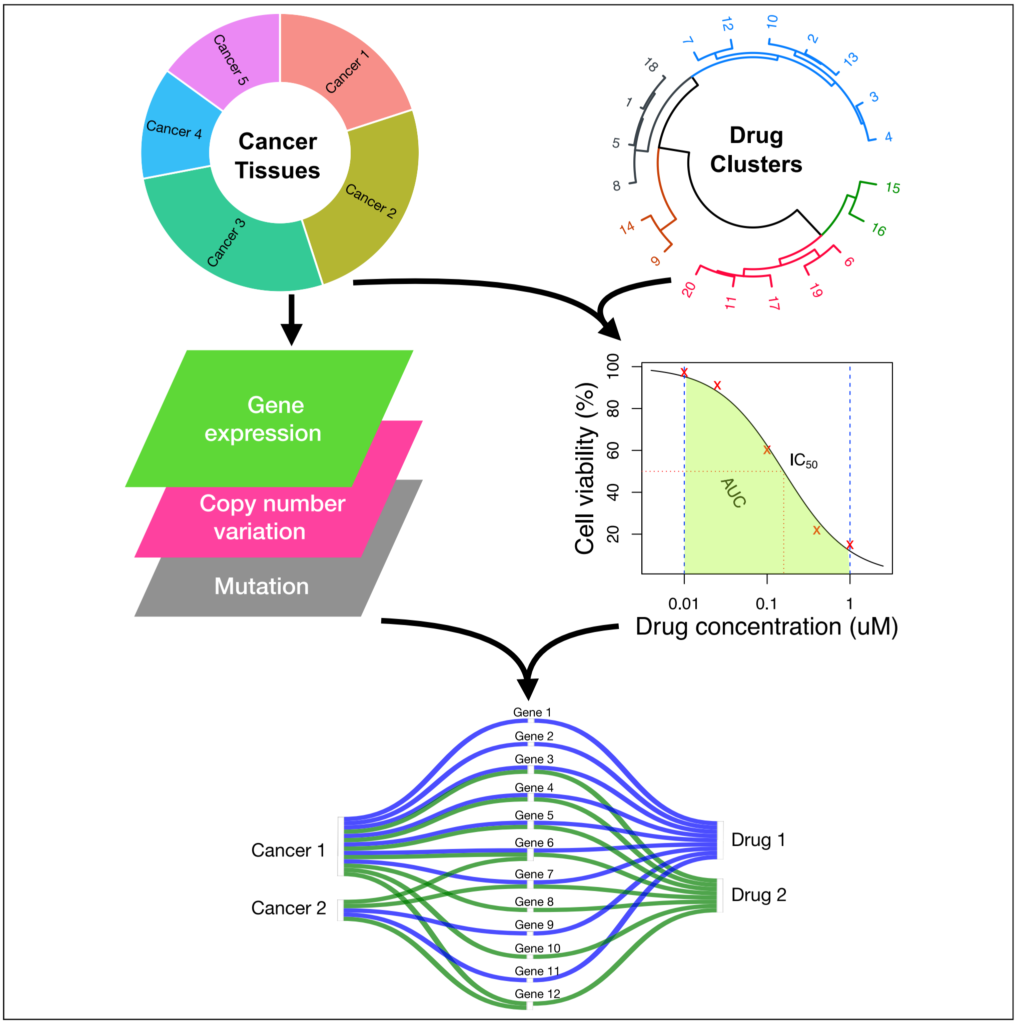
iScience

CellPress
OPEN ACCESS

Article

Tissue-specific identification of multi-omics features for pan-cancer drug response prediction



Zhi Zhao, Shixiong Wang, Manuela Zucknick, Tero Aittokallio

manuela.zucknick@medisin.uio.no (M.Z.)
t.a.aittokallio@medisin.uio.no (T.A.)

Highlights

Pan-cancer cell lines provide a test bench for exploring gene-drug relationships

Multi-omics data were integrated with pharmacological profiles for joint modeling

Mix-lasso identifies tissue-specific biomarkers predictive of multi-drug responses

Mix-lasso provides small number of stable features for drug discovery applications

Zhao et al., iScience 25, 104767
August 19, 2022 © 2022 The Author(s).
https://doi.org/10.1016/j.isci.2022.104767

iScience

CellPress
OPEN ACCESS

Article

Tissue-specific identification of multi-omics features for pan-cancer drug response prediction

Zhi Zhao,[1,2] Shixiong Wang,[1] Manuela Zucknick,[2,*] and Tero Aittokallio[1,2,3,4,*]

SUMMARY

Current statistical models for drug response prediction and biomarker identification fall short in leveraging the shared and unique information from various cancer tissues and multi-omics profiles. We developed mix-lasso model that introduces an additional sample group penalty term to capture tissue-specific effects of features on pan-cancer response prediction. The mix-lasso model takes into account both the similarity between drug responses (i.e., multi-task learning), and the heterogeneity between multi-omics data (multi-modal learning). When applied to large-scale pharmacogenomics dataset from Cancer Therapeutics Response Portal, mix-lasso enabled accurate drug response predictions and identification of tissue-specific predictive features in the presence of various degrees of missing data, drug-drug correlations, and high-dimensional and correlated genomic and molecular features that often hinder the use of statistical approaches in drug response modeling. Compared to tree lasso model, mix-lasso identified a smaller number of tissue-specific features, hence making the model more interpretable and stable for drug discovery applications.

INTRODUCTION

Targeted cancer therapies have been increasingly used during the past two decades for the treatment of certain cancer types that are driven by single oncogenic proteins (Druker et al., 2001; Tsimberidou et al., 2020); for example, HER2-positive breast cancer can be treated with HER2-targeted therapeutic agents such as trastuzumab (Vogel et al., 2002). However, our knowledge of such protein-therapy relationships is currently limited to only a few well-established links between specific cancer types and oncoprotein markers that can be used as companion diagnostics in the clinic. Personalized cancer medicine aims to target and use patient-specific genomic and molecular markers that drive the cancer or resistance development, and therefore determine the patient-specific responses to the available targeted treatments. However, due to the complexity of tumor biology and between-patient heterogeneity, targeted treatments may lead to differing or even opposite effects among patients with different cancer subtypes, yet having similar genetic or molecular backgrounds (Rowbotham et al., 2018; Gambardella et al., 2020).

Cancer tissue heterogeneity is critically important for modeling the potency and selectivity of targeted drugs across cancer types (Mannheimer et al., 2019; Lloyd et al., 2021). It has been shown that a drug inhibiting the same protein target may have drastically differing effects in different cancer tissues or cancer subtypes. For instance, inhibitors of the oncogene PI3K have shown to lead to highly varied effects (e.g., no response or extreme response) across different cancer types (Stewart et al., 2019). Therefore, a more systematic modeling of drug efficacy and identification of predictive markers beyond the target proteins requires simultaneous analysis of pharmacogenomic data from multiple cancer types or subtypes. In particular, there is a need for predictive models that can accurately estimate the tissue-specific effects on drug response profiles through integrating still rather limited sample sizes of heterogeneous cancer (sub)types, with the aim of predicting multi-drug responses by simultaneously taking into account the most relevant genomic and molecular features in a pan-cancer setting.

There are a number of publicly available data resources for large-scale pharmacogenomic screens, which include hundreds of cancer cell lines from multiple cancer types, treated with hundreds of drugs and characterized at baseline (before treatment) with multiple omics data such as gene expression, copy number variation, and point mutations (Barretina et al., 2012; Seashore-Ludlow et al., 2015). In these screens, drug response is summarized based on a dose-response curve to quantitatively score how effectively

[1]Institute for Cancer Research, Department of Cancer Genetics, Oslo University Hospital, Norway

[2]Centre for Biostatistics and Epidemiology (OCBE), Faculty of Medicine, University of Oslo, Norway

[3]Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Finland

[4]Lead contact

*Correspondence: manuela.zucknick@medisin.uio.no (M.Z.), t.a.aittokallio@medisin.uio.no (T.A.)

https://doi.org/10.1016/j.isci.2022.104767

the drug inhibits cell growth, for example, using the half-maximal inhibitory concentration (IC_{50}) or the area under the drug dose-response curve (AUC). Even though the cell line models cannot capture all the variability seen in patient tumors, these large-scale data resources provide great opportunities for estimating or even predicting of drug efficacy in a pan-cancer setting and for the development of novel statistical models for this task. However, the heterogeneous nature of the pharmacogenomic data poses challenges for predictive drug response modeling. These challenges include multivariate responses involving drug-drug similarities and frequent missing values, heterogeneous and partly unknown cancer tissues and subtypes (e.g., hidden sub-groups), high-dimensional genomic features with gene-gene correlations, and heterogeneous multi-omics profiles.

A number of statistical and machine learning models have been developed in the past years for predicting drug responses (see e.g., Ballester et al. (2022); Sharifi-Noghabi et al. (2021); Adam et al. (2020)). These models are often designed for making accurate predictions, either within a single tissue (Costello et al., 2014) or using a tissue-agnostic approach (Barretina et al., 2012), and most of the models cannot deal with missing data and other technical variability present in the high-throughput studies. Furthermore, while emphasizing accurate predictions, many of the models lack effective feature selection options, making such black-box models less practical for biological studies or clinical applications. Previously, Kim and Xing (2012) proposed tree-guided group lasso (tree lasso) for multi-response regression that leverage a hierarchical structure over multiple response variables to select relevant covariates from high-dimensional features. However, tree lasso cannot deal with heterogeneity between multiple sample groups. Huang et al. (2020) developed Tissue-guided LASSO (TG-lasso) for integrating cancer tissue of origin with genomic profiles. However, the TG-lasso pipeline repeats the analysis in each tissue type, rather than jointly modeling multiple cancer types.

To address these limitations, we developed a tissue-specific lasso model that takes advantage of our IPF-tree-lasso (Tree-guided group lasso with Integrative Penalty Factors) to capture drug-drug similarities and deal with heterogeneous high-dimensional multi-omics data (Zhao and Zucknick, 2020). For short, we call our approach mix-lasso, where the mix refers to both mixed models, a mix of multi-omics data sources, and a mix of multiple cancer types. In comparison to the existing models, the newly developed mix-lasso considers the predictive contributions of heterogeneous cancer types by borrowing the methodology from varying-coefficient mixed models (Hoover et al., 1998). To leverage pan-cancer information of the same genomic or molecular features, the tissue-specific effects are taken into account by grouping effects using the elastic net penalty (Zou and Hastie, 2005), which enables robust selection of sparse sets of multi-omics features (or markers) most predictive of drug responses across cancer types. In contrast to many other statistical models, mix-lasso can effectively deal with unmeasured drug responses, which are missing-at-random in the high-throughput screens, to make full use of the drug response profiles of each cancer type. The optimization applies the smoothing proximal gradient (SPG) method, similarly to the tree lasso (Kim and Xing, 2012), which is used as a reference comparison model in our study.

RESULTS

The Cancer Therapeutics Response Portal (CTRP) v2 is a database of large-scale cancer cell line drug screens (Seashore-Ludlow et al., 2015; Basu et al., 2013). In CTRP v2, the responses of 481 drugs are profiled across 860 cell lines from 24 primary tumor types. The genomic and molecular information of the cell lines originates from the Cancer Cell Line Encyclopedia (CCLE, Barretina et al. (2012)), including genome-wide measurements of mRNA expression, DNA copy number variation, and DNA point mutations. In the response modeling, we made use of the log2 intensity values for the genome-wide mRNA expression data (Affymetrix Human Genome U133 Plus 2.0 arrays), log2 ratio values for the genome-wide copy number variation (Affymetrix SNP Array 6.0), and binary values for the gene point mutations of selected cancer gene loci measured using mass spectrometric genotyping (OncoMap platform). We used the following criteria to preselect subsets of the drugs, cell lines, and genomic features.

