SCIENTIFIC REPORTS
nature research

Check for updates

OPEN

Feature selection strategies for drug sensitivity prediction

Krzysztof Koras[1], Dilafruz Juraeva[2], Julian Kreis[2], Johanna Mazur[2], Eike Staub[2] & Ewa Szczurek[1]✉

Drug sensitivity prediction constitutes one of the main challenges in personalized medicine. Critically, the sensitivity of cancer cells to treatment depends on an unknown subset of a large number of biological features. Here, we compare standard, data-driven feature selection approaches to feature selection driven by prior knowledge of drug targets, target pathways, and gene expression signatures. We asses these methodologies on Genomics of Drug Sensitivity in Cancer (GDSC) dataset, evaluating 2484 unique models. For 23 drugs, better predictive performance is achieved when the features are selected according to prior knowledge of drug targets and pathways. The best correlation of observed and predicted response using the test set is achieved for Linifanib ($r = 0.75$). Extending the drug-dependent features with gene expression signatures yields the most predictive models for 60 drugs, with the best performing example of Dabrafenib. For many compounds, even a very small subset of drug-related features is highly predictive of drug sensitivity. Small feature sets selected using prior knowledge are more predictive for drugs targeting specific genes and pathways, while models with wider feature sets perform better for drugs affecting general cellular mechanisms. Appropriate feature selection strategies facilitate the development of interpretable models that are indicative for therapy design.

The ability to predict a response of a specific cancer type to a therapy is one of the main goals in precision medicine. Considering molecular features of cancer cells is crucial for mitigating heterogeneity and for tailoring the therapy to specific patients[1]. The emergence of large scale high-throughput screening studies[2–6] have allowed researchers to develop computational models for drug response prediction from molecular profiles of human cancer cell lines or drug properties[7,8]. Although the inconsistencies and limitations of cell line data have been raised and extensively studied[9–12], these resources remain a vital tool for development of such models.

Arguably, the desired quality of computational models of drug response is not only their predictive performance, but also interpretability. To evaluate candidate drug efficacy on a specific patient's tumor, many approaches apply black-box algorithms with a set of highly dimensional features as input. In clinical practice, the capability of extracting such high-volume data from patient's material is limited. Thus, there is a growing need of proper identification of concise, limited subset of features, or biomarkers, that are most informative of drug response. Therefore, strong emphasis should be put on feature selection approaches for drug sensitivity prediction. Despite its paramount importance, no systematic assessment of feature selection strategies in the task of drug response prediction was so far performed.

The problem of drug response prediction has been approached by a wide spectrum of linear and non-linear machine learning algorithms, including regularized linear regression, k-nearest neighbors (KNN), support vector machines and random forests[13–18]. Multitask learning was proposed to improve drug sensitivity prediction by pooling information learned for different drugs[15,19]. Finally, a number of kernel-based multi-view and multi-task models were introduced for drug sensitivity[20–22]. Although these approaches show very good predictive performance, they suffer from low interpretability. As a remedy, a multi-task learning approach based on a Bayesian model for collaborative filtering was proposed[23], which allows for identifying general interactions between features of the drugs with features of the cell lines. For example, it gives insights in the form of "activation of pathway Y will confer sensitivity to any drug targeting protein X". This approach, however, does not directly address the crucial need of identifying biomarkers for specific drugs.

For that aim, data-driven, automatic techniques of feature selection were applied[17,22,24]. Generally, the problem of identifying the optimal subset of features is intractable[25]. Data-driven feature selection thus proceeds either as a heuristic search over the space of feature combinations, or is embedded directly in the learning algorithm

[1]Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland. [2]Merck Healthcare KGaA, Translational Medicine, Department of Bioinformatics, Darmstadt, Germany. ✉e-mail: szczurek@mimuw.edu.pl

by imposing sparsity of parameters associated with the features via regularization. Although these methods can achieve good predictive performance and deal with the curse of high data dimensionality[17], feature importance estimates and selection might not always be accurate and stable, especially in vastly high-dimensional data and in the presence of correlation between features[26]. Stability selection was proposed to mitigate this problem when regularized regression is applied[27], but it still comes without the guarantee to choose the most biologically relevant predictive features.

Drug prediction approaches largely differ with respect to the type of features that they model. Among the molecular data feature types which characterize the cancer cell lines, gene expression was assessed as the most informative, with remaining types such as mutation or copy number data bringing limited predictive power[13,14]. Accordingly, genome-wide gene expression is the most common choice in the case of models utilizing single data type[7,17,22,23,28]. Other studies reported that in some cases gene expression alone might not be sufficient, especially in a cancer- or drug-specific setting[29,30]. Importantly, expanding the feature space related only to cancer cell lines' biology with drug-related properties was shown to improve predictive performance[15,21–23,30]. The predictive drug-specific features may be related to their chemical properties, such as compound structure[15,30], their known primary targets or pathway activation[21,22]. Recently, multiple methods based on deep learning have emerged, showing promising results in the application to drug sensitivity prediction[31]. The published neural network architectures range from common stacks of fully connected layers[32] to more sophisticated architectures involving residual and convolutional networks[33–35]. Furthermore, methods employing autoencoders[36,37] and variational autoencoders[38] have been proposed. Due to their complicated, non-linear structure, neural networks may suffer from the lack of interpretablity, including difficulties in assessment of feature importance. However, methods from the growing field of explainable artificial intelligence can help to mitigate this problem[39].

Here, we utilize the knowledge regarding drug targets and their mode of action to select plausible features describing the cancer cell lines. This drug-related prior knowledge is thus used to directly limit the initial feature space, rather than first expanding it and next using data-driven selection techniques to narrow it down. We argue that this approach for feature selection in combination with common regression techniques can provide a simple and highly interpretable model without losing the predictive performance characteristic for models starting from high-dimensional data. In fact, the direct utilization of prior knowledge is the number one strategy recommended for feature selection according to the classics in machine learning[25]. It was however, never exploited in the task of drug response prediction. We assess this methodology in a systematic fashion for a broad spectrum of anti-cancer compounds, integrating multiple data types and comparing the results to the baseline models utilizing genome-wide gene expression data and data-driven feature selection techniques. On top of that, we evaluate gene expression signatures as the means of dimensionality reduction of the transcriptomics data and evaluate their predictive power in this context. This comprehensive analysis pin-points a set of drugs for which easily interpretable, informative, small sets of features can be identified.

Results

Modeling workflow. In order to comprehensively evaluate different feature selection strategies, we devised the following workflow (Fig. 1). We first extracted the sensitivity data for each particular drug and corresponding screened cell lines along with their biological features: gene expression, coding variants, copy number variation (CNV) and tissue type (see Methods for the details of the analyzed dataset). We then employed each of the feature selection approaches, which can be divided into two categories: biologically driven and automatic, data-driven selection methods. We considered different biologically driven feature selection strategies, depending on the type of prior knowledge used to define them. In the first approach, we narrowed the initial feature set by including only the features corresponding to drug's direct gene targets (shortly only targets, OT feature set). In the second, we considered the union of the direct target genes and the drug's target pathway genes (pathway genes, PG feature set). Finally, we additionally extended the only targets features and the pathway genes features with gene expression signatures, resulting in two more feature sets (OT + S and PG + S). For a baseline model we considered all available, 17737 gene expression features, referred to as the genome-wide model (GW). For data-driven feature selection we applied two techniques to the baseline gene expression feature set: stability selection (GW SEL EN) and random forest feature importance estimation (GW SEL RF). See Methods for more detailed description of the feature selection approaches. After the feature selection step, we fed the resulting data into elastic net (EN) or random forest (RF) algorithms and evaluated the predictive performance on the test set (Fig. 1). This modeling process was performed independently for each drug.

