Model ensembling as a tool to form interpretable multi-omic predictors of cancer pharmacosensitivity

Sébastien De Landtsheer1, Apurva Badkas1, Dagmar Kulms2, Thomas Sauter1\*

1: Department of Life Sciences and Medicine, University of Luxembourg, Belvaux, Luxembourg

2: Dresden affiliation XXX

\*: Corresponding author: [thomas.sauter@uni.lu](mailto:thomas.sauter@uni.lu)

Abstract

The determination of the optimal treatment for individual patients diagnosed with cancer is the major goal of personalized oncology, a rapidly growing field of modern medicine. One important aspect is the accurate prediction of the response of cancer cells to various chemotherapies, as patients frequently fail to respond adequately to first-line therapies, due to the inherent variability of cancers in a given tumor type. It is expected that the molecular characteristics of the cancer cells (genomic, transcriptomic, etc.) contain enough information to retrieve specific signatures, allowing to form accurate predictions based solely on these multi-omic data. Ideally, these predictions should be explainable to clinicians, in order to be integrated in the patients care. While a number of computational methods have been developed over the years, very few have been assessed in a clinical setting, and none has been integrated in standard cancer care. We propose a machine-learning framework based on ensemble learning to integrate multi-omic data and predict chemosensitivity to an array of commonly used compounds. We trained a set of classifiers on the different parts of our dataset to produce omic-specific features, and subsequently trained a random-forest classifier on these features to predict chemosensitivity. We used the CCLE dataset, comprising multi-omic and chemosensitivity measurements for hundreds of cell lines, to build the models, and we validated our results using a cross-validation strategy. Our results show superior performance to the state-of-the-art for several compounds (ROCAUC > 80% for 7 out of 23 compounds), belonging to different compound classes, and across the most frequent cancer types. Furthermore, the relative simplicity of our approach allows us to easily examine which features have a larger importance in the models and identify new putative markers of chemosensitivity. We propose several interpretable models based on small subsets of [transcriptional] markers with the potential to be a useful tool in a personalized oncology setting, by helping clinicians to link the characteristics of the tumors to their sensitivity to chemotherapeutic compounds, and ultimately to the clinical response in patients.

1. Introduction

Despite major breakthroughs over the past few decades, cancer remains the second leading cause of death worldwide [1]. One of the many reasons for this is the difficulty for clinicians to prescribe the right treatment for each patient. Cancer is the result of uncontrolled growth of a single, genetically damaged cell within the body. In some conditions, this clonal expansion invades other organs and impairs their normal functioning, eventually causing death. Historically, tumors have been characterized by their histological and clinical presentation, as well as their tissue of origin. However, cancer is highly heterogeneous, even within a certain tissue, and with similar clinical features. This degree of heterogeneity is itself highly variable: a number of hematological malignancies are defined by precise chromosomal alterations, for example the reciprocal translocation t(9;22)(q34;q11) resulting in the chimeric BCR-ABL protein in virtually all cases of chronic myeloid leukemia [2], while many different driver mutations are implicated in the most common tumor types, especially melanomas [3] and lung adenocarcinomas [4]. This high heterogeneity is even more reinforced by the fact that many tumors present defects in the DNA-repair mechanisms, thereby further increasing their mutational load. The cancer hallmarks [5], a set of phenotypic capabilities shared by all tumors and central to their emergence and evolution towards malignancy, have been shown to be highly polygenic, while the main cancer genes are pleiotropic [6], and are found to be mutated across tumor types (for example TP53, lost or inactivated in many cancers [7] or BRAF-activating mutations, found across melanomas, colon adenocarcinomas, gliomas, and many others[8]). While dozens of chemotherapeutic compounds, cytotoxic or targeted, have been approved over the past decades, every single one will only be efficacious in a subset of cancer patients. Following the identification of molecular targets and their functions, targeted therapies were designed to interfere with a specific protein, either via a small molecule (typically a tyrosine kinase inhibitor) or an antibody [9], [10]. Accordingly, patients diagnosed with cancer are stratified by the presence or absence of specific oncogenic mutations, or overexpression of specific proteins upon immunohistochemical observation, in addition to tissue of origin of the tumor and type-specific staging. However, this strategy has a limited effectiveness, as a number of patients do not adequately respond to the treatment they are prescribed [11], and sometimes no actionable mutation can be found [12]. It is therefore necessary to expand the arsenal of stratification tools available to clinicians to better identify the drugs or combinations to which their patients are most likely to respond, both to increase treatment efficacy and decrease unnecessary side-effects. Ultimately, the goal of personalized oncology is to be able to treat each cancer patient based on the unique array of characteristics of their tumors, and in the context of their germline genomes and clinical histories.

