PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

# Network-guided prediction of aromatase inhibitor response in breast cancer

**Matthew Ruffalo[1], Roby Thomas[2], Jian Chen[2], Adrian V. Lee[2], Steffi Oesterreich[2], Ziv Bar-Joseph[1,3]***

**1** Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States of America, **2** Women's Cancer Research Center, Department of Pharmacology and Chemical Biology, UPMC Hillman Cancer Center, Magee Womens Research Institute, Pittsburgh, Pennsylvania, United States of America, **3** Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States of America

* zivbj@cs.cmu.edu



OPEN ACCESS

**Citation:** Ruffalo M, Thomas R, Chen J, Lee AV, Oesterreich S, Bar-Joseph Z (2019) Network-guided prediction of aromatase inhibitor response in breast cancer. PLoS Comput Biol 15(2): e1006730. https://doi.org/10.1371/journal.pcbi.1006730

**Editor:** Christina S. Leslie, Memorial Sloan-Kettering Cancer Center, UNITED STATES

**Received:** May 23, 2018

**Accepted:** December 19, 2018

**Published:** February 11, 2019



**Data Availability Statement:** All relevant data are within the manuscript and its Supporting Information files.

**Funding:** This work was supported in part by the National Science Foundation (grant number DBI-1356505 to Z.B.J.), by the U.S. National Institute of Health (grants 1U54HL127624 to Z.B.J. and 1F32CA216937 to M.M.R.) and by the Pennsylvania Department of Health (Health Research Nonformula Grant (CURE) Awards to Z. B.J.). The funders had no role in study design, data

## Abstract

Prediction of response to specific cancer treatments is complicated by significant heterogeneity between tumors in terms of mutational profiles, gene expression, and clinical measures. Here we focus on the response of Estrogen Receptor (ER)+ post-menopausal breast cancer tumors to aromatase inhibitors (AI). We use a network smoothing algorithm to learn novel features that integrate several types of high throughput data and new cell line experiments. These features greatly improve the ability to predict response to AI when compared to prior methods. For a subset of the patients, for which we obtained more detailed clinical information, we can further predict response to a specific AI drug.

## Author summary

Breast cancer is the second most common type of cancer in women, with an incidence rate of over 250,000 cases per year, and breast cancer cases show significant heterogeneity in clinical and omic measures. Estrogen receptor positive (ER+) tumors typically grow in response to estrogen, and in post menopausal women, estrogen is only produced in peripheral tissues via the aromatase enzyme. Inhibition of aromatase is often an effective treatment for ER+ tumors, but aromatase inhibitor therapy is not effective for all tumors, and causes of this heterogeneity in response are largely not known. In this work, we present a feature construction and classification method to predict response to aromatase inhibitor therapy. We use network smoothing techniques to combine tumor omic data into predictive features, which we use as input to standard machine learning algorithms. We train predictive models using clinical data, including high-quality clinical data from UPMC patients, and show that our method outperforms previous approaches in predicting response to aromatase inhibitor therapy.

collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

## Introduction

A number of recent large efforts have focused on collecting genomic data from tumors. While these datasets led to several successful studies and insights, in many cases the clinical data available for patients enrolled in these studies is incomplete. This makes it hard to use such datasets for predicting tumor specific outcomes and tailoring treatments to individuals.

To develop accurate methods for for predicting treatment responses we need both, a comprehensive genomic dataset profiling the individuals being studied and accurate complimentary clinical information. To date, methods that used the former (detailed genomic data) usually were unable to use the latter for a significant number of individuals while methods that only relied on clinical information are limited in their ability to distinguish between tumor responses [1].

Consider, for example, the genomic data that is part of The Cancer Genome Atlas (TCGA, [2]). Several methods have used this data to study general questions related to cancer biology and prognosis. Examples include methods to identify molecular targets for cancer therapy [3], enhancement / creation of general prognostic classification systems [4–6], *de novo* pathway identification via identification of mutually exclusive mutations [7] and identification of genes implicated in cancer via combinations of different data types [8]. In contrast, most efforts for predicting response to specific treatments have been limited to much smaller datasets, usually focused only on specific pathways or classes of mutations, and often only relying on *in vitro* (cell line) experiments which have limited clinical utility [9–11]. Indeed, in many cases a key challenge researchers face when trying to predict such specific response is the lack of detailed and well-curated clinical data to supplement the high throughput molecular data in the large databases.

Here we focus on response to aromatase inhibitors (AIs), which block the conversion of androgen to estrogen and thus lower systemic estrogen. AIs show superior efficacy for the treatment of postmenopausal ER+ breast cancer compared to tamoxifen [12]. Despite the significant reduction of recurrence, resistance is common, and remains a tremendous clinical and societal problem. Mechanisms of resistance are very heterogenous [13], and it is currently not possible to accurately predict response for specific AI treatments. Thus, methods for predicting tumor specific AI responses are urgently needed, especially given availability of choices of endocrine therapy, their potential side effects, and recent findings that extended endocrine treatment benefits a subset of patients [14].

To predict AI response we developed computational methods to construct network smoothed features based on breast cancer genomic data from the Cancer Genome Atlas (TCGA) and combined these with manually curated clinical data for a subset of patients in TCGA that were treated at the University of Pittsburgh Medical Center (UPMC). Many previous approaches have been developed to integrate multiple types of omic data using a variety of techniques: multiple kernel learning [15–18], joint matrix factorization [19, 20], latent variable models [21, 22], and other network-based data integration methods [23, 24], though most of these methods have drawbacks in treatment-specific prediction tasks. Such methods are typically either unsupervised, and therefore intended for general-purpose clustering and stratification of patients, or sacrifice genomic / clinical interpretability.

The UPMC clinical data included information on the treatment patients received, its effectiveness and the outcomes. The genomic data we used included sequence variations, expression changes and cell line drug responses all smoothed using general protein-protein interaction networks. We used the clinical and genomic features to predict treatment response and overall survival. Overall we show that by combining genomic and clinical attributes we can obtain high accuracy and predicting cancer survival and slightly improve this accuracy

when incorporating functional cell line data. For the more challenging task of predicting treatment outcome we show that both, the addition of the cell line data and the improved clinical data, leads to greater accuracy and improves upon prior methods.

## Results

To predict response to aromatase inhibitors we combine high throughput gene expression and sequencing data with detailed clinical data for individual patients (Materials and methods). While using the expression, mutation and interaction data directly in a prediction framework can sometimes lead to useful results, as we show below in many cases such data is too sparse to provide useful features given the (relatively small) number of patients. Thus, a major challenge in the construction of successful prediction methods is learning useful summary features from the high throughput data. Here we use network smoothing to combine expression, mutation, protein interaction and drug response data across tumors and cell lines. These networks are then converted to PCA components which can be computed for each tumor and summarize the tumor expression and mutation information using the protein interaction network. These components, together with the clinical data are then used as features for several different classifiers we tested. Using these features and the labels obtained from the clinical records, we perform cross-validation experiments to examine our ability to predict non-response to aromatase inhibitor treatment.

Our feature construction workflow is shown in Fig 1a, which demonstrates the combination of genomic data into predictive features for UPMC samples, TCGA samples, and LINCS cell lines. Fig 1b shows the availability of features for each data source used in our analysis.

While the specific genes mutated in a pathway may vary between tumors, they are likely close in the graph representing the interaction network. A graph smoothing method tries to find mutated pathways by allowing information to propagate along edges in the graph and then finding sub-graphs (a collection of connected nodes) that are weighted highly. These subgraphs represent a set of genes that are mutated in several tumors and so are useful for predicting clinical outcomes for this set of patients. Smoothing is a general strategy and here we use it to combine several different types of genomic data including mutation and expression changes in tumors and drug response profiling in cancer cell lines. To summarize the smoothed networks in a few components (features) we perform PCA decomposition on the matrix obtained across the tumors for each data type (mutations, expression and drug targets). See Materials and methods for complete details.

