bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Deep Learning and Association Rule Mining for Predicting Drug Response in Cancer

[1*]Konstantinos Vougas, [2]Thomas Jackson, [1]Alexander Polyzos, [3,4]Michael Liontos, [5]Elizabeth O. Johnson, [6,3]Vassilis Georgoulias, [2]Paul Townsend, [7,8,9]Jiri Bartek and [3,2,1*]Vassilis G. Gorgoulis

1. Biomedical Research Foundation of the Academy of Athens, 4 Soranou Ephessiou St., Athens, GR-11527, Greece
2. Faculty Institute of Cancer Sciences, University of Manchester, Manchester Academic Health Science Centre, Wilmslow Road, Manchester, M20 4QL, UK
3. Molecular Carcinogenesis Group, Department of Histology and Embryology, School of Medicine, National & Kapodistrian University of Athens, 75 Mikras Asias Str, Athens, GR-11527, Greece
4. Oncology Unit, Department of Clinical Therapeutics, School of Medicine, National and Kapodistrian University of Athens, Athens, Greece
5. Laboratory for Education & Research in Neurosciences (LERNs), Department of Anatomy, School of Medicine, National & Kapodistrian University of Athens, 75 Mikras Asias Str, Athens, 11527, Greece
6. Laboratory of Tumor Cell Biology, School of Medicine, University of Crete, Heraklion, Crete, Greece
7. Genome Integrity Unit, Danish Cancer Society Research Centre, Strandboulevarden 49, Copenhagen DK-2100, Denmark
8. Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University, Hn ěvotínská, Olomouc 1333/5 779 00, Czech Republic

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

9. Science for Life Laboratory, Division of Translational Medicine and Chemical Biology, Department of Medical Biochemistry and Biophysics, Karolinska Institute, Stockholm SE-171 77, Sweden

* To whom correspondence should be addressed:

Konstantinos Vougas, E-mail: kvougas@bioacademy.gr; or kvougas@gmail.com;

Tel.: +30-210-6597129; Mob.:+30-6977631030

and Vassilis G. Gorgoulis, E-mail: vgorg@med.uoa.gr;

Tel.: +30-210-7462352

Abstract

A major challenge in cancer treatment is predicting the clinical response to anticancer drugs for each individual patient. For complex diseases, such as cancer, characterized by high inter-patient variance, the implementation of precision medicine approaches is dependent upon understanding the disease process at the molecular level. While the 'omics' era provides unique opportunities to dissect the molecular features of diseases, the ability to apply it to targeted therapeutic efforts is hindered by both the massive size and diverse nature of the 'omic' data. Recent advances with Deep Learning Neural Networks (DLNN), suggests that DLNN could be trained on large data sets to efficiently predict therapeutic responses. We present the application of Association Rule Mining (Market Basket Analysis) in combination with Deep Learning to integrate and extract knowledge in the form of easily interpretable rules from the molecular profiles of 689 cancer cell lines and predict pharmacological responses to 139 anti-cancer drugs.

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The proposed algorithm achieved superior classification and outperformed Random Forests which currently represents the state-of-the-art classification process. Finally, the *in silico* pipeline presented introduces a novel strategy for identifying drug combinations with high therapeutic potential. For the first time, we demonstrate that DLNN trained on a large pharmacogenomic data set can effectively predict the therapeutic response of specific drugs in specific cancer types, from a large panel of both drugs and cancer cell lines. These findings serve as a proof of concept for the application of DLNN to predict therapeutic responsiveness, a milestone in precision medicine.

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Predicting the clinical response to therapeutic agents is a major challenge in cancer treatment. Ultimately, the ability to generate genomic informed personalized treatment with high efficacy is dependent upon identifying molecular disease signatures and matching them with the most effective therapeutic interventions. While the advent of multiple platforms providing "omic" data permits scientists to dissect the molecular events that are known to drive carcinogenesis [1] and alter major downstream processes, such as gene expression [2], effectively translating the growing wealth of 'omic' data into a personalised treatment strategy required by precision medicine, has been daunting and without noteworthy success [3].

The successful identification of effective anti-cancer drugs has been primarily hindered by the lack of reliable preclinical models. Although cancer cell lines do not have the complexities of cancers, as they evolve in their normal tissue environment [4], new technologies provide a wealth of information regarding the enormous degree of genomic heterogeneity across cancers, helping us to define the role that this diversity may play in understanding the wide variation in treatment responses [5]. In light of this newfound understanding of the significance of cancer genomic heterogeneity, recent efforts have resulted in large panels of cancer cell lines with information defining their genetic constitution, molecular markers and responsiveness to therapeutic compounds. This information allows us to identify the genomic determinants of the clinical response to specific drugs. It has been recently reported that oncogenic alterations in a large panel of tumors from 'The Cancer Genome Atlas' (TCGA) and other studies are conserved across the large panel cell-lines [6]. Three large scale cell panels containing pharmacogenomic data have been made available to the public domain: a) the 'Cancer Cell Line Encyclopedia' (CCLE) [7], b) the Genomics of Drug Sensitivity in Cancer (GDSC) [8] and c) the NCI-60 [9]. To identify predictive biomarkers, these consortia have analysed the molecular profiles of over 1000 cancer

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

cell lines and drug profiles for a large number of anticancer drugs.

The availability of these large data sets of cell-line panels along with the availability of new computational technologies has propelled a recent surge of interest to perform parallel analyses across cell lines, to extract information and define predictive biomarkers. However, while large data sets of pharmacogenomic profiles have been compiled with detailed molecular features and drug responsiveness, well validated computational approaches to define biologically relevant rules and generate algorithms that can accurately predict the responsiveness to a specific therapeutic drug are lacking. Although data mining algorithms are supposed to analyse large volumes of data and uncover hidden relationships of potential clinical significance, today's complex "omic" data sets have been proven too multi-dimensional to be effectively managed by classical Machine Learning algorithms. Deep Learning neural networks (DLNN), on the other hand, have the ability to "understand" complexity and multidimensionality and have been effectively applied in various fields (e.g. image analysis, text mining, etc.) with increased classification accuracy compared to classical computation methods [10]. 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 [11]. 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 [12], drug toxicity in the liver [13] and pharmacological properties of drugs [14], among others. Together, studies using the DLNN architecture demonstrate that it is well suited for complex biological data because it can automatically construct complex features and allows for multi-task learning [15].

We designed a bioinformatics pipeline with multiple layers of molecular profiling information by

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

utilising the publically available pharmacogenomic profiles, in order to produce an expert system that, with demonstrated efficiency, could predict pharmacological responses to a large number of drugs over a broad panel of cancer-cell-lines (Figure 1).

Specifically, we performed feature selection in the form of association rules and utilized the selected features to train multiple state-of-the art DLNN to predict pharmacological response in a blind set. The association rules are treated as a novel meta-dataset and are utilised in the form of paradigms for knowledge extraction. To our knowledge, this is the first time that the DLNN framework is systematically applied to predict drug efficacy against cancer.

Results

1. Dataset compilation: To initiate a bioinformatics pipeline to predict drug response based on molecular profiles of multiple cancer cell types, we generated a large-scale pharmacogenomics dataset for 689 cancer cell lines and 139 anti-cancer drugs (**Supplementary methods – Main Dataset**). The new pharmacogenomics compilation was achieved by merging data from CCLP and GDSC. We used GDSC [8] as our drug response data source for 139 therapeutic compounds, which provided IC-50 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) [16]. CCLP was preferred as a data source since it provides profiles on 1,015 cancer cell lines and is not limited to the mutational status of genes (1,600), as is the case with CCLE. Although NCI-60 contains the largest number of therapeutic compounds tested for pharmacologic activity, it was excluded as a data source, as the

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

number of cell-lines presented is very low compared to the other resources used. This diminishes the effectiveness of NCI-60 to serve as a preclinical platform that can, at least, partially simulate clinically relevant tumour complexity (Figure 2).

2. Association Rule Mining: 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 [7, 8]. Previous efforts to define biomarkers of drug response primarily utilize elastic net regression, a penalized linear modelling technique, to identify cooperative interactions among multiple genes and transcripts across the genome and define response signatures for each drug [16]. While efficient, this algorithm suffers certain limitations since when used for feature selection, as described in previous studies [7, 8], 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 are supervised learning algorithms not being able to evaluate all possible combinations of a large number of features against a response variable. Associations are looked for only in the context of a predefined formula, such as Drug-Response, as compared to all genes being used as features (Drug-Response ~ All-Genes). The ability to evaluate all combinations of complex feature-to-feature interactions in relation to a response variable is not feasible within the design of the supervised learning algorithm 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

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

given its molecular profile, the elastic net algorithm does not perform optimally. This is due to the fact that at the core of the elastic net algorithm lays linear regression, as opposed to non-linear classifiers, such as Kernel-based models and Random-Forests. The later have been shown to outperform the elastic net algorithm 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 anticancer drugs[17].

2a. Apriori Algorithm: To overcome the primary limitations of the elastic net algorithm for feature selection, we applied a method used by large businesses to analyse the enormous volume of transaction data and discover all possible associations between the data features, namely Market Basket Analysis or Association Rule Mining. Previous studies moved along the same lines to produce easily interpretable logical rules out of imilar pharmacogenomic datasets [6, 18] but the methods utilised in these studies were again supervised. Association Rule Mining being a totally unsupervised method records all the associations among the data features in the form of easily interpretable rules, making knowledge extraction and meta-analysis a more straightforward process. Specifically, we applied the Apriori algorithm [19] to extract statistically 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), 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 in this study is the computationally intensive requirements. While this will likely improve as computing power increases, due to hardware limitations, 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

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Apriori algorithm can be found in the **Supplementary Methods Section, Association Rule Mining - Apriori Algorithm**.

2b. Dynamic Thresholding - Separating true rules from the noise: We devised a procedure that we named Dynamic Thresholding in order to select statistically significant rules by controlling the false discovery rate (FDR) to less than 5%. Dynamic Thresholding is based on running the Apriori algorithm on a permuted version of our initial dataset (**refer to Supplementary Methods - Dynamic Thresholding**). The biological relevance of the rules generated, were examined in separate paradigms, as we show below.

3. Rule Verification

3a. Paradigms: To validate the biological relevance of our statistically significant association rules, we examined whether known predictors of drug response are present in our rule set and whether drugs of a given target are present in sensitivity-associated rules along with the given target, if mutated or over-expressed.