The classic theory of carcinogenesis postulates that mutations accumulate over time in tissue-specific stem cells until a tumor arises [13]. While it is evident that the gene products, and not the genes themselves, are responsible for the effects of mutations, oncology has relied mostly on genomic information to assess tumors, ignoring the fact the tumor cells’ phenotypes manifest via systems-level perturbations, notably of cell cycle progression, apoptosis, cell growth, and metabolism. Therefore, while cancer is fundamentally a disease of the genome, it manifests mostly within the multiple layers of the cellular regulatory mechanisms, for example signal transduction. For this reason, large-scale screenings have been performed to characterize panels of cell lines across multiple omics levels, together with measurements of chemosensitivity. The Cancer Cell Line Encyclopedia (CCLE) dataset [14] is one of the best-known of these screening efforts, contains data for more than a thousand cell lines of various cancer types and subtypes as well as high-quality multi-omics and pharmacological characterization, and has been shown to enable predictive modeling of chemosensitivity.

Key points in the application of artificial intelligence to precision oncology have been highlighted elsewhere in excellent reviews [15]–[17]. The NCI-60 cell line panel pioneered the use of a large screening to discover characteristics of cell lines indicative of chemosensitivity [18]. Modeling was first applied to the problem of predicting cell line chemosensitivity by Staunton et al. [19], originally a simple weighted voting scheme. Later, a genetic signature based on the expression of 70 marker genes was used to predict the clinical outcome of breast cancer patients [20]. Mathematical modeling was then extended to various frameworks, notably the use of kernel methods [21], regularized linear regressions such as the Elastic Net or the LASSO [22], regression and classification trees [23], matrix factorization [24], then to various neural-networks-based algorithms like Deep Learning [25]–[27] and Graph Convolutional Networks [28], [29]. A number of studies include the chemical structure of compounds as a component of their models [30], [31]. Notably, the NCI-DREAM challenge [32], which compared the predictions of 44 teams, concluded that differences in performance between the algorithms can mostly be attributed to data quality, preprocessing strategies, and choice of the reported variable, rather than the family of the method used. It also clarified that predictions based on the combinations of individual teams’ algorithms always outcompeted the best of the individual methods, showing that different methods provide complementary information.

Stacking [33] is an ensemble learning technique that consists in first training a series of classifiers (or regressors, in the general case) on labeled training data, then training a second-level generalizer which task is to learn the biases of the individual classifiers with respect to the true labels of the training set. Stacked ensembles have been shown to lower predictor bias and, in any case, produce results that are no worse than the best individual model [34].

In this paper, we form the hypothesis that while each individual omic type only contains partial signal, it is possible to combine the imperfect information gathered from each biological layer into an integrated picture of the particular tumor, and deduce the chemo-resistant or chemo-sensitive profile. Importantly, we assume that while heterogeneity between patients, and therefore between cell lines, is large, parallels can be drawn, given a large enough sample size, allowing to learn robust correlations between molecular and functional states.

1. Methods
   1. Data source

CCLE data files were downloaded directly from the DepMap portal. For transcriptomics, we used the provided file *CCLE\_RNAseq\_genes\_rpkm\_20180929.gct* containing 56202 transcripts. We did not aggregate the data at the gene level to allow for discovery of splice variants associated with functional response. For genomics, we used the file *CCLE\_MUT\_CNA\_AMP\_DEL\_binary\_Revealer.csv* summarizing the presence or absence of specific genetic features for all cell lines as a Boolean table. Because nearly half of these features were duplicates (the same mutations are present in a set of cell lines) we did not apply the cross-correlation filter (see below) on this dataset, which could have removed driver mutations if passenger mutations were, by chance, perfectly correlated. For the miRNA, we used the file *CCLE\_miRNA\_20181103.csv* containing fpkm values for 974 miRNAs. The metabolomics data consisted of profiles for 225 metabolites, determined by Liquid Chromatography Mass Spectrometry (LS-MS) in the file *CCLE\_metabolomics\_20190502.csv*. For the proteomics data, we used the file *CCLE\_RPPA\_20181003.csv* consisting of Reverse-Phase Protein Array (RPPA) measurements of 214 proteins or protein modifications. In addition, we included the estimates of pathway activity found in the file *1-s2.0-S0092867416307462-mmc6.xlsx* from the GDSC study [35] for the samples included in both studies. These pathway activities were computed from gene expression using the algorithm SPEED [36].

* 1. Preprocessing

Table 1 describes the filtering steps that were applied to each dataset. Firstly, quantitative data was log-transformed and normalized to the [0, 1] interval to facilitate modeling. We avoided the need for data imputation by removing samples and features with missing data. Then, we applied a simple feature selection scheme, by first removing a proportion of features showing low variance across samples, then removing features showing high correlation with other features. We extracted cancer type (tissue of origin) for each sample from the samples’ names. The pre-processed dataset used in further steps contained a total of 324 samples from 23 different cancer types, and 48453 features. Drug response information, in the form of the ActArea (normalized area over the drug-response curve, a proxy for cell line sensitivity which takes partial response into account, in contrast with the IC50) was collected for the 23 compounds (Sorafenib was removed as data for this drug contained missing values) and quantized into three categories: resistant (33% of cell lines with the smallest ActArea), sensitive (33% of cell lines with the largest ActArea) and intermediate (34% of cell lines). This latter stratum was excluded from subsequent modeling steps, to exaggerate the differences between resistant and sensitive cell lines. While this drug-agnostic labeling might not be the most appropriate for all compounds, might result in some degree of mislabeling and not accurately reflect the levels of chemosensitivity of cell lines in a clinical context, it has the advantage of framing the study as a simple binary classification problem on a balanced dataset, thus avoiding the need for multi-class models, over/undersampling and data augmentation, which would possibly induce more serious biases on the methodology and the interpretation of the results.

