development. *Nat Rev Clin Oncol* 9: 338–50.

Tominaga, K., Shimamura, T., Kimura, N., Murayama, T., Matsubara, D., Kanauchi, H., et al. (2016) Addiction to the IGF2-ID1-IGF2 circuit for maintenance of the breast cancer stem-like cells. *Oncogene* 36: 1276-1286.

Tran, T.P., Ong, E., Hodges, A.P., et al. (2014) Prediction of kinase inhibitor response using activity profiling, in vitro screening, and elastic net regression. *BMC Syst Biol* 8:74.

Triantaphyllou, E. (2000). *Multi-criteria Decision-Making Methods - A Comparative Study* Springer US. ISBN 978-1-4757-3157-6

Trilla-Fuertes, L., Gámez-Pozo, A., Prado-Vázquez, G., Zapater-Moros, Díaz-Almirón, M., Arevalillo, J.M., et al. (2019) Biological molecular layer classification of muscle-invasive bladder cancer opens new treatment opportunities. *BMC Cancer*. 19: 636.

Turajlic, S., Sottoriva, A., Graham, T., Swanton, C. (2019) Resolving genetic heterogeneity in cancer. *Nat Rev Genet* 20: 404-416.

Turki, T., Wei, Z., Wang, J.T. (2018) A transfer learning approach via pro- crustes analysis and mean shift for cancer drug sensitivity prediction. *J Bioinforma Comput Biol* 16: 1840014.

Tyner, J.W., Yang, W.F., Bankhead, A., et al (2013) Kinase pathway dependence in primary human leukemias determined by rapid inhibitor screening. *Cancer Res* 73: 285–296



Ulivi, P., Delmonte, A., Chiadini, E., Calistri, D., Papi, M., Mariotti, M., et al. (2015) Gene mutation analysis in EGFR wild type NSCLC responsive to erlotinib: are there features to guide patient selection? *Int J Mol Sci* 16: 747-757.

Van der Auwera, G.A., Carneiro, M., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., et al. (2013) From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics* 43: 11.10.1-11.10.33

van der Maaten, L.J.P., and Hinton, G.E. (2008) "Visualizing Data Using t-SNE". *Journal of Machine Learning Research*. 9: 2579–2605

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & van Aken, M.A.G. (2014) A Gentle Introduction to Bayesian Analysis: Applications to Developmental Research. *Child Dev*. 85: 842–860.

van't Veer, L.J., and Bernards, R. (2008) Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature* 452: 564-570.

Varmus, H., Unni, A.M., and Lockwood, W.W. (2016) How Cancer Genomics Drives Cancer Biology: Does Synthetic Lethality Explain Mutually Exclusive Oncogenic Mutations? *Cold Spring Harb Symp Quant Biol* 81: 247-255.

Wang, B., Mezlini, A.M., Demir, F., et al. (2014) Similarity network fusion for aggregating data

types on a genomic scale. *Nat Methods* 11: 333–7.

Wang, C., Liu, J., Luo, F., Tan, Y., Deng, Z., and Hu, Q.N. (2014b). Pairwise input neural network for target-ligand interaction prediction. Paper presented at: Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on.

Wang, H., Zhang, X., Vidaurre, I., Cai, R., Sha, W., and Schally, A.V. (2018) Inhibition of experimental small-cell and non-small-cell lung cancers by novel antagonists of growth hormone-releasing hormone. *International Journal of Cancer*.

Wang, K., Li, M., and Hakonarson, H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38: e164.

Wang, S., Konorev, E.A., Kotamraju, S., Joseph, J., Kalivendi, S., and Kalyanaraman, B. (2004) Doxorubicin induces apoptosis in normal and tumor cells via distinctly different mechanisms. intermediacy of H<sub>2</sub>O<sub>2</sub>- and p53-dependent pathways. *J Biol Chem* 279: 25535-43.

Wang, Y., Mei, Q., Ai, Y.Q., Li, R.Q., Chang, L., Li, Y.F., et al. (2015) Identification of lung cancer oncogenes based on the mRNA expression and single nucleotide polymorphism profile data. *Neoplasma* 62: 966-73.

Weinstein, J.N. (2012) Drug discovery: Cell lines battle cancer. *Nature* 483: 544-545.

Weiss, K., Taghi, M., Khoshgoftaar, T.M., and Wang, D.D. (2016) A survey of transfer learning. *J Big Data*, 3(1): 9.

Wu, Y., Dowbenko, D., Spencer, S., Laura, R., Lee, J., Gu, Q., et al. (2000) Interaction of the tumor suppressor PTEN/MMAC with a PDZ domain of MAGI3, a novel membrane-associated guanylate kinase. *J Biol Chem* 275: 21477-85.

Xu, N., Fang, W., Mu, L., Tang, Y., Gao, L., Ren, S., et al. (2016) Overexpression of wildtype EGFR is tumorigenic and denotes a therapeutic target in non-small cell lung cancer. *Oncotarget* 7: 3884-96.

Xu, Y., Dai, Z., Chen, F., Gao, S., Pei, J., and Lai, L. (2015). Deep Learning for Drug-Induced Liver Injury. *Journal of Chemical Information and Modeling* 55, 2085-2093.

Yamada, H., Yanagisawa, K., Tokumaru, S., Taguchi, A., Nimura, Y., Osada, H., et al. (2008) Detailed characterization of a homozygously deleted region corresponding to a candidate tumor suppressor locus at 21q11-21 in human lung cancer. *Genes Chromosomes & Cancer* 47: 810-818.

Yamada, M., Sumida, Y., Fujibayashi, A., Fukaguchi, K., Sanzen, N., Nishiuchi, R., et al. (2008) The tetraspanin CD151 regulates cell morphology and intracellular signaling on laminin-511. *FEBS J* 275: 3335-51.

Yan, Y., Zhang, L., Xu, T., Zhou, J., Qin, R., Chen, C., et al. (2013) SAMS1 is highly expressed

and associated with a poor survival in glioblastoma multiforme. PloS One 8: e81905.

Yang, F., Teves, S.S., Kemp, C.J., and Henikoff, S. (2014) Doxorubicin, DNA torsion, and chromatin dynamics. *Biochim Biophys Acta* 1845: 84-9.

Yang, Y.A., Zhang, G.M., Feigenbaum, L., and Zhang, Y.E. (2006) Smad3 reduces susceptibility to hepatocarcinoma by sensitizing hepatocytes to apoptosis through downregulation of Bcl-2. *Cancer Cell* 9: 445-457.

Yeh, J.J., Hsu, N.Y., Hsu, W.H., Tsai, C.H., Lin, C.C., and Liang, J.A. (2005) Comparison of chemotherapy response with P-glycoprotein, multidrug resistance-related protein-1, and lung resistance-related protein expression in untreated small cell lung cancer. *Lung* 183: 177-183.

Zhang, H., Liu, T., Zhang, Z., Payne, S.H., Zhang, B., McDermott, J.E., et al. (2016) Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell* 166: 755-765.

Zhang, H., Wang, D., Sun, H., Hall, R.A., and Yun, C.C. (2007) MAGI-3 regulates LPA-induced activation of Erk and RhoA. *Cell Signal* 19: 261-8.

Zhang, N., Wang, H., Fang, Y., et al. (2015) Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model. *PLoS Comput Biol* 11: e1004498.

Zhao, B., Pritchard, J.R., Lauffenburger, D.A., et al. (2014) Addressing genetic tumor heterogeneity through computationally predictive combination therapy. *Cancer Discov* 4: 166–74.

Zhao, J., Xie, X., Xu, X., and Sun, S. (2017). Multi-view learning overview. *Inf. Fusion* 38(C): 43–54.

Zheng, Y., Zhou, J., Tong, Y. (2015) Gene signatures of drug resistance predict patient survival in colorectal cancer. *Pharmacogenomics J* 15: 135–43.

Zhong, W.Z., Zhou, Q., Wu, and Y.L. (2017) The resistance mechanisms and treatment strategies for EGFR-mutant advanced non-small-cell lung cancer. *Oncotarget* 8: 71358-71370.

Zhong, Y., Jiang, L., Hiai, H., Toyokuni, S., and Yamada, Y. (2007) Overexpression of a transcription factor LYL1 induces T- and B-cell lymphoma in mice. *Oncogene* 26: 6937-6947.

### **Supplementary References**

Aas, T., Børresen, A.L., Geisler, S., Smith-Sørensen, B., Johnsen, H., Varhaug, J.E., et al. (1996) Specific P53 mutations are associated with de novo resistance to doxorubicin in breast cancer patients. *Nat Med* 2: 811-4.

Abrams, S.L., Steelman, L.S., Shelton, J.G., Wong, E.W., Chappell, W.H., Bäsecke, J., et al (2010) The Raf/MEK/ERK pathway can govern drug resistance, apoptosis and sensitivity to targeted therapy. *Cell Cycle* 9: 1781-91.

Beesley, P.W., Herrera-Molina, R., Smalla, K.H., and Seidenbecher, C. (2014) The Neuroplastin adhesion molecules: key regulators of neuronal plasticity and synaptic function. *J Neurochem* 131: 268-83.

Byron, S.A., Loch, D.C., Pollock, P.M. (2012) Fibroblast growth factor receptor inhibition synergizes with Paclitaxel and Doxorubicin in endometrial cancer cells. *Int J Gynecol Cancer* 22: 1517-26.

Canela, A., Maman, Y., Jung, S., Wong, N., Callen, E., Day, A., et al. (2017) Genome Organization Drives Chromosome Fragility. *Cell* 170: 507-521.

Crespi, A., Bertoni, A., Ferrari, I., Padovano, V., Della Mina, P., Berti, E., et al. (2015) POF1B localizes to desmosomes and regulates cell adhesion in human intestinal and keratinocyte cell lines. *J Invest Dermatol* 135: 192-201.

Haeuw, J.F., Goetsch, L., Bailly, C., Corvaia, N. (2011) Tetraspanin CD151 as a target for antibody-based cancer immunotherapy. *Biochem Soc Trans* 39: 553-8.

Hussmann, M., Janke, K., Kranz, P., Neumann, F., Mersch, E., Baumann, M., et al E. (2015) Depletion of the thiol oxidoreductase ERp57 in tumor cells inhibits proliferation and increases sensitivity to ionizing radiation and chemotherapeutics. *Oncotarget* 6: 39247-61.

Kastenhuber, E.R., Lowe, S.W. (2017) Putting p53 in context. *Cell* 170: 1062-1078.

Lacombe, A., Lee, H., Zahed, L., Choucair, M., Muller, J.M., Nelson, S.F., et al. (2006) Disruption of POF1B binding to nonmuscle actin filaments is associated with premature ovarian failure. *Am J Hum Genet* 79: 113-9.

Lee, S.J., Park, J.W., Kang, B.S., Lee, D.S., Lee, H.S., Choi, S., et al. (2017) Chronophin activation is necessary in Doxorubicin-induced actin cytoskeleton alteration. *BMB Rep* 50: 335-340.

Liu, T., Wang, S., Wang, L., Wang, J., Li, Y. (2015) Targeting CD151 by lentivirus-mediated RNA interference inhibits luminal and basal-like breast cancer cell growth and invasion. *Mol Cell Biochem* 407: 111-21.

Lovitt, C.J., Shelper, T.B., Avery, V.M. (2018) Doxorubicin resistance in breast cancer cells is mediated by extracellular matrix proteins. *BMC Cancer* 18: 41.

Lu, J.H., Shi, Z.F., Xu, H. (2014) The mitochondrial cyclophilin D/p53 complexation mediates doxorubicin-induced non-apoptotic death of A549 lung cancer cells. *Mol Cell Biochem* 389: 17-24.

Mendelsohn, A.R., Hamer, J.D., Wang, Z.B., Brent, R. (2002) Cyclin D3 activates Caspase 2, connecting cell proliferation with cell death. *Proc Natl Acad Sci USA* 99: 6871-6.

Núñez-Enríquez, J.C., Bárcenas-López, D.A., Hidalgo-Miranda, A., Jiménez-Hernández, E., Bekker-Méndez, V.C., Flores-Lujano, J., et al. (2016) Gene Expression Profiling of Acute Lymphoblastic Leukemia in Children with Very Early Relapse. *Arch Med Res* 47: 644-655.

O'Connor, M.J. (2015) Targeting the DNA Damage Response in Cancer. *Mol Cell* 60: 547-60.

Padovano, V., Lucibello, I., Alari, V., Della Mina, P., Crespi, A., Ferrari, I., et al. (2011) The POF1B candidate gene for premature ovarian failure regulates epithelial polarity. *J Cell Sci* 124: 3356-68.

Porter, A.P., Papaioannou, A., Malliri, A. (2016) Deregulation of Rho GTPases in cancer. *Small GTPases* 7: 123-38.

Rangel, R., Guzman-Rojas, L., Kodama, T., Kodama, M., Newberg, J.Y., Copeland, N.G., et al. (2017) Identification of new tumor suppressor genes in triple-negative breast cancer. *Cancer Res pii: canres.0785.2017*.

Rickardson, L., Fryknäs, M., Dhar, S., Lövborg, H., Gullbo, J., Rydåker, M., et al. (2005) Identification of molecular mechanisms for cellular drug resistance by combining drug activity and gene expression profiles. *Br J Cancer* 93: 483-92.

Roidl, A., Berger, H.J., Kumar, S., Bange, J., Knyazev, P., and Ullrich, A. (2009) Resistance to



chemotherapy is associated with fibroblast growth factor receptor 4 up-regulation. Clin Cancer Res 15: 2058-66.

Sahai, E., and Marshall, C.J. (2002) RHO-GTPases and cancer. Nat Rev Cancer 2: 133-42.

Santana-Codina, N., Carretero, R., Sanz-Pamplona, R., Cabrera, T., Guney, E., Oliva, B., et al (2013) A transcriptome-proteome integrated network identifies endoplasmic reticulum thiol oxidoreductase (ERp57) as a hub that mediates bone metastasis. Mol Cell Proteomics 12: 2111-25.

Su, B.B., Zhou, S.W., Gan, C.B., and Zhang, X.N. (2016) MiR-330-5p regulates tyrosinase and PDIA3 expression and suppresses cell proliferation and invasion in cutaneous malignant melanoma. J Surg Res 203: 434-40.

Sun, Y., Xia, P., Zhang, H., Liu, B., and Shi, Y. (2015) P53 is required for Doxorubicin-induced apoptosis via the TGF-beta signaling pathway in osteosarcoma-derived cells. Am J Cancer Res 6: 114-25.

Wang, S., Konorev, E.A., Kotamraju, S., Joseph, J., Kalivendi, S., and Kalyanaraman, B. (2004) Doxorubicin induces apoptosis in normal and tumor cells via distinctly different mechanisms. intermediacy of H<sub>2</sub>O<sub>2</sub>- and p53-dependent pathways. J Biol Chem 279: 25535-43.

