Biological Constraints Can Improve Prediction in Precision Oncology

Mohamed Omar[1], Lotte Mulder[2], Tendai Coady[1], Claudio Zanettini[1], Eddie Luidy Imada[1], Wikum Dinalankara[1], Laurent Younes[3], Donald Geman[3], and Luigi Marchionni[1, +]

[1]Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY, USA

[2]Technical University Delft, Delft, The Netherlands

[3]Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD, USA

[+]Correspondence to:

Luigi Marchionni, M.D., Ph.D.

Associate Professor of Pathology and Laboratory Medicine

Weill-Cornell Medicine

413 East 69th Street,

The Belfer Research Building, Suite 1524

New York, NY, 10021, USA

Tel: (001) 646-962-8767

Fax: (001) 212-746-8192

Email: lum4003@med.cornell.edu

ABSTRACT

Machine learning (ML) algorithms are used to build predictive models or classifiers for specific disease outcomes using transcriptomic data. However, some of these models show deteriorating performance when tested on unseen data which undermines their clinical utility.

In this study, we show the importance of directly embedding prior biological knowledge into the classifier decision rules to build simple and interpretable gene signatures. We tested this in two important classification examples— a) progression in non-muscle invasive bladder cancer; and b) response to neoadjuvant chemotherapy (NACT) in triple-negative breast cancer (TNBC) – using different ML algorithms. For each algorithm, we developed two sets of classifiers: *agnostic*, trained using either individual gene expression values or the corresponding pairwise ranks without biological consideration; and *mechanistic*, trained by restricting the search to a set of gene pairs capturing important biological relations. Both types were trained on the same training data and their performance was evaluated on unseen testing data using different methodologies and multiple evaluation metrics.

Our analysis shows that mechanistic models outperform their agnostic counterparts when tested on independent data and show more consistency to their performance in the training with enhanced interpretability. These findings suggest that using biological constraints in the training process can yield more robust and interpretable gene signatures with high translational potential.

In oncology, machine learning (ML) algorithms are used to analyze gene expression data to identify predictive or prognostic gene signatures associated with specific cancer phenotypes like tumor metastasis, progression, or therapeutic response. Some of these signatures are currently being used in clinical settings to predict the prognosis and to guide further treatment[1,2]. Notably, the process of discovery and validation of gene signatures comes with great difficulties[3]. The most striking one is the unstable performance of the discovered signatures when tested on different data than the ones used in training. The main reason for this is the great discrepancy between the number of features or genes used for prediction (tens of thousands) and the number of observations or samples (tens to hundreds). What happens is that the ML model misinterprets "noise" as "signal" and ends up memorizing all the details in the training data which in turn cannot be generalized to other datasets, this phenomenon is known as overfitting[4]. There are several approaches to reduce overfitting and variance, the most important of which is by increasing the number of samples. However, this is not always feasible in biomedical and cancer research due to financial limitations or rarity of the studied cancer type. Another solution is to use simple algorithms which are less susceptible to overfitting[5] or by adding regularization on complex models[6,7]. A third option is to reduce the data dimensionality by filtering out non-informative or redundant features or using feature selection methods[8].

In this study, we examine whether embedding prior biological knowledge in the model training process can improve the performance and consistency of the resulting gene signatures. For this purpose, we limit the training process to a pre-defined relevant biological mechanism in the form of gene pairs whose relative ordering determines the predicted class. We compare the performance of these simple mechanistic rank-based models to conventional agnostic models trained without biological consideration in two different classification cases: 1. predicting the progression of non-muscle invasive bladder cancer (NMIBC) (stage T1) to muscle-invasive disease (MIBC) (stages T2-T4); and 2. predicting the response to neoadjuvant chemotherapy (NACT) in triple negative breast cancer (TNBC). In each case, we use four different ML algorithms: K-Top Scoring Pairs (K-TSPs), Support Vector Machine (SVM), Random Forest (RF) and Extreme Gradient Boosting (XGB).

For mechanistic models, we restrict the training process to a biological mechanism relevant to the phenotype under study. For bladder cancer progression, we use feed-forward loops (FFLs) consisting of transcription factors (TFs)

and microRNAs (miRNAs) target genes as our mechanism. Transcription factors regulate the expression of their target genes through various mechanisms[9]. MicroRNAs are small, non-coding RNAs that have an important role in the post-transcriptional gene regulation through the induction of mRNA degradation[10]. They also regulate the transcription of their target genes by setting a response threshold to transcriptional induction[11]. Current evidence shows that both TFs and miRNAs regulate the expression of common target genes and the expression of each other through feed-back (FBLs) and feed-forward loops (FFLs)[11–14]. Moreover, other studies have identified that the interaction between miRNAs targets and TFs is involved in the progression of several cancers including bladder cancer[15–19]. For the TNBC case, we restrict the training process to a mechanism involving both Notch and MYC signaling pathways owing to the role they play in mediating cancer chemoresistance. Notch signaling pathway is involved in promoting cancer angiogenesis and epithelial-mesenchymal transition (EMT)[20]. NOTCH signaling also promotes chemoresistance in several cancers including breast cancer by inhibiting apoptosis and mediating cancer stem cells (CSC) self-renewal capacity[21]. Similarly, c-MYC promotes chemoresistance by mediating CSC self-renewal and proliferation[22,23] and also by dysregulating the expression of some ATP-binding cassette (ABC) transporters necessary for cellular drug transport[24].

