Ammad-ud-din, Muhammad; Khan, Suleiman A.; Malani, Disha; Murumägi, Astrid; Kallioniemi, Olli; Aittokallio, Tero; Kaski, Samuel

Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization

Published in:
Bioinformatics

DOI:
10.1093/bioinformatics/btw433

Published: 03/09/2016

Document Version
Publisher's PDF, also known as Version of record

Bioinformatics, 32, 2016, i455–i463
doi: 10.1093/bioinformatics/btw433
ECCB 2016

OXFORD

Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization

Muhammad Ammad-ud-din[1,*], Suleiman A. Khan[2], Disha Malani[2], Astrid Murumägi[2], Olli Kallioniemi[2,3,4], Tero Aittokallio[2,5] and Samuel Kaski[1,*]

[1]Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, Espoo 02150, Finland, [2]Institute for Molecular Medicine Finland FIMM, University of Helsinki, Helsinki 00290, Finland, [3]Science for Life Laboratory, [4]Department of Oncology and Pathology, Karolinska Institutet, Solna 17165, Sweden and [5]Department of Mathematics and Statistics, University of Turku, Turku 20014, Finland

*To whom correspondence should be addressed.

Abstract

Motivation: A key goal of computational personalized medicine is to systematically utilize genomic and other molecular features of samples to predict drug responses for a previously unseen sample. Such predictions are valuable for developing hypotheses for selecting therapies tailored for individual patients. This is especially valuable in oncology, where molecular and genetic heterogeneity of the cells has a major impact on the response. However, the prediction task is extremely challenging, raising the need for methods that can effectively model and predict drug responses.

Results: In this study, we propose a novel formulation of multi-task matrix factorization that allows selective data integration for predicting drug responses. To solve the modeling task, we extend the state-of-the-art kernelized Bayesian matrix factorization (KBMF) method with component-wise multiple kernel learning. In addition, our approach exploits the known pathway information in a novel and biologically meaningful fashion to learn the drug response associations. Our method quantitatively outperforms the state of the art on predicting drug responses in two publicly available cancer datasets as well as on a synthetic dataset. In addition, we validated our model predictions with lab experiments using an in-house cancer cell line panel. We finally show the practical applicability of the proposed method by utilizing prior knowledge to infer pathway-drug response associations, opening up the opportunity for elucidating drug action mechanisms. We demonstrate that pathway-response associations can be learned by the proposed model for the well-known EGFR and MEK inhibitors.

Availability and implementation: The source code implementing the method is available at http://research.cs.aalto.fi/pml/software/cwkbmf/.

Contacts: muhammad.ammad-ud-din@aalto.fi or samuel.kaski@aalto.fi

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The fundamental aim of personalized medicine is to design and identify individualized therapies that maximize drug efficacy while minimizing the undesirable side effects. The efficacy, however, depends on a multitude of factors, including molecular, genetic, environmental and clinical characteristics of the samples, and much of this information remains unknown. A promising research direction is to computationally learn to predict, based on the available molecular and genetic descriptions of the samples, the responses they elicit in lab when exposed to a spectrum of drugs. The learned predictors help identifying potential drug response associations, and can predict responses for a new sample.

The development of molecular and genetic models of drug response has been made possible through several recent large scale high-throughput screening efforts that profile large panels of human

cancer cell lines and drugs (Barretina *et al.*, 2012; Basu *et al.*, 2013; Garnett *et al.*, 2012). Such models open up the opportunity to study the impact of molecular characteristics on the response, increasing our understanding of cancer vulnerabilities as well as making it possible to build predictive models of drug responsiveness.

Recent advances have demonstrated that molecular and genomic features have been useful in predicting the drug responses in cell lines (Costello *et al.*, 2014; Jang *et al.*, 2014). However, a key challenge underlying predictive modeling is the small sample size and very large number of genomic features. The small sample sizes offer limited statistical power leading to high uncertainty in the predictions. The inherent heterogeneity across and within different cancer types makes robust inference even harder. In the absence of technical and practical facilities to overcome these limitations, a prospective direction is to incorporate additional prior knowledge in a biologically meaningful way to facilitate the learning process.

From the computational perspective, several methodologies have been used to predict drug response (a detailed discussion in Section 2). A key constituent of inferring the molecular and genetic model is the ability to effectively integrate multiple side-data views (also called as side-data sources) for prediction of the drug responses. Methods commonly referred to as multiple kernel learning MKL (Gönen and Alpayd, 2011) can extract the common signal from multiple side-data views, effectively yielding an increased signal-to-noise ratio in the parameter space and are currently the state of the art in drug response prediction (Costello *et al.*, 2014). Multi-task learning makes it possible to learn a predictive model for all of the drugs jointly (multi-task) making it possible to gather statistical evidence across multiple drugs (Baxter, 2000).

In this study, we introduce component-wise multiple kernel learning (MKL) into the recent kernelized Bayesian matrix factorization (KBMF) method (Gönen *et al.*, 2013). The proposed model solves the prediction task by gathering evidence from multiple side-data views, selectively, for each of the output variable group. This formulation is particularly useful in drug response prediction, for taking into account multiple side-data views. It need not assume the same views to be relevant to all drugs as earlier methods, but instead predictions can be based on different views for different groups of drugs.

The multiple side-data views can be generated based on the prior biological knowledge; in the paper we use the pathways that are linked to the known primary targets of the drugs. By systematically utilizing this type of prior knowledge through kernelized Bayesian matrix factorization with the component-wise MKL approach, we hypothesize that pathway-drug response associations can be learned which are more informative for response prediction, and additionally are better interpretable for understanding drug action mechanisms.

1.1 Contributions

In this paper, we present a novel approach for improving accuracy of predicting drug responses and elucidating the underlying pathway-drug response associations. Specifically, our contributions are two-fold:

1. Methodologically we extend the current state-of-the-art model kernelized Bayesian matrix factorization (KBMF) with component-wise multiple kernel learning (MKL). The extension can be seen as multi-task learning by task factorization, however with selective data integration. Here the key assumption is that component-wise MKL allows the method to better use prior biological knowledge (pathways) input as multiple side-data views.
2. We introduce a way for incorporating prior biological knowledge, in the form of pathways, for modeling pathway-drug response associations. Instead of using a single side-data view for the genomic features, we present pathway-based groups of features as multiple side-data views. Here the key assumption is that informed grouping of the features introduces additional structure and knowledge that is valuable for prediction of particular drug groups.

We first demonstrate the model's predictive abilities on a synthetic dataset. We then substantiate the significantly better performance of our approach on predicting drug responses in two large publicly available cancer datasets. In addition, we validate the *in silico* predictions of our model with lab experiments on an in-house Acute Myeloid Leukemia (AML) cell line panel. Finally, we examine the inferred associations between drug responses and pathways in the larger dataset, demonstrating a mechanism for elucidating drug action mechanisms.