(Suppl methods)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | Sample Completeness threshold | Feature Completeness threshold | Feature Variance threshold | Cross-correlation threshold | Number of features used for modeling |
| Transcriptomics | 0.9 | 0.9 | 0.5 | 0.75 | 23398 |
| Genomics | 0.99 | 0.99 | 0.5 | NA | 23135 |
| Proteomics | 0.9 | 0.9 | 0.2 | 0.9 | 165 |
| miRNomics | 0.95 | 0.95 | 0.5 | 0.75 | 631 |
| Metabolomics | 0.95 | 0.95 | 0.5 | 0.75 | 67 |
| Cell types | NA | NA | NA | NA | 23 |
| Pathways | 0.95 | 0.95 | 0 | 0.9 | 11 |
|  |  |  |  |  |  |

Table 1. Characteristics of the data pre-processing steps for each omic type. Sample completeness threshold: samples with a proportion of missing data across features higher than (1 - threshold) were removed. Feature completeness threshold: features with a proportion of missing data across samples higher than (1 - threshold) were removed. Feature variance threshold: variance was calculated for each feature independently across samples and a proportion (threshold) of the most variable features were removed. Cross-correlation threshold: pairwise Pearson correlation coefficients p were computed for all pairs of features within a dataset. For each pair with p > threshold, the feature with the lowest average correlation with the rest of the dataset was retained. NA: not applicable.

* 1. Stacking methodology

The following nested cross-validation procedure was used to build the classifiers for each drug. In the first step, the dataset was split into a “training” set (90% of samples) and a “test” set (10% of samples). The “training” set was then split further into a “training A” set (81% of samples) and “training B” (9% of samples). Then, First-level algorithms (see Methods) were trained independently on the “training A” set of samples, using in turn each one of the seven omic layers, to form a prediction of the probability of class membership (sensitive or resistant) of each sample. These trained models were then used to predict the class of the samples in the “training B” set. This procedure was repeated over 10 non-overlapping splits of the “training” set, producing quantitative predictions for each sample in the “training” set, as well as for the “test” set (using in that case algorithms trained on the whole “training” set). These probabilities of class memberships were then used to train a second-level Random Forest using the “training” predictions to form a combined prediction of class membership for the samples in the “test” set. This complete procedure was repeated 10 times in order to produce a prediction for every sample in the dataset while avoiding data leakage.

* 1. Algorithms

Random Forests is a learning method, first formalized by Breiman [37] in which multiple decision trees are constructed independently from bootstrapped instances of the dataset and a random selection of the original features. The individual trees are grown using the Gini impurity at each split, and the prediction of the random forest for a specific instance is the mode (for classification problems) or the average (for regression problems) of the predictions of every decision tree in the ensemble. We used forests of 100 trees, selecting for each one a fraction sqrt(F) of the F features.

AdaBoost is a learning method, related to random forests, in which an ensemble of weak learners (in our case decision stumps, which are decision trees of depth 1) is built sequentially in such a way that at each step, the subsequent learner is trained on the original dataset but the weights of correctly and incorrectly classified samples are modified to help the new learner focus more on the errors of the previous learners [38]. This ‘boosted’ learner improves the classification performance of the ensemble, and predictions for specific instances are computed as the mode or average of the predictions of the constituent learners. We used this algorithm with 50 decision stumps and a learning rate of 1.

XGBoost is also a learning method based on boosted learners, and includes multiple features, notably automated handling of missing data, stochastic and regularized gradient boosting, and overall better memory and CPU utilization, resulting in this algorithm being one of the best out-of-the-box tools for machine-learning. It has been introduced by Tianqi Chen [39]. We used XGBoost with a learning rate of 0.3, a maximum depth of the constituent trees of 6, no subsampling of samples or features, and values of 0 and 1 for the L1 and L2 regularization terms, respectively.

ExtraTrees is also a learning algorithm based on boosted learners, with random subsampling of features but without bootstrapping of the samples. In contrast with other boosting algorithms, it chooses the decision threshold at each split at random from the empirical distribution of each feature rather than the value that maximizes information gain [40]. We used this algorithm with 100 constituent trees and a fraction sqrt(F) of the F features for each tree.

Logistic regression is a widely used linear learning algorithm that models the logarithm of the odds of a binary variable as a linear combination of one or more independent predictors. This regressor can be used as a classifier by setting a threshold of probability for class assignment. We used logistic regression with a L2 regularization term of value 1 to limit overfitting.

Ridge regression is a variant of logistic regression which does not include a L2 regularization term, but only a L1 regularization term in order to induce sparsity in the choice of the independent predictors in the case of ill-posed problems, for example in case of multiple collinearity. We used a ridge classifier with a L1 term of 1, and a tolerance of 1e-3.

Elastic Net regression combines the L1 and L2 regularization terms in a unique optimization problem, thereby aiming for balance between accuracy of the prediction, and sparsity of the predictors as well as their coefficients. We used this classifier with equal weighting of the L1 and L2 penalties and a tolerance of 1e-3.