Wang, Y., Mei, Q., Ai, Y.Q., Li, R.Q., Chang, L., Li, Y.F., et al. (2015) Identification of lung

cancer oncogenes based on the mRNA expression and single nucleotide polymorphism profile data. *Neoplasma* 62: 966-73.

Wu, T., Dai, Y. (2017) Tumor microenvironment and therapeutic response. *Cancer Lett.* 387: 61-68.

Wu, Y., Dowbenko, D., Spencer, S., Laura, R., Lee, J., Gu, Q., et al. (2000) Interaction of the tumor suppressor PTEN/MMAC with a PDZ domain of MAGI3, a novel membrane-associated guanylate kinase. *J Biol Chem* 275: 21477-85.

Yamada, M., Sumida, Y., Fujibayashi, A., Fukaguchi, K., Sanzen, N., Nishiuchi, R., et al. (2008) The tetraspanin CD151 regulates cell morphology and intracellular signaling on laminin-511. *FEBS J* 275: 3335-51.

Yang, F., Teves, S.S., Kemp, C.J., and Henikoff, S. (2014) Doxorubicin, DNA torsion, and chromatin dynamics. *Biochim Biophys Acta* 1845: 84-9.

Zhang, H., Wang, D., Sun, H., Hall, R.A., and Yun, C.C. (2007) MAGI-3 regulates LPA-induced activation of Erk and RhoA. *Cell Signal* 19: 261-8.

## FIGURE LEGENDS

**Figure 1. The landscape of computer sciences:** its “russian-doll”-like organization and its relationship with big data. For terminology explanation and further reading see **Table 1** and accompanying references.

**Figure 2. Machine learning algorithms comprise an extensive “universe” of application models** (reproduced with permission from Dr Jason Brownlee: <https://machinelearningmastery.com/faq/single-faq/how-do-i-reference-or-cite-a-book-or-blog-post>). For a highlight on the most prominent machine learning applications and their pros and cons, see **Table 2**.

**Figure 3. Key steps for building *in silico* models for drug response.** These models comprise three steps: i) opting the input data set, ii) selecting the appropriate algorithm (see **Table 2** and **Figure 2** for a highlight on machine learning algorithms) and training it to build a prediction model, and iii) testing of the algorithm in unseen data sets. Resources of big input data can be from cell lines, animal model or clinical cohorts and type of information include o variety of “omics” or clinical data such as gene copy numbers, gene expression, gene mutations, epigenetic changes, protein expression, pharmacological responses, survival and others.

**Figure 4. Neural Network Architecture. (a)** Comparison of prediction performance of Deep Neural Networks, an advanced form of neural networks, against other learning algorithms in relation to continuously increasing amount of “big-data” [reproduced with permission from Dr

Andrew Y. Ng (<https://medium.com/syncedreview/andrew-ng-offers-ai-for-everyone-eac04877773d>; <https://medium.com/syncedreview/andrew-ng-warns-of-centralized-ai-power-47a44a462c8>]. (b) The organization of neurons in multi-layered networks. (c) The single neuron as a unit.

**Figure 5. Schematic representation of the study design and bioinformatics pipeline.** (a) Dataset: the full data set was constructed using the GDSC and CCLP databases (see also **Supplemental Figure 1**). (b) Model construction: *Association Rule Mining* (ARM) was used to generate testable hypotheses of genes associated with sensitivity or resistance to specific drugs (left panel). (c) Validation: our models were validated computationally and in a variety of *in vitro* experimental settings.

**Figure 6. Association Rule Mining (ARM) basic interest metrics.** There are three basic metrics to describe the power and significance of the rules generated by ARM. Rules are in the form of  $A \Rightarrow B$ . The feature A is considered to be the Left Hand Side (LHS) of the rule while the feature B the Right Hand Side (RHS). Support is the frequency of the rule occurrence in the total dataset. Confidence is the frequency of rule occurrence in the cases of the dataset fulfilling the left hand side of the rule. Lift is the factor by which, the co-occurrence of A and B exceeds the expected probability of A and B co-occurring, had they been independent. Details are presented in **Supplemental Materials** section.

**Figure 7. Representative output from the interactive Rshiny web application:** <https://compbio.nyumc.org/drugs/>, confirming prior-knowledge on the Shieldin-PARPi

association.

**Figure 8. Unbiased k-means cluster of top significant rules associated with the ERK-MAPK signalling pathway.** (a) Group-wise Association Rules visualization by k-means clustering  $k=50$  of the 1000 1-way rules with the largest support, for the sensitivity state of drugs targeting the ERK-MAPK signalling pathway. (b)  $IC_{50}$  heatmaps of drugs targeting the ERK-MAPK signalling pathway for melanoma versus non-melanoma cell lines and for cell lines carrying mutated versus wild-type BRAF.

**Figure 9. Unbiased k-means cluster of top significant rules associated with the PI3K signalling pathway.** (a) Group-wise Association Rules visualization by k-means clustering  $k=50$  of the 1000 1-way rules with the largest support, for the sensitivity state of drugs targeting the PI3K signalling pathway. (b) Zoom-ins of the ID1 and PTEN clusters presented in section-a. (c)  $IC_{50}$  heatmaps of drugs targeting the PI3K signalling pathway for cell lines over versus under-expressing ID1 and for cell lines carrying wild-type versus mutated PIK3CA.

**Figure 10. Association Rules related to genes associated with PARP inhibitors response.** Application of the current pipeline on information recall from 96 literature-derived and experimentally verified genes associated with response to PARP inhibitors; Monte Carlo simulation analysis for randomness evaluation.

**Figure 11. Validation of ID1 as a biomarker for responsiveness to PI3K-targeted therapies.** (a) Apriori data mining process generated rules linking ID1 suppression with PI3K

chemosensitivity. **(b)** Sustained expression of p21<sup>WAF/Cip1</sup> in Li-Fraumeni p53-deficient cells has tumor promoting ability (Galanos et al., 2016). Upon prolonged p21<sup>WAF/Cip1</sup> expression the antitumor barrier of senescence is “bypassed”, generating “escaped” clones with aggressive and chemo-resistant features along with high ID1 expression levels (Galanos et al., 2016). Morphological features and senescence detection using SenTraGor<sup>TM</sup>, a novel staining marker (Evangelou et al., 2017), in induced and escaped Li-Fraumeni-p21<sup>WAF/Cip1</sup> Tet-ON cells (Scale bar: 20  $\mu$ m). **(c)** Combined PI3K inhibition and ID1 silencing decreased drug resistance of Li-Fraumeni p21<sup>WAF/Cip1</sup> escaped cells. Drug response curves for the PI3K inhibitors CAL-101 and ZSTK474 in the escaped Li-Fraumeni- p21<sup>WAF/Cip1</sup> cells, before and after ID1 genetic silencing, and soft agar colony formation assay. Increased sensitivity is denoted by the left pointing red arrow showing leftward shift of dose response curve. ID1 siRNA targeting efficiency was verified by quantitative real time-RT-PCR and immunoblot analysis. (see details in **Supplementary Materials** section) DOX: doxocyclin, \* denotes  $p < 0.05$

**Figure 12. CDC6 overexpression as an indicator of resistance to MAPK pathway inhibitors.**

**(a)** The Apriori data mining process generated three rules linking CDC6 overexpression with resistance to MAPK (Mitogen-Activated Protein Kinase) inhibition. **(b)** Resistance to inhibitors is based on mutations that either render the MAPK pathway insensitive to treatment or reactivate alternative components of the signaling route bypassing the inhibitory block. **(c)** A CDC6-inducible normal cellular model that recapitulates all stages of cancer development (Komseli et al., 2018). **(d)** Whole genome sequencing analysis in escaped versus OFF HBEC-CDC6 Tet-OFF cells (human bronchial epithelial cells) from three independent biological replicates demonstrated acquisition of p.R55T (c.G164C) mutation in exon 3 of MAP2K3. **(e)** Immunoblot

(IB) analysis of total and phosphorylated p38 MAPK in non-induced, one day induced and escaped HBEC-CDC6 Tet-OFF cells. **(f)** Histogram depicting the significantly increased resistance ( $p < 0.05$ ) of escaped (Esc) from senescence HBEC-CDC6 cells, with high levels of CDC6, to the MEK1/2 inhibitor PD98059, relatively to the non-induced (OFF) cells with very low CDC6 levels (see IB in panel c). Non-induced (OFF) and Esc HBEC-CDC6 cells were incubated for 24h with 25 $\mu$ M PD98059. (see details in **Supplementary Materials** section) \* denotes  $p < 0.05$

**Figure 13. Overlap of Association Rules with other frameworks.** Overlap of Association Rules of the current study with GDSC, CCLE and CTRP.

**Figure 14. U-shaped curve demonstrating drug response or behaviour.**

### Supplementary Figure legends

**Supplementary Figure 1. Description of full data set and summary of main data matrix.** **(a)** Tissue of origin of the 1001 cell lines of the data-set. **(b)** Summary of the main data matrix containing tissue of origin, mutation status, gene expression, copy number variation and drug response information for the 1001 cancer cell lines (see “Data Availability” in Supportive material section). **(c)** Description of each data type used, including source, number of features and levels.

**Supplementary Figure 2. Relationships between metrics obtained through association rule mining.** **(a)** Scatter plots presenting relation between confidence and support for 10,000 1-way

rules based on top support, confidence and lift. **(b)** Scatter plots presenting relation between confidence and support for 100,000 2-way rules based on top support, confidence and lift.

**Supplementary Figure 3. Representative output from the interactive Rshiny web application: <https://compbio.nyumc.org/drugs/>, confirming prior-knowledge on the Shieldin-PARPi association.**

**Supplementary Figure 4. Validation of novel predicted gene-targets, identified by the ARM pipeline, that affect sensitivity or resistance to Doxorubicin.**

**4a. Scheme and timeline of experiments.** **(a1)** Experimental workflow of siRNA silencing and drug treatment. Timelines for **(a2)** dose response curve generation and **(a3)** soft agar colony formation following treatments with corresponding drugs and siRNAs (see details in **Supplementary Materials** section).

**4b. Experimental validation of novel predicted gene-targets, identified by the ARM pipeline, that affect sensitivity (2) or resistance (3) to Doxorubicin.** **(b1)** Schematic representation of the gene selection algorithm. From the total of 1.326.251 found rules, 989.163 gene expression associated ones were employed. **(b2)** Fold changes in IC<sub>50</sub> levels, determined from dose response curves performed with MTT-assay (**Supplementary File 3 - Doxorubicin\_IC50**), for the cell lines A549, H1299, MCF7 and Saos-2 treated with Doxorubicin in combination with silencing of *MAGI3*, *POF1B*, *PDIA3*, *CD151* and *NPTN* (genes conferring sensitivity) relative to the IC<sub>50</sub> levels of the cells when treated with the drug alone, and drug plus control siRNA (Ctl siRNA) **(2i)**. Cell viability of A549, H1299, MCF7 and Saos-2 cells treated with: 1) Ctl siRNA, 2) *POF1B*, *MAGI3*, *PDIA3*, *CD151* and *NPTN* siRNA, respectively, 3) Ctl siRNA plus Doxorubicin,



and 4) gene silencing plus Doxorubicin (2ii). **(b3)** Fold changes in IC<sub>50</sub> levels, determined from dose response curves performed with MTT-assay (**Supplementary File 3 - Doxorubicin\_IC50**), for the cell lines A549, H1299, MCF7 and Saos-2 treated with Doxorubicin in combination with silencing of *TP53*, *CTCF*, *CCND3*, *ARHGDIB* and *ZCCHC7* (genes conferring resistance) relative to the IC<sub>50</sub> levels of the cells when treated with the drug alone, and drug plus Ctl siRNA (3i). Cell viability of A549, H1299, MCF7 and Saos-2 cells treated with: 1) Ctl siRNA, 2) *TP53*, *CTCF*, *CCND3*, *ARHGDIB* *ZCCHC7* siRNA, respectively, 3) Ctl siRNA plus Doxorubicin, and 4) gene silencing plus Doxorubicin (3ii). **Note:** H1299 and Saos-2 cell lines were not treated with si-*TP53* because they are *TP53*-null.

**4c. Efficacy of genetic silencing in A549, H1299, MCF7 and Saos2 cells of genes conferring sensitivity (1) or resistance (2) to Doxorubicin treatment (see Supplementary Figure 4b). (1)**

Real time, quantitative (RT)-PCR analysis of *POF1B*, *MAGI3*, *PDIA3*, *CD151* and *NPTN* mRNA expression levels before and after RNA silencing in A549, H1299, MCF7 and Saos-2 cells, and representative immunoblot analyses in A549 cells. **(2)** Real time, quantitative (RT)-PCR analysis of *TP53*, *CTCF*, *CCND3*, *ARHGDIB* and *ZCCHC7* mRNA expression levels before and after RNA silencing in A549, H1299, MCF7 and Saos-2 cells, and representative immunoblot analyses in A549 cells. **Note:** H1299 and Saos-2 cell lines were not treated with si-*TP53* because they are *TP53*-null.

**4d. Doxorubicin (Dox) dose response curves in the A549, NCI-H1299, MCF7 and Saos-2**

**cells. (1)** Dose response curves in the A549, NCI-H1299, MCF7 and Saos-2 cell lines after treatment with Doxorubicin (Dox) alone or with control siRNAs cells to estimate the corresponding IC<sub>50</sub> values. **(2)** Representative confirmatory dose response curves after silencing each gene (*MAGI3*, *POF1B*, *PDIA3*, *CD151*, *NPTN*) that confers sensitivity in selected cell lines.

Increased sensitivity is denoted by the left pointing red arrow showing leftward shift of dose response curve. **(3)** Representative confirmatory dose response curves after silencing each gene (*TP53*, *CTCF*, *CCND3*, *ARHBD1B* and *ZCCHC7*) that confers resistance in selected cell lines. Increased resistance is depicted by the right pointing red arrow showing rightward shift of dose response curve. **Note:** H1299 and Saos-2 cell lines were not treated with si-*TP53* because they are *TP53*-null.

**4e. Soft agar colony formation assays in the A549, NCI-H1299, MCF7 and Saos-2 cell lines after treatment with Doxorubicin (Dox) alone or with siRNAs against (1) genes conferring sensitivity *MAGI3*, *POF1B*, *PDIA3*, *CD151*, *NPTN*, (see Supplementary Figure 4b) and (2) genes conferring resistance *TP53*, *CTCF*, *CCND3*, *ARHBD1B*, *ZCCHC7* (see Supplementary Figure 4b). Note:** H1299 and Saos-2 cell lines were not treated with si-*TP53* because they are *TP53*-null. \* denotes  $p < 0.05$ , ctl-siRNA: control siRNA (see details in **Supplementary Materials** section)

**Supplementary Figure 5. (a.) CDC6 overexpression is a poor prognostic factor in common human malignancies.** Log-rank (Mantel-Cox) survival analyses, with Bonferroni correction, were performed to assess the association of CDC6 overexpression with survival of patients in four common human malignancies (lung, pancreatic and prostate adenocarcinomas, along with breast carcinomas). CDC6 mRNA expression levels were obtained from mRNA microarrays. CDC6 mRNA expression levels and patients' survival status were extracted from METABRIC.