Models with genome-wide features have larger feature sets and more samples than the models with biologically-driven features. The median numbers of input features are 3 and 387 for only targets and pathway genes feature sets, respectively (Fig. 2a). The input features are further expanded by including 128 gene expression signatures. In the case of methods based on automated feature selection, the optimal number of features, k, is shown. The median k values are 70 and 1155 for random forests and for stability selection, respectively. All foregoing values constitute a drastic decrease in comparison to the number of 17737 genome-wide input features.

The number of samples for each drug also slightly differs for only targets and pathway genes feature sets, since for some cell lines the coding variants or CNV information are not available (Fig. 2b). This results in a lower number of samples for models with biologically driven features, with the median of 849 for only targets and 818 for pathway genes feature sets, compared to 876 for genome-wide expression features.

Drug response distributions are different across compounds, tend to have low variance for drugs targeting specific genes and pathways and high variance for drugs targeting general cellular mechanisms. The area under the dose-response curve (AUC; Methods) measures the overall drug

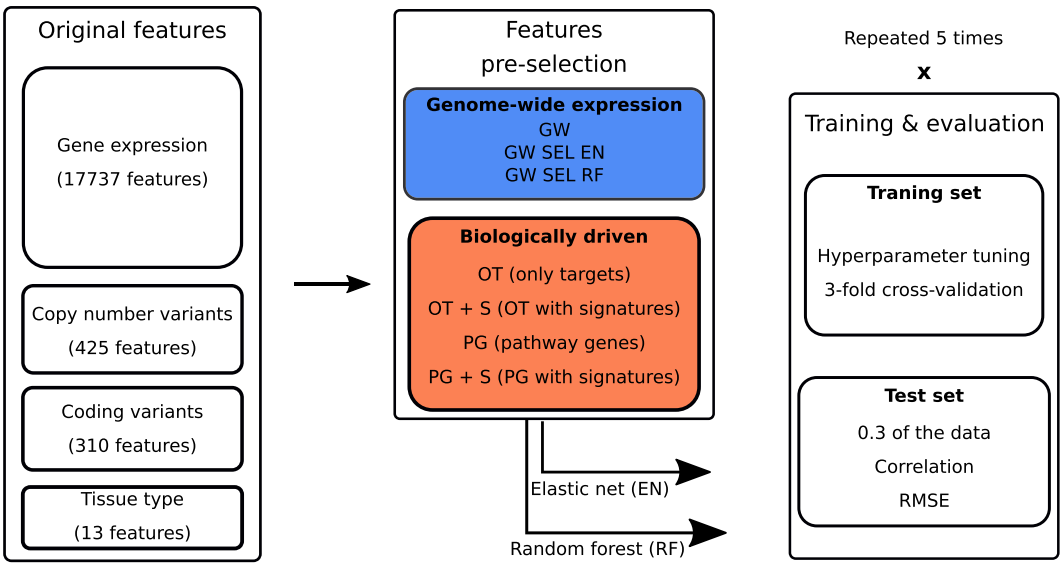

Figure 1. Flowchart describing the modeling framework for a single compound. Abbreviations: GW – genome-wide, PG – pathway genes, OT – only targets, EN – elastic net, RF – random forest, SEL – automated feature selection, S – gene expression signatures. For every feature space, we performed modeling separately for each drug. We randomly split the corresponding data into training and test set, with 0.3 of the data included in the test set. We used 3-fold cross-validation on the training data for hyperparameter tuning and evaluated the best model on the test set. The whole modeling process was repeated five times with different training/test set data splits.

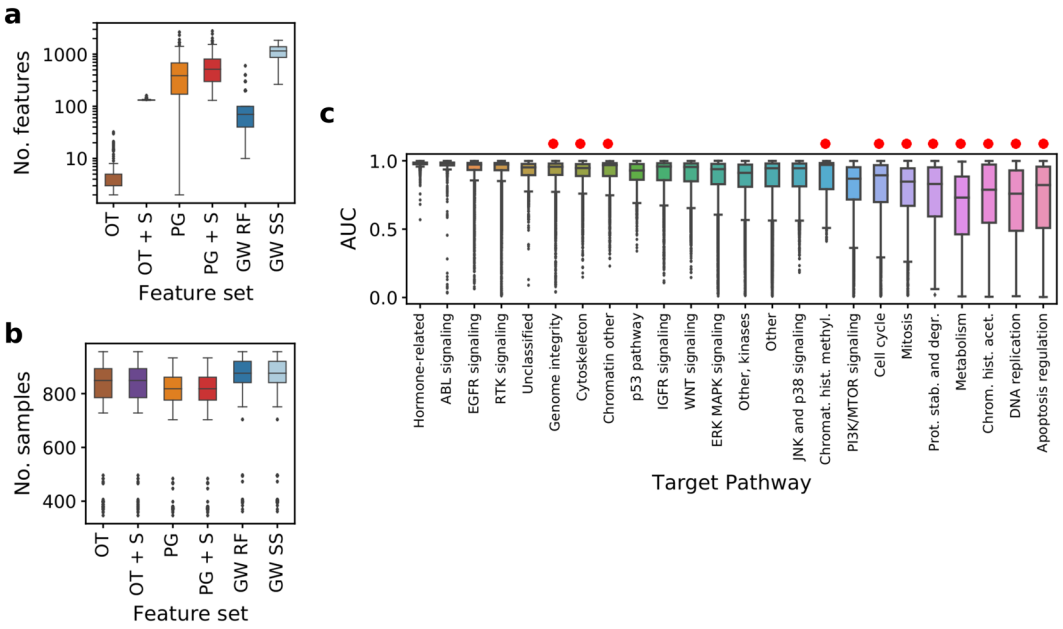

Figure 2. Models' properties and response variable grouped by target pathways. (**a**) Number of input features across compounds in different methods. For genome-wide models, number of features was 17737 for each drug. Vertical axis uses log scale. (**b**) Number of samples across compounds in different methods. Abbreviation SS refers to stability selection (Methods). (**c**) AUC values grouped by target pathway of the drug, raw data from GDSC. Target pathways are sorted by interquartile range of the AUC values. Pathways corresponding to more general cell mechanisms are marked with red dots. See Fig. 1 for abbreviations.

efficacy, with lower values corresponding to stronger efficacy. The distribution of this metric varies significantly among compounds with different target pathways (Fig. 2c). The median AUC value per target pathway ranges from 0.98 for hormone-related drugs to 0.73 for compounds targeting metabolism pathways. The smallest variation of AUC is observed for drugs targeting the hormone-related pathways. The largest AUC variation is observed for the apoptosis regulation pathway. The AUC for drugs targeting general mechanisms, such as DNA replication or metabolism, tends to have larger variance, which means their sensitivity is easier to model.