Support-Vector Machines are learning algorithms that map, explicitly or implicitly, a set of labeled data to a multi-dimensional space and find the hyperplane with the largest possible margins that separates best the data, as assessed by the Euclidian distance of the closest datapoints to the hyperplane. Such a maximum-margin classifier is guaranteed to minimize the generalization error of the classifier. We used a linear kernel with a L2 regularization value of 1, parameter shrinkage [41], and a tolerance of 1e-3.

k-Nearest-Neighbor is a learning algorithm where class predictions are formed locally from the class membership of a number k of similar labeled examples [42]. It has the appreciable properties of having a single parameter (k) and not necessitating any training. We used this algorithm with a fixed k of 5 and used Euclidian distance (Minkowski metric with p=2).

In our experiments, we used the AUROC (Area Under the Receiver-Operating Curve) as the optimized metric by all classifiers.

* 1. Data availability

All data associated with this publication, as well as necessary Python code to reproduce the results, is available at the following address: https://github.com/sysbiolux/DeepOncoAI

1. Results
   1. Overall analysis

In this study, we sought to evaluate the performance of stacked classifiers (Random Forests) for the task of discriminating the most sensitive cell lines from the least sensitive in the CCLE database of drug response profiles, based on the predictions of first-level learners (both tree-based and regression-based), themselves trained independently on specific molecular features of the cell lines: genomics, transcriptomics including miRNomics, proteomics, metabolomics, as well as the tissue of origin and 11 pathway-level features. The complete pre-processed dataset comprised 48453 features for 324 cell lines.

We generated quantitative predictions by applying a two-step ten-fold nested cross-validation scheme and used them to compute the Area Under the Receiver Operating Characteristic curve (AUROC) for each drug-specific classifier. The results are presented for the 7 classifiers with AUROC > 0.75 in Figure 1. AUROC values for the remaining 16 classifiers ranged from 0.509 to 0.721. Suppl Fig S1 Shows the results for the complete set of 23 compounds.

A picture containing text, map, sky

Description automatically generated

Fig1: ROC curves for the 7 classifiers with the highest performance

Furthermore, we retrieved the feature importances from the classifiers, with the hypothesis that the predictive signal in each omic type might be best recovered by certain types of algorithms, but also drug-specific. We computed the average feature importance for each combination of omic and first-level classifier across the 10 folds. The results are presented in Figure 2. A clear separation between a branch containing the 7 compounds for which excellent results were obtained and the others can be observed. Also visible are three main branches of features: one containing 12 combinations of omic/algorithm with the highest contributions and enriched in transcriptomics datasets, another containing 14 combinations with very low contributions and grouping all combinations using the k-nearest neighbors and ridge regression algorithms, and a third one containing the remaining combinations with intermediate contributions. The same results are presented in Suppl Figure 2 in the form of individual bar plots for each compound.

Fig2. Clustergram of the average feature importance of the different combinations of omic types and predictive algorithms. The dendrograms were computed using the UPGMA algorithm and Euclidian distance. RPPA: proteomics; RNA: transcriptomics; DNA: genomics; MIRNA: micro-RNAs; TYPE: cell type of origin; META: metabolomics; PATHWAYS: SPEED pathway activities; RFC: Random Forest classifier; ET: Extra-Trees classifier; XGB: XGBoost classifier; Ada: AdaBoost classifier; EN: Elastic Net classifier; Ridge: Ridge regression classifier; KNN: k-nearest neighbors classifier.

Chart, timeline

Description automatically generated

Then, we retrieved the feature importances of the underlying first-level models, or the weights in the case of regression-based algorithms, and computed the average rank of each feature across the 100 sub-folds, separately for each compound. Figures SXXX show the distribution of ranks for the 30 features with the lowest average ranks each omic type, algorithm, and compound.

* 1. Type-specific analysis

Because the cell type of origin of a tumor is nearly always known, we sought to estimate the performance of the classifiers on specific cancer types, with the two caveats that, by subsampling our balanced dataset, we introduce a degree of imbalance in the sample, and that many of the 23 cancer types are represented only by a very low number of cell lines. We therefore report the balanced accuracy (BA) which is the average of specificity and sensitivity, by cell type and drug in Figure 4. The values are omitted when the total number of cell lines is inferior to 10.

Chart, treemap chart

Description automatically generated

Fig3: Heatmap showing the balanced accuracy of the drug-specific classifiers on the main cancer cell types