### Combined prediction model outperforms individual constructed features

We first tested our method on the set of 590 breast cancer TCGA samples that were either prescribed aromatase inhibitors, or were not considered for this type of treatment given their ER status, for which we assigned artificial "non-response" labels (see Materials and methods, Classification). Figure A in S1 Text shows the univariate predictive performance of individual features we used based on ROC AUC metric, showing the top 20 and bottom 20 features sorted by AUC. As can be seen, while none of the features provide very high accuracy on their own (the best single feature is the mean across all genes of min{protein targets of arimidex, smoothed differential expression} with an AUC of 0.81), several features are still informative in isolation. Overall, the best single features are those using the PCA decomposition of the expression data and those that combine expression and drug target information (protein targets of aromasin and gene targets of estrogen receptors). We also see that the drug targets features from LINCS (Methods) related to Arimidex are only weakly informative which may

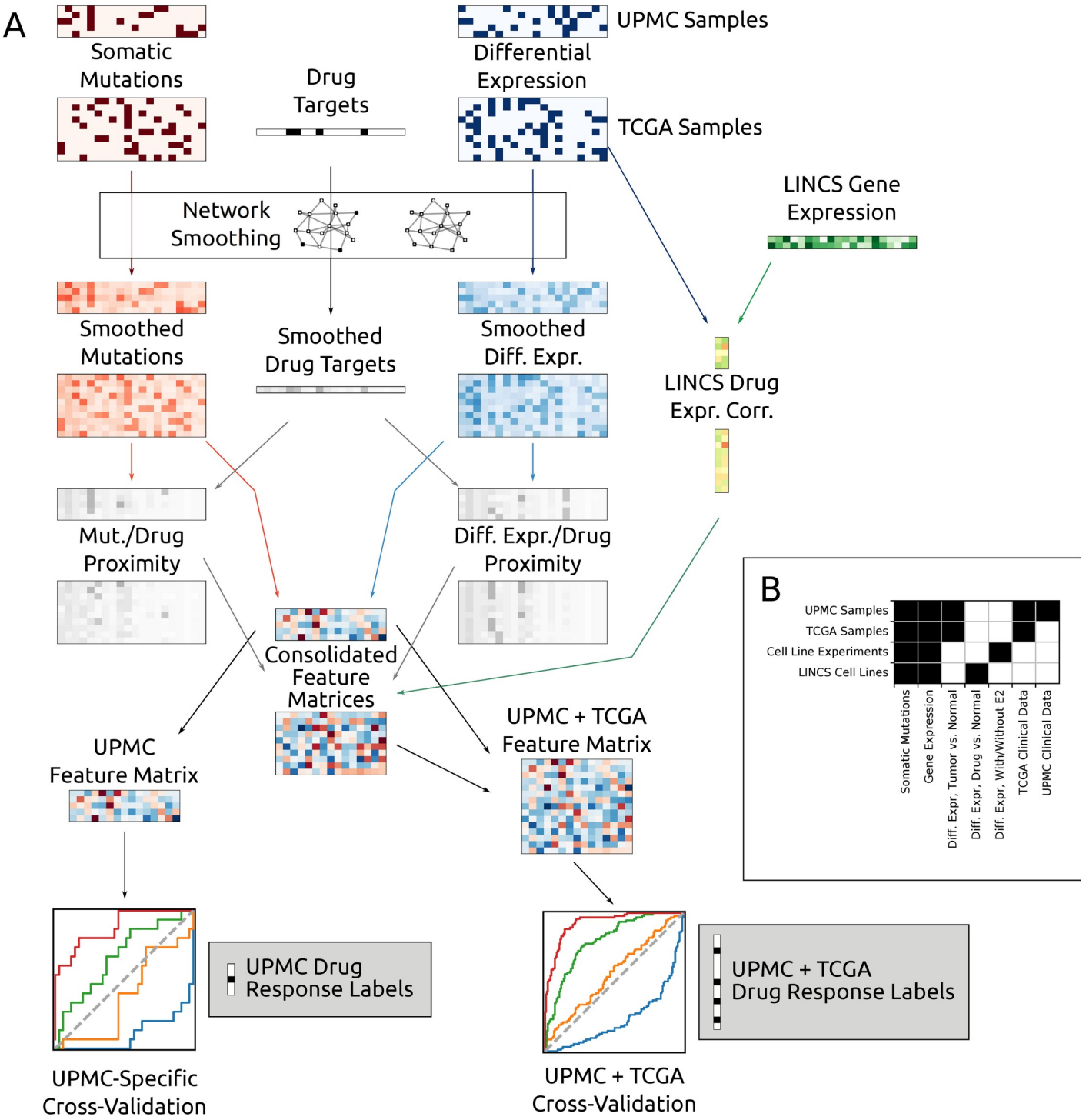


**Fig 1.** **(a)** Flowchart of our general classification approach, showing the network smoothing procedure applied to multiple data types: somatic mutations, differentially expressed genes, and protein targets for a particular drug. Smoothed mutations, differential expression, and drug targets are combined into network proximity measures by computing the element-wise minimum of the smoothed scores. Correlation is computed between LINCS expression profiles and tumor gene expression measurements. UPMC and TCGA samples are handled identically for most of the analysis pipeline until performing cross-validation: UPMC samples are used both in isolation and in combination with TCGA samples. **(b)** shows feature availability for data types used in this analysis.

https://doi.org/10.1371/journal.pcbi.1006730.g001

indicate that the specific cell line used for this drug (HA1E, kidney) is not enough for extracting general drug response profile for Arimidex.

We next trained classifiers using all features to predict general response to aromatase inhibitors. Fig 2a shows cross-validation performance in prediction of aromatase inhibitor response, using probabilistic SVM and Random Forest (RF) classifiers, the top two performing methods among those we tested (See Supplement for the performance of the other classifiers). We see that both classification methods lead to high mean ROC AUC (0.91), demonstrating the advantage of integrating several different types of features. While both methods performed equally well, it is much easier to interpret the RF results and so we focus on these results below. Fig 2b shows the importance of each feature used by RF (using the `scikit-learn` package [25]). Again, features that combine tumor expression changes with drug target information seem to be the most useful including features based on estrogen receptor targets and targets of aromasin. We also find a high scoring feature that combines tumor mutation information with estrogen receptor target information. Specific genes contributing to these high scoring PCA features are plotted in Figures B, C, and D in S1 Text. These genes include *TP53* whose mutations were identified as most significant contributors for the top feature, in line with a recent study by Gellert *et al.* [26], in which *TP53* mutations were associated with poor response in tumors treated with AIs. Other top genes included *CDH1*, which is involved in cancer progression and metastasis [27], *JUN*, a transcription factor implicated in cell proliferation and angiogenesis in invasive breast cancer [28], and *KLK4*, which codes for a kallikrein protein that is overexpressed in prostate cancer [29] and is associated with an epithelial-mesenchymal transition-like effect in prostate cancer cells [30].

## Cell line results improve prediction performance

To obtain additional data for improving the ability of our method to predict tumor response to aromtase inhibitors we performed cell line experiments. In these experiments we grew a selection of ER+ and control ER- breast cancer cell lines in serum estrogen for 5 days and then either kept the estrogen-containing serum for an additional 5 days, or switched to serum free media, thus mimicking the removal of estrogen. Growth measure results for these cells are presented in Fig 3. As can be seen, for several cells there are significant differences with and without serum estrogen. Since, unlike for the patients, we only have genomic data and no clinical information for these cells, we developed a joint prediction method by combining tumor and cell line derived classifiers (Materials and methods, Combining cell line and patient derived classifiers). The joint prediction combined the predictions of the two separate classifiers (tumor and cell line based) by learning a weight for each of them. As expected, given the small number of the cell lines tested compared to the number of patients (13 vs. 590), the weight assigned to the cell line predictor was lower (median $\gamma = 0.0276$, mean $\gamma = 0.0286$ across all leave-one-out cross-validation folds for random forest classifiers). As can be seen in Figure Q in S1 Text, with curves in the legend sorted by AUC, the addition of cell line information slightly improves cross-validation performance (though this difference is not visible when limiting AUC values to two decimal places).

## UPMC clinical data improves prediction of anastrozole response

The cross-validation results presented above correspond to an overall prediction of whether a tumor responds to an aromatase inhibitor. In the "real world", patients receive one out of three AIs, and within our cohort Arimidex (anastrozole) was the most frequently prescribed drug. Given some differences in mechanism of action, side effects and efficacies [31–33], we