- Every cancer type must have at least 15 cell lines for tissue-specific modeling, which was earlier considered large enough sample size for comparison of pan-cancer and tissue-specific prediction models (Lloyd et al., 2021).
- A part of the selected drugs must have a completely measured response sub-matrix across cell lines, which was used for a direct comparison between mix-lasso, that allows for missing responses, and tree lasso, that does not allow missing responses.

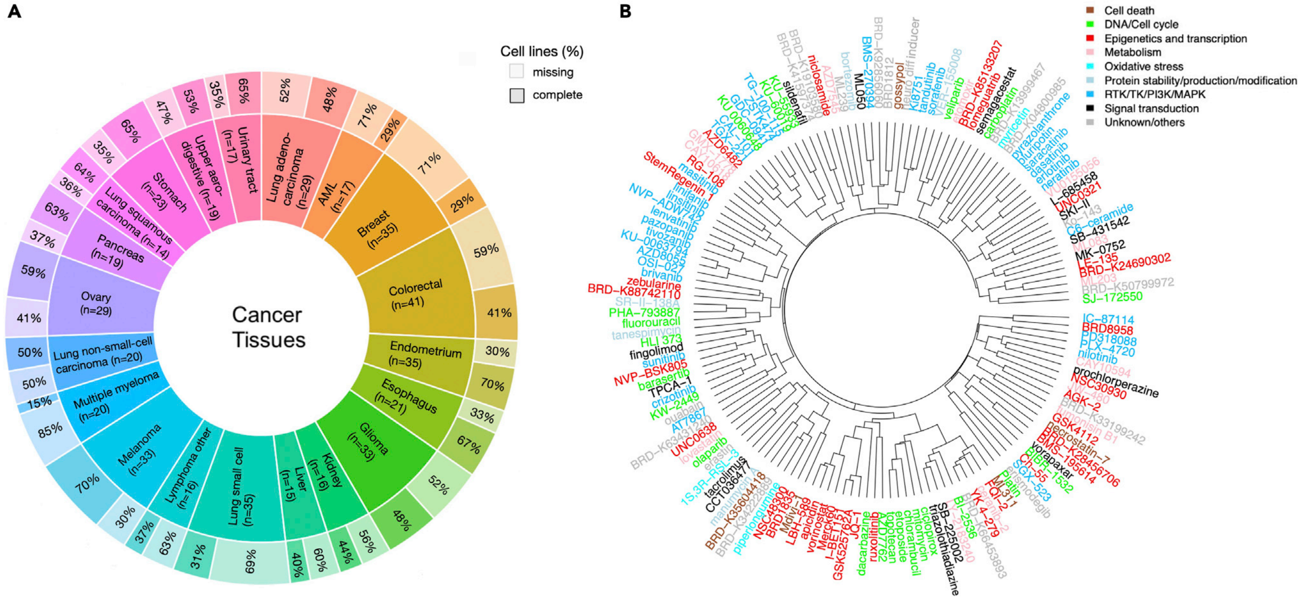


Figure 1. Distribution of the 20 cancer types and a hierarchical clustering of the drugs from CTRP

(A) The number of cell lines from each cancer type/tissue is shown in parentheses. The label "complete" denotes the portion of cell lines with complete drug response data, and "missing" the portion with some missing values in the drug response data.

(B) Clustering of the drug responses was carried out using hierarchical agglomerative complete-linkage clustering with Pearson correlation similarity measure for drug responses (AUC profiles over the cell lines). The 147 drugs shown are the common drugs shared across the repeated analyses used in the feature selection analyses below (among 10 repetitions). The colors indicate the mode of action (MoA) classes of the drugs (see Figure S1.1 for an alternative clustering of the drugs based on their structural similarity).

- Among gene expression (GEX) features, we selected the most variable features, without any missing data, such that ≥ 50% of the cumulative variance is included, similar to our previous work (Zhao and Zucknick, 2020).
- Among copy number variation (CNV) features, we selected the most variable features, without any missing data, such that ≥ 50% of the cumulative variance is included, similar to our previous work (Zhao and Zucknick, 2020).
- Mutation (MUT) features include all genes that harbor deleterious single point mutations from CCLE and have pathogenic mutation scores according to COSMIC (Catalog of Somatic Mutations In Cancer) (Tate et al., 2018), without any missing data, by following Barretina et al. (2012) and Garnett et al. (2012).

To construct the hierarchical structure of drug-drug similarity for mix-lasso's IPF-tree penalty (see Equation 4 in STAR Methods), we preselected a complete response dataset from the total 481 drugs by excluding about half of the drugs with missing values. Since mix-lasso can deal with missing data, we included as many cell lines as possible. Finally, the above preselection criteria led to ~ 200 drugs (Note that this is not an exact number, because we randomly split the cell lines into learning and validation datasets 10 times, where in every run the learning dataset forms a different complete drug response matrix with around 200 cell lines and around 200 drugs. Our analysis results focus on 147 common drugs across the 10 repeats.) and 473 cell lines from 20 cancer types as our pharmacogenomic profiling dataset. We used higher resolution histologic subtypes of primary lung tumor, due to strong heterogeneity within the primary lung cancers. The multi-omics profiling data included a total of 2069 GEX features, 8127 CNV features, and 175 MUT features preselected as input data for modeling.

Figure 1A shows the distribution of the 473 cell lines across 20 cancer types (Note that lung squamous cell carcinoma has 14 cell lines, which is fewer than the threshold of 15 in our criteria, because a few cell lines without complete genomic data were removed.), with median proportion of missing values across cancer types 51% (range 30% ~ 85%). Less than half of the cell lines have complete drug response data, making the

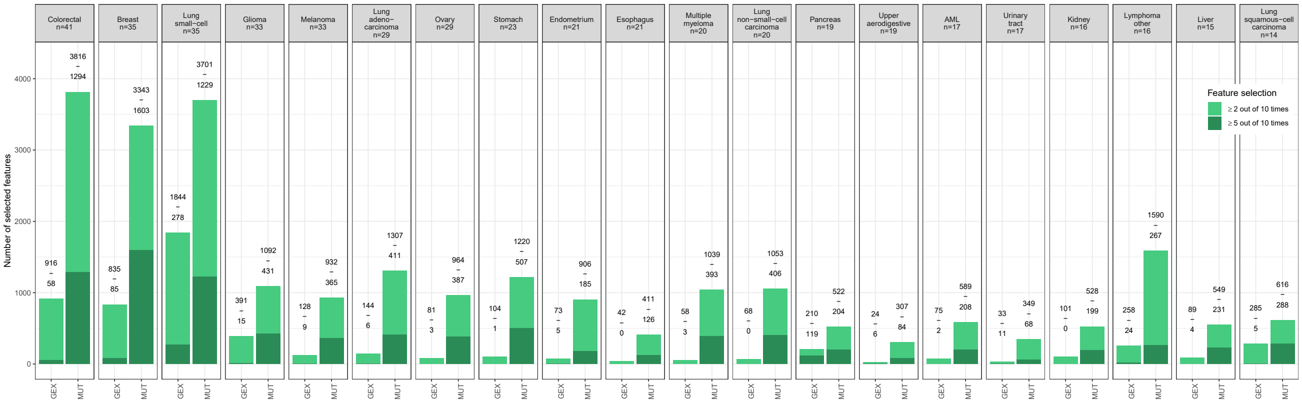


Figure 2. Feature selection with mix-lasso across the $m_{\text{drugs}} = 147$ common drugs w.r.t. the two omics data sources (GEX and MUT) and 20 cancer types (the columns)

The total numbers of features for individual data sources were $m_{\text{drugs}} \times p_{\text{GEX}} \approx 3.0 \times 10^5$ and $m_{\text{drugs}} \times p_{\text{MUT}} \approx 2.6 \times 10^4$. The two values above each bar show the numbers of features selected (i.e. nonzero coefficients; note that each feature can be counted multiple times if it was selected as a predictor for multiple drugs across the 20 cancer types). The light green (or dark green) bar shows the numbers of features selected when one of the model coefficients was nonzero at least 2 (or at least 5) out of 10 repetitions. The cancer types are ordered by sample size from colorectal to lung squamous cell carcinoma.

prediction task relatively challenging. To quantify the drug response outcome, we used the area under the drug dose-response curve (AUC) according to recent guidelines Sharifi-Noghabi et al. (2021). Figure 1B shows that the clustering of the drug responses based on the similarity of the AUC profiles across the cell lines only partly corresponds to the mode of action (MoA) classes of the drugs.

Prediction accuracy and feature selection performance

For the model evaluation, the cell lines of each cancer tissue type were split into 75% for training data and 25% for validation data. We randomly split the data into the two parts 10 times for investigating the stability of the selected features and for evaluating the average prediction performance for drug responses. A genomic or molecular feature was determined to be selected by a model if its estimated regression coefficient was nonzero at least 2 (or 5) times out of the 10 repeats. Note that we counted multiple times whether the same gene was selected as a predictor for multiple drugs, since there are both common and drug-specific predictive features among the three omics data sources (i.e., gene expression, copy number variation, and point mutations).

When using all three omics data sources, the copy number variation features did not contribute markedly to the overall drug response predictions with mix-lasso (Figure S1.2). After removing the copy number variation data, mix-lasso improved its overall prediction accuracy across the 147 drugs, w.r.t. Root Mean Squared Error (RMSE) and had similar Pearson and Spearman correlations, whereas tree lasso remained at similar level of overall prediction accuracy (Table S1). Figures S1.3 and S1.4 show the overlapping GEX and MUT features, selected by mix-lasso, when modeling either three or two omics data (i.e. removing copy number variation features); the relatively small portion of unique features identified only when modeling all the three omics data (highlighted in red) indicates that modeling of the two omics data captured most of the predictive signal from the three omics data, and suggests that the additional features selected from the GEX and MUT data compensated for the effect of the missing CNV features in the mix-lasso model. Therefore, we only use the two omics data sources, i.e., gene expression and point mutations, in the following analyses.

Figure 2 shows the feature selection performance across the 20 cancer types by the mix-lasso model. Interestingly, the point mutations were more commonly selected for overall drug response prediction by mix-lasso, even if the total number of potential GEX features was more than 10-fold higher than that of MUT features. When comparing the two frequency criteria of feature selection (i.e., ≥ 2 and ≥ 5 out of 10 times), mutation features were also more stably selected than the gene expression features, as measured by the Lance-Williams distance (16.9 for GEX vs. 9.8 for MUT) (The Lance-Williams distance measures a distance between two vectors of the numbers of selected features for one omics data source over the cancer tissue types based on criteria ≥ 2 and ≥ 5 out of 10 times, i.e., two vectors x_i and y_i $(i = 1, \cdots, n)$ have distance

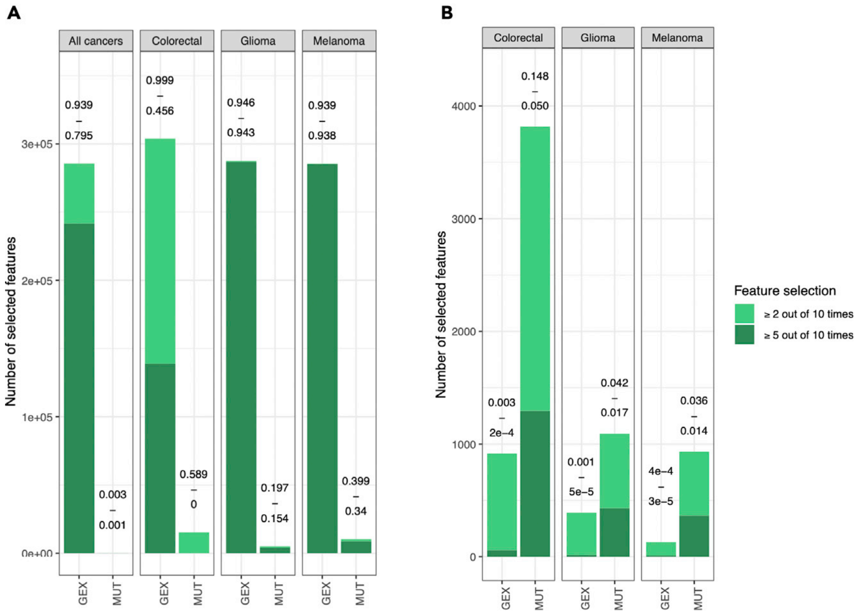


Figure 3. Feature selection with tree lasso and mix-lasso

(A) Feature selection with tree lasso across the $m_{\text{drugs}} = 147$ common drugs, either by modeling all cancer tissue types jointly (left panel) or by separately modeling each cancer type where possible (only cancer types with enough samples with complete drug response data could be used for tree lasso modeling). The two values above each bar show the estimated model sparsity (i.e., $\frac{\text{number of selected features}}{\text{number of all potential features}}$). The light green (or dark green) bar shows the numbers of features selected when one of the model coefficients was nonzero at least 2 (or at least 5) out of 10 repetitions.