BA was found to be highly dependent of the compound and of the cell type of origin of the tumor. For example, in the case of PD-0325901 (Mirdametinib, an investigational MEK inhibitor [43]), high performance was achieved in the cases of colorectal cancer (BA=1.0 for 13 sensitive and 1 resistant cell line), lung adenocarcinoma (BA=0.91 for 17 sensitive and 25 resistant cell lines), and hematopoietic tumors (BA=0.94 for 16 sensitive and 13 resistant cell lines). In contrast, performance for skin cancers (melanomas) reached only a BA of 0.5 for 15 sensitive and 2 resistant cell lines. In the case of AZD6244 (Selumetinib, another MEK inhibitor approved for neurofibromatosis type I and pediatric neurofibromas [44]), the largest performance was found for breast tumors (BA=0.96 for 1 sensitive and 12 resistant cell lines), while performance was much more modest for other cancer types. Classifiers for Paclitaxel showed remarkable performance on ovarian cancer (BA=0.88 for 4 sensitive and 7 resistant cell lines), a cancer type for which this drug is often part of the first-line treatment [45], and melanoma (BA=0.82 for 5 sensitive and 10 resistant cell lines), although this latter cancer type is more rarely treated with cytotoxic compounds. Other notable large performances are the ones of two classifiers on pancreatic cell lines: ZD-6474 (Vandetanib [46], a VEGFR/EGFR inhibitor) scoring BA=0.83 for 5 sensitive and 9 resistant cell lines and Sorafenib [47], a large-spectrum kinase inhibitor (BA=0.96 for 6 sensitive and 9 resistant cell lines), the RAF/VEGFR2 inhibitor RAF265 [48] for ovarian cancer (BA=0.85 for 9 sensitive and 2 resistant cell lines), and the EGFR inhibitor Erlotinib [49] for breast cancer (BA=0.85 for 5 sensitive and 10 negative cell lines). Supplementary Table SXXX shows the performance of the classifiers for all drugs [long table 23\*types]

* 1. Predictive biomarkers

Our analysis of the importance of the individual features in the different omic-specific datasets indicated that many alterations, including expression of specific genes or phosphoproteins, was reliably utilized by the different first-level algorithms to build their predictions. Independently for each compound and each omic type, we ranked the features according to their importance, which we calculated either, for tree-based algorithms, as the proportion of internal nodes using this feature, or in the case of regression-based algorithms, as the absolute value of the coefficients. We collected these ranks over the 100 folds of the cross-validation scheme, and Figure XXX shows the distributions of these ranks for a selection of compound/omic/algorithm combinations. The complete results are available in Suppl Fig XXX.

For Panobinostat, the largest contributions were from the SVM and Logistic classifiers, trained on transcriptomics data (Fig XXX). These classifiers ranked the same four transcripts as the most informative features: AC138623.1 (ZNF141 pseudogene (ref here?)), AC011242.6 (a pseudogene transcribed from the reverse strand of the PLEKHH2 gene (ref here but genomeviewer, no paper?)), ZNF215 (ref here?), and SFMBT2 (ref). The same four features were also picked by the RFC algorithm.

The main contributing algorithms and dataset were the same for Paclitaxel, and both SVM and Logistic models pointed to a high importance of LEPREL2 () and MAGEA6 (), as well as SLFN11 () and RCOR2 ().

The main contributions for Irinotecan were the AdaBoost and Extra-Trees algorithms, trained on the transcriptomic data. These, as well as other algorithms trained on the same dataset, highlighted SLFN11 (), hnRNPA1 (), hnRNPCP1 (), DAAM1 ( (ref Mei2020+Aspenstrom2006)) as well as two pseudogenes: AC008427.2, also called MFFP2 () and RP11-177C12.1.

In the case of Lapatinib, the most contributing datasets were transcriptomics and proteomics, analyzed with SVC (or SVM) and Extra-Trees, respectively. GPX3 (), DYRK3 (), ADORA1 (), and STYL1 () were among the low-ranking transcriptomics features, while analysis of the proteomic features pointed to Claudin7 (), E-Cadherin (), and Rab25 (). Notably, most algorithms recovered either EGFR/HER1 or HER2 among their most important features.

Erlotinib appeared as the exception, in having the proteomics as the top-contributing dataset, paired with the Elastic-Net algorithm. The most important features in this case appear to be P-Cadherin (), EGFR, as well as Shc\_pY317 and RSK1-2-3.

For Mirdametinib (PD-0325901), the main contribution came from the transcriptomics dataset, through the Elastic Net algorithm. The features with the lowest average rank were ETV4 and ETV5 (), as well as SPRY2 () and TOR4A (). We also noted the presence of CMTM7 () among the features consistently ranked low by several algorithms.

In the case of Selumetinib (AZD6244), the main contribution was the Logistic algorithm, trained on the transcriptomics dataset. The most important feature in this dataset-algorithm pair, as well as in others, appears to be CMTM7 (), as well as ETV4, S100A4 (), SPRY2/4 and TRPV2 ().

Furthermore, among the top-predictors for the other 16 classifiers with inferior performance, we noticed that a number of genes in the transcriptomics datasets were consistently picked up by various algorithms and seemed to be correlated with response, for a variety of compounds. These genes are MAGEA6, NQO1 and LEPREL2, already mentioned, as well as FAM21B () and PTEN () for Sorafenib, HERC5 () and CHRNB1 () for RAF265, and SIAH3 () for AEW541. PLX4720 (a BRAF inhibitor related to vemurafenib) was the only compound for which the genomic information was the most informative. Unsurprisingly, the BRAFV600E mutation was consistently the feature with the lowest rank for this compound. In the case of PHA665752 and AZD5030, the main contribution to the final classifier were from the miRNA dataset and evidenced the low rank of several microRNAs: miR130a, let-7c, miR1307, miR425, miR222, miR223, and miR34a, among others. The classifiers for Nutlin-3 () relied mostly on the proteomics dataset and the Elastic Net or Logistic algorithms, and pointed to Bax, VAV1, Annexin1 and p21 as top features. In addition, predictions for Nilotinib () and PF2341066 (), of intermediate performance, relied mostly on the cell type, and valued the hematopoietic origin of the tumor cells as the most important factor to predict chemosensitivity.