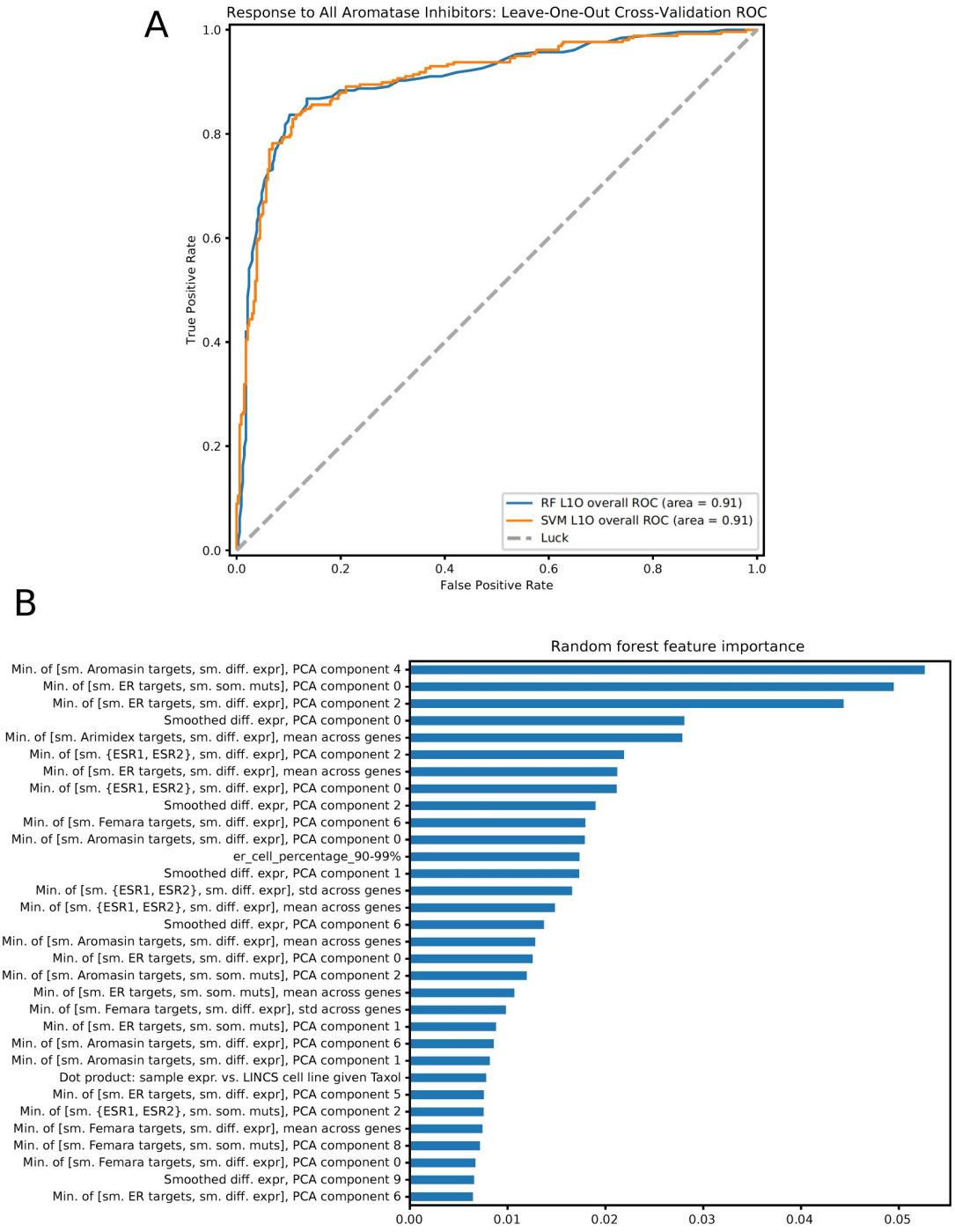


**Fig 2.** **(a)** Leave-one-out cross-validation results for prediction of non-response to all aromatase inhibitors, using random forests and probabilistic support vector machines. **(b)** Feature importance from the random forest cross-validation results, showing which constructed features contribute most to the random forest fit. Features prefixed with "Min." denote elementwise minimum of pairs of matrices, *e.g.* smoothed ("sm.") drug targets of Arimidex and smoothed binary differential expression as shown in the first feature listed. "sm. {ESR1, ESR2}" denotes network proximity to the *ESR1* and *ESR2* genes. Sample×gene matrices are collapsed across genes in various ways to produce feature values for samples: mean or standard deviation across all genes, or through PCA decomposition. Categorical clinical features are represented with one-hot encoding, and are shown as "feature name_column name", *e.g.* "er_cell_percentage_90-99%".

https://doi.org/10.1371/journal.pcbi.1006730.g002

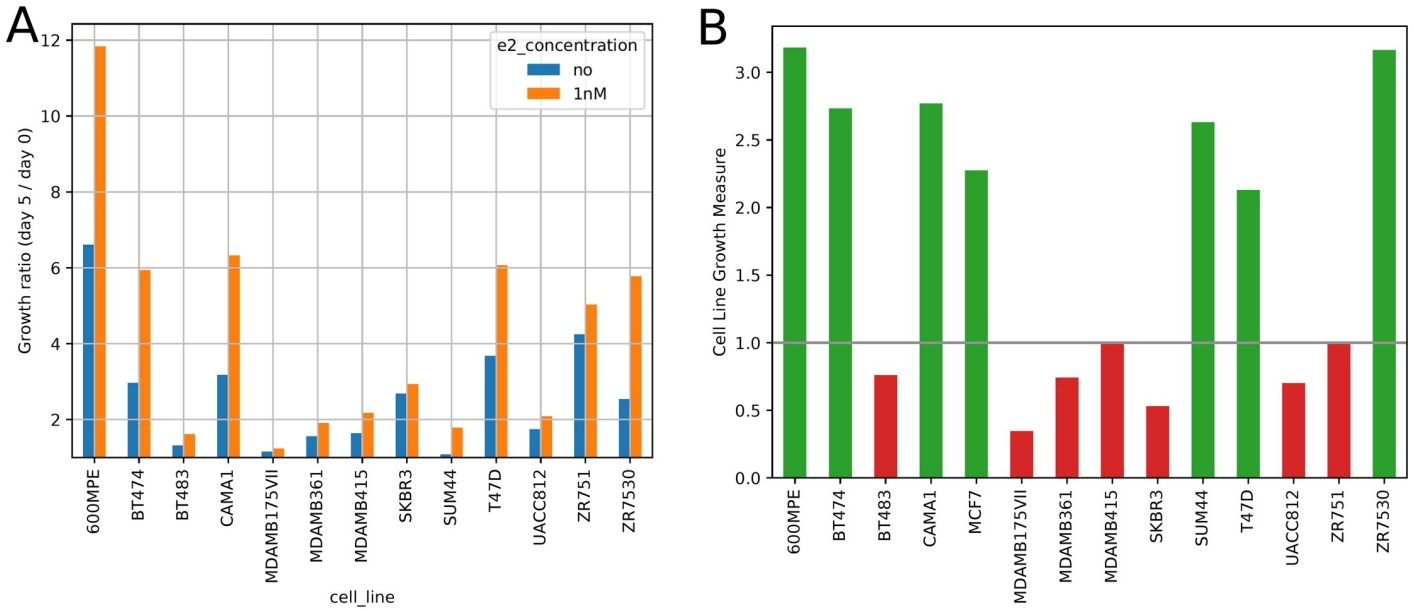


**Fig 3. Cell line growth data. (a)** shows cell line growth ratios (day 5 cell count / day 0 cell count), with and without 1nM serum estrogen. **(b)** shows the cell line growth measure defined in Eq 6. Threshold 1.0 was used to denote cell lines as responsive (green) or non-responsive (red).

https://doi.org/10.1371/journal.pcbi.1006730.g003

next used our method to predict response to Arimidex. Results are shown in Fig 4a. While this is a much more challenging prediction task than overall response to AIs (reflected in the decreased overall accuracy) the results still show the predictive power of the features that we compute. These results can be further improved with better clinical data.

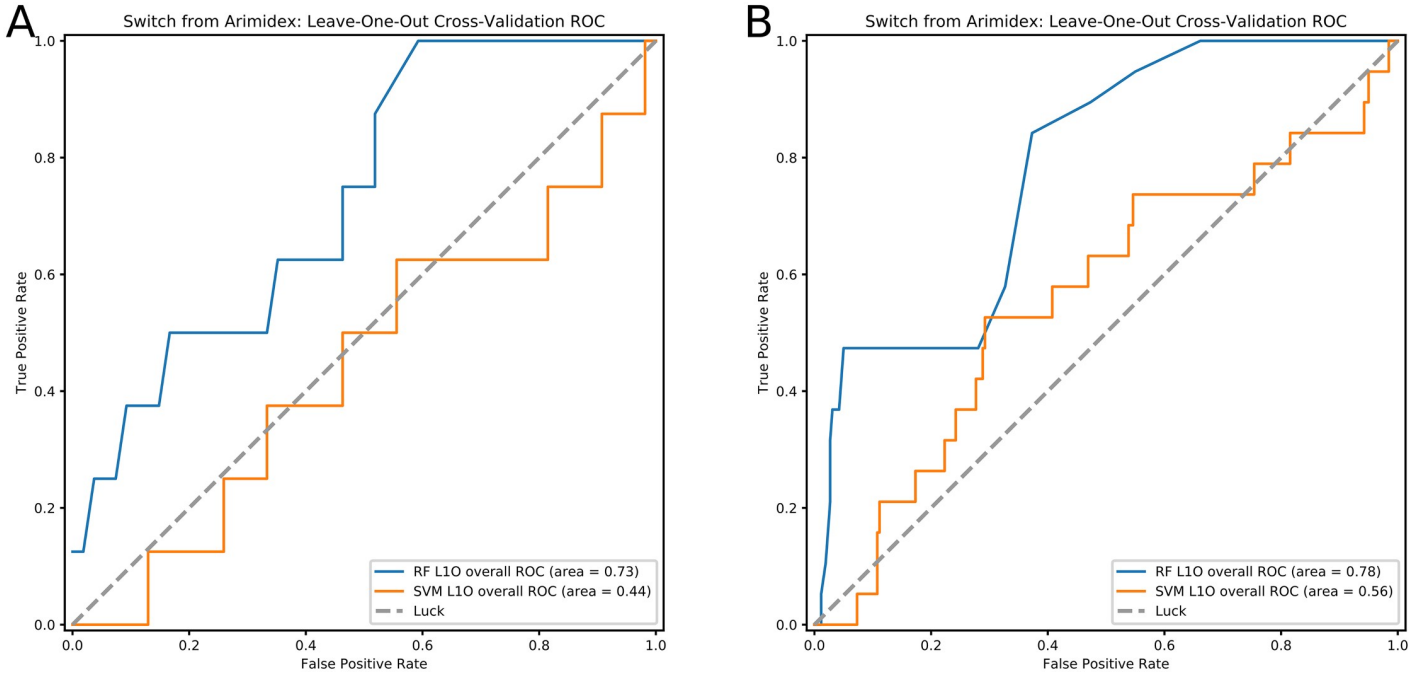


**Fig 4. Cross-validation prediction results for non-response to anastrozole. (a)** shows results for non-response to anastrozole with all available TCGA samples, and **(b)** shows prediction results restricted to UPMC patients.