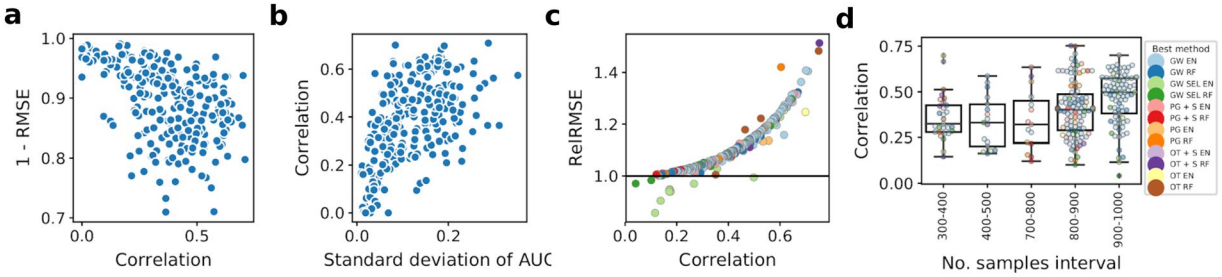

Figure 3. Predictive performance for all of the analyzed drugs. (**a**) 1 - RMSE versus correlation per drug, obtained by elastic net using genome-wide gene expression data as predictors. For 1-RMSE, higher values correspond to better performance. (**b**) Correlation versus standard deviation of true AUC for all cell lines screened for a given drug, correlation obtained by genome-wide elastic net. (**c**) RelRMSE versus correlation obtained by the best model for a given drug. Higher values of RelRMSE correspond to better performance and improvement over a dummy model, which predicts average AUC. Each point represents a single drug. For each of them, corresponding best performance was determined using correlation as a metric. Colors represent models with feature set that obtained the best performance for a given drug. Horizontal line at 1 represents the baseline RelRMSE score. Most of these correlations are statistically significant (test based on Student's t-distribution at 0.05 significance level, Fig. S1). (**d**) Distribution of per-drug predictive performance grouped by per-drug number of available samples. Colors represent models with feature set that obtained the best performance for a given drug. See Fig. 1 for model abbreviations.

The per-drug results show the importance of comparing to a dummy model and that different feature selection strategies are best suited for different drugs. Since root mean squared error (RMSE) measures the level of model error, and correlation measures the model agreement with the test set, both large (1-RMSE) and high correlation should coherently indicate a high model performance. The negative relation between (1 - RMSE) quantity and correlation, however, confirms the fact that raw RMSE is not a good metric for performance comparison between compounds (Fig. 3a; Methods). Instead, correlation achieved by the model increases with the modeled AUC variance (Fig. 3b).

Both these facts support that relative root mean squared error (RelRMSE; ratio of the RMSE obtained by a dummy model to the RMSE obtained by the analyzed model; see Methods) is a better performance measure than raw RMSE (Fig. 3c). Indeed, RelRMSE grows with the correlation. Importantly, for some drugs, the best performing models fail to achieve the baseline RelRMSE score of 1 or are very close to 1 (Fig. 3c). Further inspection of these models reveals that they can capture only the mean AUC, since the modeled AUC distribution does not have enough variation. In total, there were 19 of such compounds and these were excluded from further analysis.

It is apparent from Fig. 3c, that for most of the drugs, the best suited method is modeling using genome-wide features and elastic net. However, this is not the case for compounds with the top corresponding modeling performances, as the two best correlation scores are achieved by models with biologically driven feature space. These two compounds are Dabrafenib and Linifanib, both with correlation of 0.75, for models with feature spaces: only targets genes with gene expression signatures and only targets genes, respectively. In terms of performance, they are followed by Trametinib (correlation 0.71) and Alectinib (correlation 0.70), both scores being achieved by genome-wide methods. In general, as we consider more top performances, the frequency of genome-wide methods among them increases, although they are not as highly represented when looking at the small group of absolute best scores.

The considered set of drugs is diversified in terms of available data (Fig. 3d). The bigger number of samples leads to better predictive performance, as more training data mitigates the overfitting effect, especially in high-dimensional setting. However, there is a significant spread in performance among drugs with similar number of samples, implicating that available data is not a single factor explaining the differences in performance.

The difference in predictive performance of biologically driven versus genome-wide models is small, despite using significantly less input features. In general, genome-wide feature set combined with elastic net (GW EN) emerges as the best model with the median correlation of 0.39 (Fig. 4a). However, models with biologically driven feature spaces perform very similarly, (excluding only targets (OT) approaches), with the best median correlation of 0.37 produced by models employing target pathway genes features combined with gene expression signatures and elastic net (PG + S EN). Furthermore, the difference in median performance was negligible between genome-wide random forest (GW RF, with 17737 features) and genome-wide random forest with automated selection (GW SEL RF, with 70 features on average). This suggests that for many compounds, most gene expression features do not have significant power in predicting drug response. The spread in performance (defined as the difference between the maximum and the minimum value) reaches over 0.6 for all of the methods, suggesting that each drug should be approached individually in terms of modeling.

The standard, genome-wide model achieves the best performance for over half of considered drugs (Fig. 4b). However, for many of these cases the correlation difference between the best genome-wide model and the best model with biologically driven features is not significantly large, with the median of only 0.034 (Fig. 4c). The reverse is also true, with median correlation difference between the best biologically driven model and the worse

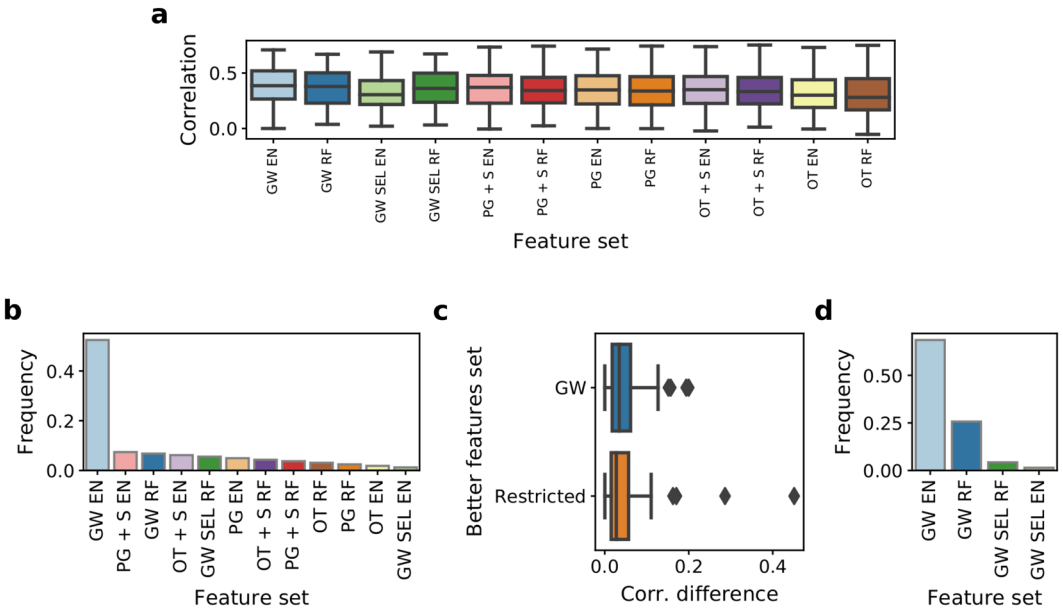

Figure 4. Frequencies of all applied methods among best models per drug. (**a**) Correlation of AUC predictions with the true AUC values in the test set across compounds in methods with different feature spaces. Results are shown for 175 drugs which were common across all applied models. (**b**) Model frequencies for compounds for which all methods were applied. (**c**) Differences in correlation between best model per drug overall and best model from the other class. Two cases are shown – genome-wide and biologically driven feature sets. (**d**) Model frequencies among best models for compounds where models with biologically driven could not have been applied. See Fig. 1 for abbreviations.