**(b.) The R55T mutation of MAP2K3.** As the R55T mutation of MAP2K3 was also observed in colon cancer we investigated its possible role in the functionality of the particular kinase. We attempted to create a theoretical model by using several available crystal structures of

homologous MAPK kinases (MAPKKs) as templates. However, in all cases R55 was unambiguously mapped on a disordered region of the kinase N-terminal lobe preceding  $\beta$  sheet 1. As a result, the mutation site and its precise topology could not be inspected within the context of a consistent homology model. Yet, there are specific indications that the positioning of R55 within the N-terminal region of MAP2K3 may be of pivotal role to the functionality of the particular kinase as a regulator of signal transduction cascades. Indeed, a number of short linear motifs have been associated in the past with regulatory properties for MAPK kinases. Such patterns have been reported to be involved in a diverse range of functions including both inactivation through the formation of autoinhibitory dimmers, like in the case of the closely related MAP2K6 or, conversely, the establishment of protein-protein interactions that can greatly increase affinity for downstream kinases, therefore facilitating more efficient phosphorylation and, consequently, ensuring higher activation levels as well as selectivity over isoforms (**Enslen et al., 2000; Chang et al., 2002; Kragelj et al., 2015; Min et al., 2009**). Those regulatory N-terminal sequence patterns include the relatively infrequent 'arginine stacks' (**Min et al., 2009**) and several categories of specificity-determining docking sites of downstream target proteins (D motifs) (**Enslen et al., 2000**). They are comprised in most of the described cases by adjacent basic residues and, as already mentioned, they have been found to drastically affect both the activity of the specific kinases as well as the activation state of their downstream targets (**Holland & Cooper, 1999**). For example, activation by different MAPKKs of specific isoforms of p38 kinase is strongly dependent upon the presence of a particular 18-residue long docking motif on the MAPKK N-terminal domain that confers the desired selectivity over the untargeted p38 isoforms (**Enslen et al., 2000**). As a result, it is reasonable to expect that the R55T mutation on MAP2K3 would possibly have a non-negligible effect on the overall functionality of the

enzyme, either with respect to its self-regulatory dynamics or regarding its activity as an effector that regulates downstream proteins such as p38. Although R55 could not be identified as a component of the docking sites of MAP2K6 and MAP2K3 $\beta$  (Enslen et al., 2000; Chang et al., 2002) or on the arginine stack motif of MAP2K6 (Min et al., 2009), the possibility that it comprises an essential part of a regulatory domain cannot be ruled out. Indeed, its spatial proximity with structural determinants on the N-terminal region that are important for kinase function such as the active site and the Glycine-rich loop could possibly justify a significant contribution of the particular residue to the stabilization and subsequent dynamics of the kinase. Whereas additional studies are needed to further clarify the structural and dynamical role of the effect the aforementioned mutation has on MAPK signaling, this finding could offer a starting point for introducing a hypothesis that the observed over-activation of p38 kinase (**Figure 12e**) can be approached as a regulatory perturbation of MAP2K3 caused by the altered dynamics of the R55T mutant that triggers aberrant activation of its downstream kinase.

**(c.) Dose response curves for the MEK1/2 inhibitor PD98059 in the HBEC-CDC6 Tet-ON cellular system.** Rightward shift in dose response curve (red arrow) in escaped relative to non-induced HBEC-CDC6 Tet-ON cells (Komseli et al., 2018), denoting resistance of these malignant counterparts to the inhibitory effect of the MEK1/2 inhibitor PD98059. \* denotes  $p < 0.05$  (see details in **Supplementary Materials** section)

**Table 1.** Terminology description and further reading.

Term	Description	
<b>Algorithm</b>	Set of instructions (performed in a stepwise manner) used to solve a class of problems or perform a computation, in the fields of mathematics and computer science	Jam
<b>Algorithm parameters</b>	The parameters set for an algorithm like k (number of clusters) or the input data	Nelder a
<b>Artificial intelligence</b>	The scientific domain aiming to give the computer systems the ability of learning, reasoning and self-correction	Broo A
<b>Batch effect removal</b>	The removal of technical variations from data that introduce systematic bias between groups of examined samples	Luc
<b>Bayesian inference</b>	A statistical method that updates the probability for a hypothesis as more data become available to the model	var
<b>Bias</b>	How different is the correct value we originally wanted to predict with our model, from the average prediction of our model	<a href="https://to">https://to</a>
<b>Big data</b>	Collection of very large information used in computational analyses to reveal patterns, trends, and associations (> 1TB information)	Kleppm The B
<b>Classification</b>	Is a supervised learning process based on an algorithm that categorizes the output into a limited set of values	Jam
<b>Clustering</b>	Unsupervised machine learning process used to group a set of objects, based on similarity (see also <b>Table 2</b> )	T
<b>Computer science</b>	Multidisciplinary field that studies computers and computational concepts	Broo A
<b>Cost function</b>	A measure of how badly a machine learning model behaves	<a href="https://to">https://to</a>
<b>Data mining</b>	Process of unveiling hidden patterns from enormous data sets using methods of statistics, database systems and machine learning	Bishop, C
<b>Feature selection</b>	The process in statistics and machine learning in which a subset of relevant features/variables is selected in order to be used in the model construction	Jam
<b>General linear models</b>	Under this term are any statistical linear models in the form of $y = ax+b$ (see also <b>Table 2</b> ), where $x$ =input, $y$ =output	Nelder a
<b>Imputation</b>	Replacing of missing data with substituted values	Jam
<b>Independent evaluation</b>	Test, after training, of a candidate model to accurately predict response on unseen settings	Bishop, C
<b>Iterative rule-based approach</b>	Rule based process that starts from all the samples in the cohort proceeding to a subset of samples and is executed until there are no features fulfilling the requirements to further divide the subset of samples into groups	
<b>Kernelized regression</b>	A non-parametric technique in statistics to estimate the conditional expectation of a random variable	Hend

<b>k-fold cross-validation (KF-CV)*</b>	A nested cross-validation technique where the dataset is split into k groups with the k-1 groups used as the training set and the remaining group as the test set	
<b>Machine learning</b>	Scientific discipline that uses algorithms and statistical tools to perform tasks without instructions but based on patterns and deductions	Bishop, C
<b>Matching</b>	The establishment of a link between separate data records that are related to the same entity	https://
<b>Metrics of performance</b>	The metrics used in order to evaluate the performance of a machine learning model (AUC, Accuracy etc)	Bishop, C
<b>Model fit</b>	The process of training a model to accurately represent the data trend	Jam
<b>Model generalisation</b>	When a trained machine learning model maintains its predictive power in blind datasets.	Dietterich
<b>Multi-task learning</b>	Concurrent solving of multiple tasks with shared use of commonalities and differences across these tasks	
<b>Multi-view learning</b>	The integration of data from multiple sources	
<b>Network-based data representations</b>	The representation of data via graphs, whose vertices represent data points (entities) and the edges represent relationships between pairs of those data points	Wang, data
<b>Normalization</b>	Is a data pre-processing technique, the goal of which is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values with the goal of integration for model training and inference.	Milligan
<b>Overfitting/ Overtraining</b>	When a model implements noise and fluctuations from the training set as real data for learning.	Dietterich
<b>Pattern recognition</b>	A procedure of recognizing patterns and regularities in data processed in machine learning.	Bish
<b>Regularization</b>	Process based on penalization that prevents the model becoming too complex and flexible, in order to avoid overfitting	Dietterich
<b>Sample stratification</b>	Sampling from a data set which can be separated into non-overlapping subgroups.	https://ar
<b>Supervised</b>	Machine learning category in which the algorithm receives as input labeled data points (see also <b>Table 2</b> )	Libbr
<b>Testing phase</b>	Part of the machine learning process where the algorithm performance after training is evaluated on a new data set not used in the training phase	Libbr
<b>Training phase</b>	Part of the machine learning process where the algorithm is provided with a large data set, processes it and builds a model	Libbr
<b>Transfer learning</b>	The term has dual different uses: i) in ensemble learning methods, it involves taking the results from one model to improve the results of another ii) inclusion of more than one features in training data, while only one of these features is used in testing data	
<b>Underfitting/ Undertraining</b>	When a model can neither learn the training data nor generalize to new data.	Dietterich
<b>Unsupervised</b>	Machine learning category in which the algorithm receives as input unlabeled data points (see also <b>Table 2</b> )	Libbr

<b>Variance</b>	Is an indication of how much our model can be generalized on new data other than the ones it was trained on	<a href="https://to">https://to</a>
<b>Weighted model</b>	Methods used in MCDA* applications for evaluating a number of alternatives in terms of a number of decision criteria	Trian Methods

**Table 2.** A highlight of machine learning algorithms used in drug response prediction.

of Algorithm	Algorithm Name	Brief Description	Pros	Cons	Reference
Supervised – Linear	Linear Regression	Is the statistical method that assumes the relationship between a single predictor value X and a quantitative response Y is linear	<ul style="list-style-type: none"> <li>- Very simple</li> <li>- Efficient solution for most simple problems</li> </ul>	<ul style="list-style-type: none"> <li>- Only models linear relationships</li> <li>- Sensitive to over-fitting when number of features &gt;&gt; number of samples</li> </ul>	James et al., (2011) Springer ISBN 978144194614-7138
	Support Vector Machines (SVMs)*	Is a classification algorithm that has an input of vectors that are non-linearly mapped to a very high dimensional feature space, and finds the optimal separating hyperplane for those data	<ul style="list-style-type: none"> <li>- Convex Optimization ensures that the solution reached is the global minimum.</li> <li>- Very fast</li> </ul>	<ul style="list-style-type: none"> <li>- Cannot model non-linear systems</li> <li>- Cannot handle many features and therefore needs extensive feature engineering as a pre-processing step</li> </ul>	Cortes and Vapnik (1995) Machine Learning, 20: 273–285
Supervised – L1-Penalisation	Ridge Regression	Is a statistical method close to least squares that uses penalisation when finding coefficient estimates. This method keeps all the initial predictors in the final model	<ul style="list-style-type: none"> <li>- It avoids overfitting and can be applied even when number of features is larger than number of data</li> <li>- It does not lose information like Lasso because it does not completely eliminate the features</li> <li>- Usually delivers better performance than the Lasso when highly correlated features are present</li> </ul>	<ul style="list-style-type: none"> <li>- Cannot be used as a feature selection tool</li> </ul>	James et al., (2011) Springer ISBN 978144194614-7138
	Lasso regression	In contrast to ridge regression, lasso yields “sparse” models that include only a subset of the initial values.	<ul style="list-style-type: none"> <li>- It avoids overfitting and can be applied even when number of features is larger than number of data</li> <li>- It can do feature selection</li> <li>- Very fast training and inference</li> </ul>	<ul style="list-style-type: none"> <li>- Unstable feature selection process. On different bootstrapped data, the selected features can vary significantly.</li> <li>- Feature selection is not easily interpretable.</li> </ul>	James et al., (2011) Springer ISBN 978144194614-7138
	Elastic Net	The elastic net is a regularized regression method that combines the penalisation used in the lasso and the ridge regression methods	<ul style="list-style-type: none"> <li>- All the advantages of Lasso and Ridge</li> </ul>	<ul style="list-style-type: none"> <li>- Complex model hyperparameter optimisation</li> </ul>	Hui and Hastie. Journal of the Royal Statistical Society B: 301–321
Supervised – Non linear	Naive Bayes	Is a probabilistic machine learning classifier based on Bayes theorem	<ul style="list-style-type: none"> <li>- Computationally efficient</li> <li>- Simple to implement</li> <li>- Works equally well with both linear and non-linear data</li> </ul>	<ul style="list-style-type: none"> <li>- Relies on the assumption that features are independent and will produce poor results if this assumption is false</li> </ul>	Maron, M.E. (1991) Journal of the ACM 44: 404–417.
	Decision Trees	Is a machine learning tool that uses a graphical representation of events/decisions composed of nodes, branches and endpoints.	<ul style="list-style-type: none"> <li>- Easily interpretable.</li> <li>- Especially good in handling categorical features</li> <li>- Computationally efficient</li> </ul>	<ul style="list-style-type: none"> <li>- Prone to overfitting</li> </ul>	Breiman, et al. (1984) Chapman Hall/CRC ISBN 9780412048418 C4841
	Neural Networks	Is a system inspired by biological neural networks. It consists of an input layer, a hidden layer and an output layer. Each layer contains nodes called neurons that are fully connected to the neurons of the next layer. Neurons transmit signals through their connections just like the biological paradigm.	<ul style="list-style-type: none"> <li>- Can capture complex non-linear relationships between features</li> <li>- No feature selection or feature engineering is required. This automatically happens in the hidden layer.</li> </ul>	<ul style="list-style-type: none"> <li>-Tendency to overfit unless techniques such as dropout are used</li> <li>-It requires large amount of data to reach maximum performance</li> <li>-Computationally expensive training</li> </ul>	McCulloch and Pitts (1943). Bulletin of Mathematical Biophysics. 5: 115–133