2 Related work in drug response prediction

The computational task underlying personalized medicine is to predict drug responses on new cancer cell lines, given a set of cancer cell lines for which some measurements of drug responses are observed.

A common approach is to use the mean of the observed responses as predictions for the unobserved (unseen) drug responses (used as baseline method here). Another well-known supervised approach uses the genomic and molecular features of the cell lines (as input side-data) and the observed drug responses to learn a predictive model of the drug responses (Jang *et al.*, 2014). The available molecular and genomic features range from gene expression to copy number and point mutations for the cancer cell lines, respectively (Barretina *et al.*, 2012; Garnett *et al.*, 2012).

Another widely used approach is the quantitative structure-activity relationship *(QSAR)* analysis which uses chemical and structural properties (often called as descriptors) of the drugs and the observed responses to learn a predictive model to infer the unobserved responses. The descriptors vary from 2D fingerprints to spatial characteristics and physiochemical features of the drugs (Myint and Xie, 2010; Perkins *et al.*, 2003; Shao *et al.*, 2013). Recently, an advanced approach has been proposed that learns a joint predictive model of the observed drug responses by combining both the genomic features of the cell lines and descriptors of the drugs (Ammad-ud din *et al.*, 2014; Cichonska *et al.*, 2015; Cortés-Ciriano *et al.*, 2015; Menden *et al.*, 2013; Zhang *et al.*, 2015).

Previous studies have used linear as well as non-linear methods. Linear methods including multivariate linear regression, partial least squares (PLS) and principal component regression (PCR) are the most prominent. Sparse linear regression has been well studied for identifying potential features predictive of drug responses by enforcing elastic net regularization techniques (Barretina *et al.*, 2012; Chen *et al.*, 2015; Garnett *et al.*, 2012).

Nonlinear drug response analysis including kernel method, neural networks and random forests have also been studied (Cichonska *et al.*, 2015; Cortés-Ciriano *et al.*, 2015; Menden *et al.*, 2013; Sutherland *et al.*, 2004; Yamanishi *et al.*, 2012; Zhang *et al.*, 2015).

In particular, Costello *et al.* (2014) proposed Bayesian multi-task multiple kernel learning (BMTMKL) to predict drug responses on new human breast cancer cell lines. The BMTMKL method uses a kernelized regression approach that combines multi-task and multi-view learning (i.e. learning from multiple side-data views) with

Bayesian inference to estimate the model parameters. Their results showed that modeling nonlinearities in the data was an essential attribute to predict drug responses. However, the model makes the simplifying assumption that the predictions are based on a single underlying component.

Alternatively, matrix factorization models integrating side-data views have also been studied in drug response analysis. The main idea behind these methods is to jointly factorize the side-data views and output matrix to find a better low-dimensional latent representation (components) for both rows and columns of the output matrix. To this end, Zhou *et al.* (2012) proposed kernelized probabilistic matrix factorization (KPMF), a low-rank matrix factorization method that uses Gaussian process priors with covariance matrix on side-data view. While the method can explain tasks with multiple components, it is, however, limited to a single kernel for each side and therefore, is unable to learn from multiple side-data views.

Recently, a kernelized Bayesian matrix factorization (KBMF) extending kernelized matrix factorization with fully Bayesian inference, combining multiple side-data views to jointly factorize the output matrix has been proposed (Gönen *et al.*, 2013; Gönen and Kaski, 2014). With side-data views encoded as kernel functions, the main idea is to project each kernel onto a low-dimensional component space, where they are combined with the kernel weights to get a composite component space of the output matrix. The KBMF method has been studied in various applications ranging from drug-target to drug response predictions (Ammad-ud din *et al.*, 2014; Gönen, 2012). However, KBMF integrates multiple side-data views assuming that a source is either relevant for all tasks or none, failing to identify component-specific dependencies between the side-data views and the output matrix.

3 Methods

3.1 Kernelized Bayesian Matrix Factorization (*cw*KBMF)

We introduce a novel extension of the state-of-the-art kernelized Bayesian matrix factorization method to model the complex associations between a large number of side-data views and the latent component space of the output matrix. This new formulation of kernelized Bayesian matrix factorization (KBMF) allows component-wise multiple kernel learning (MKL), referred to as *cw*KBMF for brevity. *cw*KBMF is characterized by the ability to comprehensively model the associations that allow two advancements: (i) improve the predictive power of the model; and (ii) identify the component-specific latent dependencies for interpreting the associations.

The model is defined for the factorization of a given matrix $\mathbf{Y} \in \mathbb{R}^{N_x \times N_z}$, using known sets of P_x side-data views for the rows and P_z side-data views for the columns. In order to represent non-linear associations, similarities between samples in the side-data views are encoded as input kernel matrices $\{\mathbf{K}_{x,m} \in \mathbb{R}^{N_x \times N_x}\}_{m=1}^{P_x}$ and $\{\mathbf{K}_{z,n} \in \mathbb{R}^{N_z \times N_z}\}_{n=1}^{P_z}$. Here matrices are denoted by capital letters, with the subscript ($_x$ or $_z$) indicating the corresponding side of the model. All equations are formulated, however, with corresponding scaler entities denoted by non-capital letters, with the superscript denoting the row index and the last subscript representing the column index (i.e. $a_{x,s}^i$ denotes the entry at [row i, column s] of matrix $\mathbf{A}_x$). Without compromising the generalizability, the rest of this article focuses on multiple side-data views in the rows only.

The model is specified as a low-rank factorization of the matrix $\mathbf{Y}$ such that the latent representations $\mathbf{H}_x \in \mathbb{R}^{N_x \times R}$ and $\mathbf{H}_z \in \mathbb{R}^{N_z \times R}$ are learned jointly from $\mathbf{Y}$ and the $\mathbf{K}_{x,m}$, $\mathbf{K}_{z,m}$ side-data views. This is achieved by an interplay of two elements. First, each of the $\{\mathbf{K}_{x,m}\}_{m=1}^{P_x}$ kernels is transformed to a lower dimensional subspace $\{\mathbf{G}_{x,m} \in \mathbb{R}^{N_x \times R}\}_{m=1}^{P_x}$ through a common projection matrix $\mathbf{A}_x \in \mathbb{R}^{N_x \times R}$. The low-rank transformations of the kernels are combined using multiple kernel learning to compute the latent matrix factors $\mathbf{H}_x$.