Paradigm 1: We demonstrate that over-expressed NAD(P)H dehydrogenase 1 (NOQ1) and MDM2, a p53 inhibitor, which are known predictors of sensitivity for the drugs 17-AAG (Tanespimycin) and Nutlin-3, respectively [20, 21], are present in our rule-set (Supplementary Table 1 - 1-way). Additionally, the ABCB1 gene whose product is the protein Multidrug-Resistance-1

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(MDR1), was found in our rule set to be linked with resistance to multiple drugs only when is over-expressed, while when suppressed it is linked only with sensitivity (Supplementary Table 1 - 1-way**).**

*Paradigm-2:*Two k-mean clustering schemes (see Supplementary Methods) of the 1000 rules with the largest support (k=50) for the sensitivity response-state of drugs associated with (a) the BRAF gene (Targets: BRAF & MEK inhibitors) (Supplementary Table 2 – Drug_Targets) and (b) the PIK3CA gene (PIK3CA, AKT & mTOR inhibitors) (Supplementary Table 2 – Drug_Targets) revealed that the mutated BRAF and PIK3CA genes were both present in the respective rule-clusters, as the most frequently occurring features (Figure 3, Supplementary Table 1- 1-way). In other words, cell-lines bearing the BRAF and PIK3CA mutation showed a non-random high-frequency trend of sensitivity to the respective drugs targeting these gene products. Additionally, the BRAF-clustering revealed that the melanoma cell-lines were highly sensitive to BRAF and MEK inhibitors, a "prediction", which can be verified in the literature with studies showing that combined BRAF and MEK inhibition is, actually, one of the most effective treatments for melanomas[22]. We subsequently repeated the two aforementioned clustering schemes from the top-lift 1000 rules for BRAF and PIK3CA (Supplementary Figure 1, Supplementary Table 1- 1-way). The BRAF top-lift clustering again revealed melanoma sensitivity to BRAF and MEK inhibitors (Supplementary Figure 1a), while the PIK3CA top-lift clustering scheme revealed a significant and unique association of chronic myeloid leukaemia with sensitivity to the drug AZD8055, which is a dual inhibitor of mTORC1 and mTORC2 (Supplementary Figure 1b). This "prediction" is confirmed in the literature, where AZD8055 was found to be effective in all patient samples, and within the six most selective drugs in all cases [23].

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The rules clustering also revealed that AZD8055 would be effective in lymphoblastic leukaemia along with several other drugs, namely Rapamycin, MK-2206 and GDC-0941 (Supplementary Figure 1b). Particularly, MK-2206, an Akt inhibitor, showed significant potency in treating T-cell acute lymphoblastic leukemia [24], whereas GDC-0941, a pan class I PI3K inhibitor, was found to be effective against primary T-ALLs from wild-type and Kras(G12D) mice [25].

Paradigm-3: In order to mine the association rules for tissue specific drug sensitivity and resistance we created two subsets, both containing tissue type at the Left Hand Side (**LHS**) of the rules and drugs' sensitivity and resistance response-status at the Right Hand Side (**RHS**) of the rules. These subsets were clustered and visualized, as described above (k=50) (Figure 4, Supplementary Table 1 - 1-way). A pattern that we immediately note is the presence of multiple drug-resistance rules in small-cell lung carcinoma (SCLC) (Figure 4a, Supplementary Table 1 - 1-way), a non-frequent lung cancer type (15% of total cases) [http://www.cancer.gov/types/lung/hp/small-cell-lung-treatment-pdq]. Patients with SCLC respond very well to front line chemotherapy, but early clinical recurrence are due to tumours which are unresponsive to different anticancer agents. By examining the sensitivity clustering visualisation (Figure 4b, Supplementary Table 1- 1-way), we observe that SCLC presents sensitivity to only two therapeutic substances, namely ABT-263 (BCL-family inhibitor) and KIN001-135 (IKK Inhibitor) (Supplementary Table 2). Sensitivity of SCLC cells to ABT-263 is confirmed in the literature [26], through the promotion of cell apoptosis.

Paradigm 4: In the Tissue of origin – Drug Sensitivity view (Figure 4b, Supplementary Table 1-1-way) we note that the 5 rules with the highest support in melanoma indicate sensitivity to 3

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

MEK inhibitors (CI-1040, RDEA119 and PD0325901) and to 2 BRAF inhibiotrs (SB590885 and PLX4720) (Supplementary, Table 1 - 1-way). Today, the Standard of Care for BRAF mutant melanoma patients is a combination of BRAF and MEK inhibition [27]. With our algorithm, when we accessed information related to the drug response of melanoma cell-lines from the 2-way rules (Supplementary Table 1 - 2-way), we observe that confidence and lift increases in all rules that in addition to Tissue=melanoma, also contain 'BRAF=Mut' at the **LHS** and Sensitivity to BRAF inhibitors at the **RHS,** confirming the fact that sensitivity to BRAF inhibitors is increased by the presence of the BRAF mutation.

*Paradigm 5***:** The following two paradigms 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 SCLC and identify potential points of intervention.

Paradigm 5a: With the 1-way rules (Supplementary Table 1 - 1-way), we observe statistically significant resistance to Sunitinib (a multi-kinase inhibitor) with a lift-value of 2.46 in 10 out of 46 Small Cell Lung Cancer (SCLC) cell-lines (22%). With the 2-way rules (Supplementary Table 1 - 2-way), we note that Growth hormone releasing hormone (GHRH) over-expression greatly increases the lift-value to 6.1, since resistance to Sunitinib is observed in 7 out of 13 SCLC cell-lines over-expressing GHRH (54%). Inhibiting GHRH activity using antagonists wields high anti-tumor activity by impending cell proliferation [28]. Furthermore GHRH activity has been linked to drug-resistance in triple negative breast cancer [29]. Moreover, the 2-way rules indicate that GHRH over-expression significantly contributes to the resistance of SCLC to CGP-082996 and XMD8-85, which are CDK4 and ERK5 inhibitors, respectively. By including interactions in

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

association rule mining, we are able to infer in this particular example that GHRH antagonists could be potentially used in combination therapy schemes with specific chemotherapeutic agents for the effective treatment of SCLC.

Paradigm 5b: With the 1-way rules (Supplementary Table 1 - 1-way), we observe statistically significant resistance to Obatoclax-Mesylate, a BCL-family inhibitor, with a lift-value of 2.87 in 14 out of 46 SCLC cell-lines (30.4%). With the 2 -way rules (Supplementary Table 1 - 2-way), we note that SMAD3 down-regulation greatly increases the lift-value to 8.4, since resistance to Obatoclax-Mesylate is observed in 8 out of 9 SCLC cell-lines under-expressing SMAD3 (88.9%). SMAD3 is known to promote apoptosis through transcriptional inhibition of BCL-2 [30]. 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 part of the resistance mechanism of SCLC to BCL-family inhibitors, by highlighting a unique molecule that presents high mechanistic relevance to BCL-inhibition.

4. Prediction of Drug-Response

4a. Train & Test datasets for Machine Learning: To predict drug response through machine learning, we split the main data-set into two subsets, referred to as training set and test set and consisting of approximately 2/3 and 1/3 of the main data-set, respectively. The detailed description of the sets construction is presented in the relevant Supplementary Methods section -

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Prediction of drug-response. (Supplementary Figure 2)

4b. Deep Learning: For our drug response classification framework, we applied DLNN [11] enhanced by Bagging Ensemble Learning[31]. Although its performance has not been tested in drug response prediction, we selected the Deep Learning Framework because it has redefined the state-of-the-art in many applications ranging from image recognition to genomics [11]. In particular we chose to use the open-source DLNN framework provided by H2O.ai (http://www.h2o.ai/). This is a cluster ready framework, which allows for the machine learning part of our pipeline to be readily deployable to a high performance-computing environment. In order for machine learning to be able to perform well on the blind-set (test-set), it is critical to select only the most relevant features for training the classifier, for example features that are highly correlated with drug response. If a large number of irrelevant features are used for training, the classifier will be trained on noise, and although it will produce excellent results on the training set, it will perform poorly on the blind-set. This problem is referred to as over-fitting and in the application of 'omics' information, where the number of features (in our case, gene expression, mutation status, etc) vastly outnumbers the total number of cases, over-fitting is inevitable.

Previous reports have used elastic-net (as discussed above) [32] as a feature reduction technique. To address this challenge and select the most relevant features, we utilised the rule-set generated by Association Rule Mining, as performed on the training-set alone. (Supplementary Table 3). The deep learning classifiers (one classifier per drug and per drug-response-state) were trained on the training set using as training features the ones selected by the Association Rule Mining procedure and as response the respective drug-response. Each classifier's performance was then assessed on a blind-set (the test-set), where we provided only the specific features upon which the classifiers

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

predicted the drug response, which was then compared to the actual drug-response value. We repeated the classification task utilising Random Forests, as the classification framework, which produced top prediction performance [17], allowing us to compare the performance of the DLNNs with a state-of-the-art framework. Random Forests is a highly adaptive tree-based machine learning tool, that has been applied for prediction and classification for genomic data, and unsupervised learning [33]. The classification performance in all cases was quantified through ROC-Curve analysis. Technical details are presented in the respective Supplementary Methods Section.

By using the genes involved in the association-rules as features and the DLNN as a machine-learning framework, we constructed classifiers that would be able to predict whether a cell-line would be sensitive or resistant to a given drug based on its molecular profile. In agreement with a previous study [6], since this is a dataset comprised of many cancer cell-lines from different tissues of origin, we observe that the vast majority of predictive features are gene expression levels. Additionally we noted that the information of tissue of origin significantly improved the prediction performance. To evaluate the classification efficiency of our classifiers, we applied a series of metrics, namely Area Under the Curve (AUC) of the Receiver Operating Characterstic (ROC)-curve, Sensitivity, Specificity, Accuracy, Positive Predictive Value (PPV), Negative Predictive Value (NPV) and False Positive Rate (FPR). The results for all our classifiers are reported in Supplementary Table 4. According to the widely accepted AUC-based classification quality grading scale, classifiers that produce AUCs 0.90 - 1 are considered excellent, 0.80 - 0.90 are good, 0.70 - 0.80 are fair, 0.60 - 0.70 are poor classifiers, while classifiers with an AUCs below 0.6 are considered failed or random classifiers [34]. Out of a total of 278 classification tasks,

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

corresponding to 139 drugs each with two responses (sensitivity and resistance), our pipeline produced 276 classifiers (see Supplementary Methods - Prediction of drug-response). Out of the 276 trained and tested DLNN classifiers, approximately 1% were excellent, 17% good, 54% fair, 24% poor and 4% random classifiers as opposed to 0%, 10%, 47%, 33% and 9% for the Random Forests using the AUC classification quality grading scale (Figure 5a, Supplementary Table. 4). More specifically, of the Random Forest classifiers an equal to or better than a fair quality grade (AUC>0.7) was achieved by 57% of the classifiers, whereas this quality grade (e.g. equal to or better than fair) was achieved by 72% of the DLNN classifiers, indicating a superior performance of DLNN over Random Forest classification quality. Of the remaining performance metrics, Random Forests outperformed the DLNNs only in sensitivity. On the other hand, DLNNs were more specific, more accurate, with higher Positive Predictive Value and lower False Positive Rate than the respective Random Forests classifiers, indicating the overall superiority of the DLNN classifiers (Figure 5b, Supplementary Table 4).

5. Drug-Clustering

5a. Drug combinations against multiple targets belonging to inter-linked or overlapping signalling cascades are strong candidates for presenting synergistic effects [35]. Our aim was to create a clustering scheme based solely on the presence of specific genes derived from the rule-set connected to a specific drug-response. To this end, we produced two individual clustering schemes: 1) one for drug-sensitivity (Figure 6, Supplementary Figure 3) and 2) one for drug resistance (Supplementary Figure 3). All dendrograms can be accessed in HTML format in the

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

folder '/Figures/Dendrograms/' at the Github repository (see Supplementary Methods - Data Availability). For details on clustering refer to Supplementary Methods - Drug-clustering.