(B) Feature selection with mix-lasso for the same tissue types and $m_{\text{drugs}} = 147$ common drugs. Note: the y-axis scale is different between the two panels, while the percentages above the bars are more comparable between the two models.

$\sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|}$. The smaller the Lance-Williams distance, the more similar are the two vectors.). This indicates that the mix-lasso selects a rather stable set of mutation features for drug response prediction, which are likely to provide more practical biomarker panels, compared to gene expression markers that are more challenging to use as companion diagnostics in clinical practice.

In comparison with mix-lasso, tree lasso resulted in much denser models (i.e., a more complex model with more selected features) for drug response prediction, which may become less practical in translational applications, where sparse models with fewer selected features are preferred. When modeling the complete dataset with all cancer types, tree lasso selected 95% and 79% of the GEX features based on the criteria "≥ 2 out of 10 times" and "≥ 5 out of 10 times", respectively (first panel Figure 3A). Similarly, dense models were estimated with tree lasso when separately modeling individual cancer types, which was possible for the three largest cancer types (colorectal, glioma, and melanoma; Figure 3A), which had sufficient sample size ($n > 15$ in complete drug response data) for tree lasso modeling. Notably, tree lasso selects 94% of all gene expression features for glioma and melanoma with both of the selection criteria. In contrast, mix-lasso results in reasonably sparse models for gene expression and mutation features (Figure 3B). Taken together, these results demonstrate that the mix-lasso model is able to identify sparse and robust subsets of tissue-specific genomic and molecular features for multi-drug response prediction in a pan-cancer setting.

Prediction accuracy across cancer tissues and MoA classes

To investigate in more detail the prediction performance of mix-lasso and tree lasso using all cancer types, we evaluated the rank correlation between measured AUCs and predicted AUCs for each drug and each cancer type using the two models (Figure S1.5). Interestingly, mix-lasso and tree lasso showed rather complementary prediction accuracy across the cancer types; for instance, mix-lasso predicted accurately more drug responses in colorectal (Figure S1.5a) and ovarian cancer (Figure S1.5g), whereas tree lasso made more accurate response predictions for a number of drugs in stomach (Figure S1.5h) and lung squamous cell carcinoma cancer (Figure S1.5s). Moreover, the accuracy of neither of the methods was dependent on

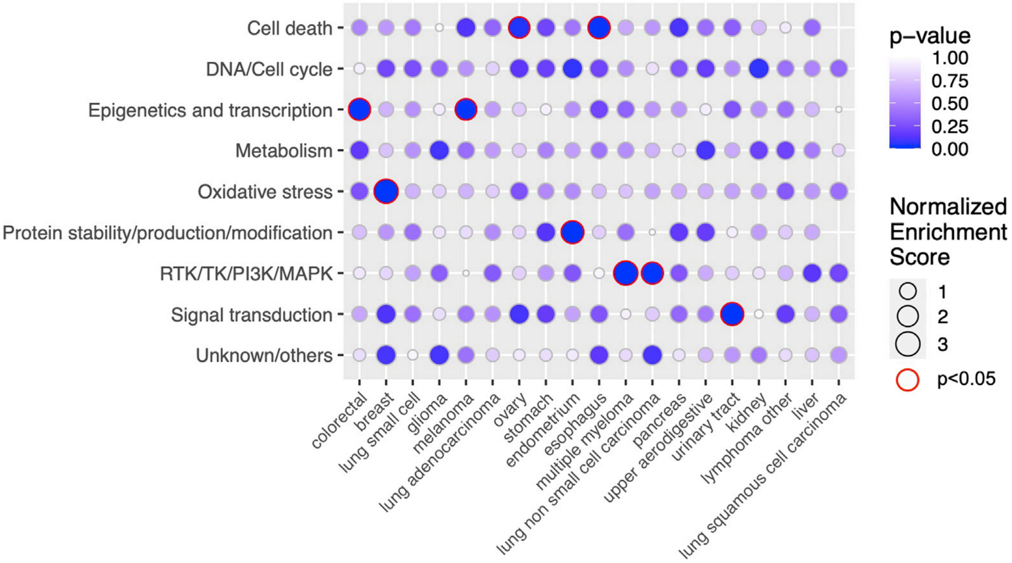


Figure 4. Drug set enrichment analysis across individual cancer tissue types and 9 drug MoA classes, where the 147 drugs were ranked by prediction accuracy of mix-lasso (Spearman correlation)

The size of a circle corresponds to a normalized enrichment score (Kolmogorov-Smirnov statistic). No circles mean missing p-value because of not enough variation in the predicted Spearman's ρ. Due to the relatively small number of drugs in the enrichment analysis, the false discovery rate was not controlled.

whether predicting targeted or non-targeted drugs, rather the methods showed again rather complementary prediction accuracies across the cancer types and drug classes (Figure S1.6). These results suggest that the two models compensate each other in the sense that when tree lasso performs poorly, the mix-lasso makes accurate predictions, and vice versa (Figure S1.7).

To investigate potential differences in the prediction accuracies across mode of action (MoA) drug classes, we grouped the drugs into 9 broad MoA classes based on their targets or inhibition mechanisms. We then applied drug set enrichment analysis (Napolitano et al., 2016) to investigate whether the mix-lasso can predict well certain classes of drugs for specific cancer tissue types (Figure 4). For example, we observed that epigenetic and transcription modulating drugs can be predicted accurately for colorectal cancer and melanoma ($p < 0.05$, Figure 4). Previous studies have also suggested that compounds that act epigenetically or modulate transcription may provide potential therapies for these cancer types (Patnaik and Anupriya, 2019; Jung et al., 2020; Strub et al., 2020; Giunta et al., 2021; Garcia-Gomez et al., 2021). As a specific example, we chose RG-108 and JQ-1 from this drug class and investigated their selected predictive features in the two cancer types (Figure 5). Even though the target proteins of RG-108 and JQ-1 were not selected by the mix-lasso model, the selected features listed in Figure 5A are connected to the drugs' target activity via Gene Ontology (GO) set enrichment analysis (Figures 5C and 5D).

For instance, it is known that the TGF-β-SMAD pathway (Wotton, 2012; Papageorgis et al., 2010; Bai and Xi, 2018) and transcription factor binding (Figure 5C) are closely related to the RG-108 target DNA methyltransferase. Similarly, the targets of JQ-1, i.e., the BET family of bromodomain proteins, are active players in transcription and epigenetics, and they can promote cancer cell proliferation and survival. In Figure 5D, enriched molecular functions, such as DNA binding and transcription factor binding, are closely related to the function of BET proteins as direct transcriptional regulators, and molecular functions of receptor tyrosine kinases and tyrosine kinases also reflect the role of BET proteins in mediating the transcription from various signals that promote cell proliferation. Furthermore, JQ-1-related proteins are also enriched in chromatin-associated cellular components, e.g., chromosome and telomeric regions (Figure S1.8), which also relate to the function of BET proteins in transcription and epigenetics. These results indicate that non-target proteins or other proteins in the target pathways predict the responses of RG-108 and JQ-1 in a tissue-specific manner.

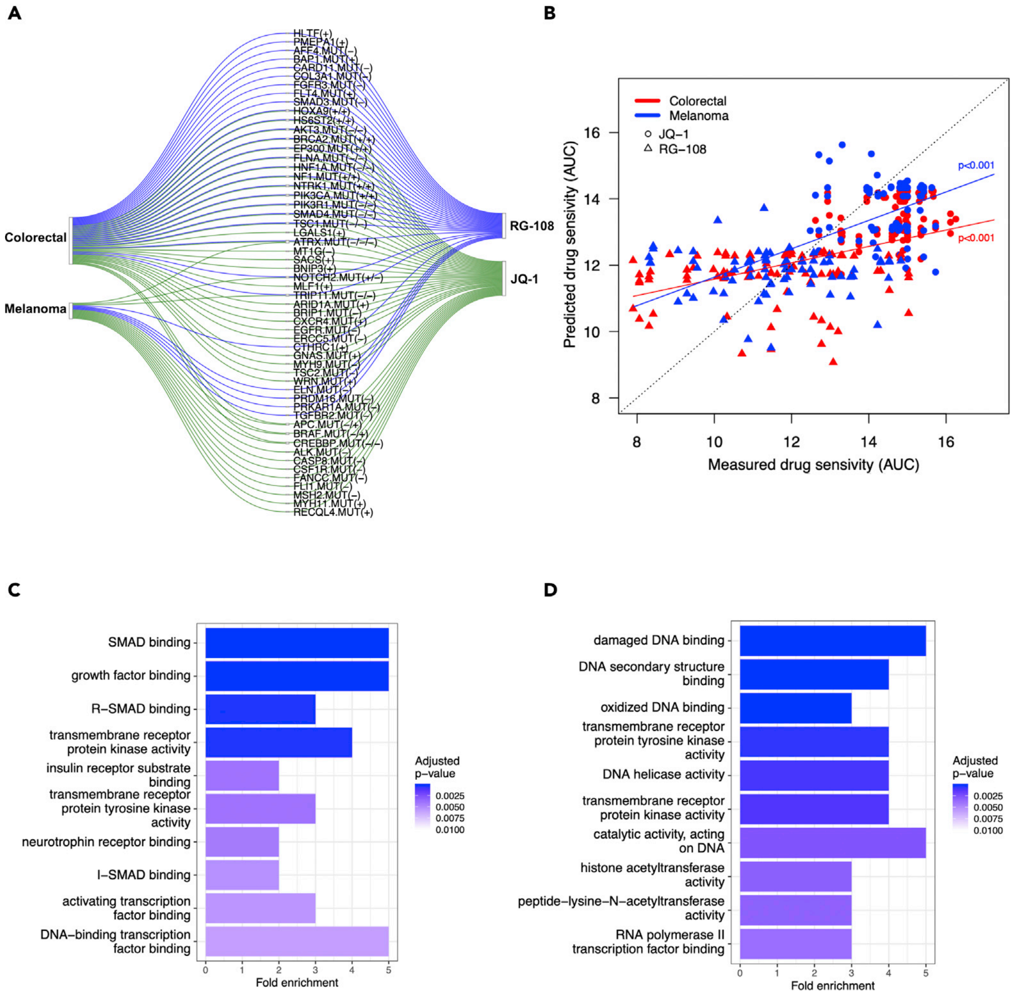


Figure 5. Set of genes predictive of drug responses and their enrichment analysis

(A) Example drugs RG-108 and JQ-1 and their identified genomic (MUT) and molecular features (GEX) linked to the two selected cancer types by mix-lasso. The selected genes were based on feature selection criteria "≥ 2 out of 10 times". Gene names with ".MUT" indicate MUT features and the rest are GEX features. "+" or "-" indicates positive or negative effect. The multiple signs correspond to distinct gene-cancer-drug response relationships (i.e., the number of connections in the sankey diagram). Note that the negative effect of "NOTCH2.MUT" corresponds to drug JQ-1 in melanoma, and the negative effect of "BRAF.MUT" corresponds to drug JQ-1 in colorectal cancer.