The *cw*KBMF method is formulated in a Bayesian setting using conjugate priors, where $\mathcal{N}(\cdot;\mu,\Sigma)$ denotes the normal distribution with mean μ and covariance Σ, while $\mathcal{G}(\cdot;\alpha,\beta)$ is the gamma distribution with the parameters, shape α and scale β. The matrix factorization is formulated as

$$y_j^i | h_{x,i}, h_{z,j} \sim \mathcal{N}(y_j^i; h_{x,i}^\top h_{z,j}, \sigma_y^2) \; \forall (i,j)$$

where $i = 1 : N_x$ and $j = 1 : N_z$ denote the samples and σ_y the noise. Here $h_{x,i}$ and $h_{z,j}$ are vectors of length R, the number of components, and represent the low-dimensional factors of the samples in Y.

Our extension formulates this factorization with the novel component-wise MKL and has the distributional assumptions:

$$\eta_{x,m}^s \sim \mathcal{G}(\eta_{x,m}^s; \alpha_\eta, \beta_\eta) \qquad \forall (m,s)$$

$$e_{x,m}^s | \eta_{x,m}^s \sim \mathcal{N}(e_{x,m}^s; 0, (\eta_{x,m}^s)^{-1}) \qquad \forall (m,s)$$

$$h_{x,i}^s | \{e_{x,m}^s, g_{x,m,i}^s\}_{m=1}^{P_x} \sim \mathcal{N}\left(h_{x,i}^s; \sum_{m=1}^{P_x} e_{x,m}^s g_{x,m,i}^s, \sigma_h^2\right) \qquad \forall (s,i)$$

where superscript $s = 1 : R$ denotes the components. The novel advancement of this formulation is in learning the latent components $\mathbf{H}_x$ as a combination of kernel-specific components $\{\mathbf{G}_{x,m} \in \mathbb{R}^{N_x \times R}\}_{m=1}^{P_x}$ while segregating between kernels that are component-specific and those which are shared across all components. This is achieved by introducing component-specific kernel weights $\mathbf{e}_{x,m}^s \in \mathbb{R}^{P_x \times R}$ that control the activity of each kernel in each component. This extension makes it possible for the method to effectively learn the underlying structure for identifying the associations between kernels and components. The method can also be viewed as combination of component-wise multiple kernel learning and matrix factorization. The $\eta_{x,m}^s$ defines an element-wise prior for the kernel-weights $e_{x,m}^s$, making it possible to effectively switch off some of the weights in a component-wise fashion.

Finally, the dimensionality reduction of the model has the distributional assumptions:

$$\lambda_{x,s}^i \sim \mathcal{G}(\lambda_{x,s}^i; \alpha_\lambda, \beta_\lambda) \qquad \forall (i,s)$$

$$a_{x,s}^i | \lambda_{x,s}^i \sim \mathcal{N}(a_{x,s}^i; 0, (\lambda_{x,s}^i)^{-1}) \qquad \forall (i,s)$$

$$g_{x,m,i}^s | a_{x,s}, k_{x,m,i} \sim \mathcal{N}(g_{x,m,i}^s; a_{x,s}^\top k_{x,m,i}, \sigma_g^2) \qquad \forall (m,s,i)$$

where a joint A_x matrix projects each of the kernels to a low-dimensional representation. The hyper-parameters $\alpha_\lambda, \beta_\lambda, \alpha_\eta, \beta_\eta, \sigma_g, \sigma_h, \sigma_y$ can be used to express prior knowledge about the data-generating process, or set to uninformative values (as in this paper).

The model is formulated with conjugate priors and variational approximation is used to perform model inference. The computational complexity of the model is $\mathcal{O}(R\max(N_x^3, N_z^3) + R\max(P_x^3, P_z^3))$ which is faster than standard pair-wise kernel approaches (Ben-Hur and Noble, 2005) and slower only linearly with a factor of R in $\max(P_x^3, P_z^3)$, in comparison to original KBMF formulation. The model achieves a run time to the tune of minutes for reasonably sized datasets (≈5 minutes of wall clock time on a standard computer, for a single cross validation fold on the largest data studied in this manuscript).

Table 1. Data used in the drug response predictions

Datasets	Cell lines	Drugs	Genes	Primary Targets	Views
GDSC	124	124	13 321	60	72 (71 pathways, 1 other genes)
CTRP	66	63	18 988	58	26 (25 pathways, 1 other genes)

3.2 Publicly available datasets and preprocessing

We used two publicly available cancer datasets to model drug response associations in this study.

***Genomics of Drug Sensitivity in Cancer*:** The first data come from Genomics of Drug Sensitivity in Cancer (GDSC) project initiated by Wellcome Trust Sanger Institute version release, June 2014 (Yang *et al.*, 2013). The data comprised of 124 human cancer cell lines and 124 anti-cancer drugs, for which complete drug response measurements are available and the response range is consistent with earlier publications (Garnett *et al.*, 2012; Menden *et al.*, 2013). Drug response measurements are summarized as log IC_{50} values (micro molar concentration of a drug required to inhibit 50% of the cell growth) obtained by curve fitting through the 9-point dose response data. The cell lines are annotated with tissue type, and drugs with their primary therapeutic targets.

***Cancer Therapeutic Response Portal*:** The second data originate from Cancer Theraupetic Response Portal (CTRP) version release v1 2013, (Basu *et al.*, 2013) by Broad Institute summarizing area-under-concentration-response curve (AUC) values from 8-point dose response data measured on human cancer cell lines. For our case study, we focus on the set of 66 cell lines and 63 anti-cancer drugs, whose AUC values were observed without missing values. The molecular profiles for the cell lines was obtained from Cancer Cell line Encylopedia CCLE (Barretina *et al.*, 2012). As in GDSC, the cell lines are annotated with tissue type and gene expression, while drugs with their primary therapeutic targets.

As the input data, we used the baseline gene expression values of all the cell lines quantizing the number of transcripts expressed in a cell. These measurements characterize the genome-wide molecular profiles that may be indicative of the response patterns.

***Prior Biological Knowledge*:** In order to incorporate prior biological knowledge, we used a selected set of pathways and gene sets from Molecular Signature database MSigDB (Liberzon *et al.*, 2011). Specifically, we extracted the C2CP and C6 collections of pathways and genesets from MSigDB, respectively. C2CP contains pathways compiled from online pathway databases, biomedical literature, published mammalian gene expression studies and MYC target gene database. C6 gene sets denote oncogene signatures of cellular pathways which are often dis-regulated in cancer. These oncogene signatures are computed using microarray data from NCBI GEO and from profiling experiments involving perturbation of known cancer genes. For simplicity in the rest of the paper, we use a common term for both of the collections: *pathways*.