When we examine the two circular dendrograms, it is clear that they bear no-resemblance to each other meaning that the differentiated genes in the cell-lines that are resistant to a specific drug are diverse from the differentiated genes in the cell-lines that are sensitive to the same drug. To confirm this observation, we created a table describing the number of genes involved in sensitivity and resistance for each drug (Supplementary Table 5, 'Genes overlap'). We were then able to calculate the overlap between all possible combinations as well as whether the magnitude of the overlap was random or over/under represented (p-value)(Supplementary Table 5, 'p-values'). Given the size of the sensitivity and resistance gene-sets for each drug, this calculation was accomplished by Monte-Carlo simulation, as described in the Supplementary Methods. This calculation highly correlates with the produced dendrograms, as drug-response-states, that cluster together, present statistically significant over-represented gene-set overlaps (Supplementary Table 5), and allowed us to assign a p-value to each dendrogram relationship. We isolated all of the above information related to the sensitivity and resistance states only of each drug and report them in Supplementary Table 5 ('comparison sens & res per drug'). We note a very low overlap among the gene-sets involved in the sensitivity and resistance states of any given drug, indicating that the pathways involved in sensitivity and resistance for any given drug are diverse which is in perfect agreement with our prediction strategy, where we use different models from predicting sensitivity and resistance for each drug. When we closely examined the sensitivity dendrogram clustering structure, we note that it was highly relevant to the drugs' target (Figure 6), indicating that drugs with the same target tend to cluster in close proximity to one another. Additionally, the sensitivity-status clustering appears to capture the broader relationships among the drugs. For

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

instance, we observe that there is a branch populated not only by BRAF but also by MEK inhibitors which practically belong to the same pathway (Figure 6)[36]**.** Likewise, we also note that PI3K, AKT and mTOR inhibitors were clustered together (Figure6). In contrast, when we examine the resistance dendrogram we note that clustering is less relevant to the drug targets in comparison to the sensitivity dendrogram meaning that the molecular cascades implicated in drug resistance are diverse from the ones that are being targeted by the drug. To determine whether drug clustering translates to highly correlated activity of closely clustered drugs (sensitivity dendrogram) across the cell-lines, we extracted the drug-to-drug rules from our total rule-set (Supplementary Table 6 – Sens_Sens) which examines whether sensitivity responses of the cell lines to certain drugs are correlated to others. We clearly observe that the rules with the largest support had their Lift and Confidence values inversely correlated to the clustering distance in the dendrogram. This means that drugs which are connected in these rules tend to cluster closer in the sensitivity denrogram **(Figure 6b),** implying that the corresponding genes are involved mechanistically in producing a drug sensitive environment.

5b. Suggestion of a Rule for the determination of drug partners with high potency: Based on drug clustering, we propose a drug-pair selection strategy for combination therapy using the following rule: Candidates with a high probability for presenting synergistic effect are those that: a) target different molecules, b) are located close together in the sensitivity dendrogram with their proximity also confirmed in the drug-to-drug association rules and c) cluster far-away from each other in the gene-based resistance dendrogram. This rule is based on the fact that drug sensitivity is dependent upon proximal gene-signalling cascades, while at the same time it is difficult for a cancer cell to develop resistance to both drugs, since their resistance mechanisms

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

are diverse, as indicated by their distance from each other on the resistance dendrogram.

5c. Paradigms supporting the Drug-Partner Rule

Paradigm-1: In a recent study, evidence was presented that SCLC which is resistant to ABT-263, a Bcl-family inhibitor, can be effectively sensitized with the addition of AZD8055, a mTOR inhibitor [26]. They provide evidence of significant tumour regression of an ABT-263 resistant SCLC patient-derived xenograft model after receiving combination therapy with both drugs. Examination of the sensitivity dendrogram, shows that ABT-263 lies in close proximity (on same main branch) to AZD8055, along with GDC0941 and AZD6482, which are PI3K inhibitors, hence part of the same pathway (PI3K/AKT/mTOR). Additionally, ABT-263 and AZD8055 show distant clustering on the resistance dendrogram, as predicted by our drug-partner rule (Supplementary Figure 3).

Paradigm-2: Based on the drug-partner rule paradigm, Bcl-family inhibitors have the potential of synergy with PI3K inhibitors. This hypothesis is supported in the literature, where it is demonstrated that combination treatment with PI3K and Bcl-Family inhibitors led to a significant increase in PARP cleavage and cell death in all Renal Carcinoma cell lines [37]. As described above, BRAF and MEK inhibitors which are closely clustered in the sensitivity dendrogram and remotely clustered in the resistance dendrogram (Figure 6, Supplementary Figure 3), have shown to produce very promising results for melanoma treatment when combined together [35].

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Discussion

We present an *in silico* pipeline that utilises a large cancer cell-line dataset (689 cell lines) with diverse genomic features (>39,000 features) and responses to a diverse number of drugs (139), to extract knowledge in the form of easily interpretable rules and then by combining these rules with the sate-of-the-art Deep Learning framework to accurately predict drug responses. We also demonstrate that prediction of sensitivity and resistance responses must by handled by different models since the genes that drive these responses are diverse. Furthermore, we suggest a strategy, based on the drug sensitivity and resistance clustering to select the most potent candidates for drug-combination therapy.

Validation of the proposed algorithm, demonstrates high classification power with superior specificity and accuracy, compared to Random Forests which is considered the current state-of-the-art process. In a personalised medicine setup, the molecular profile of the patient's tumour would be introduced in the pipeline and the later, as an expert system, would predict the drug response for a large screen of drugs allowing clinicians to select the best candidates for mono- or combination therapy. These candidates would then be tested on patient-derived primary 3D cancer cell cultures[38] and/or on xenograft models [39]. The most efficient combination would then be applied in the form of a therapeutic scheme directly on the patient with constant monitoring for administration of personalised dosing.

The power of the pipeline lies on the efficiency, expandability and ability to create easily interpretable rules of the Association Rule mining algorithm, and to the ability of Deep Learning

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

to capture the complex heterogeneity of tumors. It can be further expanded by increasing the number of cancer cell-lines, including primary cancer cell-lines, as well as by increasing the number of therapeutic agents analysed by the system. Additionally, the system allows integration of other layers of 'omic' information, including meta-genomics, proteomics, phospho-proteomics, interactomics and metabolomics that will further enhance the prediction and drug-clustering schemes. We propose that our bioinformatics pipeline is expandable and effective utilising state-of-the-art algorithms such as Association Rule Mining and Deep Learning and can effectively be applied in the rapidly expanding “omics” era for devising personalised medicine schemes, as well as for drug discovery.

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure Legends

Figure 1: Schematic representation of the study design and bioinformatics pipeline. The full data set was constructed using the GDSC and CCLP databases. Test data and training data sets were created using blocked randomisation. The use and progress of each data set is depicted by coloured arrows; full data set (red), test set (orange) and training set (green). The training and test set were used for deep learning and subsequent predicting of sensitivity and resistance. The full data set was used for clustering and dug synergy predictions. Unsupervised association rule mining was used for feature selection (from >39,000 features) of pharmacogenetic data. Significant association rules were defined by dynamic thresh-holding. Rules derived from the training data set were used for deep learning. DLNN classifiers were applied for prediction of drug sensitivity and resistance. Extracted information was validated using the test set. Genes from rules derived from the full data set were used to construct a dissimilarity matrix based on the Jaccard index. Drugs were then clustered and predictions for drug synergy made.

Figure 2: Description of full data set and summary of main data matrix. a) Tissue of origin of the 689 cell-lines of the data-set**.** b) Description of each data type used including source, number of features and levels. c) Summary of the main data matrix containing tissue of origin, mutation status, copy number variation, gene expression and drug response information for the 689 cancer cell-lines. The actual matrix is available as an R data object (MASTER_MATRIX.RData) that can be accessed from the data folder located in the Github repository (see Sup. Methods – Data Availability).

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 3: Group-wise Association Rules visualization by k-means clustering k=50 of 1000 1-way rules. Columns in the figure represent the Left Hand Side (LHS) of the rules which are molecular features (tissue of origin, mutation status, copy number variation and gene expression) while rows represent the Right Hand Side (RHS) of the rules which are drug responses. The size of the circles corresponds to the support magnitude of the rules while the color of the circles corresponds to lift values of the rules. a) Rules with the largest support, for the sensitivity state of BRAF & MEK inhibitors. b) Rules with the largest support, for the sensitivity state of PIK3CA, AKT & mTOR inhibitors.

Figure 4: Group-wise Association Rules visualization by k-means clustering k=50. Columns in the figure represent the Left Hand Side (LHS) of the rules which are molecular features (tissue of origin) while rows represent the Right Hand Side (RHS) of the rules which are drug responses. a) k-means clustering, where k=50, of all the 1-way rules containing tissue of origin and the sensitivity state of all drugs in the matrix. b) k-means clustering, where k=50, of all the 1-way rules containing tissue of origin and the resistance state of all drugs in the matrix.

Figure 5: Comparison of prediction performance between Deep Learning and Random Forests. a) Comparison of ROC-Analysis Areas Under the Curve (AUCs) achieved by Deep Learning & Random Forests. b) Comparison of Sensitivity, Specificity, Accuracy, PPV, NPV and FPR between Deep Learning & Random Forests. P-values were derived by the non-parametric two-tailed Wilcoxon test.

Figure 6: Drug sensitivity clustering based on the similarity of genes involved in the association

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

rules with the sensitivity state of each drug. a) Drug sensitivity dendrogram colour coded by drug type and with clusters of interest highlighted. Full scalable dendrograms are available in supplementary materials. b) Correlation of Jaccard-distance with rule confidence and lift for drug-to-drug association rules (sensitivity state, Supplementary Table 6 – Sens_Sens). Detailed information regarding the groups and the associated drug targets is located in Supplementary Table 2.

Supplementary Figure 1: Group-wise Association Rules visualization by k-means clustering k=50. Columns in the figure represent the Left Hand Side (LHS) of the rules which are molecular features (tissue of origin, mutation status, copy number variation and gene expression) while rows represent the Right Hand Side (RHS) of the rules which are drug responses. The size of the circles corresponds to the support magnitude of the rules while the color of the circles corresponds to lift values of the rules. a) k-means clustering, where k=50, of 1000 1-way rules with the largest lift, for the sensitivity state of BRAF & MEK inhibitors. b) k-means clustering, where k=50, of 1000 1-way rules with the largest lift, for the sensitivity state of PIK3CA, AKT & mTOR inhibitors.

Supplementary Figure 2: Training and Test set compilation. The training and test subsets were constructed by performing blocked randomization on the original data matrix. Tissue type was used as the blocking factor. Two thirds of the cell lines from each tissue type were assigned to the training set and the remaining one third to the test set. The respective pie charts represent the tissue of origin information for the training and test sets.

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Figure 3: Representation of examples verifying the suggested rule for selecting potential candidates for combination therapy. Potential candidates have different molecular targets and reside close together to the sensitivity dendrogram, presenting at the same time distant localization to the resistance dendrogram.

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods

All scripting, data-processing, statistical calculations have been performed with R-language for statistical computing [40].

1. Datasets

The dataset compilation from 'Genomics of Drug Sensitivity in Cancer' (GDSC - release 5.0) and 'COSMIC Cell Line Project' (CCLP) was created by the R script "script_make_data.R".

Tissue of origin and drug response data were obtained from:

ftp://ftp.sanger.ac.uk/pub4/cancerrxgene/releases/release-5.0/gdsc_manova_input_w5.csv.

Gene mutation data was obtained from 'CosmicCLP_MutantExport.tsv', gene expression data was obtained from 'CCLP_CompleteGeneExpression.tsv' and copy number variation data was obtained from 'CCLP_CNV.tsv'. All the aforementioned files were downloaded from

http://cancer.sanger.ac.uk/cell_lines/download.