Here, we simply embed this existing knowledge into the algorithms decision rules to build biologically relevant predictive models. We show that these models, even with a very small number of features, can have a comparable performance to agnostic models using hundreds and thousands of genes.

Results

Characteristics of the datasets

For predicting bladder cancer progression, five datasets with a total of 350 samples were used in the analysis. The training data included 260 NMIBC samples of which 49 progressed to MIBC while the testing data included 90 samples of which 18 progressed to muscle-invasive stages. For predicting the response to NACT in TNBC, the training data included pretreatment samples from 112 patients, of whom 37 achieved pathological complete response (pCR) while 75 had residual disease (RD). The testing data had pre-treatment samples from 58 patients, of whom 20 achieved pCR and 38 had RD.

Predicting bladder cancer progression

In the bootstrap approach, both K-TSPs models had a similar performance at predicting bladder cancer progression in the testing data (Figure 1). However, the mechanistic K-TSPs performance in the testing was highly consistent with that in the training data suggesting that using biological constrains together with simple algorithms can improve the consistency of performance of gene signatures. For the other three algorithms, mechanistic models showed a higher testing performance than agnostic models trained using the top DEGs (Figure 1). Notably, using pairwise comparisons derived from the top DEGs, instead of using their individual expression values, improved the performance of agnostic models and slightly reduced the gap between their training and testing predictions.

Subsequently, we investigated if increasing the number of features used for training the agnostic models can improve their performance. To this end, we re-trained the agnostic models using gene expression values from the top 100, 200, and 500 DEGs or their corresponding pairwise comparisons (50, 100, and 250 pairs) and compared their performance to mechanistic models built using FFLs. The results showed that increasing the number of features used in the training did not improve the testing performance of the agnostic compared with the mechanistic models (see Figure S1).

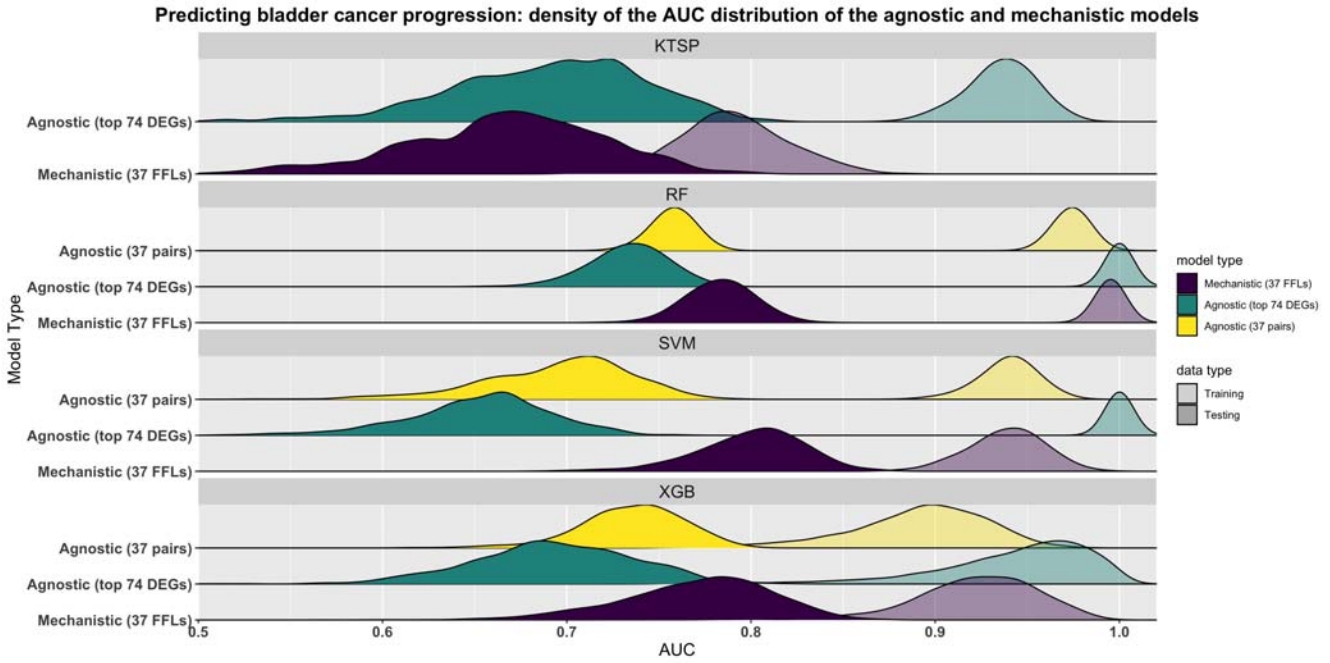

Figure 1. Performance of the agnostic and mechanistic models at predicting bladder cancer progression using the bootstrap approach. Models were trained on 1000 bootstraps of the training data (transparent colors). Mechanistic (dark violet) models were trained