3.3 Experimental setup

***Incorporating Prior Biological Knowledge*:** We focused the analysis on drug targets by, for each of the two collections, carefully selecting the subset of pathways that were directly linked to the known primary targets of the drugs. This was done by examining the correspondence between pathway names and the known primary targets of the drugs. The drug target data coming from the original annotations of GDSC and CTRP was used for this purpose. The gene expression data were then split into groups of genes, where each group represented one pathway. All the other genes which were not part of any of the target-based pathway selection, were collected in a separate single group (collectively called as '*other genes*'). When the variable groups in the data are constructed in this way, the component-wise MKL based data integration can choose what prior knowledge is useful for predicting responses. Still, no knowledge is lost as all variables have been included, and additionally allows to learn associations between other genes and the responses. The total number of groups formed per case study are listed in Table 1. We term each group with a keyword '*view*'.

Additional information about the data including the names of the cell lines, drugs, primary targets and pathways can be found from the supplementary material. The response data consist of both types of drugs: FDA approved 'drugs' and 'investigational chemical compounds'. In the paper, we use both of these terms interchangeably.

***Cross-Validation*:** We compared the performance of *cw*KBMF with several methods including $KBMF_{multi\text{-}view}$; kernelized Bayesian matrix factorization with pathway based groups, $BMTMKL_{multi\text{-}view}$; Bayesian multi-task learning with pathway based groups, $KPMF_{single\text{-}view}$; kernelized probabilistic matrix factorization without pathway based groups, $MT\text{-}LR_{single\text{-}view}$; multi-task sparse linear regression without pathway based groups and the classical Baseline; mean of the training drug response data (assuming no genomic data is available).

We performed a 5-fold cross validation procedure, where in each fold a randomly selected subset of cell lines is completely held-out (as test cell lines) and models were trained on the remaining cell lines (training data). To establish robust findings the 5-fold cross validation procedure was repeated 10 times with different random cross-validation folds.

For the kernelized Bayeisan methods (BMTMKL, KBMF and *cw*KBMF), we use uninformative priors for the projection matrices and the kernel weights. In particular, the hyperparameter values for BMTMKL are selected as $(\alpha_\lambda, \beta_\lambda, \alpha_v, \beta_v, \alpha_\gamma, \beta_\gamma, \alpha_\omega, \beta_\omega, \alpha_\epsilon, \beta_\epsilon) = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$ and for KBMF, *cw*KBMF are selected as $(\alpha_\eta, \beta_\eta, \alpha_\lambda, \beta_\lambda) = (1, 1, 1, 1)$, and the standard deviations $(\sigma_g, \sigma_b, \sigma_y)$ are set to (0.1, 0.1,1). For KPMF, the standard deviation σ_y is set to one. For the side-data views, we computed Gaussian kernels, where the width parameter σ was set in the standard way (σ = dimensionality of the side-data view). The drug response measurements were normalized to have zero mean and unit variance.

We used multi-task sparse linear regression (MT-LR) using the *glmnet* package (Friedman *et al.*, 2010). The sparse linear regression has two parameters that are to be optimized: α (elastic net mixing parameter) and λ (the penalty parameter). For each test set prediction, we performed a nested 5-fold cross validation procedure on the training data, to choose optimal values for $\alpha \in [0, 1]$ with an increment of 0.1 and λ (from 100 values). We finally selected a combination of α and λ values that gave minimum error averaged over the cross-validated folds.

***Evaluation Criteria*:** We evaluated the predictive performance of *cw*KBMF and other methods using drug-wise spearman correlation as an evaluation criterion and report an averaged correlation for each drug from 10 random repeats of the cross-validation procedure. In addition, the correlations were averaged to obtain a cumulative correlation value for each method.

4 Results and discussion

4.1 Synthetic dataset

To demonstrate that the model can infer the true associations between multiple side-data views and components, we perform the first experiment using synthetic dataset. Specifically, the *cw*KBMF method has been designed to learn the complex relationships patterns, by representing them as activity profiles of components over the views.

To this end, 100 synthetic datasets $\mathbf{Y}$ with $N_x = 100, N_z = 100$ and $R = 3$ components were generated such that each dataset was supplemented with $P_x = 10$ side-data views (encoded as kernels).

The associations between the $P_x = 10$ side-data views and $K = 3$ components were encoded such that one view is active in all the components (shared), while the rest are equally split into $K + 1$ sets, where each view is either active in one component (specific) or not active in any of the components (empty). Here the key assumption is that given the kernels for $P_x = 10$ side-data views and the output matrix $\mathbf{Y}$, the model decomposes $\mathbf{Y}$ into components while accurately learning the associations between kernels and components. In addition, 1% values in each $\mathbf{Y}$ were marked as missing data (test set) to measure the predictive accuracy of the model.

The model is run for each of the 100 datasets to learn the associations. The component-view weights e_x^s represent the activity of each view $x = 1 : P_x$ in components $s = 1 : R$. Since the model is encoded with an element-wise prior it can be effectively thresholded to illustrate component-view activity. In order to focus on the most important associations for each component, we consider the associations that are notably strong with respect to the prior (i.e. z-score $(e_x^s) > 0.67$) as active. Figure 1 (left panel) shows the resulting component-view activities inferred by the model for $P_x = 10$. The figure demonstrates that the model is able to accurately discover the component-view activities as inserted in the data (described above), up to a random permutation of the components.

Next, we measured the accuracy of the model in inferring the component-view activities as well as in predicting unobserved values in Y over a range of side-data views P_x. The associations were learned and prediction performance was evaluated for 100 datasets for each value of P_x. Figure 1 (right panel) demonstrates the accuracy of learning the associations, particularly the model performs well in discovering the shared, specific as well as empty components over the range of input views. In addition, Figure 1 (middle panel), *cw*KBMF consistently outperforms KBMF in the prediction task as well, especially when the number of views is large. As expected, the performance of the methods deteriorates as the number of views (dimensionality) increases. However, *cw*KBMF performs reasonably well over the number of views applicable to the drug-response prediction datasets in this study.

4.2 Cancer datasets

We next compare *cw*KBMF with alternatives on two case studies GDSC and CTRP, and report their predictive performance in the 5-fold cross validation procedure (described in Section 3). To evaluate the new model extension and the benefit of the principled incorporation of prior knowledge, we compare *cw*KBMF's performance to other methods in two scenarios,

1. Genomic Data + Prior Knowledge: The genomic features are divided into several side-data views based on the prior knowledge about the pathways. We represent this scenario with a subscript *multi-view* in the results and
2. Genomic Data (only): The genomic features are used as a single view and does not benefits from the prior knowledge. We denote this scenario with a subscript *single-view* in the results.