(B) Relationships between the measured drug response (area under the drug dose-response curve, AUC) and the predicted drug response using mix-lasso model in the validation data. The dotted diagonal line indicates perfect prediction of the drug response. The two colored lines indicate regression lines for the two selected cancer types, respectively, and the p-values show the significance of the regressions. Enrichment of GO molecular functions among the mix-lasso-selected genes predictive of responses to (C) RG-108 and (D) JQ-1. p-values in panels (C) and (D) were adjusted for multiple testing by controlling the false discovery rate with the Benjamini & Hochberg method.

Another example from the drug set enrichment analysis is the enrichment of RTK/TK/PI3IK/MAPK drug class in multiple myeloma (MM) and lung non-small-cell carcinoma (Figure 4). We therefore investigated IC-87114, PLX-4720, and TG-100-115 from this drug class in these two different types of cancers (Figure S1.9). IC-87114 and TG-100-115 are specific PI3Kδ/γ inhibitors, and previous studies have shown the effectiveness of targeting different PI3K isoforms in MM (Piddock et al., 2017; Sahin et al., 2014; Glauer et al., 2013; Ikeda et al., 2010) (Figure S1.9a). Even though the hierarchical clustering did not cluster the two PI3K inhibitors based on drug response (Figure 1B), their Spearman correlation is significant (Spearman's $\rho = 0.111$, $p = 0.021$). The genomic features that are linked to the responses of the two drugs are also biologically relevant (Figure S1.9a); for example, EGFR is the upstream factor of PI3Ks, and the effectiveness of combined CDK4/6 (targets of CDKN2A) and PI3K inhibition has been shown in other cancer models (O'Brien et al., 2020; Bonelli et al., 2017). The GO molecular functions of identified genes for drugs IC-87114 and TG-100-115 both include phosphatidylinositol 3′-kinase activity (Figures S3.9b and S3.8d). PLX-4720 is a BRAF inhibitor, and some of the identified features (e.g., MET, KRAS, and NF1) are either closely related to the signal initiation and transduction of the MAPK pathway, in which BRAF is involved in, or downstream effector molecules (e.g., MAX and MYC) of the RTK/MAPK pathway (Figure S1.9d).

DISCUSSION

In this study, we introduced the mix-lasso model to jointly analyze multi-omics data from *in vitro* pharmacogenomic screens in cell lines. Such integrative models are needed in the preclinical anticancer drug discovery process to evaluate the relationship between somatic variation and preclinical treatment responses, which can further guide future screening of phenotypic effects of anticancer compounds in preclinical model systems (Ballester et al., 2022). Large-scale pharmacogenomic screens have been carried out to date in hundreds of cancer cell line panels to provide insights into drug efficacy and potential molecular and genomic determinants of context-specific drug response through the omics profiling among the various cancer tissue types (Barretina et al., 2012; Seashore-Ludlow et al., 2015; Haverty et al., 2016). When analyzed using suitable statistical approaches, these rich data resources are expected to help finding biomarkers predictive of preclinical drug responses that can be followed-up in future studies.

However, there has been lack of effective approaches for tissue-specific statistical modeling and sparse feature selection in the pharmacogenomic data involving heterogeneous pan-cancer information and multiple omics data sources (Ali and Aittokallio, 2019; Adam et al., 2020). To that end, we proposed the mix-lasso model and demonstrated how it enables successful applications to datasets composed of mix of multiple cancer tissue types and multiple omics measurements. Mix-lasso provides a useful and timely modeling approach, since increasing number of cancer cell lines and associated phramacogenomic data for many tissue types are becoming available from the ongoing and emerging screening efforts, and we expect that the same integrated modeling approach can effectively integrate the current and future multi-omics data, with the aim to identify potential tissue-specific multi-omics features for further preclinical or clinical studies.

There exist only a few studies that have addressed the challenges posed by tissue-specific modeling based on pan-cancer multi-omics pharmacogenomic data (Mannheimer et al., 2019; Huang et al., 2020; Naulaerts et al., 2020; Lloyd et al., 2021). However, these previous studies have analyzed each cancer tissue type dataset separately, which makes it difficult to leverage the pan-cancer information and to distinguish between common and tissue-specific molecular and genomic features across multiple cancer types. Although useful statistical methodologies have also been developed in the past for complex structured data with integrated high-dimensional genomic information (see Ickstadt et al., 2018; Reel et al., 2021), most of these models lack options for capturing certain important structures in the complex data, including heterogeneity across sample groups. Many of the existing models cannot either deal with missing data, which is inherent to large-scale screens.

In many real-world applications, there exist heterogeneous sample groups that have opposite effects of the same features on treatment responses (e.g., in sub-groups of cancer patients). For instance, Figure 5A shows that BRAF mutant gene has negative effect on the response of JQ-1 in colorectal cancer, which has been reported earlier Nakamura et al. (2017), while having a positive effect on the response of a structurally similar compound I-BET151 in melanoma (Figures 5B and S1.10), which has been shown in a previous study Gallagher et al. (2014). In such cases, the proposed mix-lasso model was shown to improve both the identification of relevant features and prediction of treatment response in comparison with a reference tree lasso model in the simulation studies (Figure S2.1). If all the samples are relatively homogeneous, then there is no need for group/tissue-specific feature selection, and the standard tree lasso or IPF-tree-lasso model is sufficient.

Our results in the real-world CTRP data demonstrated that mix-lasso provides more interpretable feature selection results in terms of much fewer number of selected genes, with different features selected for different cancer tissue types and more stable feature selection results, compared to tree lasso, which selected almost all the gene expression features and had less stable mutation feature selection (Figure 3). In particular, a small number of stably selected point mutations can be expected to lead to practical companion diagnostics in translational applications, compared to gene expression levels that are often more difficult to use in clinical practice. Although mix-lasso resulted in highly sparse models with only a few selected genes, it still predicted accurately the responses of specific classes of drugs for many cancer tissue types (Figures 4 and S1.5a–S1.5s). This is partly because the selected genes were shown to be related to the target pathways or other MoA mechanisms of the predicted drugs.

Limitations of the study

Our selection criteria for the CTRP dataset might lead to biased results, because some cancer types do not have many cell lines, which may limit statistical power; missing drug responses might be not truly missing-at-random; and some of the filtered gene expression and mutation features with low variance might turn out to be important for drug response modeling. Similar to many other drug response prediction models (Ali and Aittokallio, 2019; Adam et al., 2020; Koras et al., 2020), mix-lasso was not able to make effective use of copy number variation information to predict drug responses. This might be because neighboring copy number variation features share strong correlations, and since copy number variation is often anticorrelated with point mutations (Iorio et al., 2016), making it difficult to distinguish their predictive contributions. A possible extension of mix-lasso is to employ a fused-lasso penalty for copy number variation features (Cheng et al., 2018). A related limitation of the current mix-lasso model is its limited capability to capture the exact relationships between the predictive features across different omics data sources. This could be addressed in the future studies by further employing group-lasso penalties corresponding to correlated features across different omics data sources, for example, grouping effects of GEX, CNV, and MUT of the same gene by penalizing $\|\beta_{GEX} + \beta_{CNV} + \beta_{MUT}\|_{\ell_2}$. Any prior knowledge of correlated features within one omics data source can also be addressed in the same way.

In addition to gene expression, mutations, and copy number variation, the mix-lasso model is also applicable to a broader set of omics data sources. Since drug response and resistance is known to be determined by complex genetic and epigenetic factors, it will be important to include other types of multi-omics input data, including protein modifications (Ali et al., 2018), gene isoforms (Safikhani et al., 2017), metabolite profiling (Daemen et al., 2015), and even microbiome data, once such data become available for multiple tissue types. Another way to improve drug response prediction and to search for response-predictive biomarkers would be to use protein-target and pathway information already in the feature selection process (Koras et al., 2020; Ben-Hamo et al., 2020). While patient tumors are beyond the scope of multi-drug testing, mix-lasso should be easily applicable to drug profiling in patient cells (*ex vivo*, Letai et al. (2022)) and animal models (*in vivo*, Nguyen et al. (2021)).

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- METHOD DETAILS
  - ○ Prediction problem and model formulation
  - ○ Mix-lasso model
  - ○ Optimization of mix-lasso
  - ○ Missing drug response data
  - ○ Benchmarking simulation study

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.104767.

ACKNOWLEDGMENTS

This work was supported by grants from Helse Sør-Øst (grant 2020026 to T.A.), the Norwegian Cancer Society (216104 to T.A.), the Radium Hospital Foundation, the Academy of Finland (grants 326238, 340141, 344698, 345803 to T.A.), Cancer Society of Finland, the European Union’s Horizon 2020 research and innovation program (grant 'RESCUER' 847912 to M.Z.), and the Innovative Medicines Initiative two Joint Undertaking of the European Union's Horizon 2020 research and innovation program and EFPIA and JDRF INTERNATIONAL (grant 'imSAVAR' 853988 to M.Z.). The authors thank Dr. Thomas Fleischer, Dr. Yanping Yin, and Dr. John Zobolas for fruitful discussions.

AUTHOR CONTRIBUTIONS

M.Z. and T.A. conceptualized and designed the study. Z.Z. designed the model and the computational framework, conducted the data analysis, and drafted the paper. S.W. contributed to the interpretation of the findings. S.W., M.Z., and T.A. edited the paper.

DECLARATION OF INTERESTS

The authors declare no competition of interests.

Received: April 25, 2022
Revised: June 28, 2022
Accepted: July 11, 2022
Published: August 19, 2022

REFERENCES

Adam, G., Rampášek, L., Safikhani, Z., Smirnov, P., Haibe-Kains, B., and Goldenberg, A. (2020). Machine learning approaches to drug response prediction: challenges and recent progress. NPJ Precis. Oncol. *4*, 19. https://doi.org/10.1038/s41698-020-0122-1.

Ali, M., and Aittokallio, T. (2019). Machine learning and feature selection for drug response prediction in precision oncology applications. Biophys. Rev. *11*, 31–39. https://doi.org/10.1007/s12551-018-0446-z.

Ali, M., Khan, S.A., Wennerberg, K., and Aittokallio, T. (2018). Global proteomics profiling improves drug sensitivity prediction: results from a multi-omics, pan-cancer modeling approach. Bioinformatics *34*, 1353–1362. https://doi.org/10.1093/bioinformatics/btx766.

Bai, J., and Xi, Q. (2018). Crosstalk between TGF-β signaling and epigenome. Acta Biochim. Biophys. Sin. *50*, 60–67. https://doi.org/10.1093/abbs/gmx122.

Ballester, P.J., Stevens, R., Haibe-Kains, B., Huang, R.S., and Aittokallio, T. (2022). Artificial intelligence for drug response prediction in disease models. Briefings Bioinf. *23*, bbab450. https://doi.org/10.1093/bib/bbab450.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., et al. (2012). The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603–607. https://doi.org/10.1038/nature11003.

Basu, A., Bodycombe, N., Cheah, J., Price, E., Liu, K., Schaefer, G., Ebright, R., Stewart, M., Ito, D., Wang, S., et al. (2013). An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. Cell *154*, 1151–1161. https://doi.org/10.1016/j.cell.2013.08.003.

Ben-Hamo, R., Jacob Berger, A., Gavert, N., Miller, M., Pines, G., Oren, R., Pikarsky, E., Benes, C.H., Neuman, T., Zwang, Y., et al. (2020). Predicting and affecting response to cancer therapy based on pathway-level biomarkers. Nat. Commun. *11*, 3296. https://doi.org/10.1038/s41467-020-17090-y.

Bonelli, M.A., Digiacomo, G., Fumarola, C., Alfieri, R., Quaini, F., Falco, A., Madeddu, D., La Monica, S., Cretella, D., Ravelli, A., et al. (2017). Combined inhibition of cdk4/6 and pi3k/akt/mtor pathways induces a synergistic anti-tumor effect in malignant pleural mesothelioma cells. Neoplasia *19*, 637–648. https://doi.org/10.1016/j.neo.2017.05.003.