More specifically, with respect to the molecular profiling data we included mutational status for 19426 genes, copy-number-variation status for the exons of 3744 genes and gene-expression status for 16444 genes. The gene mutation status is a factor consisting of 1 level, namely "Mut" that corresponds to all single point mutations apart from the silent ones. The copy-number-variation status is a factor consisting of two levels, namely "Gain" and "Loss" for gains and losses, respectively, while the gene-expression status is a factor that also consists of two levels ("over" and "under") that correspond to z-scored gene expression levels greater and lower than two standard deviations from the mean, respectively. Finally the drug status is a factor consisting

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of two levels, namely "Resistant" and "Sensitive" that correspond to z-scored IC-50 levels greater and lower than one standard deviation from the mean, respectively. The R-Data object containing the matrix is stored in the file MASTER_MATRIX.RData.

2. Association Rule Mining - Apriori Algorithm

To provide insights regarding the way the algorithm works we provide an example. Gene expression of Gene-A in our dataset has two levels, "over" and "under". The **Apriori algorithm** will generate two features out of Gene-A gene expression, namely Gene-A=over & Gene-A=under. The rules come in the form of A => 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**). For the scope of the current study, we only kept the rules containing drug sensitivity features on the **RHS.** The algorithm can also be utilized to mine for more complex association rules containing interactions on the **LHS** in the form of A , B => C which is a two-way interaction, being able to go as deep as the data-set and the computational resources permit.

There are three basic metrics utilized by the algorithm in order to describe the power and significance of the rules. These metrics are **Support**, **Confidence** and **Lift**. **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 **LHS** of the rule. Finally **Lift** is a measure of significance. For the simple rule A => B, $t = \frac{p(A \cap B)}{p(A) \times p(B)}$, which, based on probability theory will be equal to 1 if the features A & B are independent. For dependent features the value of **Lift** will be greater than 1 and the value being proportional to the power of the association. In order to run the **Apriori** algorithm, the user has to define minimum support and confidence values below

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

which all rules are discarded, plus the number of allowed interactions in the **LHS**. We initially ran the algorithm by setting a minimum support and confidence of 0.58%, corresponding to just 4 out of the total of 689 cell-lines allowing for no interactions (1-way: A =>B), which is the minimum our computational resources permitted. Finally we ran the Apriori algoritm at minimum support and confidence levels of 1.02% (due to limitations in computational resources) allowing for one interaction (2-way: A + B => C).

*Dynamic Thresholding:*The Apriori algorithm was ran on a permuted version of our initial dataset (MASTER_MATRIX_PERMUTED.RData), which was produced by randomly shuffling each individual column of the dataset. The permutated matrix was produced with the script script_make_data.R. We initially ran the algorithm with the aforementioned support and confidence values on the permuted dataset and we determined the Lift threshold that would control the false discovery rate at less than 5%. We noted, however, that for each different set of support and confidence values belonging to our actual rules, there was a different lift threshold for FDR<5% if the Apriori algorithm had run on the permuted dataset with that set of support and confidence values as the minimum support and confidence parameters of the algorithm respectively. We therefore adjusted our thresholding determination with a method we call **Dynamic Thresholding**. Specifically, for every unique set of support and confidence values, we ran the Apriori algorithm on the permuted dataset using these values as the minimum support and confidence required by the algorithm, and we then determine the Lift threshold for which FDR=5%. After the completion of that process we evaluated each one of our actual rules based on its Lift value; if above of the specific threshold, the rule was accepted as significant, otherwise it was rejected. Both 1-way and 2-way rules where filtered keeping only the significant

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

rules (FDR<5%). The significant rules are available in Supplementary Table 1. The rule-set constitutes a novel meta-data-set, which can be utilized for knowledge extraction as per the paradigms that follow in the current text.

The implementation of the Apriori & Dynamic Thresholding algorithms can be found in the script 'script_dynamic_thresholding.R'.

Group-wise rule visualization: The group wise Association Rules visualization presented in the current study utilizes k-means clustering in order to visualize data with high dimensionality and high scarcity and are described in detail in Hahsler et al. (2011)[41] and is implemented in script 'script_rules_visualize.R'.

3. Prediction of drug-response

Training & Test Sets: The training and test sets were created by the R script "script_make_data.R". The original z-scored gene expression levels were restored to the total matrix and the Training and Test subsets were constructed by performing blocked randomization on the original matrix. The blocking factor was the tissue type, and two thirds of the cell lines from each tissue type were randomly assigned to the Training sets and the remaining one third to the Test set. The ratios were always rounded in favor of the Training set. If there were only two cases for a particular tissue type then they were evenly split between the Training and Test sets and if there was only one case, it was assigned only to the Training set. Additionally, the gene-expression factors were replaced with the original z-transformed gene expression levels. The Training set consisted of 458 and the test set of 231 cell-lines (TRAIN_GE_NUM.Rdata, TEST_GE_NUM.Rdata, Supplementary Figure 2). In both sets there were several cell-lines

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

lacking gene-expression information. These cell lines were removed. The final version of the Training and Test sets utilized for Deep Learning consisted of 444 and 228 cell-lines respectively.

Feature Selection: The Apriori – Dynamic Thresholding algorithm, as described above, ran on the Training set alone at minimum support and confidence levels of 0.58%, in order to produce rules having no feedback from the Test set used for measuring the classification performance (Supplementary Table 3). For every drug and for each different drug response (Sensitive or Resistant) the genes present in the respective relevant rules having Support values greater than the support-values 1st quantile level were grouped and used as drug-state-specific feature subset along with the information on the tissue of origin (Supplementary Table. 4) for training as many individual classifiers. The 1st quantile condition was used because it provided better predictions as measured from intra-training-set k-fold cross-validation utilizing the ROC-curve AUC as the performance metric. For the total of the 139 drugs, each one having two states, Sensitive and Resistant, 278 classifiers would have been required. However two drug-state pairs, namely Nilotinib and WH-4-023 resistance, were not found in any rules, and therefore, there were no available features in order to train any classifiers. This resulted in the training of 276 classifiers for the DLNN and Random Forests frameworks.

Deep Learning: Deep Learning Neural Networks (DLNN) were constructed using the H2O.ai platform [http://www.h2o.ai/] each consisting of 3 hidden layers with 100 neurons in each layer using Maxout with Dropout as the activation function and class balancing. Further parameters for the DLNNs were: number of epochs=200, input dropout ratio=0.1, hidden dropout ratio=0.3.

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Internal performance metrics were acquired using 3-fold cross-validation. Deep Learning is implemented in the script 'script_h2o_deep_learning_predictor_with_features_random_sampling_GE_NUM.R'. The H2O.ai platform was selected because it provides a cluster-ready framework for immediate and on-demand scaling-up.

Each DLNN was utilized in a bagging-ensemble learning scheme, where for each feature subset multiple training rounds were performed. In each Training round, if the feature subset number was greater than 200, only a maximum of 200 features were randomly selected. If the feature subset number was lower than 200 then 3 training rounds were performed, each containing 75% of the feature subset, which was for each time was randomly selected. At the end of each Training round, the DLNN was asked to predict the probabilities for the Test set. The Test-set predicted probabilities from each Training round were averaged to produce the final Test-set predicted probabilities (row-wise, hence for each Test-set cell-line) using a weighted averaging scheme, the weight being the ROC-curve area under the curve (AUC) calculated from the Training step of each round based on 3-fold cross-validation. After the completion of the Test-set prediction, the classification performance was measured by calculating the Area Under the Curve (AUC) of the ROC-curve, Sensitivity, Specificity, Accuracy (ACC), Positive and Negative Predictive Values (PPV & NPV) and False Positive Rate (FPR) of the prediction by utilizing the ROCR-package [42] (Supplementary Table 4 – 'Deep Learning'). For all the calculations of the aforementioned metrics apart from the AUC, the selected class-discriminating threshold was the one maximizing the Matthews correlation coefficient [43].

Random Forests: Random Forest classifiers were constructed again using the H2O.ai platform.

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Each classifier consisted by number of trees equal to half the number of features utilized for training [44]. Class balancing and 5-fold cross validation was used. The exact parameters can be found in 'script_h2o_deep_learning_predictor_with_features_random_sampling_GE_NUM.R'. For Random Forests we utilized exactly the same bagging scheme as for the DLNNs. The classification performance was measured as described in the previous section (Supplementary Table 4 – 'Random Forests').

We did not use any further optimisation steps such as forward selection or backward elimination for model optimisation and hyper-parameter optimization to avoid over-fitting and hence report unbiased results.

4. Drug-clustering

Clustering were based upon the genes involved in the sensitivity or resistance state of each drug as extracted from the 1-way rules of the aforementioned Apriori algorithm (Supplementary Table 5 - Genes). The top 100 rules ranked by support and top 100 rules ranked for Lift for each drug and response, were combined. These rules were then converted into a binary matrix (data/Binary Matrix.RData @ GitHub) where 1 denotes the presence of a rule and 0 denotes the absence (Binary Matrix.RData). From this matrix a dissimilarity matrix was calculated using the vegdist function and the Jaccard index [45] from the R package 'vegan'[46]. Hierarchical cluster analysis was then performed using the hclust function from the R package stats using the average clustering method[40]. The resulting cluster denrodrograms are displayed in a circular format using D3: Data-Driven Documents (Bostock et al, 2011) [47]. The html versions of the three

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

dendrograms can be accessed thourgh GitHub '/Figures/Dendrograms'.

Monte-Carlo Simulation: Drug Resistance – Sensitivity gene-set overlap: For every drug and each drug state (Sensitive or Resistant) the number of genes participating in statistically significant association rules was measured (Supplementary Table 5 - Genes). The gene-set overlap for all the combinations of the drug/drug-state pairs was also measured (Supplementary Table 5 – 'Genes overlap'). The probability that an observed overlap between two drug/drug-state pairs was due to chance alone was evaluated with 100 rounds of Monte-Carlo simulation. More particularly, for each round a number equal to the number of genes participating in the statistically significant association rules for each drug/drug-state was randomly sampled from the total pool of genes participating in all the significant association rules (18216 genes). The random sampling was weighted by the frequency of occurrence of each gene in the sum of the significant association rules; hence a gene participating in numerous rules will have a greater probability of being picked in comparison to a gene participating in just a few rules. At the end of each round, the overlap between the randomly sampled gene-sets of the particular drug/drug-state pair under examination was recorded. At the end of the 100-round Monte-Carlo simulation, the distribution of the 100 measured overlaps (which was found to be normal by the Kolmogorov-Smirnov test for normality) was utilized to calculate the p-value of the actual overlap between the particular drug/drug-state pair (Supplementary Table 5 – 'p-values'). This p-value represents the probability of the actual overlap to belong to the distribution of the randomly generated overlaps; hence the actual overlap being due to chance alone. If the actual overlap is located at the far right side of the random distribution the overlap is characterized as over-represented and statistically significant; hence non-randomly relevant. In contrast, if the

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

actual overlap is located at the far left side of the random distribution the overlap is characterized as under-represented and statistically significant; hence non-randomly distant (scripts/ script_measure_gene_overlaps_of_drugStates.R @ GitHub).