Figures 2 show the predictive performances of all the methods on the GDSC (left) and CTRP (right) datasets. *cw*KBMF outperforms its competitors for both scenarios. Even though the differences in performances are rather small, the predictive performance obtained by *cw*KBMF for both scenarios is found to be significantly higher than the others ($P < 0.05$; one-sided paired Wilcoxon Sign-Rank test corrected for multiple testing, Supplementary Table S3) in GDSC dataset respectively. Similarly in CTRP dataset, the predictive performance obtained by *cw*KBMF for both scenarios is also significantly higher than the others ($P < 0.05$; one-sided paired Wilcoxon Sign-Rank test, corrected for multiple testing, Supplementary Table S6). In the GDSC and CTRP datasets, the maximum predictive performance of *cw*KBMF is achieved with 10 and 20 components, respectively. However, in-case of multiple maxima a practical choice could be to prefer solutions with smallest-R in the interest of simpler representations. We chose these components and discuss a detailed comparison of the predictions of *cw*KBMF with other methods.

In GDSC dataset, $cw\text{KBMF}_{\text{single-view}}$ outperforms Baseline, $\text{MT-LR}_{\text{single-view}}$ and $\text{KPMF}_{\text{single-view}}$ ($P < 0.05$; one- sided paired

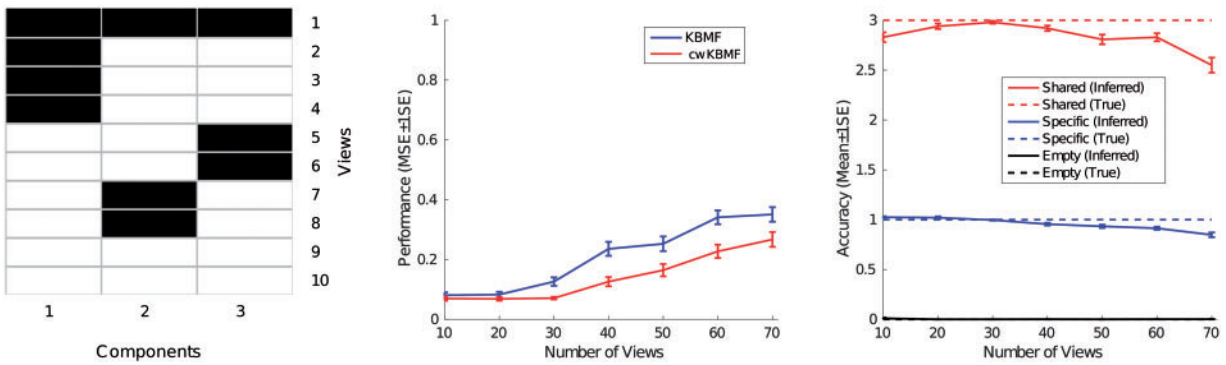

Fig. 1. Identification of component-view activities and predictions on synthetic dataset. Method abbreviation: *cw*KBMF, kernelized Bayesian matrix factorization with component-wise MKL; KBMF, kernelized Bayesian matrix factorization. Left: The component-view activities learned by the model. Black indicates that a view is active in a component while white represents not-active. Middle: the mean squared error (MSE) of predictions, averaged over 100 datasets at each point (and bars denoting 1-standard error of the mean (1SE)). The performance is indicative of the models ability to discover the underlying structure of the data. Right: The accuracy of the model to discover component-view associations. The true and the inferred, averaged accuracy of the associations from 100 datasets are marked for shared, specific and empty component, along with 1SE. The figure demonstrates the models ability to accurately discover the component-view associations

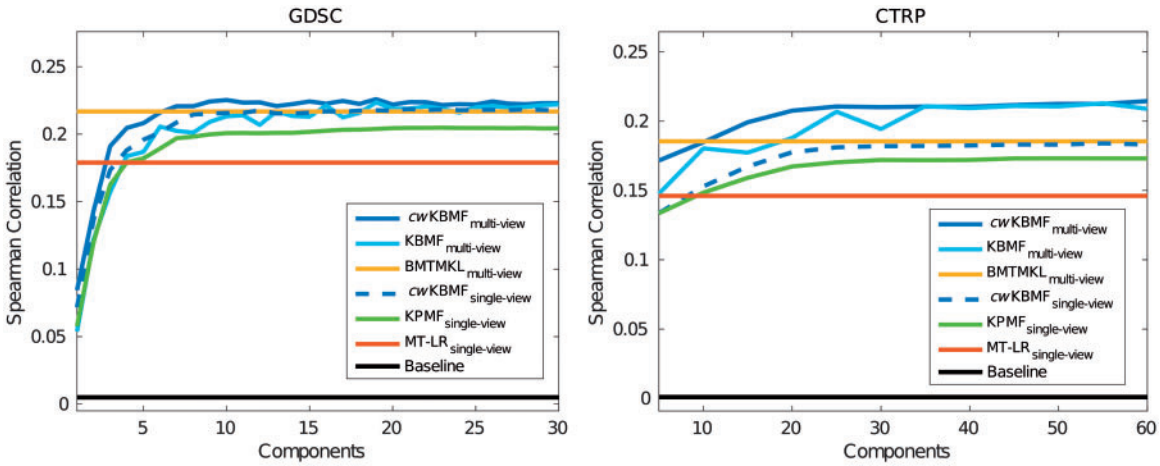

Fig. 2. Prediction performances (Spearman correlation) averaged over drugs with a 5 fold cross-validation procedure repeated 10 times. GDSC dataset (left) and CTRP dataset (right). Method abbreviation: *cw*KBMF, kernelized Bayesian matrix factorization with component-wise MKL; KBMF, kernelized Bayesian matrix factorization; BMTMKL, Bayesian multi-task MKL; KPMF, kernelized probabilistic matrix factorization; MT-LR, multi-task sparse linear regression; Baseline, mean of the training data. The predictive performance obtained by *cw*KBMF for both scenarios is found to be significantly higher than the others ($P<0.05$; one-sided paired Wilcoxon Sign-Rank test corrected for multiple testing, Supplementary Tables S3 and S6)

Wilcoxon Sign-Rank test, for comparing Spearman correlations). $cw\text{KBMF}_{\text{multi-view}}$ outperforms $\text{KBMF}_{\text{multi-view}}$ and Baseline methods ($P < 0.05$; one-sided paired Wilcoxon Sign-Rank test, Supplementary Fig. S1). Although $cw\text{KBMF}_{\text{multi-view}}$ performance is better than $\text{BMTMKL}_{\text{multi-view}}$ (averaged Spearman correlations 0.2253 and 0.2167, respectively), the difference is not statistically significant ($P = 0.11$; one-sided paired Wilcoxon Sign-Rank test).