https://doi.org/10.1371/journal.pcbi.1006730.g004

The discussion so far focused on all TCGA breast cancer samples. For a subset ($n$ = 151) of these patients we also have high-quality, manually-curated patient data, allowing us greater accuracy in identifying and predicting clinical outcomes (a detailed list of the clinical variables we extracted for this cohort is available on the supporting website). We thus examined the difference in performance between training and testing on all TCGA patients and using only University of Pittsburgh/UPMC patients ($n$ = 62 given anastrozole, and $n$ = 89 which were ER– or given any aromatase inhibitor). Results from this analysis are shown in Fig 4b and Figure I in S1 Text. While we do not observe a large performance difference when predicting response to all aromatase inhibitors (indicating that TCGA clinical features for such analysis are likely good enough), we see a larger improvement when predicting of response to anastrozole alone (ROC AUC 0.73 for UPMC samples compared to 0.70 for all TCGA samples). This indicates that accurate information about the specific drugs used for each patient, switching between drugs and responses and side effects, all present in the UPMC curated data but not in the TCGA data, can greatly help automated methods for feature construction in personalized medicine analysis.

## Comparison with other methods

To evaluate the usefulness of the features we constructed for this prediction task we first compared the results of using these features to methods that only use the measured expression and sequence data [34–39]. For this we constructed a "naïve" feature set, consisting only of somatic mutations, differential expression, and binary indicator columns for the clinical features (Methods). We repeat our cross-validation analysis using this feature set, using the "raw" binary features, and using the top 8, 32, 128, and 512 components from PCA decomposition/transformation of this matrix. Results are shown in Fig 5, Figure E in S1 Text and Figure G in S1 Text for all aromatase inhibitors, and Figures F, H, and R in S1 Text for anastrozole. We see that performance of these 'naïve" feature is comparable for the "all aromatase inhibitor" case (leave-one-out ROC AUC 0.90 for binary features, max 0.90 for PCA decomposition, vs. 0.91 for our constructed feature set), while it is significantly lower for the more challenging task of predicting response to anastrozole (leave-one-out ROC AUC 0.59 for binary features, 0.62 for PCA decomposition, vs. 0.70 for our constructed feature set). We also repeated the University of Pittsbugh/UPMC-only analysis with this 'naïve" feature set, and again note a large drop in performance (with ROC AUC dropping from 0.70 to 0.44 for anastrozole). See Figures J and K in S1 Text for full results.

We also compared our method to prior methods that used either a network based approach to analyze mutation data [40], relied on mutually exclusive mutations [7] for prognosis classification, or combined disparate network similarity measures across networks [23]. Results are presented in Fig 5 and Figure R in S1 Text. In general we find that such methods, which only use mutation information, do not perform as well as our methods that integrate several different types of data including expression and drug targets.

We have also compared our method to prior methods that are specifically focused on predicting AI response. Turnbull *et al.* [41] performed feature selection on gene expression data resulting in four genes which were used in a decision tree framework to predict tumor response to AIs. We reimplement their decision tree classifier (Supporting Methods) and used it for response predictions. Results are shown in Fig 5 for all aromatase inhibitor prediction, and Figure R in S1 Text for specific prediction of anastrozole non-response. The low accuracy of this method is likely due to the specific parameters used in the original study that are likely not appropriate for a the larger dataset studied in this paper. See Supporting Results for more details. Reijm *et al.* [42] developed an eight-gene classification system for prediction of AI

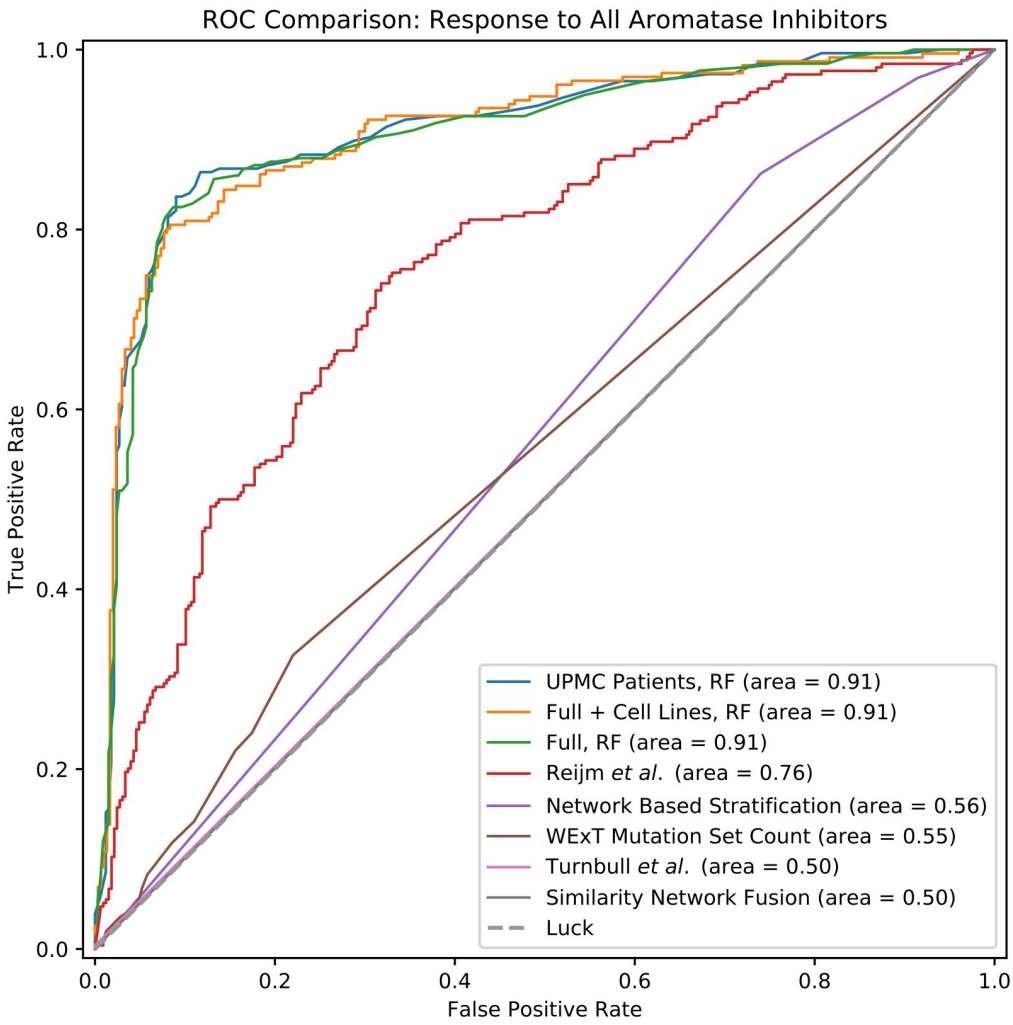


**Fig 5. Performance comparison between multiple prediction strategies, for prediction of aromatase inhibitor non-response.**

https://doi.org/10.1371/journal.pcbi.1006730.g005

response, and presented *t*-statistic values for association of these genes with tumor response. Though there are conceptual differences between *t*-statistic values and logistic regression coefficients, we nonetheless can use these *t*-values to produce continuous predictions of tumor response with log-fold gene expression data (Supplementary Methods). Results for this method are presented in Fig 5 and Figure R in S1 Text. We see reasonable performance for prediction of aromatase inhibitor response (ROC AUC 0.76) though substantially worse performance for anastrozole-specific response (ROC AUC 0.53).

## Discussion

We combined clinical and high throughput patient data with additional cell line experiments to predict tumor response to aromatase inhibitors. We developed methods for constructing PCA features by smoothing interaction networks overlaid with expression, mutation and drug target data. Our clinical data consisted of abundant (though less accurate) data for all TCGA patients and from a more detailed curated dataset for a subset of 151 patients in this set.