Finally, we noticed that long non-coding RNAs frequently appeared among the top-50 features retrieved by most algorithms in the transcriptomics database. [More here]

* 1. Comparisons with single data types

We applied our modeling pipeline to the individual parts of the CCLE dataset, in an effort to compare the performance of stacked classifiers drawing from the complete dataset with the

* 1. External validation

Published datasets[50]:

Suppl Table2: across drugs best algo: perf per cancer type ok (prec-recall-sens-spec-auroc-mcc-f1)

1. Discussion

Here we describe an analysis pipeline, comprising an ensemble learner (Random Forests) trained on the predictions of a set of machine-learning algorithms, themselves trained separately on the various omic datasets of the CCLE database. We separated the cell lines, for each of 23 compounds, into three equal-sized categories: sensitive, intermediate, and resistant, and applied nested cross-validation to classify sensitive versus resistant cell lines. Our results indicate that for seven compounds (three cytotoxic: paclitaxel, irinotecan and panobinostat and four targeted: mirdametinib, selunetinib, erlotinib, lapatinib) we are able to predict the membership of cell lines to these two categories, across cancer types, with remarkable performance. Nevertheless, the performance of our classifiers varied with cell type:

{applicability}

[Argument of predict versus explain, Breimann, Kleinberg] The nature of this type of modeling is non-interventional.

. [word on why it does not work for some]

[results in line with Dong2015]

{limitations} When evaluating the predictive performance of our models, it is important to remember that a third of the cell lines (not necessarily the same across compounds) have been excluded from the dataset, as they displayed an intermediate level of chemosensitivity which could decrease the ability of our models to form accurate predictions on the more extreme phenotypes. Therefore, while it can be said that there exists a strong correlation between our predictions and the measured sensitivity of cell lines to the tested compounds, this ‘intermediate’ class of cell lines, likely to display a mix of molecular characteristics from both sensitive and resistant cells, cannot be classified using our method. Future works should focus on addressing this issue.

Several recent studies have concentrated on validating preclinical findings on patient databases, in an effort to increase the translational potential of chemosensitivity prediction algorithms (ref Hostallero2021, Prasse2022)

[most cell lines are white-race?]

[ActArea harder to learn than IC50 ref Koras2021]

[Adding network info: Wang2019, Zhang2021]

[SLFN11 known predictor: Nogales2016, Zoppoli2012, Gardner2017, Tang2015]

[Known high variability of data across studies: Geeleher2016, Safikhani2017, Xia2022]

[Previous WINTHER trial: low contribution of RNA expression (Rodon2022)]

[MASTER trial results Horak2021]

[trials metastudy: Schwaederle2015, Schwaederle2016]

[need for biomarker: Hyman2015 (BRAF), Hyman2018 (HER2)]

[Pathway signature tool rather old. Redo with state-of-the-art tools, also on different omics levels]

1. Conclusions
2. Author contributions

References:

[1] Hannah Ritchie and Max Roser, “Causes of death.” OurWorldInData.org, Dec. 08, 2022. [Online]. Available: https://ourworldindata.org/causes-of-death

[2] D. Cilloni and G. Saglio, “Molecular Pathways: BCR-ABL,” *Clin. Cancer Res.*, vol. 18, no. 4, pp. 930–937, Feb. 2012, doi: 10.1158/1078-0432.CCR-10-1613.

[3] T. Zhang, K. Dutton-Regester, K. M. Brown, and N. K. Hayward, “The genomic landscape of cutaneous melanoma,” *Pigment Cell Melanoma Res.*, vol. 29, no. 3, pp. 266–283, May 2016, doi: 10.1111/pcmr.12459.

[4] R. Govindan *et al.*, “Genomic Landscape of Non-Small Cell Lung Cancer in Smokers and Never-Smokers,” *Cell*, vol. 150, no. 6, pp. 1121–1134, Sep. 2012, doi: 10.1016/j.cell.2012.08.024.

[5] D. Hanahan and R. A. Weinberg, “Hallmarks of Cancer: The Next Generation,” *Cell*, vol. 144, no. 5, pp. 646–674, Mar. 2011, doi: 10.1016/j.cell.2011.02.013.

[6] E. O. Paull *et al.*, “A modular master regulator landscape controls cancer transcriptional identity,” *Cell*, vol. 184, no. 2, pp. 334-351.e20, Jan. 2021, doi: 10.1016/j.cell.2020.11.045.

[7] A. J. Levine, “Spontaneous and inherited TP53 genetic alterations,” *Oncogene*, vol. 40, no. 41, pp. 5975–5983, Oct. 2021, doi: 10.1038/s41388-021-01991-3.

[8] L. M. Sholl, “A narrative review of BRAF alterations in human tumors: diagnostic and predictive implications,” *Precis. Cancer Med.*, vol. 3, pp. 26–26, Dec. 2020, doi: 10.21037/pcm-20-39.