Boulesteix, A.-L., De Bin, R., Jiang, X., and Fuchs, M. (2017). IPF-LASSO: integrative L_1-penalized regression with penalty factors for prediction based on multi-omics data. Comput. Math. Methods Med. *2017*, 7691937. https://doi.org/10.1155/2017/7691937.

Bradic, J., Claeskens, G., and Gueuning, T. (2020). Fixed effects testing in high-dimensional linear mixed models. J. Am. Stat. Assoc. *115*, 1835–1850. https://doi.org/10.1080/01621459.2019.1660172.

Cheng, Y., Dai, J.Y., Wang, X., and Kooperberg, C. (2018). Identifying disease-associated copy number variations by a doubly penalized regression model. Biotechnol. Adv. *74*, 1341–1350. https://doi.org/10.1111/biom.12920.

Costello, J.C., Heiser, L.M., Georgii, E., Gönen, M., Menden, M.P., Wang, N.J., Bansal, M., Ammad-ud-din, M., Hintsanen, P., Khan, S.A., et al. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. Nat. Biotechnol. *32*, 1202–1212. https://doi.org/10.1038/nbt.2877.

Daemen, A., Peterson, D., Sahu, N., McCord, R., Du, X., Liu, B., Kowanetz, K., Hong, R., Moffat, J., Gao, M., et al. (2015). Metabolite profiling stratifies pancreatic ductal adenocarcinomas into subtypes with distinct sensitivities to metabolic inhibitors. Proc. Natl. Acad. Sci. USA *112*, E4410–E4417. https://doi.org/10.1073/pnas.1501605112.

Druker, B.J., Sawyers, C.L., Kantarjian, H., Resta, D.J., Reese, S.F., Ford, J.M., Capdeville, R., and Talpaz, M. (2001). Activity of a specific inhibitor of the bcr-abl tyrosine kinase in the blast crisis of chronic myeloid leukemia and acute lymphoblastic leukemia with the philadelphia chromosome. N. Engl. J. Med. *344*, 1038–1042. https://doi.org/10.1056/nejm200104053441402.

Fan, Y., and Li, R. (2012). Variable selection in linear mixed effects models. Ann. Stat. *40*, 2043–2068. https://doi.org/10.1214/12-aos1028.

Gallagher, S.J., Mijatov, B., Gunatilake, D., Tiffen, J.C., Gowrishankar, K., Jin, L., Pupo, G.M., Cullinane, C., Prinjha, R.K., Smithers, N., et al. (2014). The epigenetic regulator I-BET151 induces BIM-dependent apoptosis and cell cycle arrest of human melanoma cells. J. Invest. Dermatol. *134*, 2795–2805. https://doi.org/10.1038/jid.2014.243.

Gambardella, V., Tarazona, N., Cejalvo, J.M., Lombardi, P., Huerta, M., Roselló, S., Fleitas, T., Roda, D., and Cervantes, A. (2020). Personalized medicine: recent progress in cancer therapy. Cancers *12*, 1009. https://doi.org/10.3390/cancers12041009.

Garcia-Gomez, A., Li, T., de la Calle-Fabregat, C., Rodríguez-Ubreva, J., Ciudad, L., Català-Moll, F., Godoy-Tena, G., Martín-Sánchez, M., San-Segundo, L., Muntión, S., et al. (2021). Targeting aberrant dna methylation in mesenchymal stromal cells as a treatment for myeloma bone disease. Nat. Commun. *12*, 421. https://doi.org/10.1038/s41467-020-20715-x.

Garnett, M.J., Edelman, E.J., Heidorn, S.J., Greenman, C.D., Dastur, A., Lau, K.W., Greninger, P., Thompson, I.R., Luo, X., Soares, J., et al. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature *483*, 570–575. https://doi.org/10.1038/nature11005.

Giunta, E.F., Arrichiello, G., Curvietto, M., Pappalardo, A., Bosso, D., Rosanova, M., Diana, A., Giordano, P., Petrillo, A., Federico, P., et al. (2021). Epigenetic regulation in melanoma: facts and hopes. Cells *10*, 2048. https://doi.org/10.3390/cells10082048.

Glauer, J., Pletz, N., Schön, M., Schneider, P., Liu, N., Ziegelbauer, K., Emmert, S., Wulf, G.G., and Schön, M.P. (2013). A novel selective small-molecule PI3K inhibitor is effective against human multiple myeloma in vitro and in vivo. Blood Cancer J. *3*, e141. https://doi.org/10.1038/bcj.2013.37.

Haverty, P.M., Lin, E., Tan, J., Yu, Y., Lam, B., Lianoglou, S., Neve, R.M., Martin, S., Settleman, J., Yauch, R.L., and Bourgon, R. (2016). Reproducible pharmacogenomic profiling of cancer cell line panels. Nature *533*, 333–337. https://doi.org/10.1038/nature17987.

Hoover, D., Rice, J.A., Wu, C.O., and Yang, L.-p. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. Biometrika *85*, 809–822. https://doi.org/10.1093/biomet/85.4.809.

Huang, E.W., Bhope, A., Lim, J., Sinha, S., and Emad, A. (2020). Tissue-guided lasso for prediction of clinical drug response using preclinical samples. PLoS Comput. Biol. *16*. e1007607–22. https://doi.org/10.1371/journal.pcbi.1007607.

Ickstadt, K., Schäfer, M., and Zucknick, M. (2018). Toward integrative bayesian analysis in molecular biology. Annu. Rev. Stat. Appl. *5*, 141–167. https://doi.org/10.1146/annurev-statistics-031017-100438.

Ikeda, H., Hideshima, T., Fulciniti, M., Perrone, G., Miura, N., Yasui, H., Okawa, Y., Kiziltepe, T., Santo, L., Vallet, S., et al. (2010). PI3K/p110δ is a novel therapeutic target in multiple myeloma. Blood *116*, 1460–1468. https://doi.org/10.1182/blood-2009-06-222943.

Iorio, F., Knijnenburg, T.A., Vis, D.J., Bignell, G.R., Menden, M.P., Schubert, M., Aben, N., Gonçalves, E., Barthorpe, S., Lightfoot, H., et al. (2016). A landscape of pharmacogenomic interactions in cancer. Cell *166*, 740–754. https://doi.org/10.1016/j.cell.2016.06.017.

Jung, G., Hernández-Illán, E., Moreira, L., Balaguer, F., and Goel, A. (2020). Epigenetics of colorectal cancer: biomarker and therapeutic potential. Nat. Rev. Gastroenterol. Hepatol. *17*, 111–130. https://doi.org/10.1038/s41575-019-0230-y.

Kim, S., and Xing, E.P. (2012). Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eqtl mapping. Ann. Appl. Stat. *6*. https://doi.org/10.1214/12-aoas549.

Koras, K., Juraeva, D., Kreis, J., Mazur, J., Staub, E., and Szczurek, E. (2020). Feature selection strategies for drug sensitivity prediction. Sci. Rep. *10*, 9377. https://doi.org/10.1038/s41598-020-65927-9.

Letai, A., Bhola, P., and Welm, A.L. (2022). Functional precision oncology: testing tumors with drugs to identify vulnerabilities and novel combinations. Cancer Cell *40*, 26–35. https://doi.org/10.1016/j.ccell.2021.12.004.

Li, Y., Wang, L., Zhou, J., and Ye, J. (2019). Multi-task learning based survival analysis for multi-source block-wise missing data. Neurocomputing *364*, 95–107. https://doi.org/10.1016/j.neucom.2019.07.010.

Lloyd, J.P., Soellner, M.B., Merajver, S.D., and Li, J.Z. (2021). Impact of between-tissue differences on pan-cancer predictions of drug sensitivity. PLoS Comput. Biol. *17*. e1008720–25. https://doi.org/10.1371/journal.pcbi.1008720.

Lv, J., and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. Ann. Stat. *37*, 3498–3528. https://doi.org/10.1214/09-aos683.

Mannheimer, J.D., Duval, D.L., Prasad, A., and Gustafson, D.L. (2019). A systematic analysis of genomics-based modeling approaches for prediction of drug response to cytotoxic chemotherapies. BMC Med. Genom. *12*, 87. https://doi.org/10.1186/s12920-019-0519-2.

Nakamura, Y., Hattori, N., Iida, N., Yamashita, S., Mori, A., Kimura, K., Yoshino, T., and Ushijima, T. (2017). Targeting of super-enhancers and mutant braf can suppress growth of braf-mutant colon cancer cells via repression of mapk signaling pathway. Cancer Lett. *402*, 100–109. https://doi.org/10.1016/j.canlet.2017.05.017.

Napolitano, F., Sirci, F., Carrella, D., and di Bernardo, D. (2016). Drug-set enrichment analysis: a novel tool to investigate drug mode of action. Bioinformatics *32*, 235–241. https://doi.org/10.1093/bioinformatics/btv536.

Naulaerts, S., Menden, M.P., and Ballester, P.J. (2020). Concise polygenic models for cancer-specific identification of drug-sensitive tumors from their multi-omics profiles. Biomolecules *10*, 963. https://doi.org/10.3390/biom10060963.

Nguyen, L.C., Naulaerts, S., Bruna, A., Ghislat, G., and Ballester, P.J. (2021). Predicting cancer drug response in vivo by learning an optimal feature selection of tumour molecular profiles. Biomedicines *9*, 1319. https://doi.org/10.3390/biomedicines9101319.

O'Brien, N.A., McDermott, M.S.J., Conklin, D., Luo, T., Ayala, R., Salgar, S., Chau, K., DiTomaso, E., Babbar, N., Su, F., et al. (2020). Targeting activated pi3k/mtor signaling overcomes acquired resistance to cdk4/6-based therapies in preclinical models of hormone receptor-positive breast cancer. Breast Cancer Res. *22*, 89. https://doi.org/10.1186/s13058-020-01320-8.

Papageorgis, P., Lambert, A.W., Ozturk, S., Gao, F., Pan, H., Manne, U., Alekseyev, Y.O., Thiagalingam, A., Abdolmaleky, H.M., Lenburg, M., and Thiagalingam, S. (2010). Smad signaling is required to maintain epigenetic silencing during breast cancer progression. Cancer Res. *70*, 968–978. https://doi.org/10.1158/0008-5472.can-09-1872.

Patnaik, S., and Anupriya. (2019). Drugs targeting epigenetic modifications and plausible therapeutic strategies against colorectal cancer. Front. Pharmacol. *10*, 588. https://doi.org/10.3389/fphar.2019.00588.

Piddock, R.E., Loughran, N., Marlein, C.R., Robinson, S.D., Edwards, D.R., Yu, S., Pillinger, G.E., Zhou, Z., Zaitseva, L., Auger, M.J., et al. (2017). PI3Kδ and PI3Kγ isoforms have distinct functions in regulating pro-tumoural signalling in the multiple myeloma microenvironment. Blood Cancer J. *7*, e539. https://doi.org/10.1038/bcj.2017.16.

Reel, P.S., Reel, S., Pearson, E., Trucco, E., and Jefferson, E. (2021). Using machine learning approaches for multi-omics data analysis: a review. Biotechnol. Adv. *49*, 107739. https://doi.org/10.1016/j.biotechadv.2021.107739.

Rowbotham, S.P., Li, F., Dost, A.F.M., Louie, S.M., Marsh, B.P., Pessina, P., Anbarasu, C.R., Brainson, C.F., Tuminello, S.J., Lieberman, A., et al. (2018). H3k9 methyltransferases and demethylases control lung tumor-propagating cells and lung cancer progression. Nat. Commun. *9*, 4559. https://doi.org/10.1038/s41467-018-07077-1.

Safikhani, Z., Smirnov, P., Thu, K.L., Silvester, J., El-Hachem, N., Quevedo, R., Lupien, M., Mak, T.W., Cescon, D., and Haibe-Kains, B. (2017). Gene isoforms as expression-based biomarkers predictive of drug response in vitro. Nat. Commun. *8*, 1126. https://doi.org/10.1038/s41467-017-01153-8.