Similarly, in CTRP dataset, $cw\text{KBMF}_{\text{single-view}}$ outperforms Baseline and $\text{MT-LR}_{\text{single-view}}$ ($P < 0.05$; one-sided paired Wilcoxon Sign-Rank test). Although $cw\text{KBMF}_{\text{single-view}}$ performance is better than $\text{KPMF}_{\text{single-view}}$ (averaged Spearman correlation 0.1776 and 0.1673, respectively), the difference is not statistically significant ($P = 0.07$; one-sided paired Wilcoxon Sign-Rank test). $cw\text{KBMF}_{\text{multi-view}}$ give better predictions than Baseline, $\text{BMTMKL}_{\text{multi-view}}$ and $\text{KBMF}_{\text{multi-view}}$ ($P < 0.05$; one-sided paired Wilcoxon Sign-Rank test, Supplementary Fig. S3).

The prediction results generalize previous findings that non-linear models improve drug response predictions (Costello *et al.*, 2014). Figure 2 clearly shows that non-linear methods are better than the linear counterpart, for predicting drug responses in both datasets.

Having established that our model outperforms existing methods in both single-view and multi-view settings, we next specifically study the advantage of using prior pathway and target knowledge. To this end, Figure 3 illustrates the improvement in performance (in % units) relative to Baseline and when genomic data is supplemented with prior knowledge.

As the first observation, the introduction of genomic data via different methods outperforms the baseline predictions demonstrating the genomic features are response predictive. Secondly, incorporating prior biological knowledge improves the prediction performance systematically over a range of methods. Third, systematically modelling the associations between pathway-based genomic profiles and drug response with *cw*KBMF outperforms the existing approaches in predicting drug responses. Specifically, in case of the GDSC dataset, using genomic data with *cw*KBMF improves the prediction performance by 21% and when the genomic data is supplemented with prior knowledge the performance is improved by 22%. Similarly, in case of the CTRP dataset, using genomic data with *cw*KBMF improves the prediction performance by 17% and when the genomic data is supplemented with prior knowledge the performance is improved by 20%. The findings also suggest that incorporating the

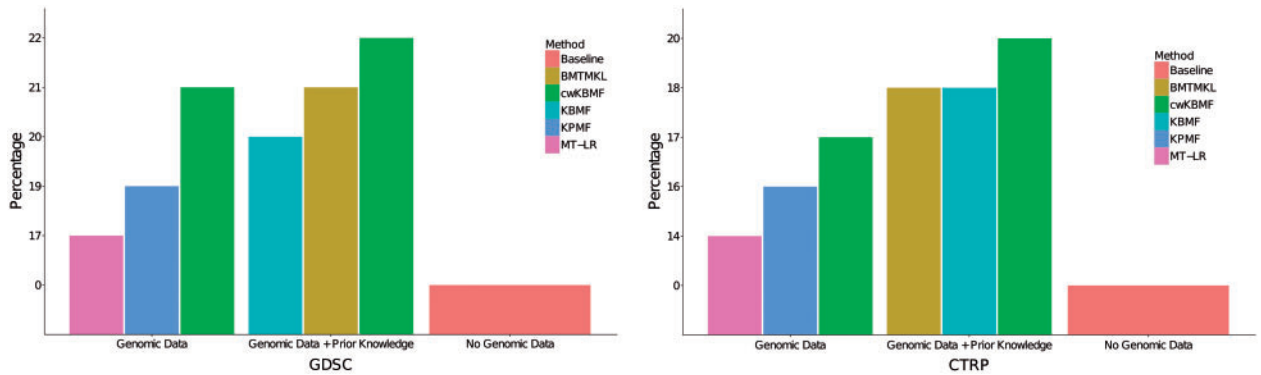

Fig. 3. Pathway-based groups of genes (prior biological knowledge) improves predictive performance. Left, GDSC dataset and Right, CTRP dataset. The height of the bar (*y*-axis) denotes the percentage increase in performance relative to Baseline, computed using the Spearman correlations averaged over drugs. On *x*-axis, the bars are grouped based on the type of learning data used, where 'Genomic Data' means that all of the genes are used as one group and 'Genomic Data + Prior Knowledge' means that all of the genes are used, grouped into several sets based on the pathway knowledge and lastly 'No Genomic Data' implies that only mean of the training drug response data is used for prediction

prior knowledge is more beneficial when the number of samples is smaller (for instance, in the CTRP dataset).

***Fully Blinded Experimental Validations*:** Finally, we experimentally validated the drug response predictions of our model using an in-house Acute Myeloid Leukemia (AML) cell line panel (Malani et al., manuscript in preparation). The model is learned, analogously to the experiments with public datasets above, using the available training drug response data. Specifically, we made drug response predictions for 8 compounds using 6 AML cell lines of which 83% measurements were not available for initial model training. To validate the predictions, an independent experiment was carried out in laboratory. The predicted drug responses were found to be correlated with the independent lab measurements (Spearman correlation 0.44 Figure 4, $p < 0.05$; compared to the distribution of correlation values obtained via randomization; Supplementary Fig. S4). This fully-blinded experimental validation confirms the predictive power of the model, and gives confidence that *in silico* predictions are fairly robust and may be used to study the spectrum of therapeutic choices.

4.3 Inferring pathway-drug response associations

The use of prior knowledge not only improves the prediction performance, but also helps to infer pathway-drug response associations by *cw*KBMF, being the first kernelized method making it possible to study such associations. We next study the pathway-drug response associations in the GDSC dataset.

We selected the model learned with 10 components based on cross-validation (as discussed in section 4) and show the pathway-drug response associations in Figure 5. A component can be characterized by the set of pathways that are active in it and the drugs whose responses they are predictive off, yielding hypotheses of pathways associated with drug responses. The hypotheses generated by all the ten components are illustrated in Figure 5, while those from the first two components are elaborated in detail below. In order to analyze target-driven effects, the components were sorted based on the consistency of the drug targets in the components.

Component 1 is characterized by EGFR/ERBB2 inhibitors lapatinib, erlotinib, BIBW2992 (afatinib) and gefitinib. On the pathway side, we found *reactome SHC1 events in EGFR signalling, reactome GRB2 events in ERBB2 signaling*, among the top 10 pathways. It is biologically meaningful that the inhibitors are related to the EGFR signaling, making it possible to inhibit the pathway activity in cancer using the EGFR inhibitors. It is also evident that signaling pathways RAS-RAF-MAPK, PI3K/AKT and JAK/STAT mediate the downstream effect of EGFR autophosphorylation, thus affecting cellular proliferation, anti-apoptosis, metastasis and tumor invasion (Whirl-Carrillo *et al.*, 2012). We give additional details of component 1 explaining the variation of EFGR responses in Supplementary Figure S2 (left). Other drugs explained by the component are aicar (target: AMPK agonist), thapsigargin (target: ATPase, Ca ++ transporting, cardiac muscle, slow twitch 2), OSU-03012 (target: PDK1/PDPK1), GSK-650394 (target: SGK3), WZ-1-84 (target: BMX) and AZD-0530 (target: SRC, ABL1). The pathways involved in mediating the downstream signaling may generate novel hypotheses for the action mechanism of these drugs.