genome-wide model 0.028. Despite a drastic reduction in feature space, the biologically driven models based either on only targets or pathways yield the best modeling performance for 23 drugs, outperforming all other models including the genome-wide approach. For further 60 drugs, the best models have feature space expanded with expression signatures. Noticeably, there are also 15 cases where data-driven feature selection helps to produce better performance with much smaller subset of the original feature set (Fig. 4b,d).

Predictive performance using different feature selection strategies depends on drugs' target pathways. Next, we investigate the general tendencies concerning which feature selection is particularly better suited for modeling drugs targeting specific pathways. To this end, we compare the overall performance of biologically driven feature selection as one group to the baseline of genome-wide features and the genome-wide features with automatic selection as another (Fig. 5a), for different target pathways. Genome-wide models achieve better performance in 15 out of 24 pathways in total, however, the difference is statistically significant in only four of them (at 0.05 significance level): DNA replication, metabolism, apoptosis regulation pathways and a group of pathways referred to as "other". This indicates that these models capture a broad mechanism of action of the corresponding drugs. Conversely, the target pathways for which the models with biologically driven features most notably outperform models with genome-wide features include ABL, IGFR and EGFR signaling pathways, although these results are not statistically significant due to small sample sizes. The models with biologically-driven features perform better also for the hormone-related pathway, but overall the modeling performance is bad in this case and we do not consider this result reliable. In summary, compounds with specific signaling target pathways seem to benefit more from the initially restricted feature space. Notably, the median number of available sample sizes for drugs targeting specific pathways is similar between the pathways (Fig. S2a) and does not affect the modeling performance (Fig. S2b). Although the number of drugs per target pathway does differ between the pathways, these differences should not affect the comparison outcome as the comparisons of model performance are made within a given pathway.

We next inspect in detail the results for distinct drugs coming from DNA replication and RTK signaling pathways, respectively (Fig. 5b,c). Among the drugs targeting the DNA replication pathway, Bleomycin, Methotrexate and SN-38 exhibit good modeling ability with the genome-wide features. However, in case of Methotrexate similar performance is achieved also by methods with biologically driven feature space, contrary to SN-38. Conversely to DNA replication pathway, among the drugs targeting the RTK signaling pathway the best result is more often produced by biologically driven features, with most noticeable cases of Linifanib and Quizartinib. In contrast, Alectinib exhibits good modeling performance exclusively with genome-wide approaches. In general, although the above described general tendencies apply, information about drug's target pathway alone seems to be insufficient to clearly tell which feature space is the most suitable for predicting its response, with the potential exception of the DNA replication pathway.

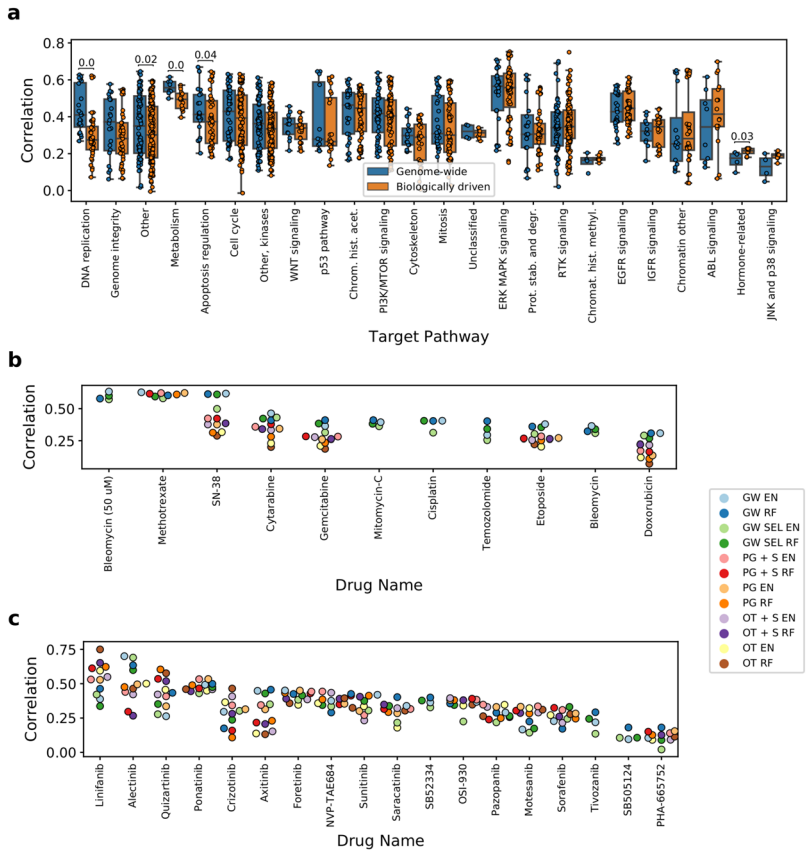

Figure 5. Predictive performance in relation to compounds' target pathway. (**a**) Correlation with the test set grouped by pathways. Methods were classified into two groups – one that uses genome-wide feature space, and one with biologically driven feature space. Numbers displayed represent p-values for the one-sided Mann-Whitney-Wilcoxon test. Lack of number means no statistical significance at 0.05 significance level. (**b**) Predictive performance for drugs with DNA replication target pathway. (**c**) Predictive performance for drugs with RTK signaling pathway. See Fig. 1 for model abbreviations.

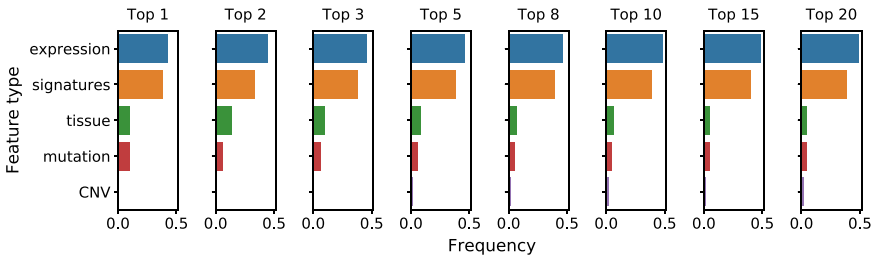

Figure 6. Frequencies of considered feature types among top k most predictive features. Feature importance coefficients were extracted from top 50 drugs in terms of modeling performance using methods with biologically driven feature space.