5. Data Availability

All scripts, data objects, figures and tables have been deposited and can be accessed at the public GitHub repository (folder: 'Vougas_DeepLearning') (https://github.com/kvougas/Vougas_DeepLearning)

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Halazonetis, T.D., Gorgoulis, V.G. & Bartek, J. An oncogene-induced DNA damage model for cancer development. *Science* **319**, 1352-1355 (2008).
2. Galanos, P. et al. Chronic p53-independent p21 expression causes genomic instability by deregulating replication licensing. *Nature cell biology* **18**, 777-789 (2016).
3. van't Veer, L.J. & Bernards, R. Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature* **452**, 564-570 (2008).
4. Weinstein, J.N. Drug discovery: Cell lines battle cancer. *Nature* **483**, 544-545 (2012).
5. Sharma, S.V., Haber, D.A. & Settleman, J. Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. *Nature reviews. Cancer* **10**, 241-253 (2010).
6. Iorio, F. et al. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* (2016).
7. Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-607 (2012).
8. Garnett, M.J. et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570-575 (2012).
9. Shoemaker, R.H. The NCI60 human tumour cell line anticancer drug screen. *Nature reviews. Cancer* **6**, 813-823 (2006).
10. Schmidhuber, J. Deep learning in neural networks: an overview. *Neural networks : the official journal of the International Neural Network Society* **61**, 85-117 (2015).
11. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-444 (2015).
12. Wang, C. et al. in 2014 IEEE International Conference on Bioinformatics and

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Biomedicine (BIBM) (2014).

13. Xu, Y. et al. Deep Learning for Drug-Induced Liver Injury. *Journal of chemical information and modeling* **55**, 2085-2093 (2015).

14. Aliper, A. et al. Deep Learning Applications for Predicting Pharmacological Properties of Drugs and Drug Repurposing Using Transcriptomic Data. *Molecular pharmaceutics* **13**, 2524-2530 (2016).

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

16. Forbes, S.A. et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic acids research* **43**, D805-811 (2015).

17. Costello, J.C. et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature biotechnology* **32**, 1202-1212 (2014).

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

19. Agrawal, R., Imielinski, T. & Swami, A. Mining association rules between sets of items in large databases. *ACM SIGMOD* **22**, 216 (1993).

20. Kelland, L.R., Sharp, S.Y., Rogers, P.M., Myers, T.G. & Workman, P. 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 (1999).

21. Muller, C.R. et al. Potential for treatment of liposarcomas with the MDM2 antagonist Nutlin-3A. *International journal of cancer* **121**, 199-205 (2007).

22. Long, G.V. et al. Combined BRAF and MEK inhibition versus BRAF inhibition alone in

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

melanoma. *The New England journal of medicine* **371**, 1877-1888 (2014).

23. Pietarinen, P.O. et al. Novel drug candidates for blast phase chronic myeloid leukemia from high-throughput drug sensitivity and resistance testing. *Blood cancer journal* **5**, e309 (2015).

24. Simioni, C. et al. Cytotoxic activity of the novel Akt inhibitor, MK-2206, in T-cell acute lymphoblastic leukemia. *Leukemia* **26**, 2336-2342 (2012).

25. Dail, M. et al. Loss of oncogenic Notch1 with resistance to a PI3K inhibitor in T-cell leukaemia. *Nature* **513**, 512-516 (2014).

26. Faber, A.C. et al. Assessment of ABT-263 activity across a cancer cell line collection leads to a potent combination therapy for small-cell lung cancer. *Proceedings of the National Academy of Sciences of the United States of America* **112**, E1288-1296 (2015).

27. Grazia, G., Penna, I., Perotti, V., Anichini, A. & Tassi, E. Towards combinatorial targeted therapy in melanoma: from pre-clinical evidence to clinical application (review). *International journal of oncology* **45**, 929-949 (2014).

28. Kiaris, H., Schally, A.V. & Varga, J.L. 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 (2000).

29. Perez, R. et al. 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 (2014).

30. Yang, Y.A., Zhang, G.M., Feigenbaum, L. & Zhang, Y.E. Smad3 reduces susceptibility to hepatocarcinoma by sensitizing hepatocytes to apoptosis through downregulation of Bcl-

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2. *Cancer cell* **9**, 445-457 (2006).

31. Breiman, L. Bagging Predictors. *Machine Learning* **24**, 123-140 (1996).

32. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society* **67**, 320 (2005).

33. Chen, X. & Ishwaran, H. Random forests for genomic data analysis. *Genomics* **99**, 323-329 (2012).

34. Metz, C.E. Basic principles of ROC analysis. *Seminars in nuclear medicine* **8**, 283-298 (1978).

35. Chen, D., Liu, X., Yang, Y., Yang, H. & Lu, P. Systematic synergy modeling: understanding drug synergy from a systems biology perspective. *BMC systems biology* **9**, 56 (2015).

36. Burotto, M., Chiou, V.L., Lee, J.M. & Kohn, E.C. The MAPK pathway across different malignancies: a new perspective. *Cancer* **120**, 3446-3456 (2014).

37. Zhu, S., Cohen, M.B., Bjorge, J.D., Mier, J.W. & Cho, D.C. PI3K inhibition potentiates Bcl-2-dependent apoptosis in renal carcinoma cells. *Journal of cellular and molecular medicine* **17**, 377-385 (2013).

38. Das, V. et al. Pathophysiologically relevant in vitro tumor models for drug screening. *Drug discovery today* **20**, 848-855 (2015).

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

40. Team, R.C. R: A language and environment for statistical computing. *R Foundation for Statistical Computing* (2016).

41. Hahsler, M. & Chelluboina, S. Visualizing Association Rules in Hierarchical Groups.

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

42nd Symposium on the Interface: Statistical, Machine Learning, and Visualization Algorithms (Interface 2011) (2011).

42. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **21**, 3940-3941 (2005).

43. Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et biophysica acta* **405**, 442-451 (1975).

44. Oshiro, T.M., Perez, P.S. & Baranauskas, J.A. in Machine Learning and Data Mining in Pattern Recognition: 8th International Conference, MLDM 2012, Berlin, Germany, July 13-20, 2012. Proceedings. (ed. P. Perner) 154-168 (Springer Berlin Heidelberg, Berlin, Heidelberg; 2012).

45. Jaccard, P. The distribution of the flora in the alpine zone. *New Phytologist* **11**, 37-50 (1912).

46. Oksanen, J. et al. vegan: Community Ecology Package. (2016).

47. Bostock, M., Ogievetsky, V. & Heer, J. IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis), 2011. (2011).

Author Contributions: K.V: study conception and design, scripting, bioinformatic analysis, results interpretation manuscript preparation and writing, T.J and A.P: scripting, data analysis and assistance in manuscript preparation, M.L: assistance in data interpretation, E.J: guidance and assistance in manuscript preparation, P.T: data interpretation and guidance and V.G: study design, data analysis and interpretation, guidance and assistance in manuscript preparation. All authors discussed the results and commented on the manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Competing financial interests

The authors declare no competing financial interests.

Figure 1

bioRxiv preprint doi: https://doi.org/10.1101/070450 This version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

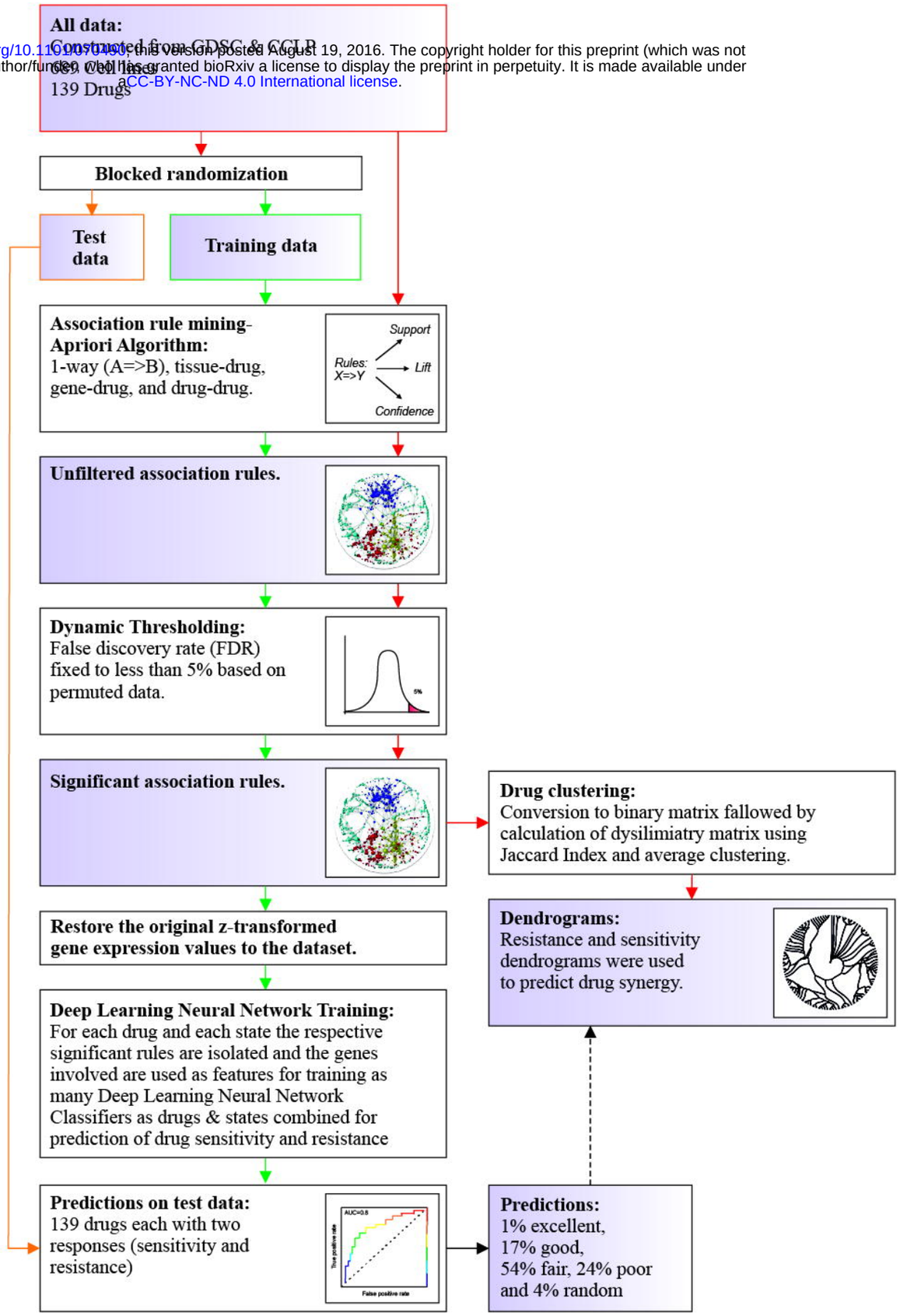
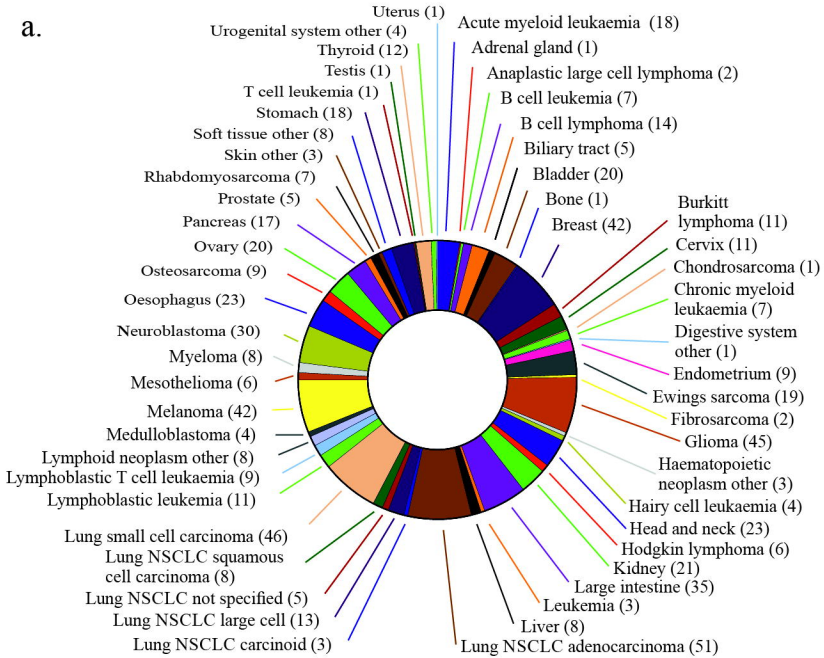


Figure 2

a.



b.