To further improve the classifiers and the labels, we analyzed electronic medical records within the University of Pittsburgh Medical Center (UPMC) system for the subset of TCGA patients that were treated at UPMC. These elements constituted data involving known breast cancer risk factors that either were not included in the TCGA data sets, or it was uncertain how the data was obtained and/or validated. This included reproductive history of patients at the time of breast cancer diagnosis, family history including both first and second degree relatives as well as other malignancy history for the patients if applicable. Data involving comorbid diseases that are common in adult populations including hypertension, diabetes, hyperlipidemia and metabolic information on patient's weight at the time of diagnoses were also obtained, as these may impact patients' breast cancer specific survival as well as overall survival. In regards to tumor biology, information from the EMR was obtained to the specific degree of hormone receptor status including H-score or percent staining as well as HER2/neu status. As we show, the models we developed provide accurate general predictions for the success of treatment with aromatase inhibitors. Focusing on treatment with a specific drug, Arimidex, we show that using the more detailed clinical data can lead to much better results when using our methods, greatly improving upon the use of naïve features and on prior methods suggested for this task.

For our labels, the analysis of EMRs allowed us to obtain the most up to date survival data. While most TCGA based analysis relies on survival data that was collected several years ago, EMRs are continuously updated and so we were able to use much more up to date information. Finally, we were able to use the EMRs to determine reasons for stopping specific therapy or drug including toxicity. Combined, the new features and improved labels, led to better performance for the challenging task of predicting response to a specific drug as we showed in Results.

The top-scoring PCA component in the random forest prediction is strongly influenced by the cell cycle gene *CCND1*, overexpression of which correlates with early cancer onset and tumor progression [43, 44]. It has been well known that proliferation is a strong predictive factor of endocrine treatment response, for example elegantly shown in a series of neoadjuvant short-term pre-surgical studies [45–47]. Another gene that ranked highly in our feature importance score was *CDH1*. *CDH1* encodes E-cadherin, a calcium-dependent cell-cell adhesion protein, that is frequently mutated in a number of tumor types, including breast cancer. E-cadherin protein is lost in up to 95% of invasive lobular breast cancer, and is one of the hallmark features of this disease, whereas it is lost in less than 5% of invasive ductal breast cancers [48]. We have previously shown that estrogen treatment of breast cancer cells results in downregulation of E-cadherin, potentially contributing to estrogen-mediated activation of migration and motility of cells [49]. In addition, we have shown that ILC cell lines with genetic loss of *CDH1* have a unique estrogen response compared to IDC cell lines [50]. Thus the results from this study further suggest a critical role for E-cadherin in response to estrogen and aromatase inhibitors.

Much prior work has focused on the prediction of response to endocrine therapy in general and aromatase inhibitors specifically [10, 51–53]. However, only a few of these studies are directly comparable to this work. Many prior studies use data types unavailable for large clinical datasets (*e.g.* proteomic data), focus on other organisms such as mice, or are restricted to cell lines only. We therefore focused on comparison to relevant work and on the analysis of the usefulness of the features constructed. As we show, for the more challenging task of predicting specific drug response our method outperforms especially when comparing the results for the more accurate the University of Pittsburgh/UPMC cohort.

Our results indicate that, while high throughput datasets are key to constructing accurate prediction methods, it is extremely important to couple these datasets with complete and

accurate clinical data. While information on drug prescription and usage is available for all individuals in the TCGA breast cancer dataset, we found several discrepancies between the more detailed UPMC data and the TCGA data for the same individual. This may indicate that data on other patients is noisy as well. We believe that our study provides a strong incentive for additional efforts aimed at curation of such clinical data.

## Materials and methods

### Data

The input to our method consists of genomic and clinical BRCA (breast cancer invasive carcinoma) data obtained from TCGA [54] and detailed clinical data for a subset of 151 University of Pittsburgh/UPMC patients (data description on supplementary website). The UPMC data provides specific treatments, reasons for changes in treatments, dates and responses for these patients. Specifically, we used the following data types:

- Somatic mutations obtained from whole-exome sequencing.
- Gene expression data, postprocessed and provided as part of the COSMIC cancer gene census [55], as both continuous log-fold change and binary differential expression status.
- LINCS [56] gene expression signatures for the cell lines treated with the drug we studied.
- Treatment data available in TCGA clinical information: drugs that each patient was prescribed, and global "responded to treatment" status for the patient's entire drug regimen.
- Treatment data parsed by University of Pittsburgh/UPMC researchers which in addition to the TCGA information mentioned above includes detailed clinical information about dates of specific events and reasons for patients who discontinued the use of a specific drug.

We focus on the three aromatase inhibitors prescribed most in BRCA patients: anastrozole (Arimidex), exemestane (Aromasin), and letrozole (Femara). To construct labels for tumors (response / non response) for each of these drugs, we examine the treatment information to identify which patients were prescribed that drug, and whether the patient discontinued that drug due to non-response. We then construct a "non-response" vector for each drug, denoting a patient as positive if the patient discontinued that drug or died during treatment with it.

### Pre-processing the omics data

We constructed a binary mutation matrix $M$, a log-fold gene expression matrix $E$, and a binary differential gene expression matrix $D$, with samples as rows and genes as columns. We use $C(A)$ to denote the set of column labels of matrix $A$, so that *e.g.* $C(M)$ is the set of genes that appear in the TCGA somatic mutation data. Similarly, we define $R(A)$ as the set of row labels of matrix $A$, corresponding to the distinct samples (individuals) present in each data set.

The mutation matrices $M$ are defined as

$$M[i,j] = \begin{cases} 1 & \text{if gene } j \text{ is mutated in sample } i, \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The COSMIC database [55] provides differential gene expression data for TCGA samples, represented as log-fold change between tumor and matched normal samples in the same tumor/tissue. The COSMIC database additionally annotates each log-fold differential expression measurement with "over", "under", or "normal" gene expression, for genes with log-fold differential expression outside $\sigma = 2$ standard deviations from the mean in each sample. We

collect the continuous log-fold gene expression measurements into a matrix $E$, and collect the normal/over/under expression status into a binary matrix $D$:

$$D[i,j] = \begin{cases} 1 & \text{if gene } j \text{ is over- or under-expressed in sample } i, \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

The TCGA BRCA data includes somatic mutations in 22,232 genes across 1,081 samples, and differential expression for 17,747 genes in 1,079 samples.

In addition to the condition specific omics data we also use general interaction datasets. We use the HIPPIE protein-protein interaction network [57, 58] (version 2.1, released 2017-07-18), which contains confidence scores for 318,757 interactions between 17,204 proteins. Additionally, we use gene expression data from the LINCS LDS-1191 assay, which contains measurements of gene expression in cell lines after gene knockouts and introduction of small molecules ("perturbagens"). We use gene expression data in cell lines given the chemotherapy agent Taxol, and the aromatase inhibitor Arimidex.

## Gene set network smoothing

As has been shown in the past, protein interaction networks provide a useful way to overcome data sparsity and noise when predicting cancer responses [4]. Here we use the network propagation/smoothing method described in Vanunu *et al.* [59] to combine omics data across patients. Given a network $G = (V, E, w)$ with $V$ as the set of proteins, $E$ as the set of their interactions, $w(u, v)$ representing the reliability of an interaction $(u, v) \in E$, and a prior knowledge vector $Y$: $V \rightarrow [0, 1]$, we compute a function $F(v)\ \forall v \in V$ that is both smooth over the network and accounts for the prior knowledge about each node.

This network smoothing process uses a normalized edge weight matrix $W'$, computed via Laplacian normalization of the edge weight matrix $W$: we first construct a diagonal matrix $\Delta$ with $\Delta[i, i] = \sum_j W[i, j]$, and compute $W' = \Delta^{-1/2} W \Delta^{-1/2}$. Given a prior knowledge vector $Y$, we then compute the smoothed vector $F$ using the iterative procedure described by Zhou *et al.* [60]. Starting with $F^{(0)} = Y$, we update $F$ at iteration $t$ as follows:

$$F^{(t)} = \alpha W' F^{(t-1)} + (1 - \alpha) Y \quad (3)$$

This procedure is repeated iteratively until convergence; namely we stop when $\|F^{(t)} - F^{(t-1)}\|_2 < \epsilon$. Note that Laplacian normalization produces a $W'$ with $|\lambda|_{\max} \leq 1$, which is required for this iterative method to converge.

When $Y$ is a binary vector, *i.e.* $Y[u] \in \{0,1\} \forall u \in V$, the value $F[v]$ for a gene $v$ in the smoothed vector $F$ naturally corresponds to a continuous measure of network proximity between $v$ and the "selected" genes $s \in S \subseteq V$ for which $Y[s] = 1$. We therefore this network smoothing method to compute scores of proximity for each gene with respect to multiple gene sets:

- For each tumor, genes in which that tumor harbors a non-synonymous somatic mutation.
- Differentially expressed genes in each tumor, from the COSMIC cancer gene census [55].
- Protein targets of breast cancer drugs, from queries to DGIdb [61].
- Estrogen receptor proteins ESR1 and ESR2.
- Genes targeted by (transcription factors) ESR1 and ESR2, as listed in the TRRUST database [62].