				-multi-dimensional feature relationships captured in the hidden layers is not interpretable	
	Deep Neural Networks	Is like the Artificial Neural Network, the only difference being that there are multiple fully connected hidden layers	- Same as Neural Networks only much more efficient due to higher number of hidden layers	-Same as Neural Networks only much more computationally expensive training	Hinton, G.E. (2006) Trends in Cognitive Sciences. 11: 42
Supervised – Tree-Ensemble	Random Forests	Is an ensemble learning method that combines a multitude of single fully grown decision trees (low bias, high variance) with randomly selected subsets of features to calculate the final result	- Top predictive performance with minimal model tuning - Provides a robust feature selection importance metric - They do not over-fit	- Computationally expensive training and inference - Low interpretability of the ensemble model	Breiman, L. (2001) Machine Learning. 32.
	Gradient Boosting Machines (GBMs)*	Is an ensemble learning method that combines a multitude of weak learners - shallow trees with high bias and low variance that are increasingly focused on hard examples in contrast to the fully grown decision trees used in random forests	- Top predictive performance equivalent or superior to Random Forests - Resistant to over-fitting	- Same as random forests plus model instability hence small changes in the training or feature set can create models of radically different performance	Friedman, J.H. (1999) The Annals of Statistics. 29: 1189-1232
Supervised - Clustering	k-means	Is a hard clustering method aiming to assign n data points to k clusters, using the mean and resulting to partitioning of the space into Voronoi cells.	- Very computationally efficient when it comes to big data - Works well with non-linear data	- k needs to arbitrarily be defined - Unstable in the sense that can create different representations based on different initializations	Nidheesh, et al. (2019) Comput Biol Med. 113: 213-221; Trilla-Fuertes (2019) BMC Cancer. 19: 636
	Hierarchical clustering	Is a method that seeks to build hierarchy clusters either through a bottom-up (Agglomerative) or a top-down (Divisive) approach.	- The tree-like structure is very informative - Results are very stable and independent of different initialisations	- Quite computationally demanding - Cannot readily identify distinct groups	Lior and Maimon (2007) Springer US, 32; Pritchard et al. (2004) Mol Biosyst 9: 1
Supervised - Dimensionality Reduction	PCA (Principal Component Analysis)*	Is a linear statistical procedure that converts a set of observations into a set of linearly uncorrelated variables called principal components	- Reduction in size of data. - It creates totally uncorrelated components	- Not computationally efficient when handling big data - Works best when original features are linearly correlated	Pearson, K. (1901) Philosophical Magazine. 2: 559-572
	t-SNE (t-distributed Stochastic Neighbor Embedding)*	Is a machine learning algorithm for non-linear dimensionality reduction and visualisation	- Works well when features are non-linearly correlated - Produces superior visualisations to PCA	- Not computationally efficient when handling big data - Underperforms unless data is strongly non-linear	van der Maaten and Hinton (2008) Journal of Machine Learning Research. 9: 257
	Deep Autoencoders	Is an unsupervised deep learning network that applies backpropagation for training with the goal to reconstruct its input	Same as deep neural networks	Same as deep neural networks	Hinton and Zemel (1994). Advances in neural information processing systems. 7: 10; Rampášek and Hájek (2019) Bioinformatics. 35: btz158
Supervised - Rule based	Association Rule Mining	Is a statistical procedure to identify association patterns in data and express them in the form of rules	- Efficient algorithm, ideal for big-data handling - Exhaustive algorithm that discovers all associations in a data-set - Generates easy to interpret rules - Can model complex multi-way relationships given a data-set of adequate size - Multi measures of significance	- If data-set is small the algorithm tends to generate false associations - Can only model AND logical associations. Cannot represent rules containing various logic handlers such as OR, NOT, XOR	Agrawal et al (1993) Proceedings of the ACM SIGMOD international conference on Management of Data. pp. 207-216

ACCEPTED MANUSCRIPT

**Table 3.** Publicly available repository panels containing big-data for building machine learning and data mining frameworks.

Features\Resource	NCI-DREAM	AstraZeneca-Sanger DREAM	NCI-60	GDSC	CCL
<b>Sample type</b>	53 breast cancer cell lines	85 cancer cell lines	59 cell lines from 9 tissue types	1124 cell lines from 29 tissue types	>1000 cell lines from 29 tissue types
<b>Number of compounds:</b>	28 compounds	910 pairwise combinations of 118 drugs	>1,500	265	24
<b>Main omics data sets</b>	Mut, CNV, Meth, GE, PR	Mut, CNV, Meth, GE	Mut, CNV, GE, Meth, PR	Mut, CNV, Meth, GE	Mut, CNV, Meth, GE, PR
<b>Number of cancers</b>	1	6	9	55	36
<b>Reference</b>	Costello et al. (2014) Nat Biotechnol 32: 1202.	Menden et al. (2019) Nat Commun 10: 2674	Shoemaker RH. (2006) Nat Rev Cancer 6: 813.	Garnett et al. (2012) Nature 483: 570.	Barretina et al. (2012) Nature 483: 600.
<b>Website</b>	<a href="https://www.synapse.org/#!Synapse:syn2785778/wiki/70252">https://www.synapse.org/#!Synapse:syn2785778/wiki/70252</a>	<a href="https://www.synapse.org/#!Synapse:syn2785778/wiki/70252">https://www.synapse.org/#!Synapse:syn2785778/wiki/70252</a>	<a href="https://discover.nci.nih.gov/cell_miner/">discover.nci.nih.gov/cell_miner/</a>	<a href="http://www.cancerrxgene.org/">http://www.cancerrxgene.org/</a>	<a href="http://www.broadinstitute.org/ccle">http://www.broadinstitute.org/ccle</a>

Mut: gene Mutation; CNV: gene Copy Number Variation; GE: Gene Expression; Meth: DNA Methylation, PR: Protein Expression, Hist: Histopathological images

**Table 4.** Performance metrics of machine learning frameworks.

Model types	Performance measure	Type of measure
<b>Regression models</b>	$R^2$	R-squared is a statistical measure that represents the proportion of the variance for a dependent variable that is explained by an independent variable or variables in the regression model under evaluation.
	Adjusted $R^2$	Similar to $R^2$ but with a penalty for increasing model complexity
	Root Mean Square Error (RMSE)*	The Root Mean Squared Error measures the square root of the average of the squared difference between the predictions and the ground truth.
	Mean Absolute Error	The Mean Absolute Error measures the average of the absolute difference between each ground truth and the predictions.
	F-Test	The F-Test compares the model to be evaluated against a model with no variables. The null hypothesis is that the model with no variables performs just as good as the model with the variables.
<b>Classification models</b>	Log Loss (Logarithmic Loss or Cross Entropy Loss)	Penalizes classifiers during prediction. It is maximal for false prediction classification.
	True Positive (TP)*	Equivalent with hit
	True Negative (TN)*	Equivalent with correct rejection
	False Positive (FP)*	Equivalent with false alarm (Type I error)
	False Negative (FN)*	Equivalent with miss (Type II error)
	Sensitivity, recall, hit rate, or true positive rate (TPR)*	True Positives over all Positives
	Specificity, selectivity or true negative rate (TNR)*	True Negatives over all Negatives
	Precision or positive predictive value (PPV)*	True Positives over True Positives plus False Positives
	Negative predictive value (NPV)*	True Negatives over True Negatives plus False Negatives
	Miss rate or false negative rate (FNR)*	False Negatives over all Positives
	Fall-out or false positive rate (FPR)*	False Positives over all Negatives
	False discovery rate (FDR)	False positives over False Positives plus True Positives
	False omission rate (FOR)	False Negatives over False Negatives plus True Negatives
	Accuracy (ACC)	True Positives plus True Negatives over all Positives plus all Negatives

	F1 Score	The harmonic mean of precision and sensitivity
	Youden's Index	A single statistic that captures the performance of a dichotomous diagnostic test
	Area under the ROC curve (AUC)	The ROC curve is plotted with TPR against the FPR where TPR is on the y-axis and FPR is on the x-axis. AUC is the area under this curve. AUC 0.5 indicates a random model whose performance is equivalent to chance. AUC 1 indicates the perfect predictive model

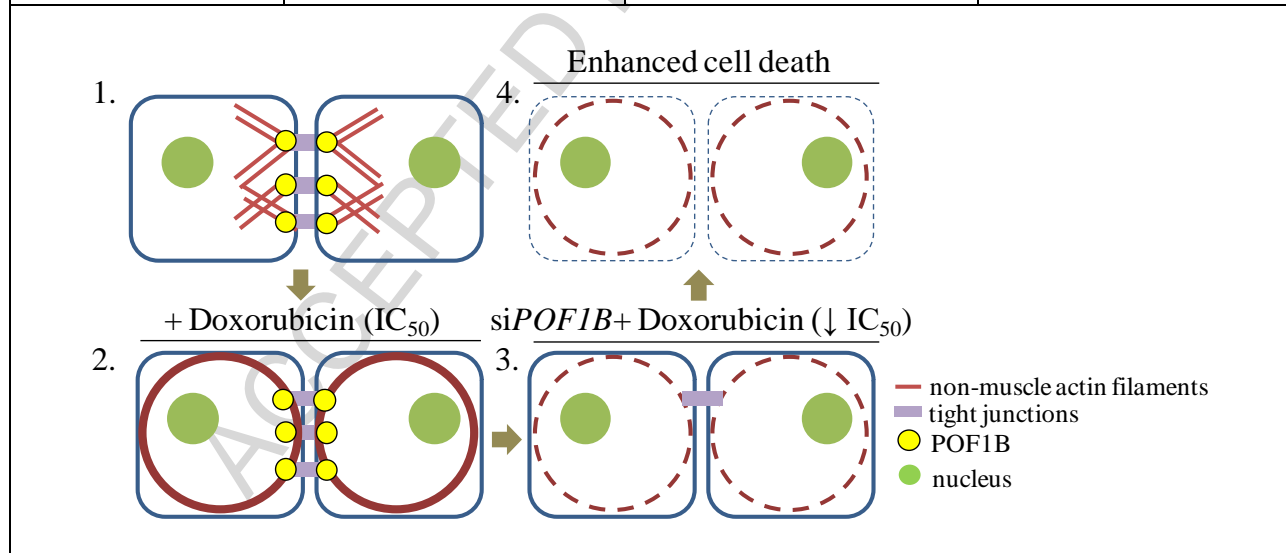
**Table 5.** Highlights of machine learning applications in oncology in chronological order

Reference	Model applied	Training set	Testing set	Outcomes
Gillet et al. (2011) Proc Natl Acad Sci USA 108: 18708–13.	BRB-ArrayTools for classification of tumor types and Hierarchical clustering analysis	NCI-60 cancer cell line panel	Primary tumors of different origin	Tendency of cell lines to cluster by anatomical origin, rather than each other, rather than by origin
Daemen et al. (2013) Genome Biol 14: R110.	1) Weighted least squares support vector machine (LS-SVM) and 2) Random Forests (RF)	Breast cancer cell lines	TCGA breast tumors for which expression (Exp), copy number (CNV) and methylation (Meth) measurements were available	AUC based sensitivity
Niepel et al. (2013) Sci Signal 6: ra84.	Partial least-squares regression to simulate signaling networks activation profile	NCI-ICBP43 breast cancer cell line collection	Breast cancer cell lines	Prediction of drug response
Byers et al. (2013) Clin Cancer Res 191: 279–90.	Hierarchical clustering and principal component Analysis (PCA)	Non–small cell lung carcinoma (NSCLC) cell lines	i) non–small cell lung carcinoma (NSCLC) cell lines ii) patients treated in the Biomarker-Integrated Approaches of Targeted Therapy for Lung Cancer Elimination (BATTLE) study.	EMT signature that predicts resistance to EGFR tyrosine kinase inhibitors
Geeleher et al. (2014) Genome Biol 15: R47.	Ridge Regression	Cancer Genome Project (CGP) cell lines	i) Docetaxel treated breast cancer patients ii) Paclitaxel treated breast cancer patients iii) Bortezomib treated myeloma iv) Erlotinib treated NSCLC	Sensitivity versus drug response prediction
Guinney et al. (2014) Clin Cancer Res 20: 265–72.	Penalized ElasticNet regression	Fresh-frozen colorectal cancer tissues analyzed for K-ras (codons 12 and 13) mutations	i) Cetuximab response: mouse xenografts and patients ii) MAP–ERK kinase (MEK) inhibition: cell lines and mouse xenografts	Prediction response to EGFR inhibitors and MEK inhibitors on RAS phenotype
Tran et al. (2014) BMC Syst Biol 8: 74.	ElasticNet regression combined with logarithmic transformation of the data	Kinase inhibitor treated cell lines	Experimental validation	Identification of specific drug response linked to drug response in cell lines
Liang et al. (2014) Int J Mol Sci 15: 11220–33.	A linear model was applied for continuous covariates along with ANOVA test for categorical covariates	Neuroblastoma cell lines and patients	Neuroblastoma patients	REST-driven transcriptional signature associated with neuroblastoma drug response
Costello et al. (2014) Nat Biotechnol 32: 1202-12.	Wining model: Bayesian efficient multiple kernel learning (BEMKL) method	Breast cancer cell lines	Cell lines	Community effort to build a state-of-the-art in drug response prediction from ‘omics’ data
Falgreen et al. (2015) BMC Cancer 15: 235.	Penalized ElasticNet regression combined with Lasso and Ridge Regression	Combined human B-cell cancer cell lines (HBCCL) with published CGP gene expression datasets	Diffuse large B-cell lymphoma (DLBCL) patients treated with CHO: cyclophosphamide (C), doxorubicin (H), and vincristine (O)	Generate resistance signatures (REGS) for sensitivity or resistance
Chen et al. (2015) Cancer Res 75: 2987–98.	PSFinder: an iterative rule–based unsupervised approach	TCGA derived high-grade serous ovarian cancer (HGS-OvCa) with platinum–taxane therapy	Separate TCGA derived high-grade serous ovarian cancer (HGS-OvCa) with platinum–taxane therapy	Classification into positive survival
Fey et al. (2015) Sci Signal 8: ra130.	Rule based modeling employing ordinary differential equations (ODEs) to simulate reactions and states of the JNK pathway	i) Neuroblastoma cells lines ii) Neuroblastoma patients	i) Neuroblastoma cells lines ii) Neuroblastoma patients iii) Zebrafish neuroblastoma model	Survival prediction and activation status of JNK pathway

Pereira et al. (2015) PLoS One 10: e0145754.	Log-binomial models combined with logistic regression models	Patients with gynecologic malignancies	Patients with gynecologic malignancies	Circulating tumor DNA as a post-treatment biomarker
Zheng et al. (2015) Pharmacogenomics J 15: 135–43.	BRB-arrayTools to perform regression analysis	Colorectal cancer cell lines with available gene expression profiles	Colorectal cancer clinical cohorts	Contribution of drug-related genes to patient response
Menden et al. (2019) Nat Commun 10: 2674	Ensemble models	Cancer cell lines	Cancer cell lines and PDX models	Community effort for computational stratification predicting synergistic drug response and biomarkers
Chiu et al. (2019) BMC Med Genomics 31: 12:18.	DLNN	Cancer cell lines from CCLE & GDSC, clinical samples from TCGA	clinical samples from TCGA (33 cancer types)	Drug response prediction

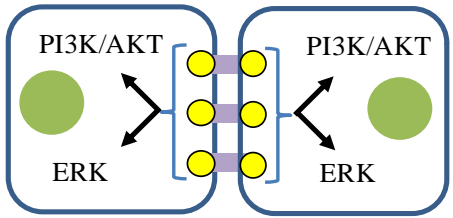
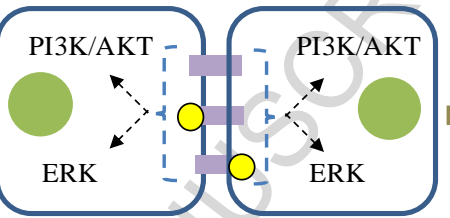
**Table 6.** Potential mechanism of action following genes silencing that confers **sensitivity** to doxorubicin treatment.