Sahin, I., Moschetta, M., Mishima, Y., Glavey, S.V., Tsang, B., Azab, F., Manier, S., Zhang, Y., Maiso, P., Sacco, A., et al. (2014). Distinct roles of class i pi3k isoforms in multiple myeloma cell survival and dissemination. Blood Cancer J. *4*, e204. https://doi.org/10.1038/bcj.2014.24.

Schelldorder, J., Bühlmann, P., and van de Geer, S. (2011). Estimation for high-dimensional linear mixed-effects models using l_1-penalization. Scand. J. Stat. *38*, 197–214. https://doi.org/10.1111/j.1467-9469.2011.00740.x.

Seashore-Ludlow, B., Rees, M.G., Cheah, J.H., Cokol, M., Price, E.V., Coletti, M.E., Jones, V., Bodycombe, N.E., Soule, C.K., Gould, J., et al. (2015). Harnessing connectivity in a large-scale small-molecule sensitivity dataset. Cancer Discov. *5*, 1210–1223. https://doi.org/10.1158/2159-8290.cd-15-0235.

Sharifi-Noghabi, H., Jahangiri-Tazehkand, S., Smirnov, P., Hon, C., Mammoliti, A., Nair, S.K., Mer, A.S., Ester, M., and Haibe-Kains, B. (2021). Drug sensitivity prediction from cell line-based pharmacogenomics data: guidelines for developing machine learning models. Briefings Bioinf. *22*, bbab294. https://doi.org/10.1093/bib/bbab294.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. J. Comput. Graph Stat. *22*, 231–245. https://doi.org/10.1080/10618600.2012.681250.

Smirnov, P., Safikhani, Z., El-Hachem, N., Wang, D., She, A., Olsen, C., Freeman, M., Selby, H., Gendoo, D.M., Grossmann, P., et al. (2016). PharmacoGx: an R package for analysis of large pharmacogenomic datasets. Bioinformatics *32*, 1244–1246. https://doi.org/10.1093/bioinformatics/btv723.

Stewart, A., Coker, E.A., Pölsterl, S., Georgiou, A., Minchom, A.R., Carreira, S., Cunningham, D., O'Brien, M.E., Raynaud, F.I., de Bono, J.S., et al. (2019). Differences in signaling patterns on pi3k inhibition reveal context specificity in kras-mutant cancers. Mol. Cancer Therapeut. *18*, 1396–1404. https://doi.org/10.1158/1535-7163.mct-18-0727.

Strub, T., Ballotti, R., and Bertolotto, C. (2020). The "art" of epigenetics in melanoma: from histone "alterations, to resistance and therapies". Theranostics *10*, 1777–1797. https://doi.org/10.7150/thno.36218.

Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., et al. (2018). Cosmic: the catalogue of somatic mutations in cancer. Nucleic Acids Res. *47*, D941–D947. https://doi.org/10.1093/nar/gky1015.

Tsimberidou, A.M., Fountzilas, E., Nikanjam, M., and Kurzrock, R. (2020). Review of precision cancer medicine: evolution of the treatment paradigm. Cancer Treat Rev. *86*, 102019. https://doi.org/10.1016/j.ctrv.2020.102019.

Vogel, C.L., Cobleigh, M.A., Tripathy, D., Gutheil, J.C., Harris, L.N., Fehrenbacher, L., Slamon, D.J., Murphy, M., Novotny, W.F., Burchmore, M., et al. (2002). Efficacy and safety of trastuzumab as a single agent in first-line treatment of her2-overexpressing metastatic breast cancer. J. Clin. Oncol. *20*, 719–726. https://doi.org/10.1200/jco.2002.20.3.719.

Wotton, D. (2012). Tgf-β drives dna demethylation. Mol. Cell *46*, 556–557. https://doi.org/10.1016/j.molcel.2012.05.031.

Zhao, Z., and Zucknick, M. (2020). Structured penalized regression for drug sensitivity prediction. J. R. Stat. Soc. C Appl. Stat. *69*, 525–545. https://doi.org/10.1111/rssc.12400.

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. J. R. Stat. Soc. B Stat. Methodol. *67*, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Pharmacological profiling,i.e. the area under thedrug-dose response curve	The Cancer TherapeuticsResponse Portal (CTRP)	https://ocg.cancer.gov/programs/ctd2/data-portal
Multi-omics data of CTRP,i.e. gene expression (GEX), copy number variation (CNV),point mutation (MUT)	PharmacoDB	From R/Bioconductor package**PharmacoGx** directly
Software and algorithms		
R3.6.0	This study	https://www.r-project.org
mix-lasso	This study	https://github.com/zhizuio/mixlasso

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Tero Aittokallio (t.a.aittokallio@medisin.uio.no).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Data: The CTRP v2 pharmacological data are publicly available at https://ocg.cancer.gov/programs/ctd2/data-portal. The corresponding genomic data were obtained using the freely available R package PharmacoGx (Smirnov et al., 2016).
- Code: The R code for the CTRP and simulated data analysis have been made available at https://github.com/zhizuio/mixlasso_example. The R-package **mixlasso** for our mix-lasso model is available on GitHub at https://github.com/zhizuio/mixlasso.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Prediction problem and model formulation

Let us suppose drug responses are profiled in n cell lines for m drugs, hence forming a $n \times m$ drug response matrix $\mathbf{Y}$. The n cell lines originate from T cancer tissue types, or in general T sample groups (e.g., patient samples), and the t-th tissue type has n_t samples, $\sum_{t=1}^{T} n_t = n$. We further suppose that high-dimensional genomic and molecular features originate from S omics data sources (e.g., gene expression, copy number variation and point mutations), in total $p = \sum_{s=1}^{S} p_s$. The t-th cancer tissue type has a multi-omics predictor matrix

$$\left[\mathbf{X}_{(t)1},\cdots,\mathbf{X}_{(t)s},\cdots,\mathbf{X}_{(t)S}\right] = \mathbf{X}_{(t)} = \left\{x_{(t)ij}\right\},\quad \left(i = 1,\cdots,n_t; j = 1,\cdots,p_s\right)$$

where $\mathbf{X}_{(t)s}$ is a $n_t \times p_s$ matrix representing the s-th omics data source corresponding to the t-th tissue type. The full omics data $\mathbf{X} = [\mathbf{X}_{(1)};\ldots;\mathbf{X}_{(T)}]$ is constructed by stacking by rows.

To predict the multi-drug responses $\mathbf{Y}$ using the multi-omics profiling data $\mathbf{X}$, it is necessary to take into account (i) drug-drug similarities (i.e. correlations between the response variables) that may mutually support the prediction of correlated drugs, and (ii) gene-gene correlations and heterogeneity between the multi-omics data sources that include jointly or separately predictive features. We previously proposed a structured penalized regression model, IPF-tree-lasso (Zhao and Zucknick, 2020), which captures these two aspects, but not the heterogeneous contribution of multiple cancer tissue types to drug response modeling. IPF-tree-lasso builds on multivariate linear regression, and it minimizes the sum of squares of residuals (Frobenius norm) of differences between predicted and measured responses, penalized by a IPF-tree-structure penalty term:

$$\frac{1}{2mn}\|\mathbf{Y} - \mathbf{1}_n\beta_0^\top - \mathbf{X}B\|_F^2 + \mathbf{pen}_{\mathrm{IPF-tree}}(B), \qquad \text{(Equation 1)}$$

where the penalty term $\mathbf{pen}_{\mathrm{IPF-tree}}(B)$ uses integrative penalty factors to penalize different omics data sources differently (Boulesteix et al., 2017) and uses a tree-structure ℓ_1/ℓ_2-penalty (Kim and Xing, 2012) to take into account a hierarchical structure of correlations between $\mathbf{Y}$ columns, which encourages the model to identify similar sets of genomic and molecular features for drugs with similar responses.

Mix-lasso model

Here, we introduce varying coefficients into the IPF-tree-lasso model, which makes it possible to estimate tissue-specific feature effects in a pan-cancer setting. For the t-th cancer tissue dataset $\mathbf{Y}_{(t)}$ and $\mathbf{X}_{(t)}$, we estimate the tissue-specific feature effect matrix $B_{(t)}$ through a linear model

$$\mathbf{Y}_{(t)} = \mathbf{1}_{n_t}\boldsymbol{\beta}_{(t)0}^\top + \mathbf{X}_{(t)}B_{(t)} + \mathbf{E}_{(t)},$$

where $\beta_{(t)0}$ denotes the intercept vector and $\mathbf{E}_{(t)} = (\varepsilon_{(t)1}, \cdots, \varepsilon_{(t)m})$ is a noise matrix with each column $\varepsilon_{(t)\bullet} \sim \mathcal{N}(0, \sigma_\varepsilon^2\mathbb{I}_{n_t})$. However, since the drug response profiles from the same cancer tissue are often correlated, a random effect $u_t \sim \mathcal{N}(0, \sigma_u^2)$ is added to take into account the correlations between drug responses. The joint model of all cancer tissue types data becomes

$$\begin{bmatrix} \mathbf{Y}_{(1)} \\ \vdots \\ \mathbf{Y}_{(T)} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_{n_1}\beta_{(1)0} \\ \vdots \\ \mathbf{1}_{n_T}\beta_{(T)0} \end{bmatrix} + \begin{bmatrix} \mathbf{X}_{(1)}B_{(1)} \\ \vdots \\ \mathbf{X}_{(T)}B_{(T)} \end{bmatrix} + \mathbf{Z}u\mathbf{1}_m^\top + \begin{bmatrix} \mathbf{E}_{(1)} \\ \vdots \\ \mathbf{E}_{(T)} \end{bmatrix} \qquad \text{(Equation 2)}$$

where $\mathbf{Z}$ is a $n \times T$ dummy variable of the cancer tissue types, $\boldsymbol{u} = (u_1, \cdots, u_T)^\top$. We also assume the random effects u_t and noise terms $\mathbf{E}_{(t)}$ $(t = 1, \cdots, T)$ are mutually independent.

Instead of directly minimizing the (penalized) sum of squared residuals, similarly to the IPF-tree-lasso or tree-lasso, we need to maximize the penalized log likelihood function to account for the random effects in the tissue-specific IPF-tree-lasso model. The negative log likelihood of the t-th tissue corresponding to the k-th drug response variable $\boldsymbol{y}_{(t)k}$ is

$$\begin{aligned} -\ell\left(\boldsymbol{y}_{(t)k}; \beta_{(t)0}, \beta_{(t)k}, \sigma_u^2, \sigma_\varepsilon^2\right) &= \frac{n_t}{2}\log(2\pi) + \frac{1}{2}\log|V(t)| \\ &- \frac{1}{2}\left(\boldsymbol{y}_{(t)k} - \mathbf{1}_{n_t}\beta_{(t)0} - \mathbf{X}_{(t)}\beta_{(t)k}\right)^\top V_{(t)}^{-1}\left(\boldsymbol{y}_{(t)k} - \mathbf{1}_{n_t}\beta_{(t)0} - \mathbf{X}_{(t)}\beta_{(t)k}\right), \end{aligned} \qquad \text{(Equation 3)}$$

where the covariance matrix $V_{(t)}$ is $n_t \times n_t$ dimensional with diagonals $\sigma_u^2 + \sigma_\varepsilon^2$ and off-diagonals σ_u^2. The variance of the random effect σ_u^2 is a nuisance parameter, since we focus on prediction of drug responses and feature selection (i.e. estimation of feature effects), rather than correlation within a cancer tissue type. The variance σ_u^2 is not straightforward to estimate because of often limited sample sizes of each cancer tissue in practice, and computational challenges associated with simultaneous estimation of the high-dimensional feature effects. To simplify the optimization problem, we use a proxy $\tilde{V} = \mathbb{I}_n + \mathbf{Z}\mathcal{M}\mathbf{Z}^\top$ for $\mathrm{diag}\{V_{(1)}, \cdots, V_{(T)}\}$, where $\mathcal{M} = (\log n)\mathbb{I}_T$. Fan and Li (2012) proved in linear mixed effects models that the proxy matrix ensures the model selection consistency, i.e., weak oracle property of coefficient estimators in the sense of Lv and Fan (2009). A slightly different proxy with $\mathcal{M} = \frac{2}{3T}\mathbb{I}_T$ was proposed by Bradic et al. (2020), which does not result in model selection consistency, but has a slightly higher power for the fixed effects in simulations.