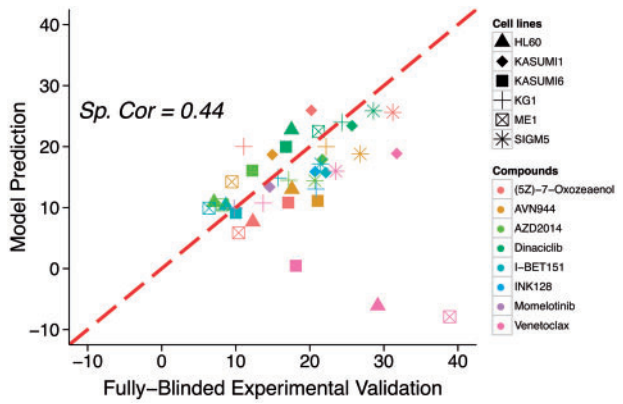

Fig. 4. Prediction of the drug sensitivity score (DSS; (Yadav *et al.*, 2014)) of 8 compounds in 6 AML cell lines. The *y*-axis shows the predictions made by the *cw*KBMF model and the *x*-axis the corresponding validations as measured in the lab. The predictions have a spearman correlation of 0.44, and the correlation increases to 0.70, if the outlier venetoclax is excluded

Component 2 is representative of MEK inhibitors RDEA119 (refametinib), PD-0325901, CI-1040 and AZD6244. Interestingly, on the pathway side, *MEK up.v1 up* is identified as one of the top pathways (shown in Fig. 5). It is biologically plausible that the drugs are connected to the up-regulation of MEK pathway, making it possible to inhibit the pathway activity in cancer using the MEK inhibitors. It is also known that MEK inhibition leads to PI3K/AKT activation (Turke *et al.*, 2012), supporting the identification of the AKT-related pathways in this component. In general, stimulation of the PI3K/AKT/mTOR cascade enhances growth, survival and metabolism of many cancer cells, and therefore PI3K/AKT/mTOR signaling pathway is a promising therapeutic target for cancer therapy. We give additional details of component 2 explaining the variation of MEK responses in Supplementary Figure S2 (right). Other drugs explained by the component are bexarotene (target: Retinioic acid X family agonist), bicalutamide (target: Androgen receptor ANDR), MG-132 (target: Proteasome), TGX221 (target: PI3K beta), Salubrinal (target: GADD34-PP1C phosphatase) and FH535. In particular FH535 primary target is unknown, however it has been shown to downregulate the activity of Wnt/β-Catenin signaling pathway (Gedaly *et al.*, 2014; Liu *et al.*, 2014). The presence of FH535 in this component suggests potential associations between FH535 response, MEK and AKT-related pathways, which could be further investigated in the lab to identify novel biomarkers for predicting FH535 responses.

The analysis conclude that pathway-drug response associations provide biologically meaningful findings. Even though these are well-known cancer-related pathways (serving as proof-of-concept positive controls), the current clinical challenge is to find the patients in which these pathways are perturbed, making it possible to select targeted treatments like MEK inhibitors individually.

5 Conclusion

We extended the KBMF method with a novel approach of component-wise MKL. In experiments with two publicly available cancer datasets, the new method showed improved predictive performance compared to other methods (including its predecessor KBMF). Additionally, we confirmed the predictive performance of the method using an in-house AML cell line panel with experimental validation, performed independently in the lab. We also showed that incorporating prior knowledge in the form of pathways helps to improve the prediction performance. We also demonstrated the usefulness of component-wise MKL, combined with prior knowledge for inferring the associations between pathways and drug responses. This way of analyzing drug responses with groups of genes (encoded in the form of pathways) may enhance our understanding of the