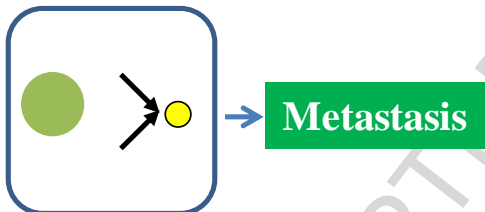
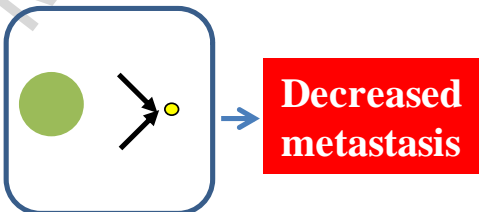
Gene	Function	Mechanism affecting sensitivity	Reference
<b><i>POF1B</i></b> (Premature Ovarian Failure Protein 1B)	Plays a key role in the organization of epithelial monolayers by regulating the actin cytoskeleton.	POF1B loss: 1. Disrupts binding of non-muscle actin filaments. 2. Abolishes tight junction localization. Thus, potentially enhances Doxorubicin mediated cytoskeleton re-organization related to cell shrinkage, detachment and apoptosis. Consequently cells develop increased sensitivity to Doxorubicin requiring lower IC <sub>50</sub> values of the drug.	Padovano V et al, J Cell Sci 2011 Lacombe A et al, AJHG 2006 Crespi A et al, J Invest Dermatol 2014 Lee SJ et al, BMB Rep 2017

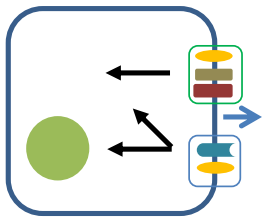
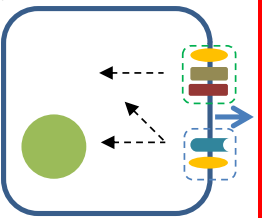


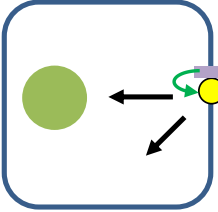
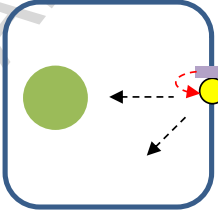
Gene	Function	Mechanism affecting sensitivity	Reference



<p><b>MAGI3</b> (Membrane-Associated Guanylate Kinase)</p>	<p>Acts as a scaffolding protein at cell-cell junctions, thereby regulating various cellular and signaling processes. Modulate the activity of ERK and AKT1 pathways.</p>	<p>Loss of MAGI3 expression disrupts activation of the PI3K/AKT and/or ERK pathways assisting Doxorubicin treatment effect (lower IC<sub>50</sub> value for Doxorubicin treatment).</p>	<p>Zhang H et al, Cell Signal 2007 Abrams SL et al, Cell Cycle 2010 Wang Y et al, Neoplasma 2015 Wu Y et al, J Biol Chem. 2000</p>
<div> <div> <p>Doxorubicin (IC<sub>50</sub>)</p>  </div> <div> <p>siMAGI3 + Doxorubicin (↓ IC<sub>50</sub>)</p>  </div> <div> <p>Decreased cell survival</p> </div> </div>			

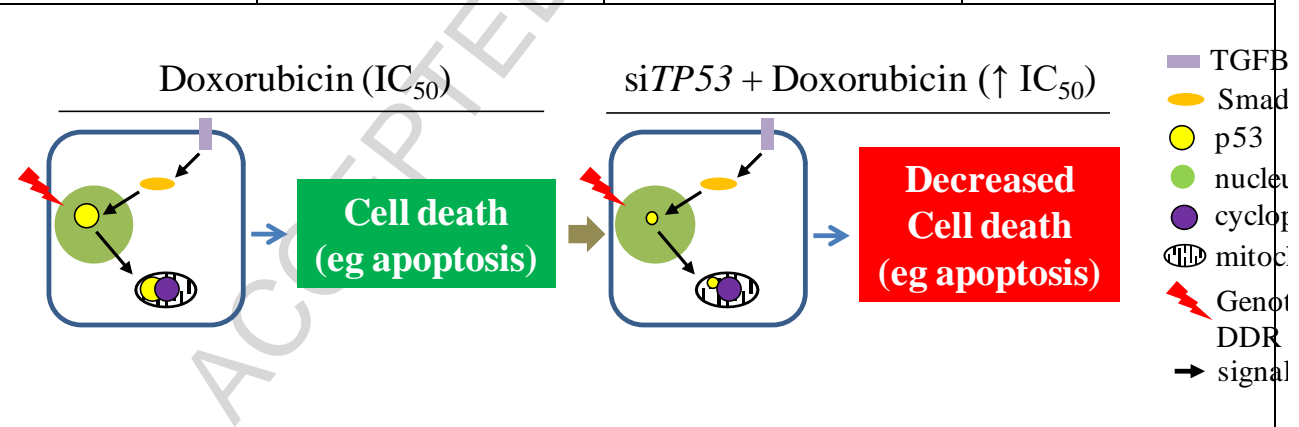
Gene	Function	Mechanism affecting sensitivity	Reference
<b><i>PDIA3</i></b> (ERp57/PDIA3: Protein disulfide isomerase family)	A phosphatidylinositol 4,5-bisphosphate phosphodiesterase type I (phospholipase C- $\alpha$ ). Catalyzes the rearrangement of -S-S- bonds in proteins. Acts in concert with calreticulin and calnexin in the folding of glycoproteins destined to the plasma membrane or to be secreted.	Functions as a hub integrating signals that mediate metastasis. Its silencing inhibits cell proliferation and increases sensitivity to ionizing radiation and chemotherapeutics. Therefore, cells develop increased sensitivity to Doxorubicin requiring lower IC <sub>50</sub> values of the drug.	Santana-Codina N et al, Mol Cell Proteom 2013 Husmann M et al, Oncotarget 2015 Su BB et al, J Surg Res 2016
<div style="display: flex; justify-content: space-around; align-items: flex-start;"> <div style="text-align: center;"> <p>Doxorubicin (IC<sub>50</sub>)</p>  </div> <div style="text-align: center;"> <p>si<i>PDIA3</i> + Doxorubicin (↓ IC<sub>50</sub>)</p>  </div> <div style="margin-left: 20px;"> <p>● PDIA3 ● nucleus → signaling</p> </div> </div>			

Gene	Function	Mechanism affecting sensitivity	Reference
<b><i>CD151</i></b> (Tetraspanin-24)	CD151 is a cell surface glycoprotein that associates strongly with the laminin-binding integrins ( $\alpha 3\beta 1$ , $\alpha 6\beta 1$ and $\alpha 6\beta 4$ ), growth factors and matrix metalloproteinases. It is involved in epithelial cell–cell adhesion.	Inhibition of CD151 affects integrin-mediated cell adhesion and signaling, resulting in sensitivity to Doxorubicin treatment (lower $IC_{50}$ value for Doxorubicin treatment). Targeting CD151 inhibits metastasis by blocking cell motility.	Yamada M et al, FEBS J 2008 Haeuw J-F et al, Biochem Soc Trans 2011 Lovitt CJ et al, BMC Cancer 2018 Liu T et al, Mol Cell Biochem 2015
<div style="display: flex; justify-content: space-around; align-items: flex-start;"> <div style="text-align: center;"> <p>Doxorubicin (<math>IC_{50}</math>)</p>  <p><b>Tumor growth Migration Invasion Metastasis</b></p> </div> <div style="text-align: center;"> <p>si<i>CD151</i> + Doxorubicin (<math>\downarrow IC_{50}</math>)</p>  <p><b>Decreased Tumor growth Migration Invasion Metastasis</b></p> </div> </div>			

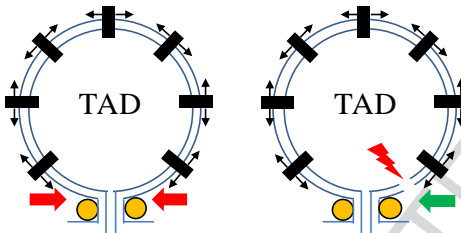
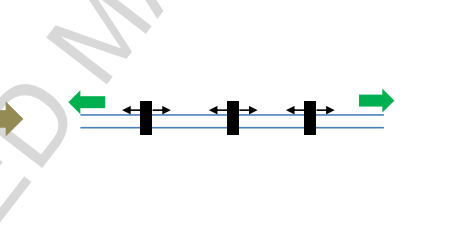
Gene	Function	Mechanism affecting sensitivity	Reference
<i>NPTN</i> (Neuroplastin)	Probable homophilic and heterophilic cell adhesion molecule. In cancer context it activates the FGFR signaling pathway, promoting neo-angiogenesis and metastasis.	FGFR inhibition synergizes with Doxorubicin treatment leading to increased sensitivity (lower $IC_{50}$ value for Doxorubicin treatment).	Beesley PW et al, J Neurochem 2014 Roidl A et al, Clin Cancer Res 2009 Byron SA et al, Int J Gynecol Cancer 2012
<div style="display: flex; justify-content: space-around; align-items: center;"> <div style="text-align: center;"> <p>Doxorubicin (<math>IC_{50}</math>)</p>  <p><b>Neo-angiogenesis Metastasis</b></p> </div> <div style="text-align: center;"> <p>si<i>NPTN</i> + Doxorubicin (<math>\downarrow IC_{50}</math>)</p>  <p><b>Decreased Neo-angiogenesis Metastasis</b></p> </div> </div>			

**Table 7.** Potential mechanism of action following genes silencing that confers **resistance** to doxorubicin treatment.

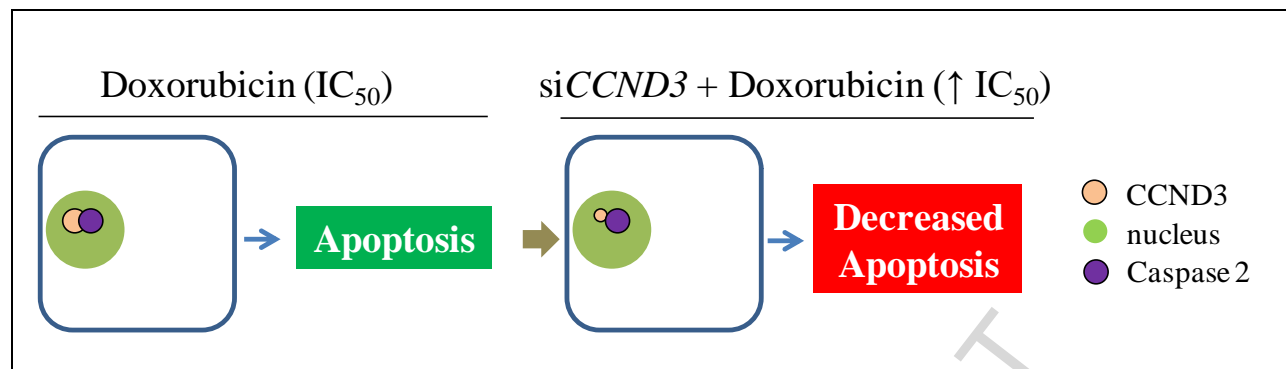
Gene	Function	Mechanism affecting resistance	Reference
<b><i>TP53</i></b> (Tumor Protein p53)	A key tumor suppressor that acts in many tumor types, inducing growth arrest, senescence or apoptosis depending on the physiological circumstances and cell type. Involved in cell cycle regulation as a trans-activator that acts to negatively regulate cell division by controlling genes required for this process.	Loss of p53 or mutation augments resistance to Doxorubicin (Dox) mediated apoptotic and non-apoptotic death. Several p53-dependent cell death inducing routes upon Dox treatment include: i) the DNA damage response (DDR) pathway, ii) the mitochondrial cyclophilin D/p53 complex, iii) p53 assisted TGF- $\beta$ /Smad3 apoptosis induction. Consequently cells develop increased resistance to Dox requiring higher IC <sub>50</sub> values of the drug.	Lu J-H et al, Mol Cell Biochem 2014 Aas T et al, Nat Med 1996 Sun Y et al, Am J Cancer Res 2015 Wang S et al, J Biol Chem 2004 O'Connor MJ, Mol Cell 2015 Negrini S et al, Nat Rev Mol Cell Biol. 2010 Kastenhuber ER, Lowe SW, Cell 2017



Gene	Function	Mechanism affecting resistance	Reference
	DNA binding protein	Evidence indicates that doxorubicin forms a complex with the DNA by intercalation of its planar rings between	

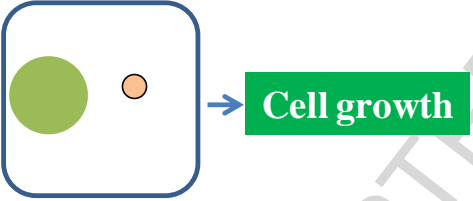
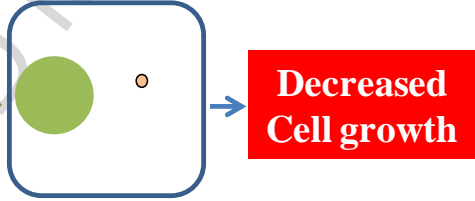
<p><b>CTCF</b> (11-zinc finger protein or CCCTC-binding factor)</p>	<p>responsible for insulator function, nuclear architecture and transcriptional control, which probably acts by recruiting epigenetic chromatin modifiers.</p>	<p>nucleotide base pairs. This intercalation generates bidirectional torsional stress on the DNA helix, which along with the Topoisomerase 2 inhibitory effect of Doxorubicin, leads eventually to DNA double strand breaks. The stress is possibly relieved upon removal of CTCF stable boundaries, thus requiring higher Doxorubicin (IC<sub>50</sub> values) to exert a similar stress induced DNA damage and cell death.</p>	<p>Yang F et al, Biochim Biophys Acta 2014 O'Connor MJ, Mol Cell 2015 Canella A et al, Cell 2017</p>
<div style="display: flex; justify-content: space-around; align-items: flex-start;"> <div style="text-align: center;"> <p>Doxorubicin (IC<sub>50</sub>)</p>  </div> <div style="text-align: center;"> <p>siCTCF + Doxorubicin (↑ IC<sub>50</sub>)</p>  </div> <div style="margin-left: auto; text-align: right;"> <p> <span style="color: yellow;">●</span> CCTCF  <span style="color: black;">■</span> Doxorubicin  <span style="color: black;">→</span> torsion stress  <span style="color: red;">⚡</span> DNA double strand breaks  <span style="color: red;">→</span> torsion stress accumulation  <span style="color: green;">→</span> torsion stress dissipation  TAD: Topological Association Domain </p> </div> </div>			