Gene expression and mutations constitute the most predictive feature types. In order to assess, which feature types are most informative of drug response, we consider such models with biologically driven feature space, which use all five available data types (Fig. 6). To make results more robust, we consider only top 50 drugs in terms of corresponding modeling performance achieved by the biologically driven feature sets, resulting in worst considered model's correlation of 0.47. Next, we extract top k most predictive features in each model and record the frequencies of particular data classes among them. Results confirm the fact that gene expression is the most predictive feature type, although mutation (coding variant) and tissue type are also important, especially for drugs designed to target specific cancer type with a particular mutation. In contrast, copy number variants seem not to incorporate much useful information. The relative effect of gene expression data increases with number of considered most predictive features, but this is expected given that this category is the most frequent of all available data types overall. Finally, the high frequency of gene expression signatures among the top predictive features implies that the signatures can act as good representatives of genome-wide information.

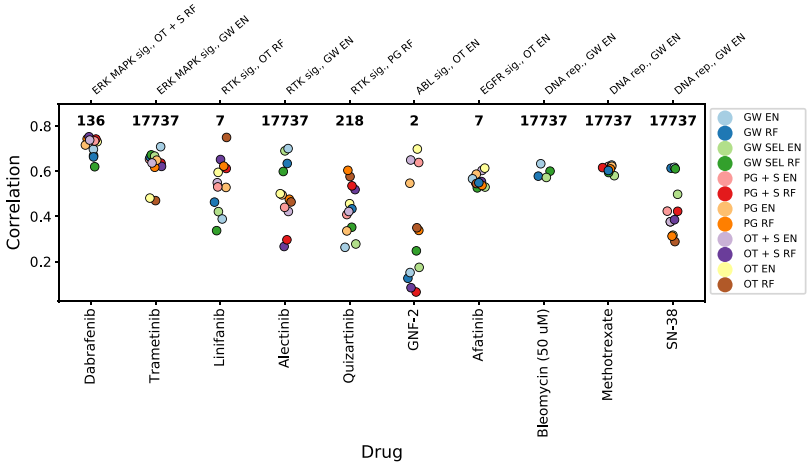

Figure 7. Results for specific compounds exhibiting good ability to model with one or all of the methods. Displayed numbers represent number of features which was used by the best performing model for a particular drug. Top horizontal axis shows compounds' target pathways along with the model which achieved the best modeling result. See Fig. 1 for model abbreviations.

Feature selection enables interpretation of the mode of action and pin-pointing biomarkers for the best modeled drugs. We further focus the analysis on ten drugs of most interest (Fig. 7), based on two simple criteria: top modeling performance achieved by all of the feature selection methods, or distinctly better performance achieved by one of the methods' class (genome-wide or biologically driven) in comparison to another. In five of those compounds the best result is produced by models with the genome-wide features, whereas another five are better modeled with biologically driven features.

From all analyzed drugs, Dabrafenib emerges as the compound which is the easiest to model. The highest correlation of 0.75 is achieved by the model combining only targets features with gene expression signatures and random forest (OT + S RF), and performance of other approaches is only slightly worse (Fig. 7). This good modeling ability with the OT + S RF features could be explained by two factors. First, the AUC distribution corresponding to Dabrafenib is well-diversified, with relatively many cell lines sensitive to treatment (Fig. 8a), which leads to better modeling performance (compare Fig. 3b). Second, the relative effects of the selected features are in excellent concordance with the Dabrafenib's pharmaceutical properties. The most predictive feature – mutation in BRAF oncogene (Fig. 8a) – and the second most predictive feature – the BRAF gene expression signature – well agree with the design of Dabrafenib as the BRAF inhibitor. Interestingly, the feature corresponding to BRAF gene expression alone ranks lower, 28 among 136 features for the best OT + S RF model and as low as 15817 among 17737 features for the GW EN model in terms of predictive power. Finally, in concordance with Dabrafenib's intended use in treatment of BRAF mutation-positive melanomas and lung cancers[40,41], the skin tissue feature is the third most predictive one for the best OT + S RF model.

In the case of Linifanib, the best result (0.75 correlation) is accomplished by using only 7 features related to the drug's targets (only targets and random forest, OT RF model), which significantly outperforms the genome-wide models (Fig. 7). Linifanib is an inhibitor of FMS-like tyrosine kinase 3 (FLT3) and vascular endothelial growth factor receptor (VEGF) tyrosine kinases, and is involved in clinical trials concerning non-small cell lung cancer (NSCLC), breast, liver, and colorectal cancer as well as leukemia[42–44]. Contrary to the Dabrafenib's example, Linifanib is one of the rare examples where good modeling results are achievable despite low standard deviation of the AUC distribution (Fig. 8b). The high correlation achieved by the OT RF model mainly comes from its ability to accurately predict lowered AUC for three outlying, sensitive cell lines. The most decisive predictive feature in this model is the expression of FLT3 gene, which exhibits high over expression in these cell lines, with much higher mean expression of 11.53 than the mean of 3.30 for all cell lines in the training set. The expression of FLT3 ranks lower (11th) among features of the genome-wide model.

Similarly to Linifanib, Quizartinib is also characterized by low variation in the treatment response (Fig. 8c), and is also an FLT3 inhibitor. Quizartinib is tested in clinical trials for acute myeloid leukemia (AML)[45]. The best biologically-driven model (pathway genes and random forest, PG RF) uses features related to genes present in drug's target pathway (218 features), and the most important feature is expression of FLT3. The accurate prediction done by PG RF model for the single outlying, responsive sample (Fig. 8c) probably arises from the over-expression of FLT3 in that cell line (11.20 value for that feature in this sample versus the mean of 3.26 for all training samples). Although expression of FLT3 also appears as the fourth most important feature in the genome-wide model, it is unable to correctly predict AUC for the responsive cell line, since the relative impact of FLT3 is much smaller. Overall, these three examples well show that feature selection can facilitate derivation of interpretable insights.