[9] P. L. Bedard, D. M. Hyman, M. S. Davids, and L. L. Siu, “Small molecules, big impact: 20 years of targeted therapy in oncology,” *The Lancet*, vol. 395, no. 10229, pp. 1078–1088, Mar. 2020, doi: 10.1016/S0140-6736(20)30164-1.

[10] D. Zahavi and L. Weiner, “Monoclonal Antibodies in Cancer Therapy,” *Antibodies*, vol. 9, no. 3, p. 34, Jul. 2020, doi: 10.3390/antib9030034.

[11] B. Gyawali, E. D’Andrea, J. M. Franklin, and A. S. Kesselheim, “Response Rates and Durations of Response for Biomarker-Based Cancer Drugs in Nonrandomized Versus Randomized Trials,” *J. Natl. Compr. Canc. Netw.*, vol. 18, no. 1, pp. 36–43, Jan. 2020, doi: 10.6004/jnccn.2019.7345.

[12] P. Horak *et al.*, “Precision oncology based on omics data: The NCT Heidelberg experience,” *Int. J. Cancer*, vol. 141, no. 5, pp. 877–886, Sep. 2017, doi: 10.1002/ijc.30828.

[13] C. O. Nordling, “A New Theory on the Cancer-inducing Mechanism,” *Br. J. Cancer*, vol. 7, no. 1, pp. 68–72, Mar. 1953, doi: 10.1038/bjc.1953.8.

[14] J. Barretina *et al.*, “NIH Public Access of anticancer drug sensitivity,” vol. 483, no. 7391, pp. 603–607, 2012, doi: 10.1038/nature11003.The.

[15] F. Azuaje, “Artificial intelligence for precision oncology: beyond patient stratification,” *Npj Precis. Oncol.*, vol. 3, no. 1, p. 6, Feb. 2019, doi: 10.1038/s41698-019-0078-1.

[16] R. Rafique, S. M. R. Islam, and J. U. Kazi, “Machine learning in the prediction of cancer therapy,” *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 4003–4017, 2021, doi: 10.1016/j.csbj.2021.07.003.

[17] F. Firoozbakht, B. Yousefi, and B. Schwikowski, “An overview of machine learning methods for monotherapy drug response prediction,” *Brief. Bioinform.*, vol. 23, no. 1, p. bbab408, Jan. 2022, doi: 10.1093/bib/bbab408.

[18] J. N. Weinstein *et al.*, “An Information-Intensive Approach to the Molecular Pharmacology of Cancer,” *Science*, vol. 275, no. 5298, pp. 343–349, Jan. 1997, doi: 10.1126/science.275.5298.343.

[19] J. E. Staunton *et al.*, “Chemosensitivity prediction by transcriptional profiling,” *Proc. Natl. Acad. Sci.*, vol. 98, no. 19, pp. 10787–10792, Sep. 2001, doi: 10.1073/pnas.191368598.

[20] L. J. van ’t Veer *et al.*, “Gene expression profiling predicts clinical outcome of breast cancer,” *Nature*, vol. 415, no. 6871, pp. 530–536, Jan. 2002, doi: 10.1038/415530a.

[21] M. Gönen and A. A. Margolin, “Drug susceptibility prediction against a panel of drugs using kernelized Bayesian multitask learning,” *Bioinformatics*, vol. 30, no. 17, pp. i556–i563, Sep. 2014, doi: 10.1093/bioinformatics/btu464.

[22] M. J. Garnett *et al.*, “Systematic identification of genomic markers of drug sensitivity in cancer cells,” *Nature*, vol. 483, no. 7391, pp. 570–575, 2012, doi: 10.1038/nature11005.

[23] P. Geeleher, N. J. Cox, and R. S. Huang, “Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines,” *Genome Biol.*, vol. 15, no. 3, pp. 1–12, 2014, doi: 10.1186/gb-2014-15-3-r47.

[24] C. Suphavilai, D. Bertrand, and N. Nagarajan, “Data and text mining Predicting Cancer Drug Response using a Recommender System”, doi: 10.1093/bioinformatics/bty452.

[25] Y.-C. Chiu *et al.*, “Predicting drug response of tumors from integrated genomic profiles by deep neural networks,” *BMC Med. Genomics*, vol. 12, no. S1, p. 18, Jan. 2019, doi: 10.1186/s12920-018-0460-9.

[26] T. Sakellaropoulos *et al.*, “A Deep Learning Framework for Predicting Response to Therapy in Cancer,” *Cell Rep.*, vol. 29, no. 11, pp. 3367-3373.e4, Dec. 2019, doi: 10.1016/j.celrep.2019.11.017.

[27] Y. Chang *et al.*, “Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature,” *Sci. Rep.*, vol. 8, no. 1, pp. 1–11, 2018, doi: 10.1038/s41598-018-27214-6.

[28] J. Kong *et al.*, “Network-based machine learning in colorectal and bladder organoid models predicts anti-cancer drug efficacy in patients,” *Nat. Commun.*, vol. 11, no. 1, p. 5485, Oct. 2020, doi: 10.1038/s41467-020-19313-8.