For each aforementioned gene set $S$, we construct a binary prior knowledge vector $Y_S$:

$$Y_S[s] = \begin{cases} 1 & \text{if } s \in S \cap V, \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

We then perform network propagation on the vector $Y_S$, producing a vector $F_S$. Note that not all genes in the set $S$ are necessarily included in the protein interaction network, and therefore the vectors $Y$ for *e.g.* somatic mutations in a tumor can differ from rows of the somatic mutation matrix $M$.

For somatic mutations and differential expression, we then collect the smoothed vectors into "propagated" matrices $M_P$ and $D_P$, with $R(M_P) = R(M) = R(D_P) = R(D)$ and $C(M_P) = C(D_P) = V$. Intuitively, the propagated matrices $M_P$ and $D_P$ contain the per-sample binary vectors of $M$ and $D$ smoothed over the network. In biological terms, each row of these matrices represents the network proximity of each gene product to mutated and differentially expressed genes in that sample. Consequently, as illustrated in Fig 1, the columns of these matrices provide propagated mutation and differential expression profiles for each gene product across all samples, indicating the proximity of the respective gene product to the products of mutated or differentially expressed genes in the respective sample.

## Network-integrated proximity features

We next combine the smoothed matrices $M_P$ and $D_P$ with the smoothed vectors of multiple gene sets $S$ as mentioned above:

- Proteins targeted by anastrozole
- Proteins targeted by exemestane
- Proteins targeted by letrozole
- Estrogen receptor proteins
- Genes targeted by estrogen receptors

Protein targets of each drug are obtained from queries to DGIdb [61], and gene targets of estrogen receptors are obtained from the TRRUST database [62].

Given one of the smoothed "target" vectors described above, denoted as $T$, we compute a new matrix $M_{P,T}$:

$$M_{P,T}[i,j] = \min\{M_P[i,j], T[j]\} \quad (5)$$

That is, for some tumor $i$ and gene $j$, the value $M_{P,T}[i,j]$ quantifies gene $j$'s proximity to both somatic mutations in tumor $i$ and the gene set $S$ represented by the smoothed vector $T$. We compute $D_{P,T}$ similarly, replacing $M_P$ with $D_P$ in Eq 5.

We use these matrices to compute features for response to treatment in tumor $i$:

- Row-wise mean, representing the total network proximity between a tumor's somatic mutations (or differential expression) and genes in a predefined set (*e.g.* targets of a specific drug).
- Row-wise standard deviation, quantifying the variance of mutational or differential expression proximity to drug targets.

In addition to these summary statistics for each tumor, we also perform PCA decomposition of these "minimum" matrices $M_{P,T}$ and $D_{P,T}$, and use the top 10 PCA components as predictive features. Plots of PCA component scores for genes are shown in Figures B, C, and D in

S1 Text—in these figures, genes are sorted by absolute value of there scores as assigned by PCA decomposition, and these absolute values are plotted as the importance of each gene for that PCA component.

## LINCS expression features

We obtained data from the LINCS project [56] L1000 LDS-1191 assay, which has profiled the gene expression of many cell lines under normal conditions, after introduction of small molecules ("perturbagens"), and under gene knockouts. We selected the experiments involving the drugs analyzed in this study and identified the DE genes for each of these treatments.

Two relevant drugs have been administered to cell lines by the LINCS consortium: Taxol (a taxane, also known as Paclitaxel and Abraxane), and the aromatase inhibitor Arimidex. Each of these two drugs were tested on a single cell line, and we create LINCS features for each tumor by combining that tumor's continuous log-fold differential expression with the expression change induced by that drug in the appropriate cell line. We compute two features for each (tumor,drug) pair:

- Correlation between that tumor's differential expression and the cell line differential expression induced by administering that drug.
- Dot product between that tumor's differential expression and the cell line differential expression induced by administering that drug.

While the two features above are conceptually similar, we note that in addition to the direction of agreement, the dot product also represents the *magnitude* of change in expression between a tumor and the cell line in question.

## Clinical feature extraction

We use the following categorical variables from the general TCGA clinical data:

- Tumor pathological stage
- Node pathological stage
- Metastasis pathological stage
- Overall pathological stage
- Histological type
- ICD-10 type
- ICD-O-3 histology
- HER2 immunohistochemistry level result
- Post-surgery margin status

We expand each categorical variable listed above into 0/1 indicator columns for use in classification methods. We additionally extract the estrogen receptor status of each tumor, used for selecting additional patients based on prior clinical knowledge.

## Classification

With the above features, we perform cross-validation experiments to assess our ability to predict response to aromatase inhibitor treatment. We examine all patients who were given any of the aforementioned drugs: 279 patients were given anastrozole, 51 were given exemestane, and

80 were given letrozole. An additional 180 samples were not considered for aromatase inhibitor therapy due to having ER– tumors, which are known not to respond to AI therapy. In this general aromatase inhibitor response prediction task, we assign a patient a "non-response" label if they were removed from any such drug for clinical reasons, or if the patient died during drug treatment. We also include prior clinical knowledge in this "all aromatase inhibitor" analysis; we integrate this prior knowledge by also computing the above features for the 180 patients who were not given an aromatase inhibitor, but who had estrogen receptor negative (ER–) tumors. These tumors are known not to respond to this type of treatment, so we assign these samples "non-response" labels. We use the features discussed above to learn various types of classifiers including logistic regression (with both L1 and L2 regularization), Random Forest and Probabilistic SVMs. For each of these methods and each setting we perform leave-one-out cross-validation.

## Cell line treatment

We performed cell line experiments to compare breast cancer cell line growth with and without the addition of estrogen. We initially grow cell cultures for 5 days, with estrogen present, simulating the initial growth of breast cancer in a patient. We then separately grow cultures with a continued supply of serum estrogen, or with replacement cell medium that lacks estrogen—the environment without estrogen simulates the introduction of an aromatase inhibitor. We measure cell line growth with and without serum estrogen after the initial growth period, and from these cell counts we compute measures of how much each cell line responded to the presence of estrogen. We performed this experiment with 12 replicates of each cell line, 6 with and 6 without estrogen after the initial growth period, and used a mixture of ER+ and ER– cell lines (details in Table B in S1 Text). We computed a growth measure for these cells as

$$\sqrt{\frac{1\text{nM E2 GR} - 1}{\text{no E2 GR} - 1}(1\text{nM E2 GR} - \text{no E2 GR})} \tag{6}$$

with "GR" denoting growth ratio with or without serum E2.

## Cell line experiment

MCF-7, BT474, BT483, CAMA1, Uacc812, ZR75-1 ZR75-30 and T47D breast cancer cell lines were purchased from American Type Culture Collection [ATCC], Manassas, VA, USA. SUM44PE was purchased from Asterand Bioscience, Detroit, MI, USA, and 600MPE cells were a gift by Dr. Rachel Schiff. For the estrogen removal experiments, the cells were kept for 5 days in IMEM supplemented with 10% charcoal stripped serum (CSS) with 1nM E2, and then plated into 96-well plates with or without 1 nM estradiol. An exception are Sum44PE cells that were kept in IMEM with 2% CSS. After 5 days, cell numbers were measured using Cell-titer Glo (Promega, Madison, WI, USA) according to the manufacturer's instructions. Luminescence was measured with GloMax® multi-Detection System (Promega, Madison, WI, USA), using a VICTOR X4 plate reader (PerkinElmer, Waltham, MA, USA). Bars represent the mean of six biological replicates ± SD. 17$\beta$-Estradiol (E2) was obtained from Sigma-Aldrich (St. Louis, MO, USA).

## Combining cell line and patient derived classifiers

We separated a random 10% of patients to use for training weights between tumor and cell line classifiers, and used the remaining 90% of patients for cross-validation analysis. In each cross-validation fold, we fit classifiers to the corresponding training set of patients, and then

used those classifiers to produce non-response predictions of the 10% of patients initially set aside. We then computed predictions for those 10% of patients using classifiers trained on cell lines, and chose the optimal convex combination of tumor and cell line predictions in the training set, producing final prediction $p = \gamma p_c + (1 - \gamma) p_p$, with $p_c$ denoting predictions from cell lines and $p_p$ denoting predictions from tumors. The validation set predictions then combines the tumor and cell line predictions via the hyper-parameter $\gamma$ tuned by cross-validation (note that in this way we can use the full set of features for tumors while still using the cell lines in the prediction algorithm).

## Supporting information

**S1 Text. Supporting information.** Descriptive statistics of the UPMC patient cohort, details of comparisons with other methods, and results of additional analyses.
(PDF)

**S1 Data. Analysis scripts and processed data (network-constructed features) used to produce the results shown in this work.**
(ZIP)

## Acknowledgments

We would like to thank Dr. David Davidson for helpful discussions. The results published here are based on data generated by TCGA. Information about TCGA and the investigators and institutions who constitute it can be found at https://cancergenome.nih.gov/.