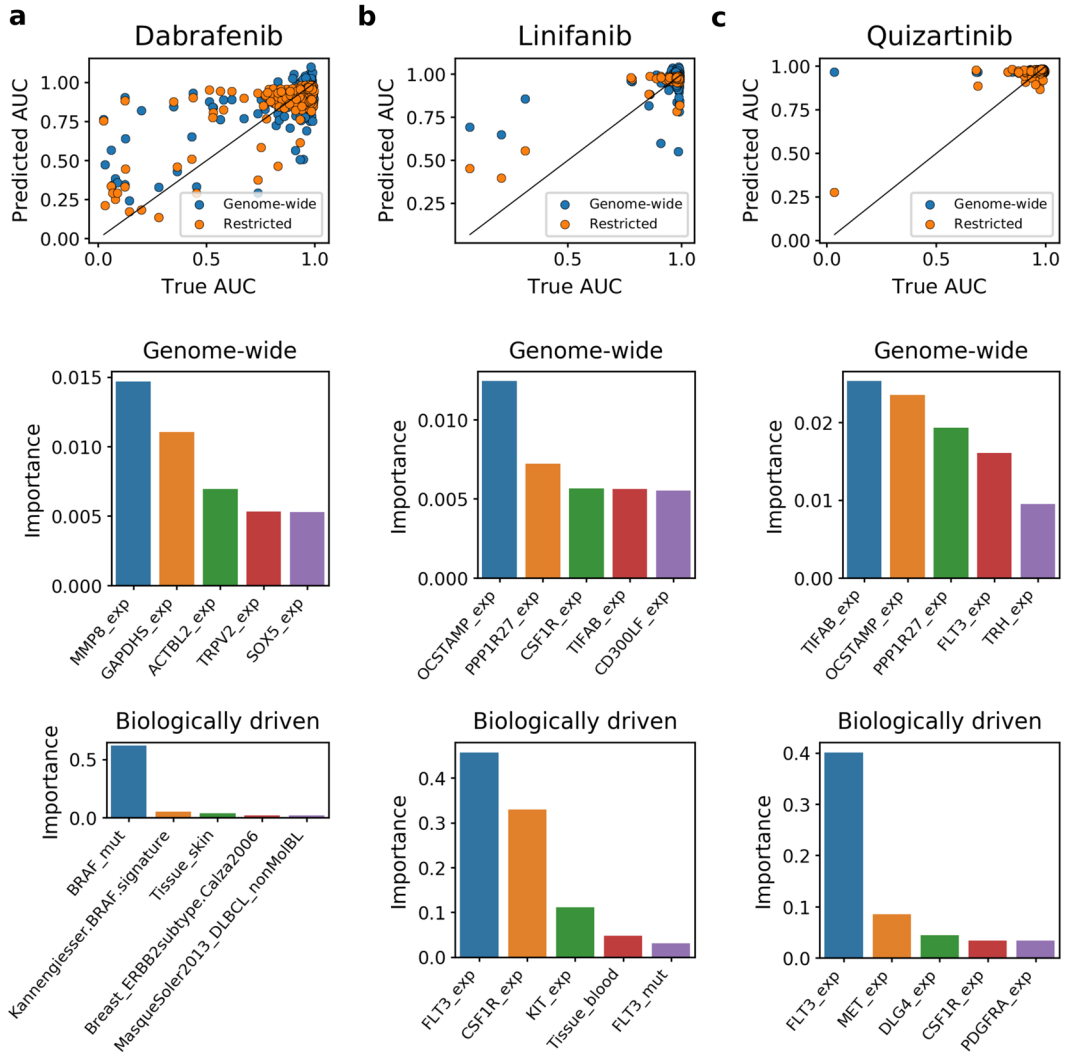

Figure 8. Predicted versus actual AUC values and most predictive features for (**a**) Dabrafenib, (**b**) Linifanib and (**c**) Quizartinib. Top panels show predicted versus actual AUC values when both biologically driven and genome-wide models were trained and tested on the same sets of samples. The biologically driven models correspond to best suited feature set for each drug: OT + S RF for Dabrafenib, OT RF for Linifanib and PG RF for Quizartinib. Middle and bottom panels present top 5 most informative features when fitting the model with genome-wide data (middle) and biologically driven feature space (bottom).

Discussion

This work, to our knowledge, is the first comprehensive analysis of feature selection strategies for drug sensitivity prediction. Previous systematic assessments[13,14] compared different modeling techniques and data types describing the cell lines, but did not comprehensively evaluate feature selection approaches. Similarly, although numerous modeling methods were developed specifically for the task of drug sensitivity prediction[7,8], they were solely optimized for predictive power and not interpretability. If feature selection was applied at all, it was not driven by pre-existing biological knowledge, but performed using standard and often not robust selection techniques such as regularization[22].

Such comprehensive feature selection assessment is needed for several reasons. First of all, both feature selection driven by pre-existing biological knowledge and data driven selection have their advantages and disadvantages. Intuitively, selecting the features using *a priori* knowledge of the drug mode of action as a guideline should improve modeling. On the other hand, it is also restricting the available information for the model, and if the prior knowledge is wrong, may result in missing important dependencies. Given the vast number of features compared to the number of samples, the models with genome-wide data as features or ones with automated feature selection are badly ill-posed and prone to over-fitting. On the other hand, they are given the advantage of a larger number of samples (resulting in higher power) and access to more information, compared to the models with biologically driven features (Fig. 2). Second, as there is no obvious recipe for choosing the feature set for a particular drug, the in-depth comparative analysis of different feature selection strategies may suggest indications for the recommended type of features for drugs depending on their mode of action or knowledge of their target

pathway. Finally, if the best performing feature set is small, each particular feature can be inspected and further evaluated as a potential biomarker for the drug.

Here, different feature selection strategies driven by prior knowledge were compared to using genome-wide feature sets and the data-driven, automatic feature selection techniques across all analyzed drugs. We identified the best suited feature set for each drug and investigated them in the context of drugs' target pathways. Finally, we evaluated the predictive power of different feature types and inspected example drug-specific models in more detail. The entire assessment workflow aimed at the identification of such strategies that could deliver highly predictive, but also highly interpretable models, bringing insights about specific drugs that are informative for their application in precision medicine.

Both Jang *et al.*[14] and the DREAM challenge[13] assessments indicated that adding the features representing mutation and copy number status on top of genome-wide expression features did not improve the overall performance of modeling drug sensitivity[13,14]. This is likely due to the fact that gene expression is sometimes already reflecting genomic changes or tissue type. In contrast, our analysis shows that additional features corresponding to mutations are often significant predictors when they are evaluated as part of smaller feature set and are not vastly outnumbered by the gene expression features (for example, in the cases of Dabrafenib, PLX-4720, Nutlin-3a, SB590885 and Pelitinib).

Our results bring important conclusions about feature selection strategies for drug sensitivity prediction. In general, the baseline genome-wide set of features or data-driven feature selection yields higher median predictive performance than biologically driven features. There are, however, multiple individual drugs, for which the feature selection driven by biological knowledge gives the best results, including models for the drugs with the top two performance scores. Moreover, feature selection driven by prior knowledge drastically reduces the number of features. At the same time, if the drop of performance in comparison with genome-wide models occurs, it is often only slight.

In addition, the presented analysis illuminates the mechanisms behind the sensitivity of different cancer cell lines to different types of drugs, suggesting which types of features should be used to model different classes of drugs. Drugs that are generally toxic or target general cellular mechanisms such as DNA replication or metabolism affect a relatively large proportion of cancer cell lines and thus have a wide response distribution. These compounds tend to be better modeled using genome-wide features, indicating that their effect on the cancer cells depends on a large spectrum of different cellular features. Conversely, for drugs targeting specific pathways, sensitivity distribution tends to be narrow, with most cells not responding at all and only a few interesting outliers of sensitive cells. For these compounds, high-level drug properties such as direct targets or target pathways allow to build highly predictive models with small numbers of interpretable features, such as Dabrafenib, Linifanib or Quizartinib. In particular, highly predictive models with an extremely low number of input features can be obtained, as in the cases of Linifanib, Afatinib, and GNF-2. Overall, this analysis shows the importance of using adequate feature selection strategies for each individual drug.

Methods

Analyzed dataset. The analyzed dataset was acquired from the Genomics of Drug Sensitivity in Cancer (GDSC)[3] database. A total of 251 compounds were included in the analysis. Each was assigned one of 24 classes of target pathways, defined by the GDSC.

The total set of samples consisted of 983 cancer cell lines originated from 13 tissue sites. The available data types for describing the cell lines included: gene expression (17737 features), coding variants (310 features), copy number variants (CNV, 425 features) and tissue type (13 features). Coding variants and copy number variants were represented as binary calls determining the presence or absence of a variant in a given gene or segment, respectively. We have dummy encoded the tissue types resulting in 13 distinct binary features for every cell line. All biological input data were acquired directly from the GDSC resource.