Gene	Function	Mechanism affecting resistance	Reference
<p><b>CCND3</b> (cyclin D3)</p>	<p>Member of the highly conserved cyclin D family, regulating cell cycle progression. It also activates Caspase 2, triggering apoptosis.</p>	<p>CCND3 silencing results in loss of sensitizing cells to apoptosis through inability to activate Caspase 2. In turn, higher Doxorubicin IC<sub>50</sub> values are required to bypass the acquired resistance to this drug.</p>	<p>Mendelsohn AR et al, PNAS 2002</p>



Gene	Function	Mechanism affecting resistance	Reference
<b><i>ARHGDIB</i></b> (Rho GDP dissociation inhibitor beta)	Regulates the GDP/GTP exchange reaction of the Rho proteins by inhibiting the dissociation of GDP from them, and the subsequent binding of GTP.	Aberrantly activated Rho proteins promote many “hallmarks” of cancer. Silencing of <i>ARHGDIB</i> facilitates activation of Rho proteins that mediate increased resistance to Doxorubicin treatment (higher IC <sub>50</sub> values).	Rickardson L et al, Br J Cancer 2005 Sahai E, Marshall CJ, Nat Rev Cancer 2002 Porter AP et al, Small GTPases 2016
<p>The diagram illustrates the mechanism of ARHGDIB in tumor growth control and resistance to Doxorubicin. It shows two scenarios:</p> <ul style="list-style-type: none"> <li><b>Doxorubicin (IC<sub>50</sub>)</b>: In this scenario, ARHGDIB (blue oval) inhibits the Rho protein (grey oval) from exchanging GDP (red dot) for GTP (green dot), keeping Rho inactive. This leads to <b>Control over Tumor growth</b> (red box).</li> <li><b>siARHGDIB + Doxorubicin (↑ IC<sub>50</sub>)</b>: In this scenario, silencing ARHGDIB allows Rho to become active by binding GTP. This leads to <b>Enhanced Tumor growth</b> (green box).</li> </ul> <p>A legend identifies the components: ARHGDIB (blue oval), GTP (green dot), GDP (red dot), and Rho (grey oval).</p>			



Gene	Function	Mechanism affecting resistance	Reference
<b>ZCCHC7</b> (zinc finger CCHC-type containing 7)	Possibly involved in deadenylation-dependent mRNA decay.	ZCCHC7 down-regulation in Acute lymphoblastic leukemia (ALL) is associated with relapse and poor survival. Its silencing in breast cancer is associated with increased cell proliferation. Therefore higher IC <sub>50</sub> Doxorubicin values are required to arrest tumor cell growth.	Nunez-Enriquez JC et al, Arch Med Res 2016 Rangel R et al, Cancer Res 2017
<div style="display: flex; justify-content: space-around; align-items: flex-start;"> <div style="text-align: center;"> <p>Doxorubicin (IC<sub>50</sub>)</p>  </div> <div style="text-align: center;"> <p>siZCCHC7 + Doxorubicin (↑ IC<sub>50</sub>)</p>  </div> <div style="margin-left: 20px;"> <p>○ ZCCHC7 ● nucleus</p> </div> </div>			

Supplementary File

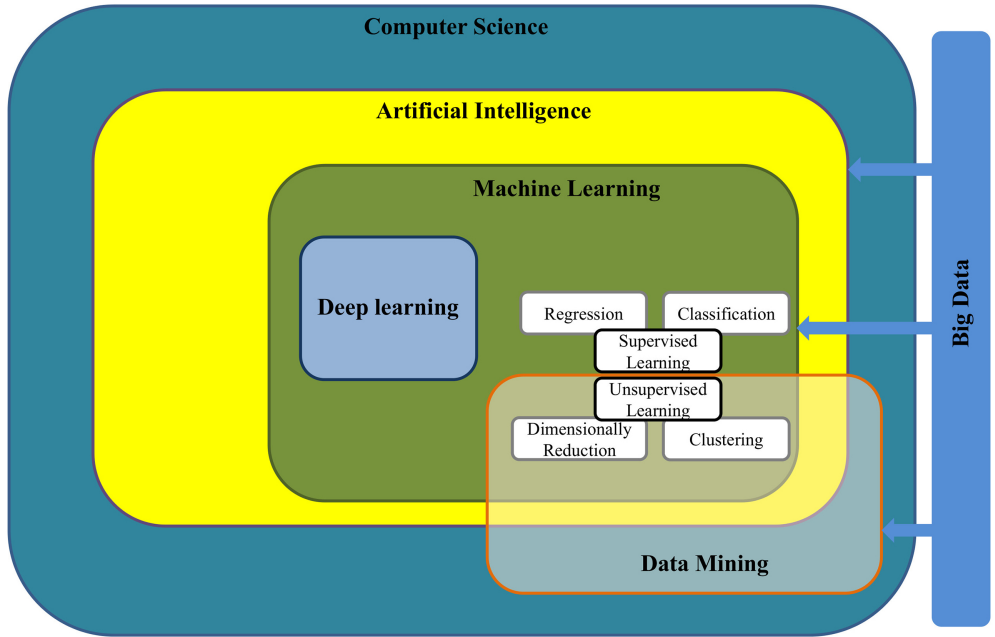


Figure 1

# The “universe” of Machine Learning Algorithms

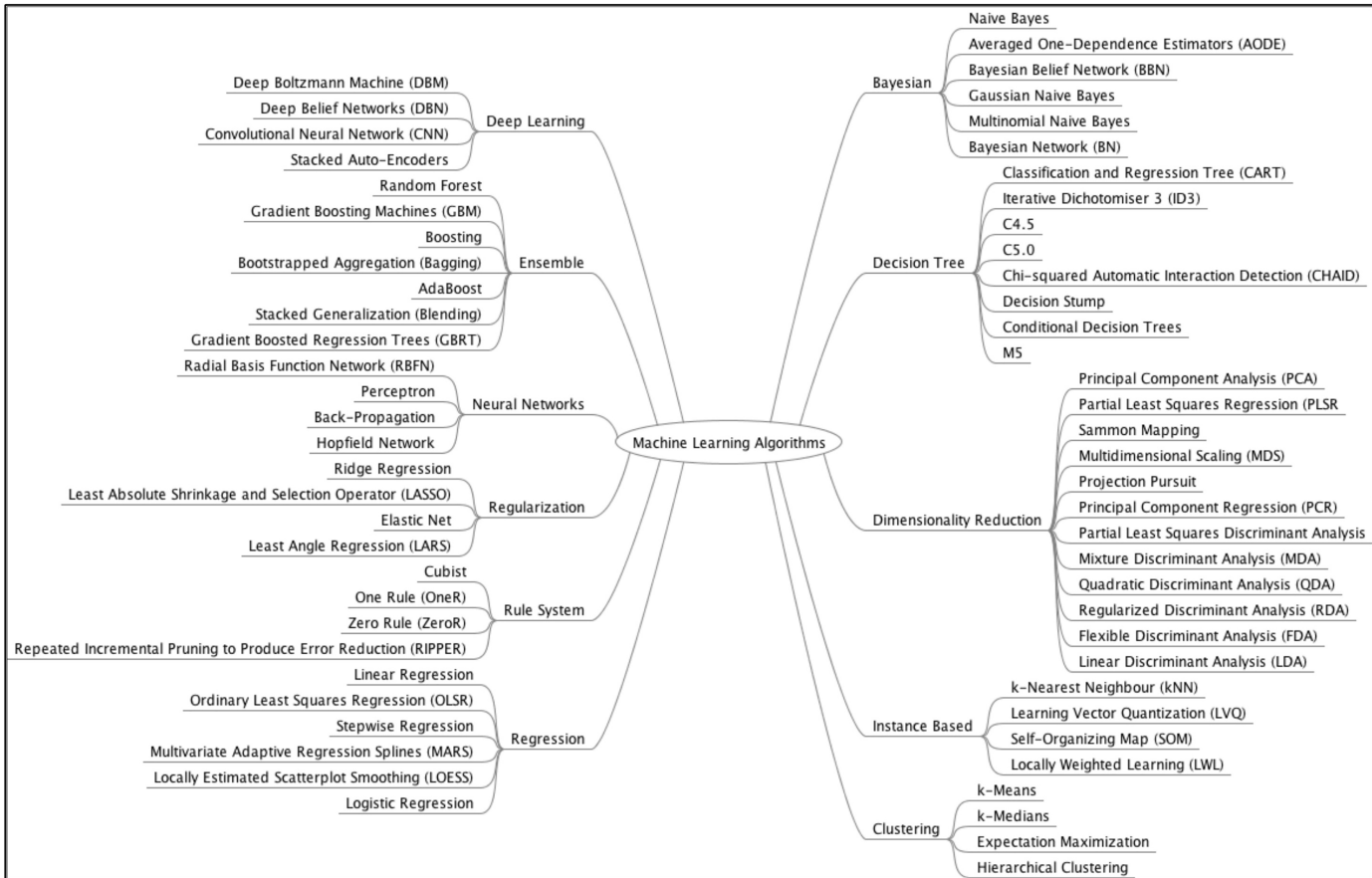


Figure 2

## Public data resources/clinical cohorts

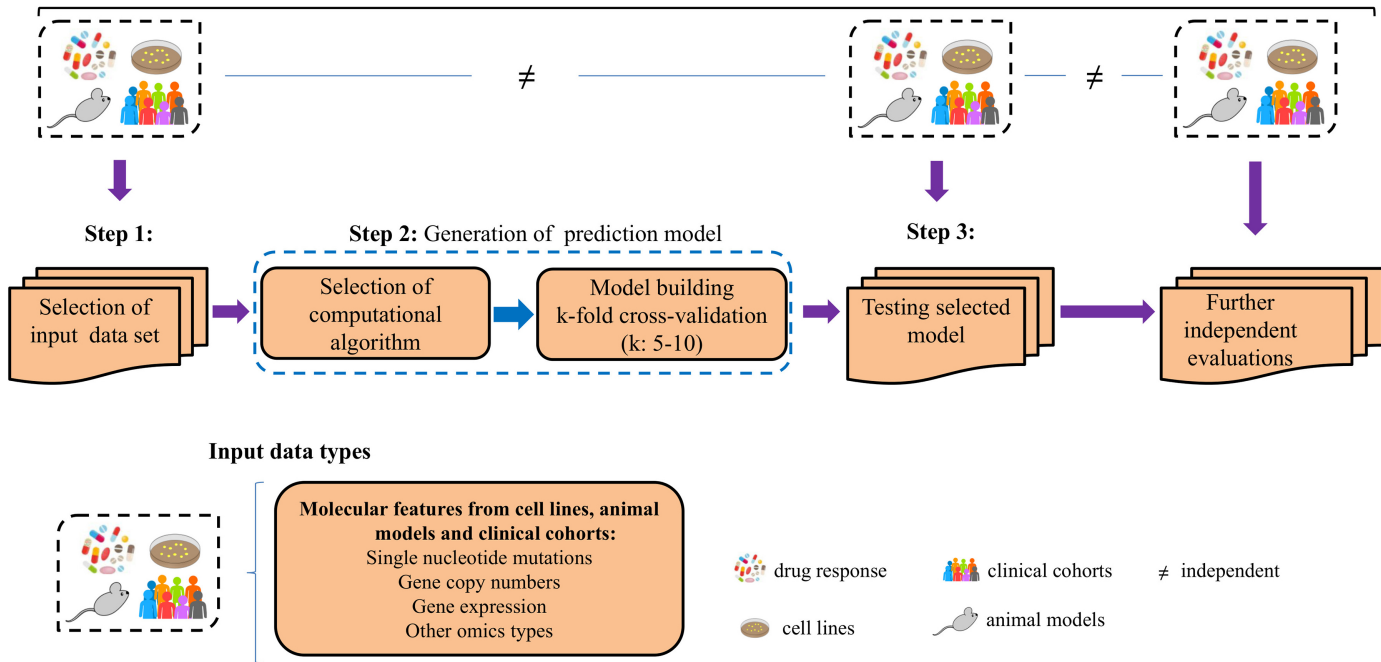
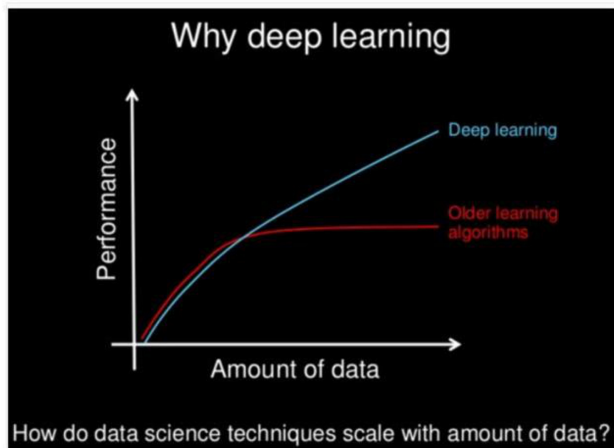
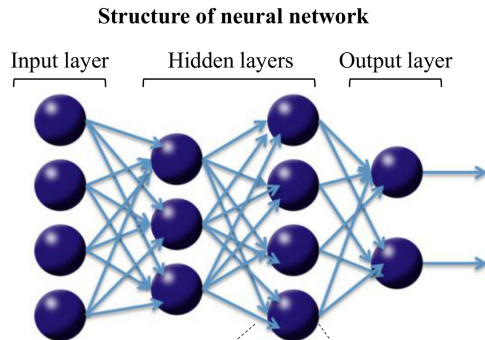


Figure 3

a.



b.



c.

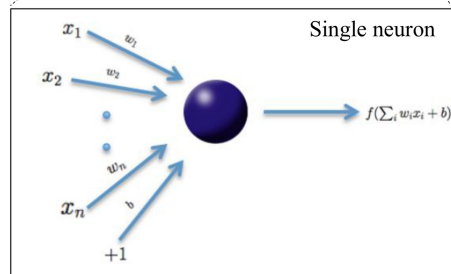
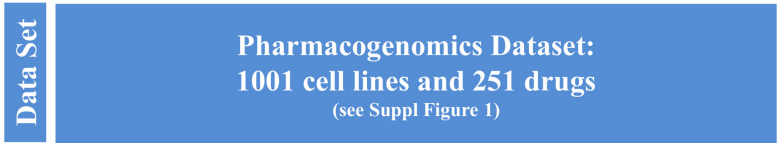


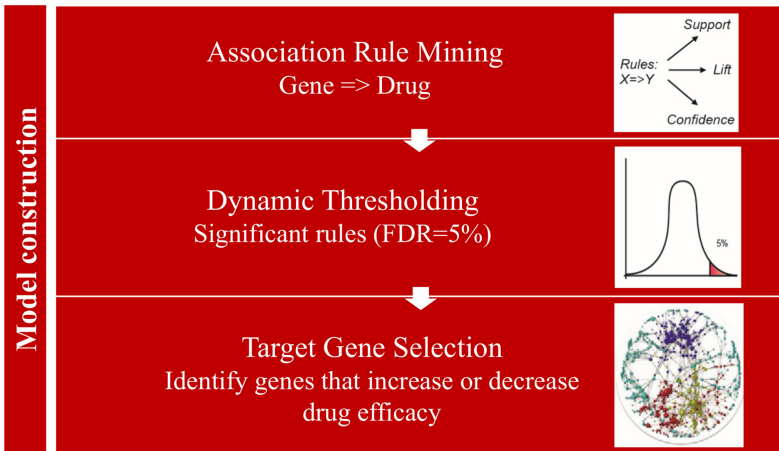
Figure 4

# ARM SCREENING PROCESS

a.



b.



c.