## Author Contributions

**Conceptualization:** Matthew Ruffalo, Roby Thomas, Adrian V. Lee, Steffi Oesterreich, Ziv Bar-Joseph.

**Data curation:** Matthew Ruffalo, Roby Thomas, Jian Chen.

**Funding acquisition:** Ziv Bar-Joseph.

**Investigation:** Matthew Ruffalo, Steffi Oesterreich, Ziv Bar-Joseph.

**Methodology:** Matthew Ruffalo, Jian Chen, Adrian V. Lee, Ziv Bar-Joseph.

**Resources:** Steffi Oesterreich, Ziv Bar-Joseph.

**Supervision:** Ziv Bar-Joseph.

**Validation:** Matthew Ruffalo, Jian Chen, Adrian V. Lee, Steffi Oesterreich, Ziv Bar-Joseph.

**Visualization:** Matthew Ruffalo.

**Writing – original draft:** Matthew Ruffalo, Ziv Bar-Joseph.

**Writing – review & editing:** Matthew Ruffalo, Roby Thomas, Adrian V. Lee, Steffi Oesterreich, Ziv Bar-Joseph.

## References

1. Ellis MJ, Ding L, Shen D, Luo J, Suman VJ, Wallis JW, et al. Whole-genome analysis informs breast cancer response to aromatase inhibition. Nature. 2012; 486(7403):353–360. https://doi.org/10.1038/nature11143 PMID: 22722193

2. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The cancer genome atlas pan-cancer analysis project. Nature genetics. 2013; 45(10):1113–1120. https://doi.org/10.1038/ng.2764 PMID: 24071849
3. Polivka J, Janku F. Molecular targets for cancer therapy in the PI3K/AKT/mTOR pathway. Pharmacology & therapeutics. 2014; 142(2):164–175. https://doi.org/10.1016/j.pharmthera.2013.12.004
4. Zhang W, Ota T, Shridhar V, Chien J, Wu B, Kuang R. Network-based Survival Analysis Reveals Subnetwork Signatures for Predicting Outcomes of Ovarian Cancer Treatment. PLoS Comput Biol. 2013; 9 (3):e1002975+. https://doi.org/10.1371/journal.pcbi.1002975 PMID: 23555212
5. Banerjee M, Muenz DG, Chang JT, Papaleontiou M, Haymart MR. Tree-based model for thyroid cancer prognostication. The Journal of Clinical Endocrinology & Metabolism. 2014; 99(10):3737–3745. https://doi.org/10.1210/jc.2014-2197
6. Ruffalo M, Husseinzadeh H, Makishima H, Przychodzen B, Ashkar M, Koyutürk M, et al. Whole-exome sequencing enhances prognostic classification of myeloid malignancies. Journal of biomedical informatics. 2015; 58:104–113. https://doi.org/10.1016/j.jbi.2015.10.003 PMID: 26453823
7. Leiserson MD, Reyna MA, Raphael BJ. A weighted exact test for mutually exclusive mutations in cancer. Bioinformatics. 2016; 32(17):i736–i745. https://doi.org/10.1093/bioinformatics/btw462 PMID: 27587696
8. Ruffalo M, Koyutürk M, Sharan R. Network-Based Integration of Disparate Omic Data To Identify "Silent Players" in Cancer. PLOS Comput Biol. 2015; 11(12):e1004595. https://doi.org/10.1371/journal.pcbi.1004595 PMID: 26683094
9. Kavuri SM, Jain N, Galimi F, Cottino F, Leto SM, Migliardi G, et al. HER2 activating mutations are targets for colorectal cancer treatment. Cancer discovery. 2015; 5(8):832–841. https://doi.org/10.1158/2159-8290.CD-14-1211 PMID: 26243863
10. Kang J, D'Andrea AD, Kozono D. A DNA repair pathway–focused score for prediction of outcomes in ovarian cancer treated with platinum-based chemotherapy. Journal of the National Cancer Institute. 2012; 104(9):670–681. https://doi.org/10.1093/jnci/djs177 PMID: 22505474
11. Ellis MJ, Perou CM. The genomic landscape of breast cancer as a therapeutic roadmap. Cancer discovery. 2013; 3(1):27–34. https://doi.org/10.1158/2159-8290.CD-12-0462 PMID: 23319768
12. Lønning PE, Eikesdal HP. Aromatase inhibition 2013: clinical state of the art and questions that remain to be solved. Endocrine-related cancer. 2013; 20(4):R183–R201. https://doi.org/10.1530/ERC-13-0099 PMID: 23625614
13. Ma CX, Reinert T, Chmielewska I, Ellis MJ. Mechanisms of aromatase inhibitor resistance. Nature Reviews Cancer. 2015; 15(5):261. https://doi.org/10.1038/nrc3920 PMID: 25907219
14. Blok E, Derks M, van der Hoeven J, van de Velde C, Kroep J. Extended adjuvant endocrine therapy in hormone-receptor positive early breast cancer: current and future evidence. Cancer treatment reviews. 2015; 41(3):271–276. https://doi.org/10.1016/j.ctrv.2015.02.004 PMID: 25698635
15. Gönen M, Alpaydın E. Multiple kernel learning algorithms. Journal of machine learning research. 2011; 12(Jul):2211–2268.
16. Speicher NK, Pfeifer N. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. Bioinformatics. 2015; 31(12):i268–i275. https://doi.org/10.1093/bioinformatics/btv244 PMID: 26072491
17. Guadagni F, Zanzotto FM, Scarpato N, Rullo A, Riondino S, Ferroni P, et al. RISK: A random optimization interactive system based on kernel learning for predicting breast cancer disease progression. In: International Conference on Bioinformatics and Biomedical Engineering. Springer; 2017. p. 189–196.
18. Nascimento AC, Prudêncio RB, Costa IG. A multiple kernel learning algorithm for drug-target interaction prediction. BMC bioinformatics. 2016; 17(1):46. https://doi.org/10.1186/s12859-016-0890-3 PMID: 26801218
19. Fujita N, Mizuarai S, Murakami K, Nakai K. Biomarker discovery by integrated joint non-negative matrix factorization and pathway signature analyses. Scientific reports. 2018; 8(1):9743. https://doi.org/10.1038/s41598-018-28066-w PMID: 29950679
20. Chen J, Zhang S. Discovery of two-level modular organization from matched genomic data via joint matrix tri-factorization. Nucleic Acids Research. 2018;. https://doi.org/10.1093/nar/gky440
21. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009; 25 (22):2906–2912. https://doi.org/10.1093/bioinformatics/btp543 PMID: 19759197
22. Mo Q, Shen R, Guo C, Vannucci M, Chan KS, Hilsenbeck SG. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. Biostatistics. 2017; 19(1):71–86. https://doi.org/10.1093/biostatistics/kxx017

23. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nature methods. 2014; 11(3):333. https://doi.org/10.1038/nmeth.2810 PMID: 24464287

24. Vitali F, Cohen LD, Demartini A, Amato A, Eterno V, Zambelli A, et al. A network-based data integration approach to support drug repurposing and multi-target therapies in triple negative breast cancer. PloS one. 2016; 11(9):e0162407. https://doi.org/10.1371/journal.pone.0162407 PMID: 27632168

25. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011; 12(Oct):2825–2830.

26. Gellert P, Segal CV, Gao Q, López-Knowles E, Martin LA, Dodson A, et al. Impact of mutational profiles on response of primary oestrogen receptor-positive breast cancers to oestrogen deprivation. Nature communications. 2016; 7:13294. https://doi.org/10.1038/ncomms13294 PMID: 27827358

27. Beavon IR. The E-cadherin-catenin complex in tumour metastasis: structure, function and regulation. Eur J Cancer. 2000; 36(13 Spec No):1607–1620. https://doi.org/10.1016/S0959-8049(00)00158-1 PMID: 10959047

28. Vleugel MM, Greijer AE, Bos R, van der Wall E, van Diest PJ. c-Jun activation is associated with proliferation and angiogenesis in invasive breast cancer. Hum Pathol. 2006; 37(6):668–674. https://doi.org/10.1016/j.humpath.2006.01.022 PMID: 16733206

29. Xi Z, Klokk TI, Korkmaz K, Kurys P, Elbi C, Risberg B, et al. Kallikrein 4 is a predominantly nuclear protein and is overexpressed in prostate cancer. Cancer Res. 2004; 64(7):2365–2370. https://doi.org/10.1158/0008-5472.CAN-03-2025 PMID: 15059887

30. Veveris-Lowe TL, Lawrence MG, Collard RL, Bui L, Herington AC, Nicol DL, et al. Kallikrein 4 (hK4) and prostate-specific antigen (PSA) are associated with the loss of E-cadherin and an epithelial-mesenchymal transition (EMT)-like effect in prostate cancer cells. Endocr Relat Cancer. 2005; 12(3):631–643. https://doi.org/10.1677/erc.1.00958 PMID: 16172196