	Tissue of origin	Gene mutation status	Gene copy number variation	Gene expression	Drug response
Source	GDSC	CCLP	CCLP	CCLP	GDSP
Features	54 Tissues of origin	19426 genes	3744 genes	16444 genes	139 drugs
Level	NA	'Mut'	'Loss' or 'Gain'	"Over" or "Under"	'Sensitive' or 'Resistant'

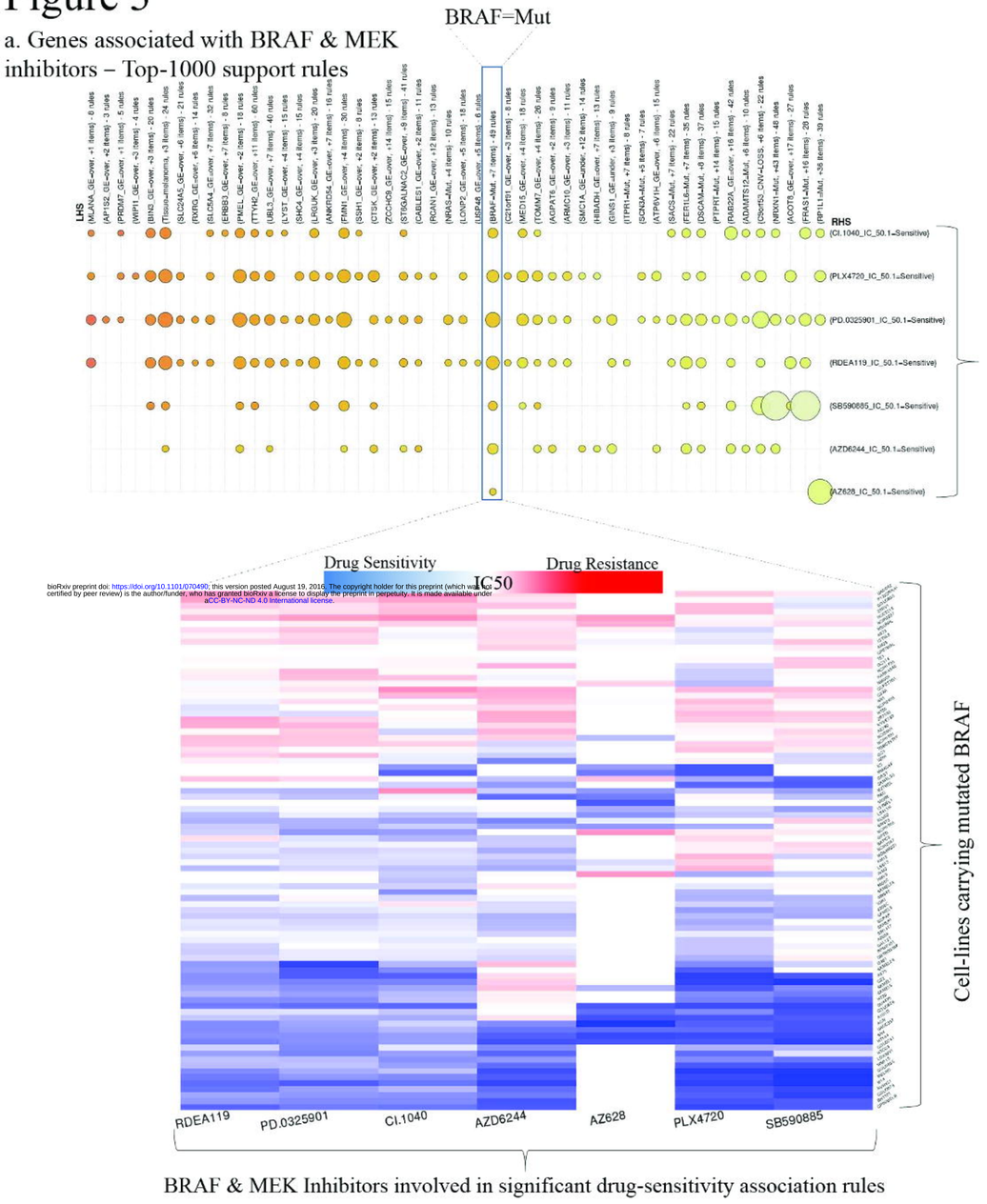
c.

689 cell lines

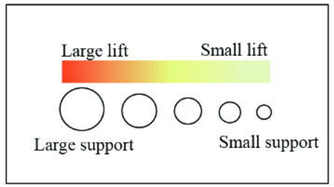
Tissue of origin	Gene mutation status (Gene 1 – Gene 19426)	Gene copy number variation (Gene 1 – Gene 3744)	Gene expression (Gene 1 – Gene 1166444)	Drug response (1-139)
Breast	Normal	Normal	Over	Normal
Cervix	Normal	Normal	Under	Normal
•••	•••	•••	•••	•••
Liver	Normal	Loss	Normal	Normal
Bladder	Normal	Normal	Normal	Sensitive