Figure 5

Basic metrics that describe the power and significance of the rules generated by **Association Rule Mining**

**Support:** is the frequency of the rule occurrence in the total dataset.

$$\text{Support} = \frac{\text{Number of transactions with both } A \cap B}{\text{Total number of transactions}} = P(A \cap B)$$

**Confidence:** is the frequency of rule occurrence in the cases of the dataset fulfilling the **LHS** of the rule.

$$\text{Confidence} = \frac{\text{Number of transactions with both } A \cap B}{\text{Total number of transactions with } A} = \frac{P(A \cap B)}{P(A)}$$

**Lift:** is a measure of significance.

$$\text{Lift} = \frac{P(A \cap B)}{P(A) \times P(B)}$$

A: left hand side (**LHS**) feature of rule in the form of  $A \Rightarrow B$

B: right hand side (**RHS**) feature of rule in the form of  $A \Rightarrow B$

Figure 6

### Prior Knowledge :

*Authors:* Mirman et al.

**Title:** 53BP1–RIF1–shieldin counteracts DSB resection through CST- and Pol $\alpha$ -dependent fill-in.

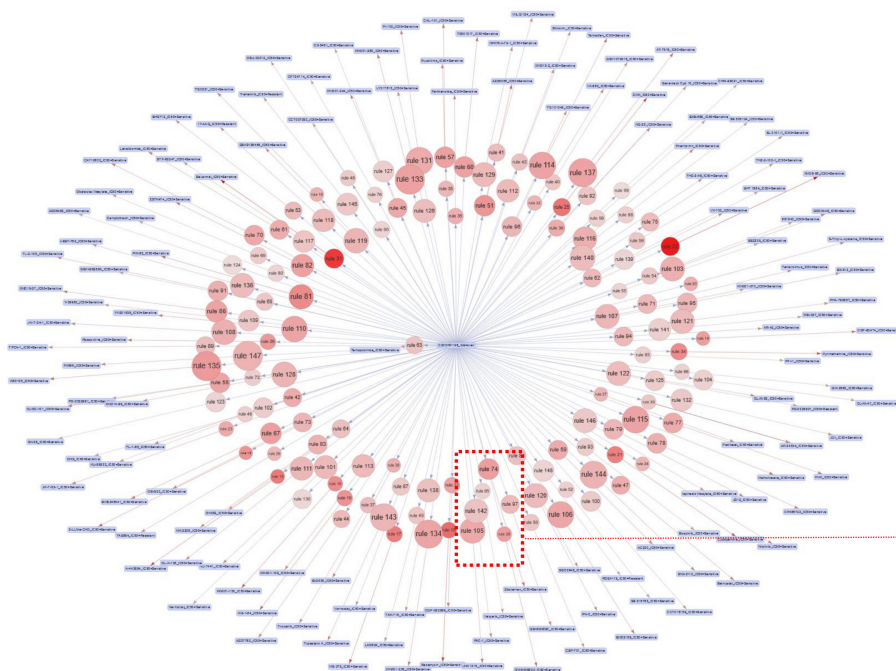
*Journal:* Nature 2018, 560(7716):112-116.

*Authors:* Noordermeer et al.

**Title:** The shieldin complex mediates 53BP1-dependent DNA repair.

*Journal:* Nature 2018, 560(7716):117-121.

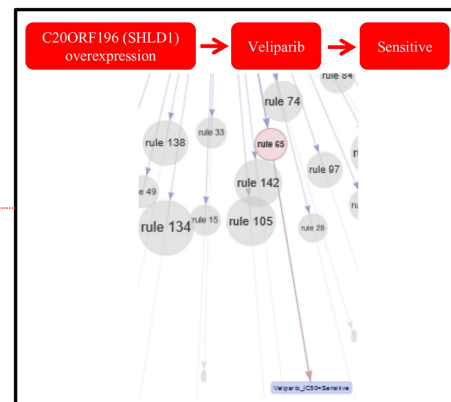
## Interaction network



**Inner** : C20ORF196 (SHLD1) overexpression

**Outer** : Veliparib IC50 Sensitivity

**rule 65** : Association (rule)



### Interaction table

Tissue	Gene	Data type	Value	Drug	Measurement	Effect	Support	Confidence	Lift	P-value
NaN	C20ORF196 (SHLD1)	GE	over	Veliparib	IC50	Sensitive	0.007992008	0.2666667	2.152688	0.0007872295