31. Geisler J. Differences between the non-steroidal aromatase inhibitors anastrozole and letrozole–of clinical importance? British journal of cancer. 2011; 104(7):1059. https://doi.org/10.1038/bjc.2011.58 PMID: 21364577

32. Zhao X, Liu L, Li K, Li W, Zhao L, Zou H. Comparative study on individual aromatase inhibitors on cardiovascular safety profile: a network meta-analysis. OncoTargets and therapy. 2015; 8:2721. https://doi.org/10.2147/OTT.S88179 PMID: 26491345

33. Ellis MJ, Suman VJ, Hoog J, Lin L, Snider J, Prat A, et al. Randomized phase II neoadjuvant comparison between letrozole, anastrozole, and exemestane for postmenopausal women with estrogen receptor-rich stage 2 to 3 breast cancer: clinical and biomarker outcomes and predictive value of the baseline PAM50-based intrinsic subtype–ACOSOG Z1031. J Clin Oncol. 2011; 29(17):2342–2349. https://doi.org/10.1200/JCO.2010.31.6950 PMID: 21555689

34. Golan T, Kanji Z, Epelbaum R, Devaud N, Dagan E, Holter S, et al. Overall survival and clinical characteristics of pancreatic cancer in BRCA mutation carriers. British journal of cancer. 2014; 111(6):1132–1138. https://doi.org/10.1038/bjc.2014.418 PMID: 25072261

35. Monteith JA, Mellert H, Sammons MA, Kuswanto LA, Sykes SM, Resnick-Silverman L, et al. A rare DNA contact mutation in cancer confers p53 gain-of-function and tumor cell survival via TNFAIP8 induction. Molecular Oncology. 2016; 10(8):1207–1220. https://doi.org/10.1016/j.molonc.2016.05.007 PMID: 27341992

36. McLaughlin JR, Rosen B, Moody J, Pal T, Fan I, Shaw PA, et al. Long-term ovarian cancer survival associated with mutation in BRCA1 or BRCA2. Journal of the National Cancer Institute. 2013; 105(2):141–148. https://doi.org/10.1093/jnci/djs494 PMID: 23257159

37. Phipps AI, Buchanan DD, Makar KW, Win AK, Baron JA, Lindor NM, et al. KRAS-mutation status in relation to colorectal cancer survival: the joint impact of correlated tumour markers. British journal of cancer. 2013; 108(8):1757–1764. https://doi.org/10.1038/bjc.2013.118 PMID: 23511557

38. Garcia-Murillas I, Schiavon G, Weigelt B, Ng C, Hrebien S, Cutts RJ, et al. Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. Science translational medicine. 2015; 7(302):302ra133–302ra133. https://doi.org/10.1126/scitranslmed.aab0021 PMID: 26311728

39. Sorbye H, Dragomir A, Sundstrom M, Pfeiffer P, Thunberg U, Bergfors M, et al. O-014High BRAF mutation frequency and marked survival differences in subgroups according to KRAS/BRAF mutation status and tumor tissue availability in a prospective population-based metastatic colorectal cancer cohort. Annals of Oncology. 2015; 26(suppl 4):iv113–iv113. https://doi.org/10.1093/annonc/mdv235.13

40. Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. Nature Methods. 2013; 10(11):1108–1115. https://doi.org/10.1038/nmeth.2651 PMID: 24037242

41. Turnbull AK, Arthur LM, Renshaw L, Larionov AA, Kay C, Dunbier AK, et al. Accurate Prediction and Validation of Response to Endocrine Therapy in Breast Cancer. J Clin Oncol. 2015; 33(20):2270–2278. https://doi.org/10.1200/JCO.2014.57.8963 PMID: 26033813

42. Reijm EA, Sieuwerts AM, Smid M, Bolt-de Vries J, Mostert B, Onstenk W, et al. An 8-gene mRNA expression profile in circulating tumor cells predicts response to aromatase inhibitors in metastatic breast cancer patients. BMC cancer. 2016; 16(1):123. https://doi.org/10.1186/s12885-016-2155-y PMID: 26892682

43. Diehl JA. Cycling to cancer with cyclin D1. Cancer Biol Ther. 2002; 1(3):226–231. https://doi.org/10.4161/cbt.72 PMID: 12432268

44. Velasco-Velázquez MA, Li Z, Casimiro M, Loro E, Homsi N, Pestell RG. Examining the role of cyclin D1 in breast cancer. Future Oncology. 2011; 7(6):753–765. https://doi.org/10.2217/fon.11.56 PMID: 21675838

45. Dowsett M, Smith IE, Ebbs SR, Dixon JM, Skene A, A'hern R, et al. Prognostic value of Ki67 expression after short-term presurgical endocrine therapy for primary breast cancer. Journal of the National Cancer Institute. 2007; 99(2):167–170. https://doi.org/10.1093/jnci/djk020 PMID: 17228000

46. Klintman M, Dowsett M. Early surrogate markers of treatment activity: where are we now? Journal of the National Cancer Institute Monographs. 2015; 2015(51):24–28. https://doi.org/10.1093/jncimonographs/lgv002 PMID: 26063881

47. Ellis MJ. Lessons in precision oncology from neoadjuvant endocrine therapy trials in ER+ breast cancer. The Breast. 2017;. https://doi.org/10.1016/j.breast.2017.06.039

48. Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, et al. Comprehensive molecular portraits of invasive lobular breast cancer. Cell. 2015; 163(2):506–519. https://doi.org/10.1016/j.cell.2015.09.033 PMID: 26451490

49. Oesterreich S, Deng W, Jiang S, Cui X, Ivanova M, Schiff R, et al. Estrogen-mediated down-regulation of E-cadherin in breast cancer cells. Cancer Res. 2003; 63(17):5203–5208. PMID: 14500345

50. Sikora MJ, Cooper KL, Bahreini A, Luthra S, Wang G, Chandran UR, et al. Invasive lobular carcinoma cell lines are characterized by unique estrogen-mediated gene expression patterns and altered tamoxifen response. Cancer Res. 2014; 74(5):1463–1474. https://doi.org/10.1158/0008-5472.CAN-13-2779 PMID: 24425047

51. Bacci M, Giannoni E, Fearns A, Ribas R, Gao Q, Taddei ML, et al. miR-155 drives metabolic reprogramming of ER+ breast cancer cells following long-term estrogen deprivation and predicts clinical response to aromatase inhibitors. Cancer research. 2016; 76(6):1615–1626. https://doi.org/10.1158/0008-5472.CAN-15-2038 PMID: 26795347

52. Martin EC, Krebs AE, Burks HE, Elliott S, Baddoo M, Collins-Burow BM, et al. miR-155 induced transcriptome changes in the MCF-7 breast cancer cell line leads to enhanced mitogen activated protein kinase signaling. Genes Cancer. 2014; 5(9-10):353–364. https://doi.org/10.18632/genesandcancer.33 PMID: 25352952

53. Huan J, Wang L, Xing L, Qin X, Feng L, Pan X, et al. Insights into significant pathways and gene interaction networks underlying breast cancer cell line MCF-7 treated with 17$\beta$-estradiol (E2). Gene. 2014; 533(1):346–355. https://doi.org/10.1016/j.gene.2013.08.027 PMID: 23978611

54. The Cancer Genome Atlas. Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490(7418):61–70. https://doi.org/10.1038/nature11412 PMID: 23000897

55. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Research. 2011; 39(suppl 1):D945–D950. https://doi.org/10.1093/nar/gkq929 PMID: 20952405

56. Duan Q, Flynn C, Niepel M, Hafner M, Muhlich JL, Fernandez NF, et al. LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. Nucleic acids research. 2014; p. gku476. https://doi.org/10.1093/nar/gku476

57. Schaefer MH, Fontaine JF, Vinayagam A, Porras P, Wanker EE, Andrade-Navarro MA. HIPPIE: Integrating Protein Interaction Networks with Experiment Based Quality Scores. PLoS ONE. 2012; 7(2):e31826. https://doi.org/10.1371/journal.pone.0031826 PMID: 22348130

58. Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. Nucleic acids research. 2016; p. gkw985. https://doi.org/10.1093/nar/gkw985 PMID: 27794551

59. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol. 2010; 6(1):e1000641. https://doi.org/10.1371/journal.pcbi.1000641 PMID: 20090828

60. Zhou D, Bousquet O, Lal TN, Weston J, Schölkopf B. Learning with local and global consistency. Advances in neural information processing systems. 2004; 16(16):321–328.

61. Griffith M, Griffith OL, Coffman AC, Weible JV, McMichael JF, Spies NC, et al. DGIdb: mining the druggable genome. Nature methods. 2013; 10(12):1209–1210. https://doi.org/10.1038/nmeth.2689 PMID: 24122041

62. Han H, Shim H, Shin D, Shim JE, Ko Y, Shin J, et al. TRRUST: a reference database of human transcriptional regulatory interactions. Scientific reports. 2015; 5.