Figure 3

a. Genes associated with BRAF & MEK inhibitors – Top-1000 support rules

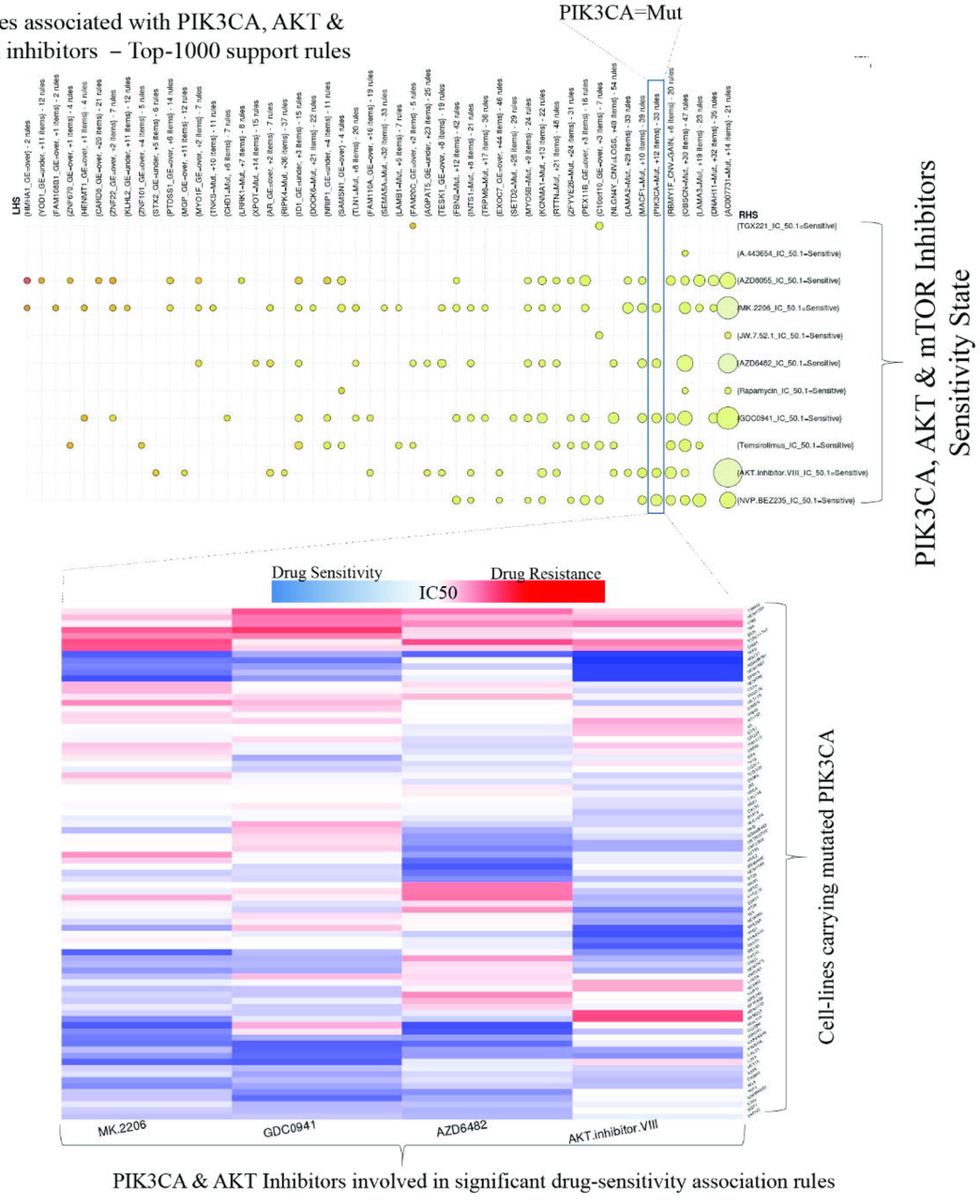


bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



BRAF & MEK Inhibitors Sensitivity State

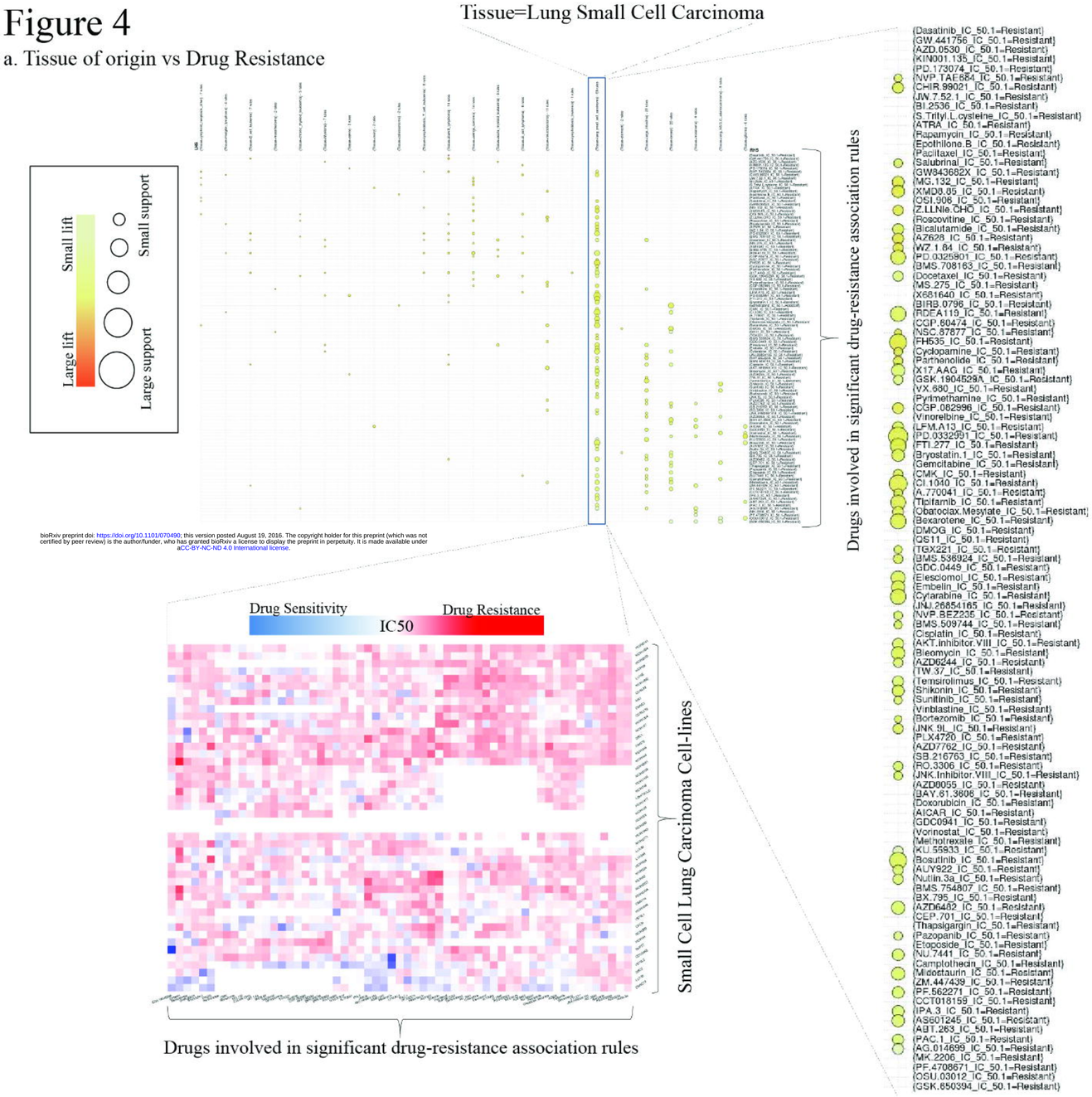
b. Genes associated with PIK3CA, AKT & mTOR inhibitors – Top-1000 support rules



PIK3CA, AKT & mTOR Inhibitors Sensitivity State

Figure 4

a. Tissue of origin vs Drug Resistance



bioRxiv preprint doi: https://doi.org/10.1101/070490; this version posted August 19, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

b. Tissue of origin vs Drug Sensitivity

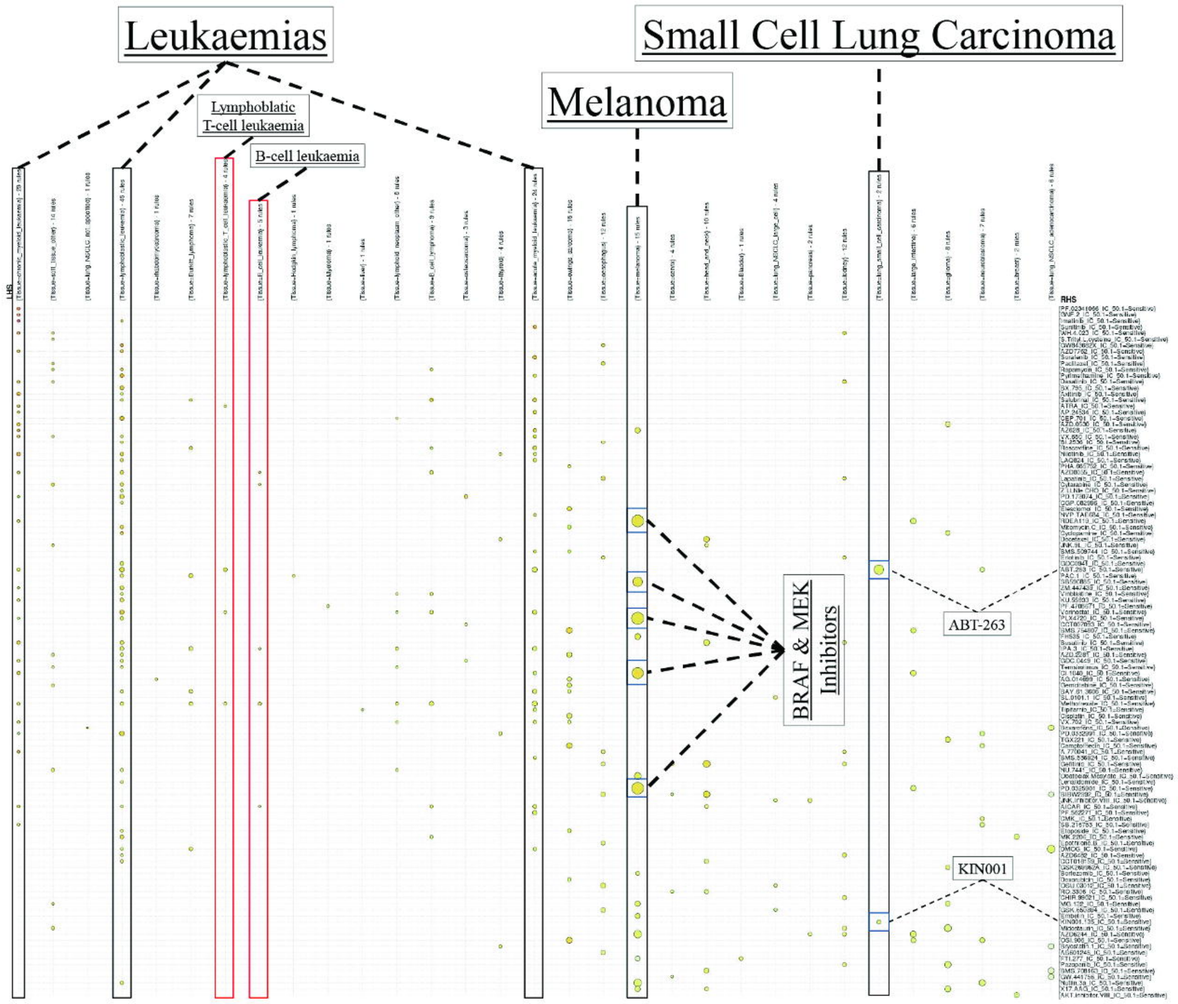
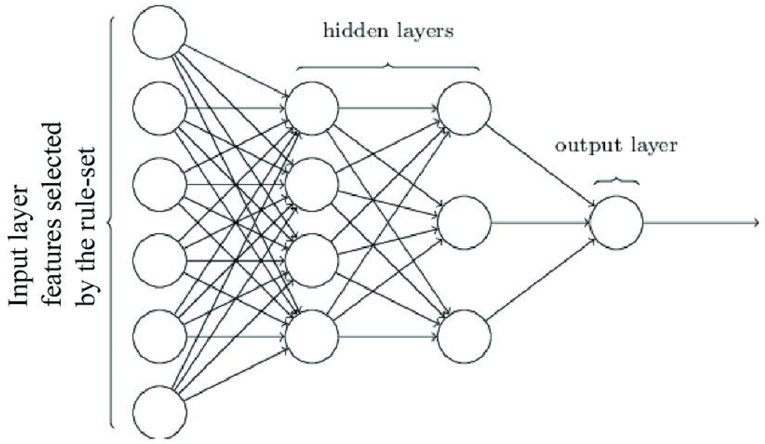


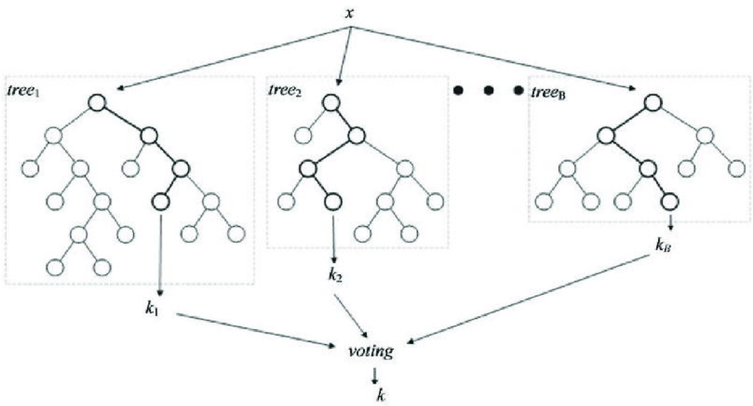
Figure 5

a

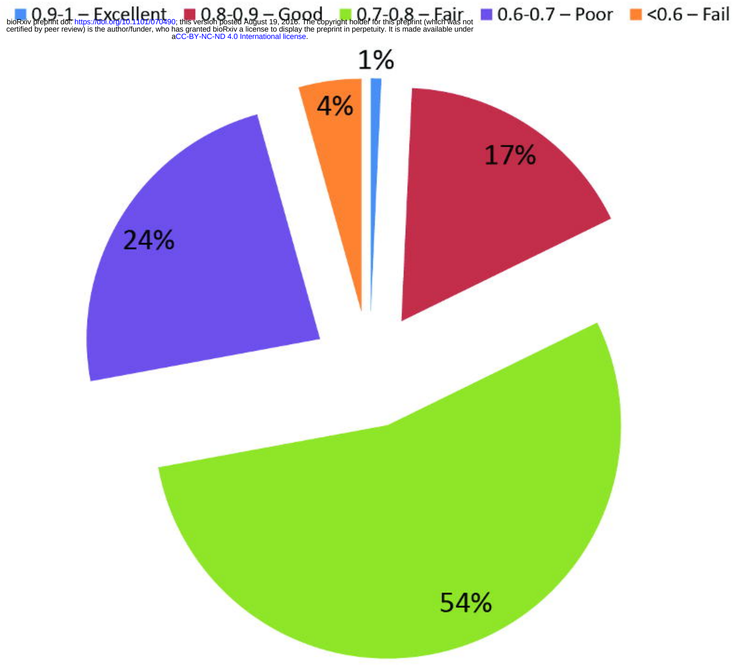
Deep Learning



Random Forests

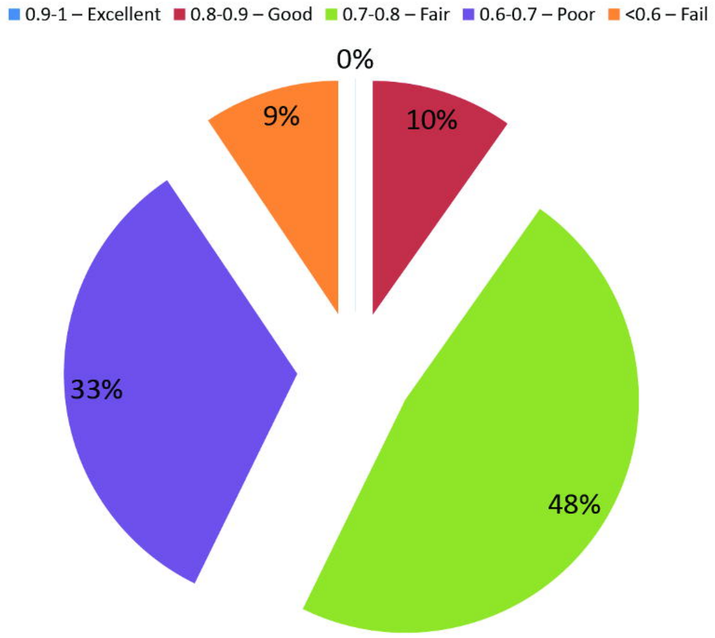


ROC-Analysis AUC



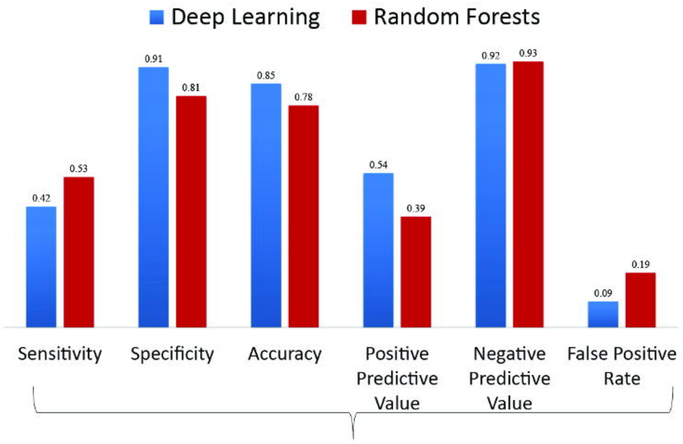
bioRxiv preprint doi: https://doi.org/10.1101/671990; this version posted August 19, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ROC-Analysis AUC