Figure 7



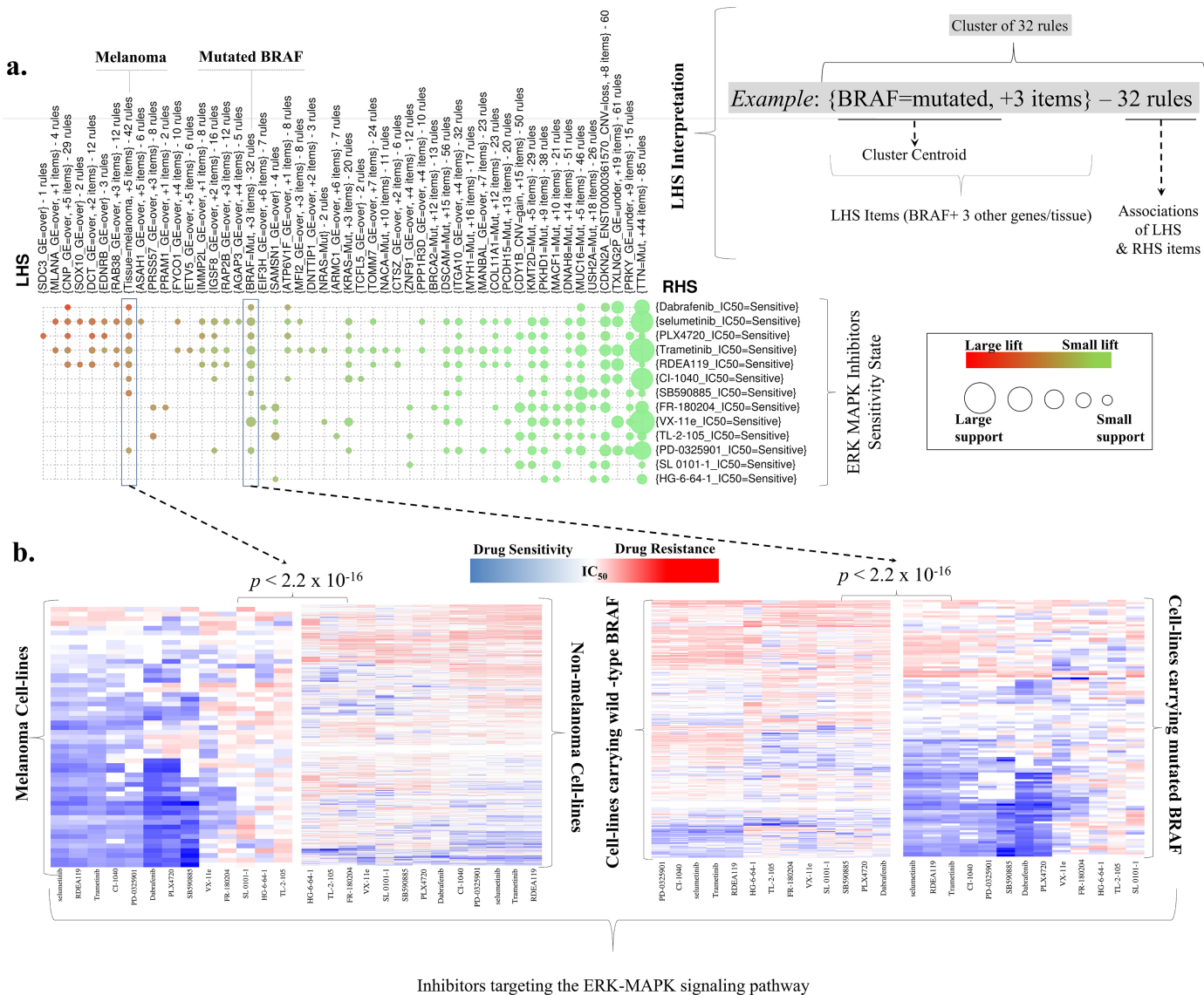


Figure 8

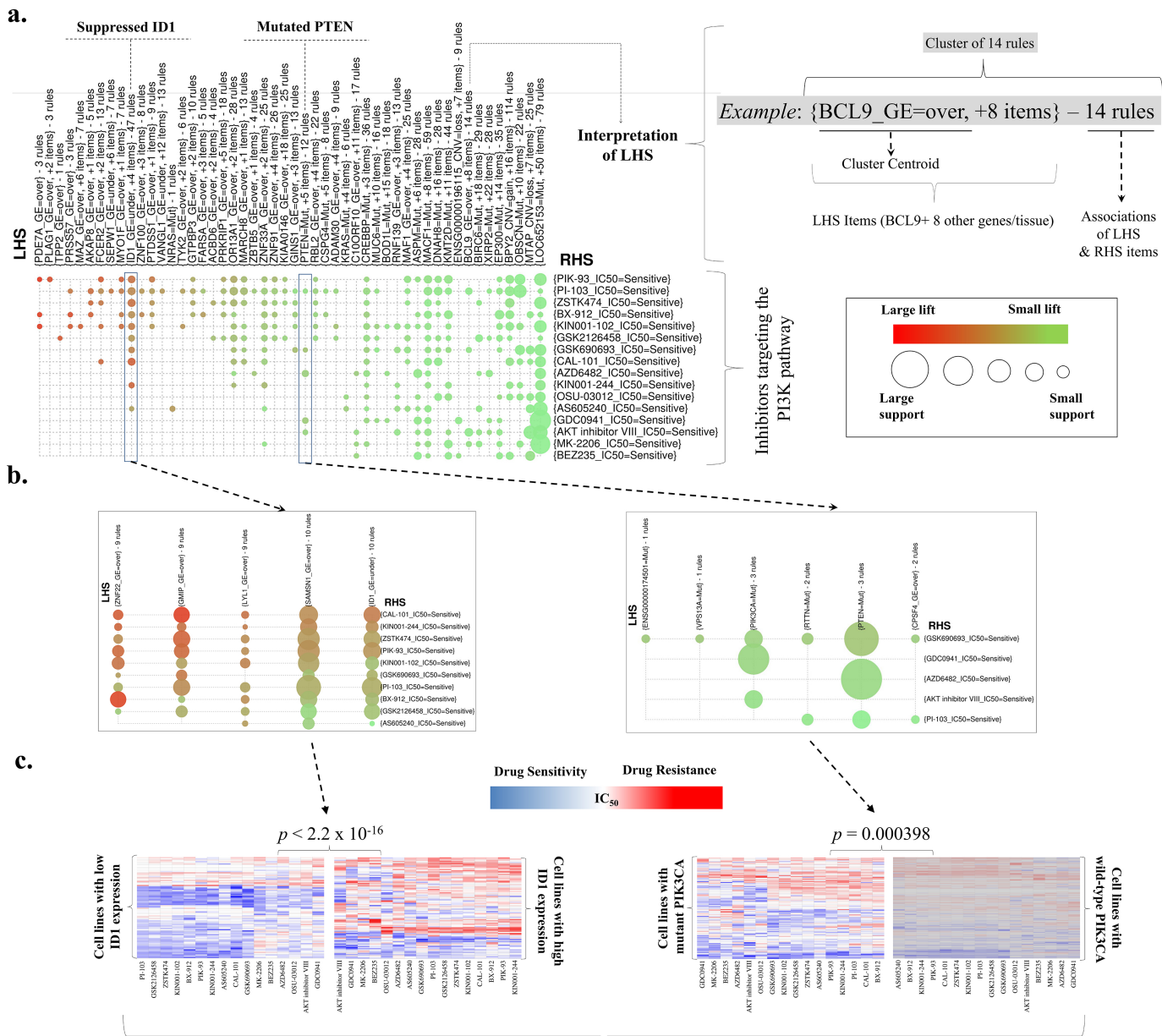


Figure 9

# PARPi Associations

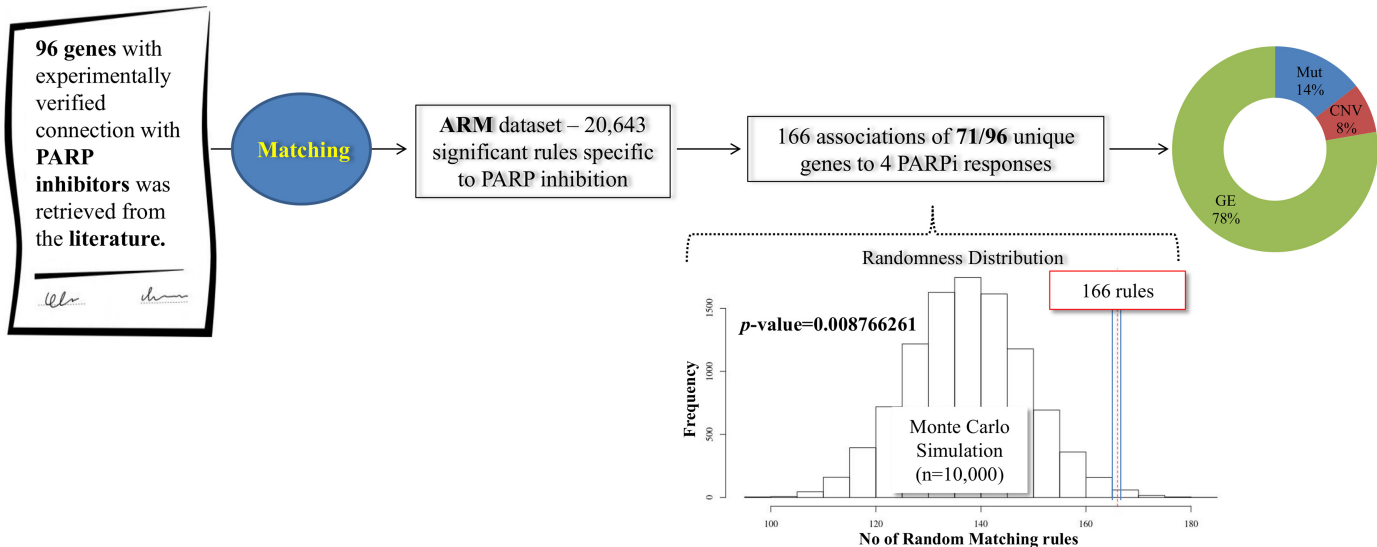


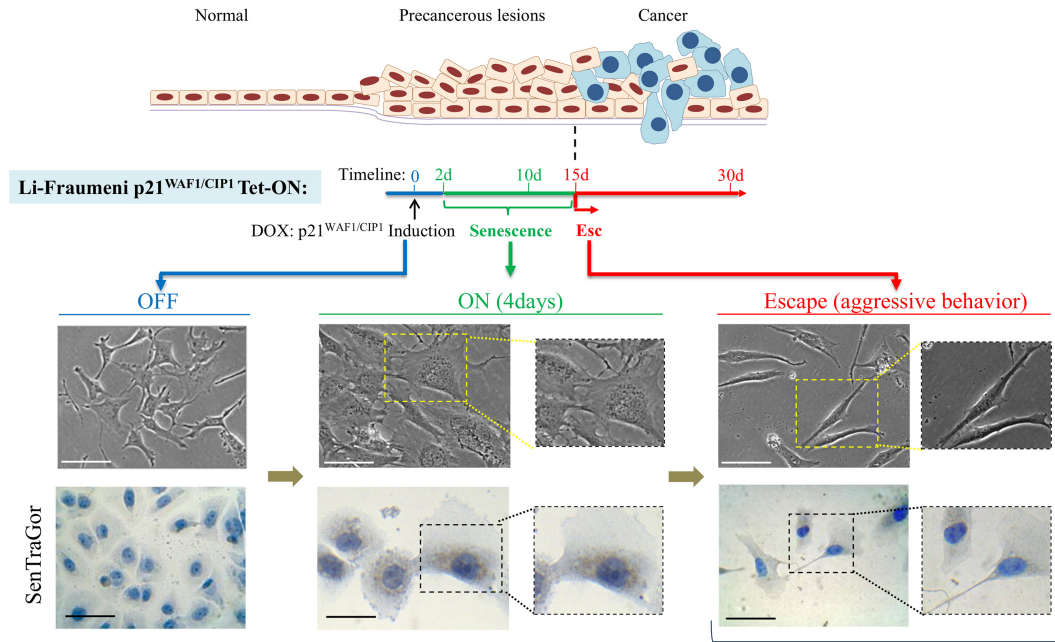
Figure 10

a.

## Rule linking ID1 suppression with PI3K chemosensitivity

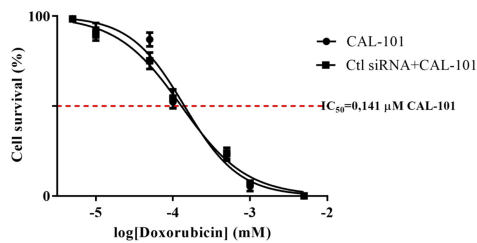
Inhibitor	Response Status	Support	Confidence	Lift	p-value	Target Pathway
CAL-101	Sensitivity	0.036963037	0.4625000	3.763923	>0,01E-05	PI3K signaling
ZSTK474	Sensitivity	0.036963037	0.4625000	3.283422	>0,01E-05	PI3K signaling

b.



c.

## CAL-101



## ZSTK474

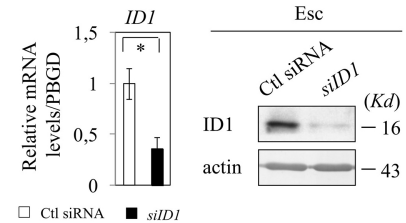
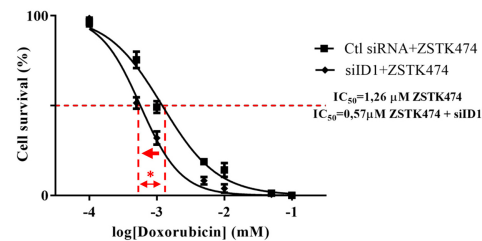
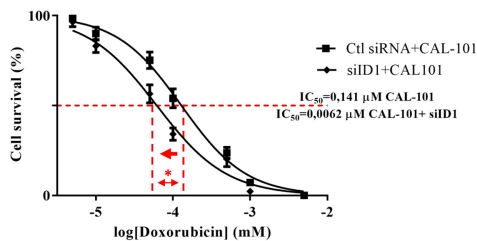
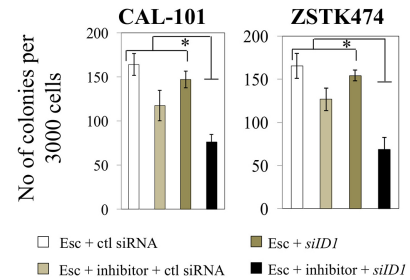
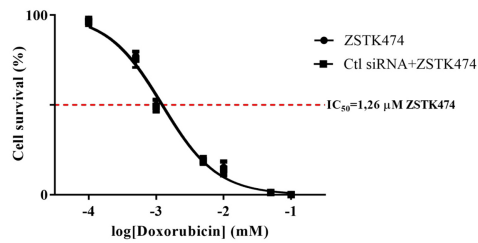


Figure 11



# Rules linking CDC6 overexpression with MAPK kinase inhibitors resistance

Inhibitor	Response Status	Support	Confidence	Lift	p-value	Target Pathway
PD-0325901	Resistant	0,00999	0,294118	2,0445	2,11E-05	MAPK/ERK signaling
Trametinib	Resistant	0,010989	0,323529	1,9991	2,03E-06	MAPK/ERK signaling
RDEA119	Resistant	0,00999	0,294118	1,692	0,0021252	MAPK/ERK signaling

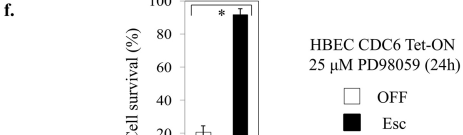
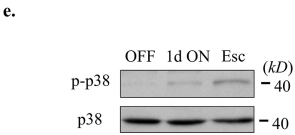
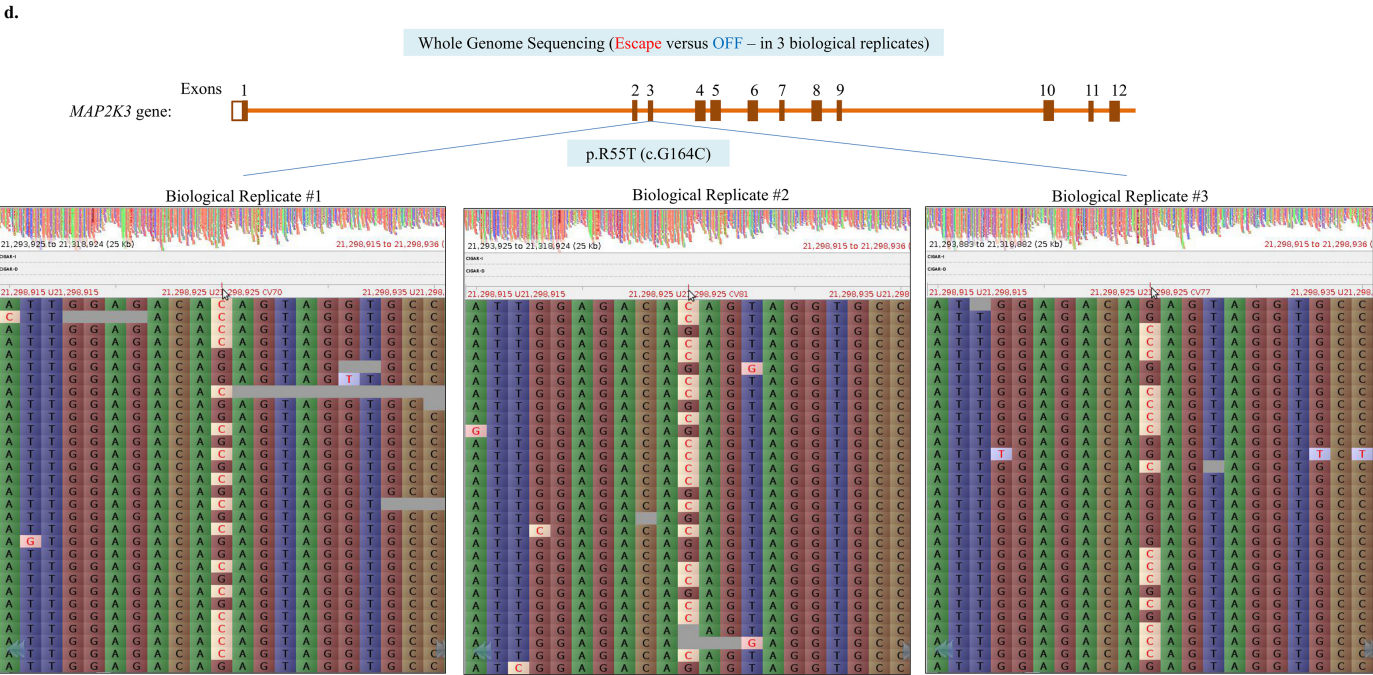
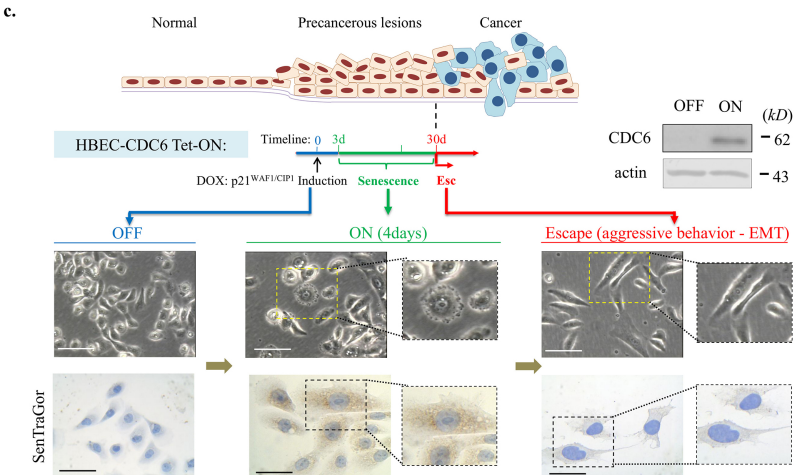
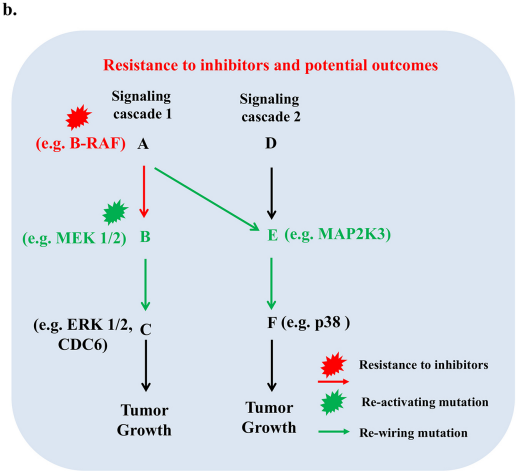


Figure 12

## Association Rule Mining (ARM) overlap with GDSC, CCLE & CTRP

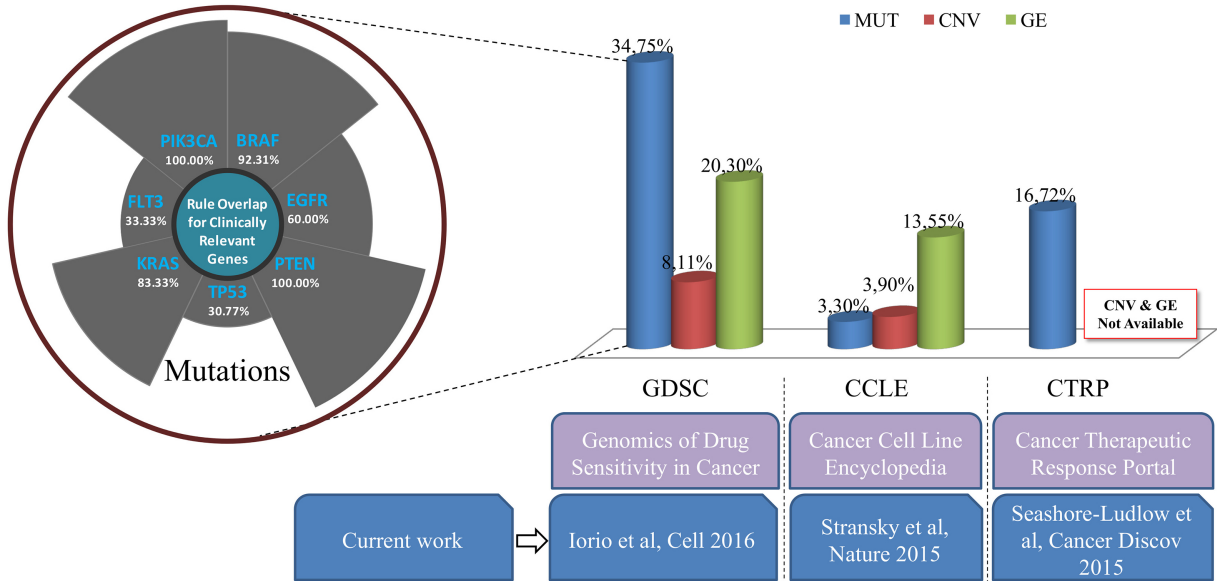


Figure 13

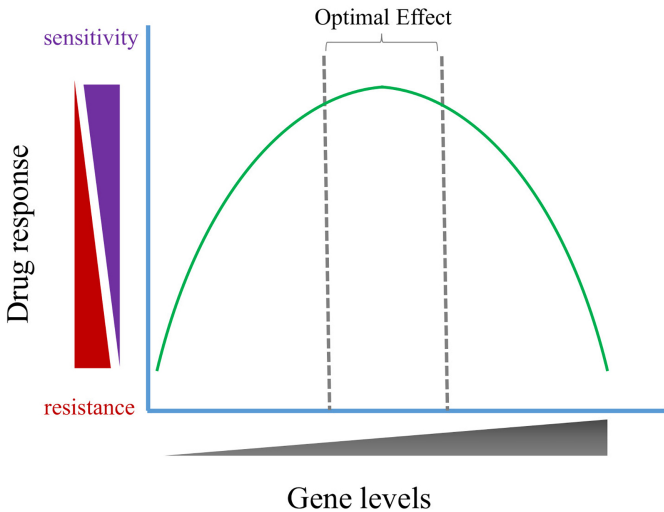


Figure 14