GDSC provides two types of metrics representing the drug efficacy: half maximal inhibitory concentration (IC50) and area under the dose-response curve (AUC). Since in our analysis we did not observe significant differences in predictive performance when using one metric in favor of the other, we picked AUC as our single target variable.

Predictive algorithms. We employed two common machine learning algorithms in order to predict the AUC values: elastic net linear regression and random forest regression. We implemented both methods using Python3 scikit-learn 0.19.2 library[46]. See Supplementary Methods for descriptions of the algorithms and implementation details.

Feature selection. With a total of 18485 biological features that can be used to describe the cancer cell lines, the analyzed dataset is very high-dimensional. In contrast, the number of samples is in the order of hundreds, which poses the danger of overfitting. This might especially be the case when considering all available genome-wide information regardless of the drug being modeled. Here, we investigate different feature selection methods to mitigate this problem. These approaches can be divided into two groups: biologically driven and automatic, data-driven selection methods.

Biologically driven feature selection. Features based only on drug targets and tissue type, shortly only targets (OT). In the most restricted feature space, we included only predictors corresponding to the direct targets of the drugs, as well as tissue type. Drug targets information was derived directly from GDSC. As an additional resource, we used DrugBank[47] database, assigning targets for 88 matched compounds. For each drug target, we included features representing the target gene's expression, coding variant and copy number variation. In the case of copy

number data, a given genetic feature was incorporated if the corresponding segment included at least one of the drug target genes. We only considered drugs with explicit gene targets annotation in GDSC or DrugBank and for which at least one feature in addition to the tissue type was available in the data. These conditions were met for 184 compounds. Applying two regression algorithms for each drug resulted in 368 separate models.

Set of features based on drug targets, tissue type, and target pathways, shortly pathway genes (PG). In this approach, we included features related to genes that belonged to the same signaling pathway as the set of target genes. Pathways information was derived from Reactome[48,49] database (version 66 accessed on October 2018). For each compound, first its target set was derived, followed by finding all pathways which included at least one of the given targets. The total set of considered genes was then computed as the union of all members of the found pathways. Lastly, corresponding gene expressions, coding variants, copy number variants and tissue types were extracted to create the final feature set. The drug targets and pathway information was available for 186 drugs, producing 372 models.

Sets of features resulting from addition of gene expression signatures, shortly OT + S or PG + S. Gene expression signatures can explain the activation level of complex biological phenomena in the investigated cell lines. Here, we refer to a gene signature as a set of genes related to a certain known biological phenomenon that can be deduced from cancer gene expression data (Supplementary Table S1). For each signature S with i genes, we calculated two scores. The first characterizes the coherent expression and the second estimates the activation level of S. Given a gene expression matrix for S in n samples ($X^{i\times n}$), the previously described coherence score (CS)[50], is calculated as the mean pairwise Pearson correlation between all columns of X. Therefore, a strong negative or positive correlation between all genes in S is indicated by CS values close to -1 and 1, respectively. The activity of S (i.e. the signature score) for each sample is calculated by first z-scoring the gene expression values across samples, followed by averaging the resulting z-scores across genes. Here, we calculated the signature scores using the cancer cell line expression data provided by GDSC. We set the threshold for a significantly coherent activation of S to $\mathrm{CS}(S) \geq 0.1$, resulting in 128 signature features. The OT + S set contains features based on target genes, signature scores and tissue type. The PG + S set contains target genes, pathway genes, signature scores and tissue type. Applying two regression algorithms for each drug resulted in 740 separate models.

Set of features based on genome-wide gene expression, shortly genome-wide (GW). Finally, we constructed a feature set based exclusively on the expression of 17737 genes as features. We evaluated this feature set for 251 drugs in total, resulting in 502 different models.

Data-driven feature selection, shortly GW SEL. In addition to feature pre-selection based on drug properties and biological relevance, we also evaluated automated feature selection algorithms in application to genome-wide expression data. We used two techniques, based on linear and non-linear methods. First, stability selection, which uses lasso regression on multiple bootstrap samples in order to choose robust features[27]. Such selected features were next passed as input for elastic net models (further referred to as *GW SEL EN*). For the second technique, feature importance estimates derived directly from random forest were used. These features were then used for random forest regression models (*GW SEL RF*). For more detailed description of both techniques, see Supplementary Methods.

Model evaluation. During cross-validation tuning, we used *Mean Squared Error* (MSE) as a scoring metric for best hyperparameters search. Although MSE is suitable for evaluation of different models within one compound, it is not reliable when comparing results across diverse drugs because of differences in corresponding AUC distributions. Furthermore, when a given target variable distribution has little variation, one can achieve a reasonably low MSE just by predicting the mean of a target variable. In order to avoid this problem and identify the models which performed well, we used *Relative Root Mean Squared Error* (RelRMSE), which is normalized in such a way that the score of 1 corresponds to a dummy model which always predicts the mean of target variable in the training data. RelRMSE is defined as the fraction of the dummy model's RMSE on the test data and the analyzed model's RMSE on the test data:

$$\mathrm{RelRMSE} = \frac{\mathrm{RMSE}_{dummy}}{\mathrm{RMSE}_{model}}, \tag{1}$$

i.e. better performance corresponds to bigger RelRMSE metric, with a baseline score of 1.

The use of RelRMSE allowed us to distinguish drugs for which predictive algorithms could not outperform the dummy model, meaning that for those compounds no actual learning occurred.

In order to make further assessments and comparisons between compounds, we used Pearson correlation coefficient with the response AUC in the test set as a performance metric. As stated in a previous section, the recorded results for each method were averaged over five modeling procedures that were performed with different data splits.

Data availability

The data analysed in this study were acquired from the Genomics of Drug Sensitivity in Cancer repository: https://www.cancerrxgene.org/ (2018) and the Reactome repository: https://reactome.org/ (2018). Code used to conduct the analysis is available at https://github.com/kkoras/feature-selection-in-cancer-drug-response.