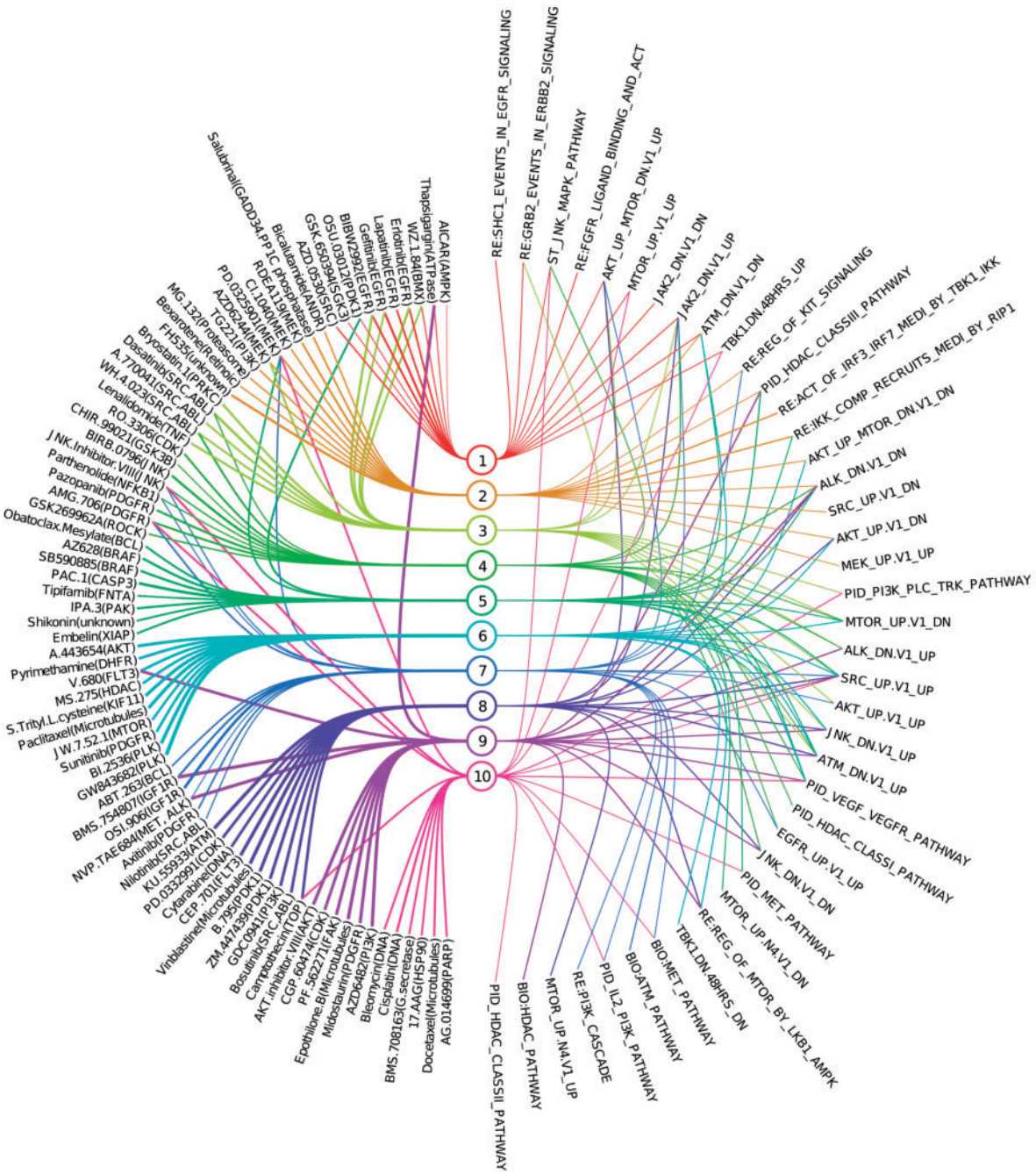

Fig. 5. Pathway-response associations decomposed into components in the GDSC dataset, visualized as an 'eye diagram' showing *cw*KBMF 10 components (circles), connecting pathways (right) and drugs (with their primary targets in parenthesis; left). The widths of the curves from the components to pathways and drugs indicate the strength of the corresponding associations. For each component, 10 drugs and 10 pathways showing the largest strength are shown

action mechanism of drugs and can potentially be used to identify novel predictive biomarkers for designing new therapies in cancer. In the future, the method could further be extended with strict sparsity assumptions for component-wise MKL, facilitating the discovery of potentially strong associations between pathways and drug responses.

Funding

This work was financially supported by the Academy of Finland (Finnish Center of Excellence in Computational Inference Research COIN; grants 295503 and 292337 to MA and SK; grants 272437, 269862, 279163 to TA), and Cancer Society of Finland (TA). We acknowledge the computational resources provided by Aalto Science-IT project and CSC-IT Center for Science Ltd.

Conflict of Interest: none declared.

References

Ammad-Ud din,M. *et al.* (2014) Integrative and personalized QSAR analysis in cancer by kernelized Bayesian matrix factorization. *J. Chem. Inf. Model.*, **54**, 2347–2359.

Barretina,J. *et al*. (2012) The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.

Basu,A. *et al*. (2013) An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell*, **154**, 1151–1161.

Baxter,J. (2000) A model of inductive bias learning. *J. Artif. Intell. Res. (JAIR)*, **12**, 149–198.

Ben-Hur,A. and Noble,W.S. (2005) Kernel methods for predicting protein–protein interactions. *Bioinformatics*, **21**, i38–i46.

Chen,B.J. *et al*. (2015) Context sensitive modeling of cancer drug sensitivity. *PloS One*, **10**, e0133850.

Cichonska,A. *et al*. (2015) Identification of drug candidates and repurposing opportunities through compound–target interaction networks. *Expert Opin. Drug Discovery*, **10**, 1–13.

Cortés-Ciriano,I. *et al*. (2015) Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel. *Bioinformatics*, **31**, btv529.

Costello,J.C. *et al*. (2014) A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol*., **32**, 1202–1212.

Friedman,J. *et al*. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw*., **33**, 1.

Garnett,M.J. *et al*. (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, **483**, 570–575.

Gedaly,R. *et al*. (2014) Targeting the Wnt/β-catenin signaling pathway in liver cancer stem cells and hepatocellular carcinoma cell lines with fh535. *PloS One*, **9**, e99272.

Gönen,M. (2012) Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics*, **28**, 2304–2310.

Gönen,M. and Alpayd,E. (2011) Multiple kernel learning algorithms. *J. Mach. Learn. Res*., **12**, 2211–2268.

Gönen,M. and Kaski,S. (2014) Kernelized Bayesian matrix factorization. *IEEE Trans. Pattern Anal. Mach. Intell*., **36**, 2047–2060.

Gönen,M. *et al*. (2013) Kernelized Bayesian matrix factorization. In: *Proceedings of The 30th International Conference on Machine Learning*. pp. 864–872.

Jang,I.S. *et al*. (2014) Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. In: *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*. p. 63. NIH Public Access.

Liberzon,A. *et al*. (2011) Molecular signatures database (msigdb) 3.0. *Bioinformatics*, **27**, 1739–1740.

Liu,J. *et al*. (2014) Fh535 inhibits the proliferation of hepg2 cells via downregulation of the Wnt/β-catenin signaling pathway. *Mol. Med. Rep*., **9**, 1289–1292.

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

Myint,K.Z. and Xie,X.Q. (2010) Recent advances in fragment-based qsar and multi-dimensional qsar methods. *Int. J. Mol. Sci*., **11**, 3846–3866.

Perkins,R. *et al*. (2003) Quantitative structure-activity relationship methods: perspectives on drug discovery. *Environ. Toxicol. Chem. Toxicol*., **22**, 1666–1679.,

Shao,C.Y. *et al*. (2013) Dependence of qsar models on the selection of trial descriptor sets: a demonstration using nanotoxicity endpoints of decorated nanotubes. *J. Chem. Inf. Model*., **53**, 142–158.

Sutherland,J.J. *et al*. (2004) A comparison of methods for modeling quantitative structure-activity relationships. *J. Med. Chem*., **47**, 5541–5554.

Turke,A.B. *et al*. (2012) MEK inhibition leads to PI3K/AKT activation by relieving a negative feedback on ERBB receptors. *Cancer Res*., **72**, 3228–3237.

Whirl-Carrillo,M. *et al*. (2012) Pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Therap*., **92**, 414.

Yadav,B. *et al*. (2014) Quantitative scoring of differential drug sensitivity for individually optimized anticancer therapies. *Scientific reports*, **4**, 5193–5202.

Yamanishi,Y. *et al*. (2012) Drug side-effect prediction based on the integration of chemical and biological spaces. *J. Chem. Inf. Model*., **52**, 3284–3292.

Yang,W. *et al*. (2013) Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res*., **41**, D955–D961.

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

Zhou,T. *et al*. (2012) Kernelized probabilistic matrix factorization: exploiting graphs and side information. In: *Proceedings of the 2012 SIAM International Conference on Data Mining*, vol. **34**, pp. 403–414.