For the purpose of drug response prediction, the random effect u_t $(t = 1, \cdots, T)$ can be predicted by the maximum a posteriori principle which is essentially its conditional mean given data and model parameters. We need this estimator for predicting a differing effect for each cancer type, since the average effect across all cancer types is zero. Similar to Schelldorder et al. (2011), we define

$$
\begin{aligned}
\tilde{u}_t &= \underset{u_t}{\operatorname{argmin}} f\left(u_t | \mathbf{Y}_1, \cdots, \mathbf{Y}_{(t)}, \beta_{(t)0}, \mathbf{B}, \sigma_u^2\right) \\
&= \underset{u_t}{\operatorname{argmin}} \frac{f\left(\mathbf{Y}_{(t)} \middle| u_t, \beta_{(t)0}, \mathbf{B}_{(t)}, \sigma_u^2\right) f(u_t)}{f\left(\mathbf{Y}_{(t)} \middle| \beta_{(t)0}, \mathbf{B}_{(t)}, \sigma_u^2\right)} \\
&= \underset{u_t}{\operatorname{argmin}} \left\{ \sum_{k=1}^{m} \frac{1}{\sigma_\varepsilon^2} \| \mathbf{y}_{(t)k} - \mathbf{1}_{n_t} \beta_{(t)0} - \mathbf{X}_{(t)} \beta_{(t)k} - \mathbf{1}_{n_t} u_t \|^2 + u_t^2 / \sigma_u^2 \right\} \\
&= \left(m \mathbf{1}_{n_t}^\top \mathbf{1}_{n_t} + \sigma_\varepsilon^2 / \sigma_u^2 \right)^{-1} \mathbf{1}_{n_t}^\top \sum_{k=1}^{m} \left(\mathbf{y}_{(t)k} - \mathbf{1}_{n_t} \beta_{(t)0} - \mathbf{X}_{(t)} \beta_{(t)k} \right),
\end{aligned}
$$

where f is the density of the corresponding Gaussian distributed variable. The $\sigma_\epsilon^2 / \sigma_u^2$ can be obtained by $\mathcal{M}^{-1}$ in the proxy matrix of Fan and Li (2012), and $\widehat{\beta}_{(t)0}$ and $\widehat{\beta}_{(t)k}$ are estimated by the Smoothing proximal gradient (SPG) method proposed in tree lasso (Kim and Xing, 2012). From this u_t is predicted by

$$
\widehat{u}_t = \left(m \mathbf{1}_{n_t}^\top \mathbf{1}_{n_t} + (\log n)^{-1} \right)^{-1} \mathbf{1}_{n_t}^\top \sum_{k=1}^{m} \left(\mathbf{y}_{(t)k} - \mathbf{1}_{n_t} \widehat{\boldsymbol{\beta}}_{(t)0} - \mathbf{X}_{(t)} \widehat{\boldsymbol{\beta}}_{(t)k} \right).
$$

The model (Equation 2) estimates multiple tissue-specific effects of each genomic and molecular feature on prediction of a particular drug response. The model also allows for grouping effects of multiple effects originating from the same feature, for example, one gene may have similar effects in multiple cancer types. For the j-th gene corresponding to the k-th drug, one needs to estimate the regression coefficients $\beta_{(1:T)jk} = (\beta_{(t)jk}, \cdots, \beta_{(T)jk})^\top$. A sparse group lasso penalty (Simon et al., 2013) is used for the grouping effect of $\beta_{(1:T)jk}$, i.e., $(1 - \alpha)\gamma\sqrt{T} \|\beta_{(1:T)jk}\|_{\ell_2} + \alpha\gamma \|\beta_{(1:T)jk}\|_{\ell_1}$, where $\gamma > 0$, $\alpha \in [0, 1]$ and $\|\beta_{(1:T)jk}\|_{\ell_q} = \left(\sum_{t=1}^{T} \left| \beta_{(t)jk} \right|^q \right)^{1/q}$ $(q \in \mathbb{N}^+)$. For ℓ_q-norm of a matrix, $\|\boldsymbol{B}\|_{\ell_q} = \left(\sum_{k=1}^{m} \sum_{j=1}^{p} \left| \beta_{jk} \right|^q \right)^{1/q}$.

Finally, the mix-lasso model has the objective function

$$
\begin{aligned}
&- \sum_{t=1}^{T} \sum_{k=1}^{m} \ell\left(\beta_{(t)0}, \beta_{(t)k}, \sigma_u^2, \sigma_\varepsilon^2\right) + \sum_{t=1}^{T} \sum_{s=1}^{S} \sum_{j_s=1}^{p_s} \lambda_s \sum_{\nu \in V_{\text{int}}} \omega_{\text{v}} \|\beta_{(t)j_s}^{G_{\text{v}}}\|_{\ell_2} \\
&+ \sum_{t=1}^{T} \sum_{s=1}^{S} \sum_{j_s=1}^{p_s} \lambda_s \sum_{\nu \in V_{\text{leaf}}} \omega_{\text{v}} \|\beta_{(t)j_s}^{G_{\text{v}}}\|_{\ell_2} + (1 - \alpha)\gamma \sum_{k=1}^{m} \sum_{j=1}^{p} \sqrt{T} \|\beta_{(1:T)jk}\|_{\ell_2} + \alpha\gamma \|\boldsymbol{B}\|_{\ell_1}.
\end{aligned}
\qquad \text{(Equation 4)}
$$

The 1st term is the sum of negative log-likelihoods in (Equation 3) over multiple sample groups. The 2nd and 3rd terms are the IPF-tree penalty, in which a tree of drug responses with a set of vertices V and groups $\{G_{\text{v}} : v \in V\}$, V consists of internal nodes V_{int} and leaf nodes V_{leaf}, and $\boldsymbol{\beta}_{(t)j_s}^{G_{\text{v}}}$ are coefficients corresponding to predictors $\mathbf{X}_{(t)j_s}$ in the s-th data source across response group G_{v} (see Zhao and Zucknick (2020) for details). If $\lambda_s = \lambda\,(s = 1, \cdots, S)$, then the 3rd and 5th terms together simplify to $(\lambda + \alpha\gamma)\|\boldsymbol{B}\|_{\ell_1}$, since $\omega_{\text{v}} = 1$ when $\nu \in V_{\text{leaf}}$ and the heights of the dendrogram are normalized. To apply the SPG method for model optimization, we smooth the penalty term $\gamma \sum_{k=1}^{m} \sum_{j=1}^{p} \sqrt{T} \|\boldsymbol{\beta}_{(1:T)jk}\|_{\ell_2}$ and the IPF-tree-lasso penalty terms involving internal nodes.

Optimization of mix-lasso

Multiple data sources of predictors can be easily transformed to an equivalent problem of one data source, see Zhao and Zucknick (2020). We here only provide details of the optimization of mix-lasso with one data source of predictors. Mix-lasso with one data source of predictors has the following objective function

$$
\begin{aligned}
&- \sum_{t=1}^{T} \sum_{k=1}^{m} \ell\left(\beta_{(t)0}, \beta_{(t)k}, \sigma_u^2, \sigma_\varepsilon^2\right) + \lambda \left\{ \sum_{t=1}^{T} \sum_{j=}^{p} \sum_{\nu \in V_{\text{int}}} \|\omega_\nu \beta_{(t)j}^{G_\nu}\|_{\ell_2} + \sum_{t=1}^{T} \sum_{j=1}^{p} \sum_{\nu \in V_{\text{leaf}}} \|\omega_\nu \beta_{(t)j}^{G_\nu}\|_{\ell_2} \right\} \\
&+ (1 - \alpha)\gamma \sum_{k=1}^{m} \sum_{j=1}^{p} \sqrt{T} \|\beta_{(1:T)jk}\|_{\ell_2} + \alpha\gamma \|\boldsymbol{B}\|_{\ell_1},
\end{aligned}
$$

where

$$
\begin{aligned}
-\ell\left(\boldsymbol{\beta}_{(t)0},\boldsymbol{\beta}_{(t)k},\sigma_u^2,\sigma_\varepsilon^2\right) &= \frac{n_t}{2}\log(2\pi)+\frac{1}{2}\log\left|\mathrm{V}_{(t)}\right| \\
&-\frac{1}{2}\left(\boldsymbol{y}_{(t)k}-1_{n_t}\boldsymbol{\beta}_{(t)0}-\mathrm{X}_{(t)}\boldsymbol{\beta}_{(t)k}\right)^{\top}\mathrm{V}_{(t)}^{-1}\left(\boldsymbol{y}_{(t)k}-1_{n_t}\boldsymbol{\beta}_{(t)0}-\mathrm{X}_{(t)}\boldsymbol{\beta}_{(t)k}\right).
\end{aligned}
$$

For the covariance matrix, we can use a plug-in proxy matrix $\tilde{V}$ suggested by Fan and Li (2012) or Bradic et al. (2020). Then we modify the smoothing proximal gradient (SPG) method proposed by Kim and Xing (2012). Combining the tree lasso penalty involving internal nodes and the grouped-tissue penalty, we have

$$
\begin{aligned}
\Omega(B):\; &= \lambda\sum_{t=1}^{T}\sum_{j=1}^{p}\sum_{\nu\in V_{\text{int}}}\omega_\nu\left\|\beta_{(t)j}^{G_\nu}\right\|_{\ell_2}+(1-\alpha)\gamma\sum_{k=1}^{m}\sum_{j=1}^{p}\sqrt{T}\left\|\beta_{(1:T)jk}\right\|_{\ell_2} \\
&= \lambda\sum_{t=1}^{T}\sum_{j=1}^{p}\sum_{\nu\in V_{\text{int}}}\omega_\nu\max_{\left\|\boldsymbol{\alpha}_j^{G_\nu}\right\|_{\ell_2}\le 1}\left(\boldsymbol{\alpha}_j^{G_\nu}\right)^{\top}\beta_{(t)j}^{G_\nu}+(1-\alpha)\gamma\sum_{k=1}^{m}\sum_{\nu\in\mathcal{K}_{(t)}}\max_{\left\|\boldsymbol{\alpha}_k^{*K_\nu}\right\|_{\ell_2}\le 1}\left(\boldsymbol{\alpha}_j^{*K_\nu}\right)^{\top}\beta_k^{K_\nu} \\
&= \sum_{t=1}^{T}\max_{\mathrm{A}\in\mathcal{Q}}\left\langle CB_{(t)}^{\top},\mathrm{A}\right\rangle+\max_{\mathrm{A}^*\in\mathcal{Q}^*}\left\langle C^*B,\mathrm{A}^*\right\rangle,
\end{aligned}
$$

where C, $\mathbf{A}$, C^* and $\mathbf{A}^*$ are

$$
C_{(\nu,i)}^{l}=\begin{cases}\omega_\nu & \text{if } l\in\ G_{V_{\text{int}}}\\ 0 & \text{otherwise}\end{cases},\ C_{(\nu,i)}^{*l}=\begin{cases}1 & \text{if } l\in\ K_{\mathcal{K}_{(t)}}\\ 0 & \text{otherwise}\end{cases},
$$

$$
\mathrm{A}=\begin{bmatrix}\boldsymbol{\alpha}_1^{G_1} & \dots & \boldsymbol{\alpha}_p^{G_1}\\ \vdots & \ddots & \vdots\\ \boldsymbol{\alpha}_1^{G_{|V_{\text{int}}|}} & \dots & \boldsymbol{\alpha}_p^{G_{|V_{\text{int}}|}}\end{bmatrix},\mathrm{A}^*=\begin{bmatrix}\boldsymbol{\alpha}_1^{*K_1} & \dots & \boldsymbol{\alpha}_m^{*K_1}\\ \vdots & \ddots & \vdots\\ \boldsymbol{\alpha}_1^{*K_1} & \dots & \boldsymbol{\alpha}_m^{*K_p}\end{bmatrix}.
$$