Received: 13 December 2019; Accepted: 6 May 2020;
Published online: 10 June 2020

References

1. Bedard, P. L., Hansen, A. R., Ratain, M. J. & Siu, L. L. Tumour heterogeneity in the clinic. *Nature* **501**, 355–364 (2013).
2. Barretina, J. *et al*. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity (vol 483, pg 603, 2012). *Nature* **492**, 290–290 (2012).
3. Benes, C. *et al*. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* **41**, D955–D961 (2012).
4. Rees, M. *et al*. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nature chemical biology* **12** (2015).
5. Seashore-Ludlow, B. *et al*. Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer Discovery* **5**, 1210–1223 (2015).
6. Basu, A. *et al*. An Interactive Resource to Identify Cancer Genetic and Lineage Dependencies Targeted by Small Molecules. *Cell* **154**, 1151–1161 (2013).
7. Azuaje, F. Computational models for predicting drug responses in cancer research. *Brief. Bioinforma.* **18**, 820–829 (2016).
8. Ali, M. & Aittokallio, T. Machine learning and feature selection for drug response prediction in precision oncology applications. *Biophysical Rev.* **11**, 31–39 (2019).
9. Haibe-Kains, B. *et al*. Inconsistency in large pharmacogenomic studies. *Nature* **504**, 389–393 (2013).
10. Stransky, N. *et al*. Pharmacogenomic agreement between two cancer cell line data sets. *Nature* **528** (2015).
11. Gillet, J.-P., Varma, S. & Gottesman, M. M. The Clinical Relevance of Cancer Cell Lines. *JNCI: J. Natl Cancer Inst.* **105**, 452–458 (2013).
12. Gillet, J.-P. *et al*. Redefining the relevance of established cancer cell lines to the study of mechanisms of clinical anti-cancer drug resistance. *Proc. Natl Acad. Sci.* **108**, 18708–18713 (2011).
13. Costello, J. C. *et al*. A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.* **32**, 1202–1212 (2014).
14. Jang, I. S., Chaibub Neto, E., Guinney, J., Friend, S. & Margolin, A. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. *Pac. Symposium Biocomputing. Pac. Symposium Biocomputing* **19**, 63–74 (2014).
15. Menden, M. *et al*. Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. *PLoS one* **8**, e61318 (2013).
16. Tran, T. P., Ong, E., Hodges, A. P., Paternostro, G. & Piermarocchi, C. Prediction of kinase inhibitor response using activity profiling, *in vitro* screening, and elastic net regression. *BMC Syst. Biol.* **8**, 74 (2014).
17. Dong, Z. *et al*. Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. *BMC cancer* **15**, 489 (2015).
18. Riddick, G. *et al*. Predicting *in vitro* drug sensitivity using Random Forests. *Bioinformatics* **27**(2), 220–4 (2011).
19. Yuan, H., Paskov, I., Paskov, H., Gonzalez, A. J. & Leslie, C. S. Multitask learning improves prediction of cancer drug sensitivity. *Sci. Rep.* **6**, 31619 (2016).
20. Cichonska, A. *et al*. Learning with multiple pairwise kernels for drug bioactivity prediction. *Bioinformatics* **34**, i509–i518 (2018).
21. Ammad-ud din, M. *et al*. Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. *Bioinformatics* **32**, i455–i463 (2016).
22. Ammad-ud din, M., Khan, S., Wennerberg, K. & Aittokallio, T. Systematic identification of feature combinations for predicting drug response with Bayesian multi-view multi-task linear regression. *Bioinformatics* **33**, i359–i368 (2017).
23. Yang, M. *et al*. Linking drug target and pathway activation for effective therapy using multi-task learning. *bioRxiv* (2018).
24. Xu, X., Gu, H., Wang, Y., Wang, J. & Qin, P. Autoencoder Based Feature Selection Method for Classification of Anticancer Drug Response. *Front. Genet.* **10**, 233 (2019).
25. Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**, 1157–1182 (2003).
26. Khaire, U. M. & Dhanalakshmi, R. Stability of feature selection algorithm: A review. *Journal of King Saud University - Computer and Information Sciences* (2019).
27. Meinshausen, N. & Bühlmann, P. Stability selection. *J. R. Stat. Society: Ser. B* **72**, 417–473 (2010).
28. Geeleher, P., Cox, N. J. & Huang, R. S. Clinical drug response can be predicted using baseline gene expression levels and *in vitro* drug sensitivity in cell lines. *Genome Biol.* **15**, R47 (2014).
29. Amin, S. *et al*. Gene Expression Profile Alone Is Inadequate In Predicting Complete Response In Multiple Myeloma. *Leukemia* **28** (2014).
30. Cortes, I. *et al*. Improved Large-Scale Prediction of Growth Inhibition Patterns on the NCI60 Cancer Cell-Line Panel. *Bioinformatics* 1–11 (2015).
31. Baptista, D., Ferreira, P. G. & Rocha, M. Deep learning for drug response prediction in cancer. *Briefings in Bioinformatics*, Bbz171 (2020).
32. Sakellaropoulos, T. *et al*. A deep learning framework for predicting response to therapy in cancer. *Cell Rep.* **29**, 3367–3373.e4 (2019).
33. Xia, F. *et al*. Predicting tumor cell line response to drug pairs with deep learning. *BMC Bioinformatics* **19** (2018).
34. Chang, Y. *et al*. Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature. *Scientific Reports* **8** (2018).
35. Oskooei, A. *et al*. PaccMann: Prediction of anticancer compound sensitivity with multi-modal attention-based neural networks (2018).
36. Chiu, Y.-C. *et al*. Predicting drug response of tumors from integrated genomic profiles by deep neural networks. *BMC Medical Genomics* **12** (2019).
37. Li, M. *et al*. DeepDSC: A Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 1–1 (2019).
38. Rampášek, L., Hidru, D., Smirnov, P., Haibe-Kains, B. & Goldenberg, A. Dr.VAE: improving drug response prediction via modeling of drug perturbation effects. *Bioinformatics* **35**, 3743–3751 (2019).
39. Samek, W. & Müller, K.-R. Towards Explainable Artificial Intelligence. *Lecture Notes in Computer Science* 5–22 (2019).
40. Hauschild, A. *et al*. Dabrafenib in BRAF-mutated metastatic melanoma: A multicentre, open-label, phase 3 randomised controlled trial. *Lancet* **380**, 358–65 (2012).
41. Khunger, A., Khunger, M. & Velcheti, V. Dabrafenib in combination with trametinib in the treatment of patients with BRAF V600-positive advanced or metastatic non-small cell lung cancer: clinical evidence and experience. *Therapeutic Adv. Respiratory Dis.* **12**, 175346661876761 (2018).
42. Linifanib. *Drugs R D* **10**, 111–122 (2010).
43. Tan, E.-H. *et al*. Phase 2 Trial of Linifanib (ABT-869) in Patients with Advanced Non-small Cell Lung Cancer. *J. Thorac. Oncol.* **6**, 1418–1425 (2011).
44. Wang, E. S. *et al*. Phase 1 trial of linifanib (ABT-869) in patients with refractory or relapsed acute myeloid leukemia. *Leukemia & Lymphoma* **53**, 1543–1551, PMID: 22280537 (2012).
45. Levis, M. Quizartinib for the treatment of FLT3/ITD acute myeloid leukemia. *Future Oncol.* **10**, 1571–1579 (2014).
46. Pedregosa, F. *et al*. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
47. Wishart, S. D. *et al*. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic acids research* **46** (2017).
48. Fabregat, A. *et al*. The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **46**, D649–D655 (2017).

49. Fabregat, A. *et al*. Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinforma.* **18**, 142 (2017).
50. Staub, E. An Interferon Response Gene Expression Signature Is Activated in a Subset of Medulloblastomas. *Transl. Oncol.* **5**, 297–IN6 (2012).

Acknowledgements
This work was supported by grant 2015/19/P/NZ2/03780 to ESz from the National Science Centre, Poland, https://www.ncn.gov.pl/?language=en. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 665778. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions
Conceived and designed the study: K.K., E.St. and E.Sz. Analyzed the data: K.K. Wrote the paper: K.K. and E.Sz. Gave critical comments: D.J., J.K. and J.M. All authors read and approved the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-65927-9.

Correspondence and requests for materials should be addressed to E.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.