b

Deep Learning vs Random Forests



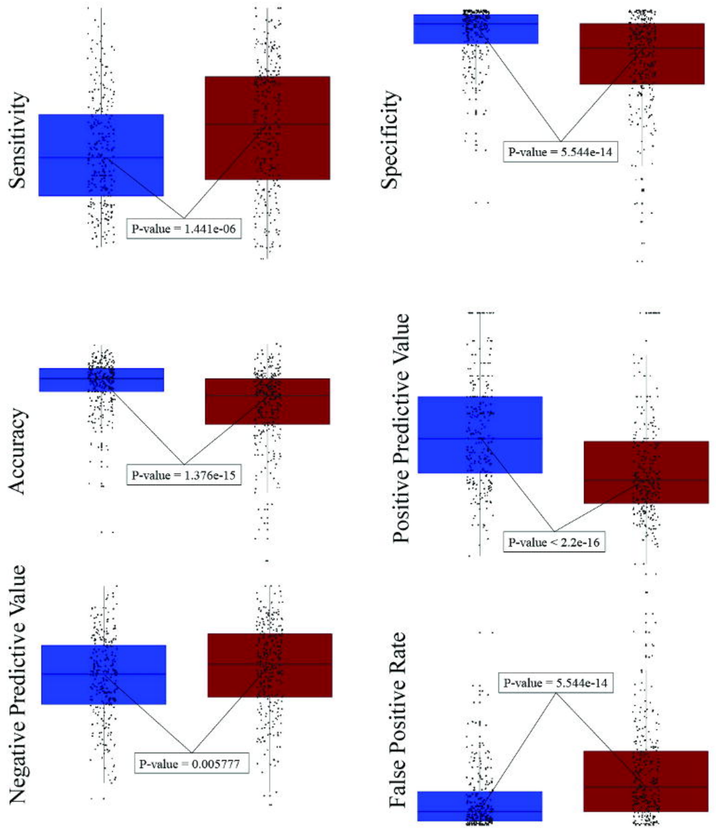
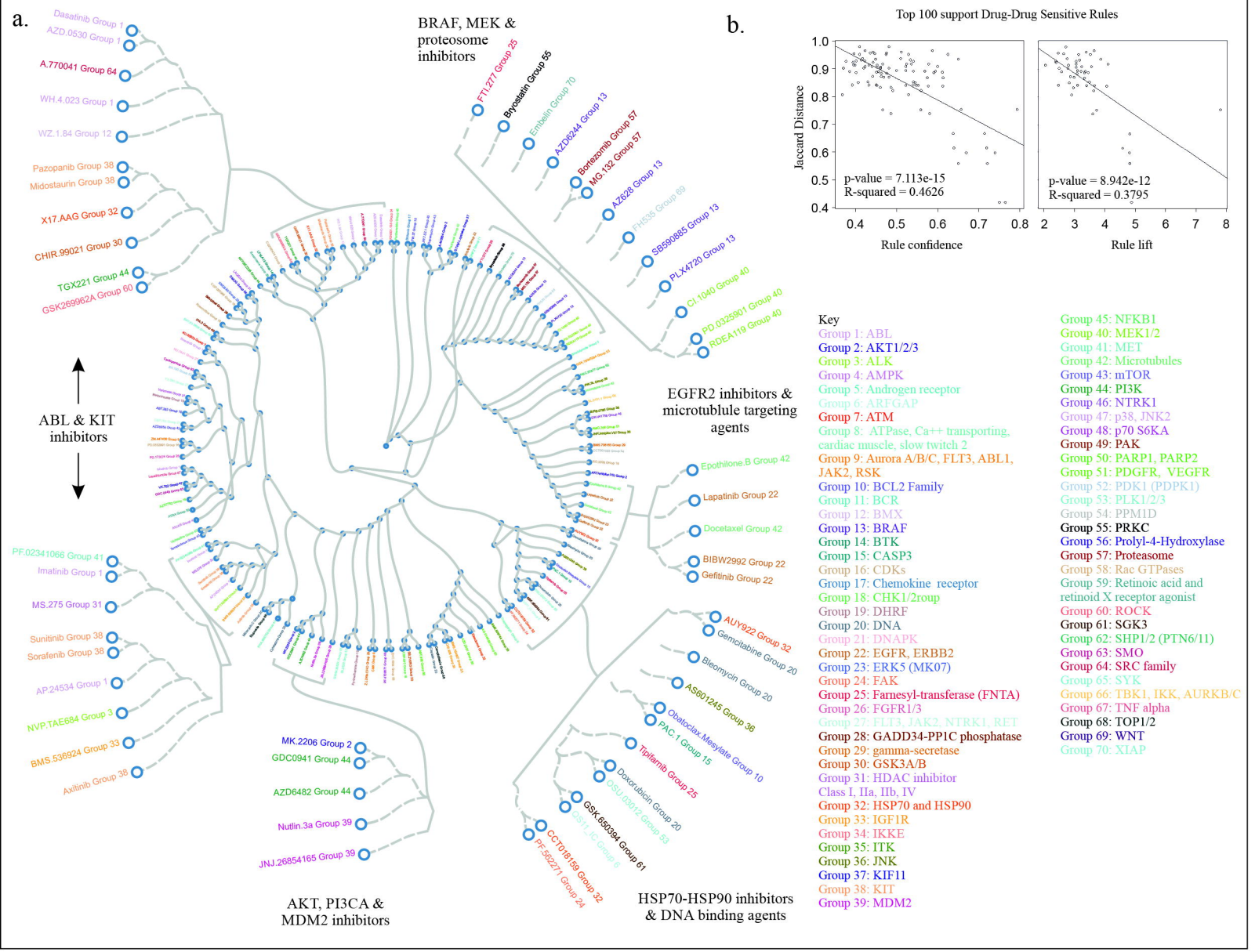


Figure 6


a.
Dasatinib Group 1
AZD.0530 Group 1
A.770041 Group 64
WH.4.023 Group 1
WZ.1.84 Group 12
Pazopanib Group 38
Midostaurin Group 38
X17.AAG Group 32
CHIR.99021 Group 30
TGX221 Group 44
GSK269962A Group 60
BRAF, MEK & proteosome inhibitors
FTI.277 Group 25
Bryostatin Group 55
Emtansin Group 70
AZD6244 Group 13
Bortezomib Group 57
MG.132 Group 57
AZ628 Group 13
FH535 Group 69
SB590885 Group 13
PLX4720 Group 13
CI.1040 Group 40
PD.0325901 Group 40
RDEA119 Group 40
ABL & KIT inhibitors
EGFR2 inhibitors & microtublule targeting agents
Epothilone.B Group 42
Lapatinib Group 22
Docetaxel Group 42
BIBW2992 Group 22
Gefitinib Group 22
PF.02341066 Group 41
Imatinib Group 1
MS.275 Group 31
Sunitinib Group 38
Sorafenib Group 38
AP.24534 Group 1
NVP.TAE684 Group 3
BMS.536924 Group 33
Axitinib Group 38
AUY922 Group 32
Gemcitabine Group 20
Bleomycin Group 20
AS601245 Group 36
Obatoclax.Mesylate Group 10
PAC.1 Group 15
Tipifarnib Group 25
Doxorubicin Group 20
OSU.03012 Group 53
GSK.650394 Group 61
CS1.1C Group 6
CCT018159 Group 32
PF.562271 Group 24
MK.2206 Group 2
GDC0941 Group 44
AZD6482 Group 44
Nutlin.3a Group 39
JNJ.26854165 Group 39
AKT, PI3CA & MDM2 inhibitors
HSP70-HSP90 inhibitors & DNA binding agents
b.
Top 100 support Drug-Drug Sensitive Rules
Jaccard Distance
1.0
0.9
0.8
0.7
0.6
0.5
0.4
p-value = 7.113e-15
R-squared = 0.4626
0.4
0.5
0.6
0.7
0.8
Rule confidence
p-value = 8.942e-12
R-squared = 0.3795
2
3
4
5
6
7
8
Rule lift
Key
Group 1: ABL
Group 2: AKT1/2/3
Group 3: ALK
Group 4: AMPK
Group 5: Androgen receptor
Group 6: ARFGAP
Group 7: ATM
Group 8: ATPase, Ca++ transporting, cardiac muscle, slow twitch 2
Group 9: Aurora A/B/C, FLT3, ABL1, JAK2, RSK
Group 10: BCL2 Family
Group 11: BCR
Group 12: BMX
Group 13: BRAF
Group 14: BTK
Group 15: CASP3
Group 16: CDKs
Group 17: Chemokine receptor
Group 18: CHK1/2roup
Group 19: DHRF
Group 20: DNA
Group 21: DNAPK
Group 22: EGFR, ERBB2
Group 23: ERK5 (MK07)
Group 24: FAK
Group 25: Farnesyl-transferase (FNTA)
Group 26: FGFR1/3
Group 27: FLT3, JAK2, NTRK1, RET
Group 28: GADD34-PP1C phosphatase
Group 29: gamma-secretase
Group 30: GSK3A/B
Group 31: HDAC inhibitor
Class I, IIa, IIb, IV
Group 32: HSP70 and HSP90
Group 33: IGF1R
Group 34: IKKE
Group 35: ITK
Group 36: JNK
Group 37: KIF11
Group 38: KIT
Group 39: MDM2
Group 45: NFKB1
Group 40: MEK1/2
Group 41: MET
Group 42: Microtubules
Group 43: mTOR
Group 44: PI3K
Group 46: NTRK1
Group 47: p38, JNK2
Group 48: p70 S6KA
Group 49: PAK
Group 50: PARP1, PARP2
Group 51: PDGFR, VEGFR
Group 52: PDK1 (PDPK1)
Group 53: PLK1/2/3
Group 54: PPM1D
Group 55: PRKC
Group 56: Prolyl-4-Hydroxylase
Group 57: Proteasome
Group 58: Rac GTPases
Group 59: Retinoic acid and retinoid X receptor agonist
Group 60: ROCK
Group 61: SGK3
Group 62: SHP1/2 (PTN6/11)
Group 63: SMO
Group 64: SRC family
Group 65: SYK
Group 66: TBK1, IKK, AURKB/C
Group 67: TNF alpha
Group 68: TOP1/2
Group 69: WNT
Group 70: XIAP