The smooth approximation to the nonsmooth penalty $\Omega(B)$ is

$$
f_\mu(B)=\sum_{t=1}^{T}\left(\max_{\mathrm{A}\in\mathcal{Q}}\left\langle CB_{(t)}^{\top},\mathrm{A}\right\rangle-\mu d(\mathrm{A})\right)+\max_{\mathrm{A}^*\in\mathcal{Q}^*}\left\langle C^*B,\mathrm{A}^*\right\rangle-\mu d(\mathrm{A}^*),
$$

and its gradient is

$$
\nabla f_\mu\left(\mathrm{B}_{(t)}\right)=\mathrm{A}_{(t)1}^{\top}C+C^{*\top}\mathrm{A}_2,
$$

where $\mathbf{A}_{(t)1}=(\boldsymbol{\alpha}_j^{G_\nu})^{\star}=S\left(\frac{\lambda\omega_\nu\beta_{(t)j}^{G_\nu}}{\mu}\right)$, $\mathbf{A}_2=(\boldsymbol{\alpha}_k^{*K_\nu})^{\star}=S\left(\frac{(1-\alpha)\gamma\sqrt{T}\beta_k^{K_\nu}}{\mu}\right)$, $S(\bullet)$ is the shrinkage operator. Note that the same $\mathbf{A}_2$ is applied to different sample groups, which induces similar gradients for effects of different sample groups.

Let the smoothing (penalized) likelihood be

$$
h(B)=-\sum_{t=1}^{T}\sum_{k=1}^{m}\ell\left(\beta_{(t)0},\beta_{(t)k},\sigma_u^2,\sigma_\varepsilon^2\right)+f_\mu(B).
$$

Its gradient w.r.t. the intercept and coefficients of the t-th sample group is

$$
\nabla h\left(\beta_{(t)0}^{\top},\mathrm{B}_{(t)}\right)=\mathrm{X}_{(t)}^{*\top}V_{(t)}^{-1}\mathrm{X}_{(t)}^{*}\begin{bmatrix}\beta_{0,t}^{\top}\\ B_{(t)}\end{bmatrix}-\mathrm{X}_{(t)}^{*\top}V_{(t)}^{-1}\mathrm{Y}_{(t)}+\begin{bmatrix}0 & 0 & 0\\ 0 & \mathrm{A}_{1,t}^{\top} & \mathrm{A}_2^{\top}\end{bmatrix}\begin{bmatrix}0\\ \lambda\mathbb{I}_pC\\ (1-\alpha)\gamma\mathbb{I}_pC^*\end{bmatrix},
$$

which is Lipschitz continuous with Lipschitz constant

$$
L_{(t)}=\lambda_{\max}\left(\mathrm{X}_{(t)}^{*\top}V_{(t)}^{-1}\mathrm{X}_{(t)}^{*}\right)+\frac{1}{\mu}\left\|\begin{bmatrix}0\\ \lambda\mathbb{I}_pC\\ (1-\alpha)\gamma\sqrt{T}\mathbb{I}_p\end{bmatrix}\right\|^2.
$$

Let $\mathrm{B}_{(t)}=\mathrm{W}_{(t)}^{(b)}-\frac{1}{L_{(t)}}\nabla h(\mathrm{W}_{(t)}^{(b)})$. By second-order Taylor approximation,

$$
h\left(\mathrm{B}_{(t)}\right)\approx h\left(\mathrm{W}_{(t)}^{(b)}\right)+\left\langle\mathrm{B}_{(t)}-\mathrm{W}_{(t)}^{(b)},\nabla h\left(\mathrm{W}_{(t)}^{(b)}\right)\right\rangle+\frac{L_{(t)}}{2}\left\|\mathrm{B}_{(t)}-\mathrm{W}_{(t)}^{(b)}\right\|_{\ell_2}^2.
$$

According to the proximal gradient method and calculating the order of subgradient, we can obtain a closed-form solution of $(b+1)$-th iterated $\boldsymbol{B}_{(t)}^{(b+1)}$

$$\boldsymbol{\beta}_{(t)0}^{\top} = \omega_{0,t}^{\top(b)} - \frac{1}{L_{(t)}}\nabla h\left(\omega_{0,t}^{\top(b)}\right),$$

$$\beta_{(t)jk} = \mathrm{sign}\left(w_{(t)jk}\right)\max\left(0, \left|w_{(t)jk}\right| - \frac{\lambda\omega_{\nu(k,t)}}{L_{(t)}}\right),$$

where $w_{(t)jk}$'s ($j = 1, \cdots, p$) are the elements of $\mathbf{W}_{(t)}^{(b)} - \frac{1}{L_{(t)}}\nabla h(\mathbf{W}_{(t)}^{(b)})$.

Missing drug response data

In practice, some of the cancer cell lines may not be treated with all the drugs, or some of the drug assays may have failed for technical reasons or been removed in the quality control phase, which results in missing data in the drug response matrix **Y**. If we can assume that the data are missing-at-random, we can make use of all the available data for multi-drug response modeling, including the cell lines and drug responses where some values are missing. We use a projection operator $\Pi(\bullet)$ to project the missing entries to zeros similarly to Li et al. (2019). In practice, calculating the residuals between responses **Y** and linear predictors $1_n\widehat{\beta}_0^{\top} + \mathbf{X}\widehat{B}$ in the penalized likelihood (Equation 1) becomes

$$\Pi\left(\mathbf{Y} - 1_n\widehat{\beta}_0^{\top} - \mathbf{X}\widehat{B}\right),$$

which only takes into account non-missing drug response data and ignores missing entries. If $\mathbf{Y}_{ik}$ is missing, $\{\Pi(\bullet)\}_{ik} = 0$; otherwise $\{\Pi(\bullet)\}_{ik} = \{\mathbf{Y} - 1_n\beta_0^{\top} - \mathbf{X}B\}_{ik}$. We note that this technique is used when calculating a Frobenius norm or quadratic form of $\mathbf{Y} - 1_n\beta_0^{\top} - \mathbf{X}B$ when optimizing the objective function (Equation 4) of mix-lasso.

Benchmarking simulation study

To evaluate the performance of the proposed mix-lasso and to compare it against a reference method, tree lasso, we simulated m response variables, n samples from T sample groups and p potential features. A comparison between tree lasso, IPF-tree-lasso and other lasso-type methods for multi-omics data was carried out in Zhao and Zucknick (2020), so we only use tree lasso as a reference method in this study. The penalty parameters of mix-lasso and tree lasso were optimized using 3-fold cross-validation among the n simulated samples, which would in real-world applications correspond to cancer cell lines or patient-derived primary samples, for example.

The simulation data of the t-th group ($t = 1, \cdots, T$) are generated by

$$\begin{aligned} \mathbf{x}_{(t)i} &\sim \mathcal{N}\left(0_{p\times 1}, \Sigma_X\right), \\ \mathbf{Y}_{(t)} &\sim \mathcal{N}\left(\mathbf{X}_{(t)}B_{(t)}, \mathbb{I}_m \otimes V_{(t)}\right), \end{aligned}$$

where Σ_X is designed in the same way as in Zhao and Zucknick (2020) to simulate correlated features, and $B_{(t)}$ is a sparse structured matrix to generate responses with tree-structure relationships; see Kim and Xing (2012) or Zhao and Zucknick (2020) for more details.

In the simulated settings, we set $m = 120$, $n = 300$, $T = 10$, $p = 1000$, Σ_X with diagonals 1 and off-diagonals of 10 diagonal blocks 0.4, $V_{(t)}$ with diagonals 1 and off-diagonals 0.5, and $B_{(t)}$ has the same tree structure as the design in Zhao and Zucknick (2020), including 1800 out of $mp = 120000$ nonzero coefficients for each sample group. In each setting, we assume 5% randomly missing drug responses. We further consider various practical settings for other parameters to mimic large-scale pharmacogenomic screens:

- **Scenario 1**: nonzero coefficients of $B_{(t)}$ ($t = 1, \cdots, T$) are **0.5.**
- **Scenario 2**: nonzero coefficients of $B_{(1)}$ and $B_{(2)}$ are **-0.5**, nonzero coefficients of $B_{(t)}$ ($t = 3, \cdots, T$) are **0.5.**
- **Scenario 3**: nonzero coefficients of $B_{(1:2)}$ are **0.4**, $B_{(3:4)}$ are **0.6**, $B_{(5:6)}$ are **0.8**, $B_{(7:8)}$ are **1.0** and $B_{(9:10)}$ are **1.2**, where $B_{(a:b)}$ represents both $B_{(a)}$ and $B_{(b)}$.
- **Scenario 4**: nonzero coefficients of $B_{(1)}$ are **-0.7**, $B_{(2)}$ are **-0.5**, $B_{(3)}$ are **-0.3**, $B_{(4)}$ are **0.2**, $B_{(5)}$ are **0.4**, $B_{(6)}$ are **0.6**, $B_{(7)}$ are **0.8**, $B_{(8)}$ are **1.0**, $B_{(9)}$ are **1.2** and $B_{(10)}$ are **1.4.**

After training the models in the simulated data, we additionally simulated $n = 300$ samples for validation of the prediction accuracy. As an evaluation metric, we calculated Spearman's ρ between each sample group (e.g., cancer tissue type) and each response variable (e.g., drug) to investigate the rank correlation between the observed responses and the model-predicted responses in the validation set. We ran 50 simulations, and for each sample group and each response variable (i.e. drug response) the Spearman's ρ was averaged over the 50 simulations. We also used another evaluation metric, Root Mean Squared Error (RMSE), for evaluating the accuracy of predicting continuous drug response levels, in addition to the ranking accuracy as evaluated by the Spearman's ρ.

In scenario 1, where the multiple sample groups share the same covariate effects, tree lasso and mix-lasso have similar prediction accuracy (Wilcoxon test $p = 0.071$; Figure S2.1a). In a more challenging scenario 2, where the first two sample groups have opposite covariate effects (i.e. negative and positive regression coefficients) compared to the other groups, mix-lasso shows much better prediction accuracy compared to tree lasso ($p < 0.001$). In scenario 3, where the covariate effects are different across the sample groups, mix-lasso has again similar prediction performance to that of tree lasso ($p = 0.533$). In scenario 4, where the heterogeneous sample groups have both positive and negative effects and varying scales, mix-lasso shows again much better prediction accuracy compared to tree lasso ($p < 0.001$). RMSE shows similar conclusions than Spearman's ρ in scenarios 2 and 4, while tree lasso outperforms mix-lasso in scenarios 1 and 3 based on RMSE (see Figure S2.2). These results indicate that mix-lasso results in better prediction performance than tree lasso in cases, where there exist heterogeneous feature effects in different sample groups, especially when there are opposite effects of the same features in different sample groups.

To evaluate the feature selection performance of the two models, we used a receiver operating characteristic (ROC) curve to investigate if the estimated coefficient of a covariate is truly nonzero or zero, compared to the ground-truth simulation model. Figure S2.1 shows that mix-lasso and tree lasso have very similar feature selection accuracy w.r.t. the area under the receiver operating characteristic curve (AUC) in scenarios 1, 3 and 4. However, similar to the prediction accuracy, mix-lasso shows a much better AUC value than tree lasso in the more challenging scenario 2, where there exist opposite effects of the same features in different sample groups. This indicates that the mix-lasso accurately identifies relevant features for drug responses across multiple tissue types, especially when there is strong heterogeneity between sample groups, e.g., if the same feature may have opposite effects in two patient groups of cancer types.