[29] L. Pu, M. Singha, J. Ramanujam, and M. Brylinski, “CancerOmicsNet: a multi-omics network-based approach to anti-cancer drug profiling,” *Oncotarget*, vol. 13, no. 1, pp. 695–706, May 2022, doi: 10.18632/oncotarget.28234.

[30] M. P. Menden *et al.*, “Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties,” *PLoS ONE*, vol. 8, no. 4, 2013, doi: 10.1371/journal.pone.0061318.

[31] N. Zhang, H. Wang, Y. Fang, J. Wang, X. Zheng, and X. S. Liu, “Predicting Anticancer Drug Responses Using a Dual-Layer Integrated Cell Line-Drug Network Model,” *PLOS Comput. Biol.*, vol. 11, no. 9, p. e1004498, Sep. 2015, doi: 10.1371/journal.pcbi.1004498.

[32] J. C. Costello *et al.*, “A community effort to assess and improve drug sensitivity prediction algorithms,” *Nat. Biotechnol.*, vol. 32, no. 12, pp. 1202–1212, Dec. 2014, doi: 10.1038/nbt.2877.

[33] D. H. Wolpert, “STACKED GENERALIZATION,” 1992.

[34] K. Matlock, C. D. Niz, R. Rahman, S. Ghosh, and R. Pal, “Investigation of model stacking for drug sensitivity prediction,” *BMC Bioinformatics*, vol. 19, no. Suppl 3, 2018, doi: 10.1186/s12859-018-2060-2.

[35] F. Iorio *et al.*, “A Landscape of Pharmacogenomic Interactions in Cancer,” *Cell*, vol. 166, no. 3, pp. 740–754, 2016, doi: 10.1016/j.cell.2016.06.017.

[36] J. R. Parikh, B. Klinger, Y. Xia, J. A. Marto, and N. Blï¿½thgen, “Discovering causal signaling pathways through gene-expression patterns,” *Nucleic Acids Res.*, vol. 38, no. suppl\_2, pp. W109–W117, Jul. 2010, doi: 10.1093/nar/gkq424.

[37] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[38] Y. Freund and R. E. Schapire, “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting,” *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997, doi: 10.1006/jcss.1997.1504.

[39] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[40] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, Apr. 2006, doi: 10.1007/s10994-006-6226-1.

[41] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, Apr. 2011, doi: 10.1145/1961189.1961199.

[42] Fix, Evelyn, “Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties,” USAF School of Aviation Medicine, Randolph Field, Technical Report 4, 1951.

[43] B. D. Weiss *et al.*, “NF106: A Neurofibromatosis Clinical Trials Consortium Phase II Trial of the MEK Inhibitor Mirdametinib (PD-0325901) in Adolescents and Adults With NF1-Related Plexiform Neurofibromas,” *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.*, vol. 39, no. 7, pp. 797–806, Mar. 2021, doi: 10.1200/JCO.20.02220.

[44] D. Casey *et al.*, “FDA Approval Summary: Selumetinib for Plexiform Neurofibroma,” *Clin. Cancer Res.*, vol. 27, no. 15, pp. 4142–4146, Aug. 2021, doi: 10.1158/1078-0432.CCR-20-5032.

[45] K. Fujiwara, K. Hasegawa, and S. Nagao, “Landscape of systemic therapy for ovarian cancer in 2019: Primary therapy,” *Cancer*, vol. 125, no. S24, pp. 4582–4586, Dec. 2019, doi: 10.1002/cncr.32475.

[46] A. Morabito *et al.*, “Vandetanib (ZD6474), a Dual Inhibitor of Vascular Endothelial Growth Factor Receptor (VEGFR) and Epidermal Growth Factor Receptor (EGFR) Tyrosine Kinases: Current Status and Future Directions,” *The Oncologist*, vol. 14, no. 4, pp. 378–390, Apr. 2009, doi: 10.1634/theoncologist.2008-0261.

[47] S. Wilhelm *et al.*, “Discovery and development of sorafenib: a multikinase inhibitor for treating cancer,” *Nat. Rev. Drug Discov.*, vol. 5, no. 10, pp. 835–844, Oct. 2006, doi: 10.1038/nrd2130.

[48] T. E. Williams *et al.*, “Discovery of RAF265: A Potent mut-B-RAF Inhibitor for the Treatment of Metastatic Melanoma,” *ACS Med. Chem. Lett.*, vol. 6, no. 9, pp. 961–965, Sep. 2015, doi: 10.1021/ml500526p.

[49] M. N. Dickler, M. A. Cobleigh, K. D. Miller, P. M. Klein, and E. P. Winer, “Efficacy and safety of erlotinib in patients with locally advanced or metastatic breast cancer,” *Breast Cancer Res. Treat.*, vol. 115, no. 1, pp. 115–121, May 2009, doi: 10.1007/s10549-008-0055-9.

[50] G. Dinstag *et al.*, “Clinically oriented prediction of patient response to targeted and immunotherapies from the tumor transcriptome,” *Med*, vol. 4, no. 1, pp. 15-30.e8, Jan. 2023, doi: 10.1016/j.medj.2022.11.001.