using the feed-forward loops mechanism. Agnostic models were trained either using the top differentially expressed genes (green) or the corresponding pairwise comparisons (yellow). Each model was evaluated on the untouched testing data (solid colors) using the Area Under the ROC Curve (AUC). Curves represent the smoothed density distribution of the AUC values and each panel corresponds to one of the four algorithms used. KTSP: K-top scoring pairs; RF: random forest; SVM: support vector machine; XGB: extreme gradient boosting; FFLs: feed-forward loops; DEGs: differentially expressed genes.

In the cross-study validation approach, we calculated the average performance of both models across the five training-testing iterations. We found that both the mechanistic and agnostic K-TSPs had a similar average AUC in the testing data while the mechanistic model outperformed its agnostic counterpart in terms of the average balanced accuracy, sensitivity and MCC (see Table 1). In alignment with the bootstrap results, the testing performance of the mechanistic K-TSPs was also highly consistent with that in the training suggesting reduced overfitting. Similar results were seen with the other three algorithms (RF, SVM, and XGB) in which the mechanistic models had a higher average performance than their agnostic counterparts (see Table 1).

Table 1. The average performance in the cross-study validation approach. Each of the five studies was used once for testing and the other four for training. Agnostic models were trained using either individual gene expression values (Agnostic genes) or their corresponding pairwise comparisons (Agnostic Pairs). Mechanistic models were trained using the FFLs mechanism.

Features	Performance metric	KTSP *		RF		SVM		XGB	
		training	testing	training	testing	training	testing	training	testing
Agnostic genes	AUC	NA	NA	1.00	**0.72**	1.00	**0.64**	0.92	**0.64**
	Accuracy	NA	NA	1.00	**0.56**	1.00	**0.60**	0.90	**0.64**
	Balanced Accuracy	NA	NA	1.00	**0.52**	1.00	**0.50**	0.92	**0.55**
	Sensitivity	NA	NA	1.00	**0.18**	1.00	**0.06**	0.94	**0.23**
	Specificity	NA	NA	1.00	**0.87**	1.00	**0.94**	0.90	**0.86**
	MCC	NA	NA	0.99	**0.05**	1.00	**0.00**	0.81	**0.09**
Agnostic pairs	AUC	0.92	**0.71**	1.00	**0.71**	0.95	**0.64**	0.96	**0.71**
	Accuracy	0.84	**0.66**	0.94	**0.60**	0.89	**0.65**	0.88	**0.64**
	Balanced Accuracy	0.85	**0.59**	0.96	**0.55**	0.90	**0.57**	0.92	**0.60**
	Sensitivity	0.86	**0.34**	1.00	**0.27**	0.91	**0.30**	0.97	**0.38**
	Specificity	0.84	**0.85**	0.93	**0.83**	0.89	**0.84**	0.86	**0.82**
	MCC	0.61	**0.19**	0.85	**0.10**	0.73	**0.13**	0.73	**0.20**
Mechanistic pairs	AUC	0.75	**0.69**	1.00	**0.73**	0.92	**0.71**	0.91	**0.76**
	Accuracy	0.68	**0.57**	0.95	**0.62**	0.87	**0.63**	0.84	**0.70**
	Balanced Accuracy	0.72	**0.62**	0.97	**0.59**	0.87	**0.64**	0.86	**0.70**
	Sensitivity	0.77	**0.56**	1.00	**0.41**	0.86	**0.53**	0.90	**0.60**
	Specificity	0.66	**0.69**	0.94	**0.77**	0.87	**0.74**	0.83	**0.81**
	MCC	0.35	**0.22**	0.86	**0.14**	0.65	**0.25**	0.63	**0.35**

* Note that for the K-TSPs algorithm, only pairs can be used for classification.

K-TSPs: K-Top Scoring Pairs; RF: Random Forest; SVM: Support Vector Machine; XGB: Extreme Gradient Boosting; AUC: Area Under the ROC Curve; MCC: Matthews Correlation Coefficient.

Predicting the response to neoadjuvant chemotherapy in TNBC

When predicting the response to NACT in patients with TNBC, the mechanistic K-TSPs models built on the NOTCH-MYC mechanism had a similar testing performance to the agnostic models built using the top 50 DEGs (Figure 2). However, the mechanistic RF, SVM, and XGB outperformed their agnostic counterparts on the testing data (Figure 2).

Additionally, the results showed that increasing the number of features used to train the agnostic models slightly improved their classification performance. Particularly, agnostic models built using the top 200 and 500 DEGs (or 100 and 250 pairs) had a higher performance than those trained with smaller number of features (see Figure S2). Furthermore, models trained on the top 500 DEGs (or 250 pairs) achieved a similar testing performance to the mechanistic models especially using the XGB algorithm (see Figure S2).

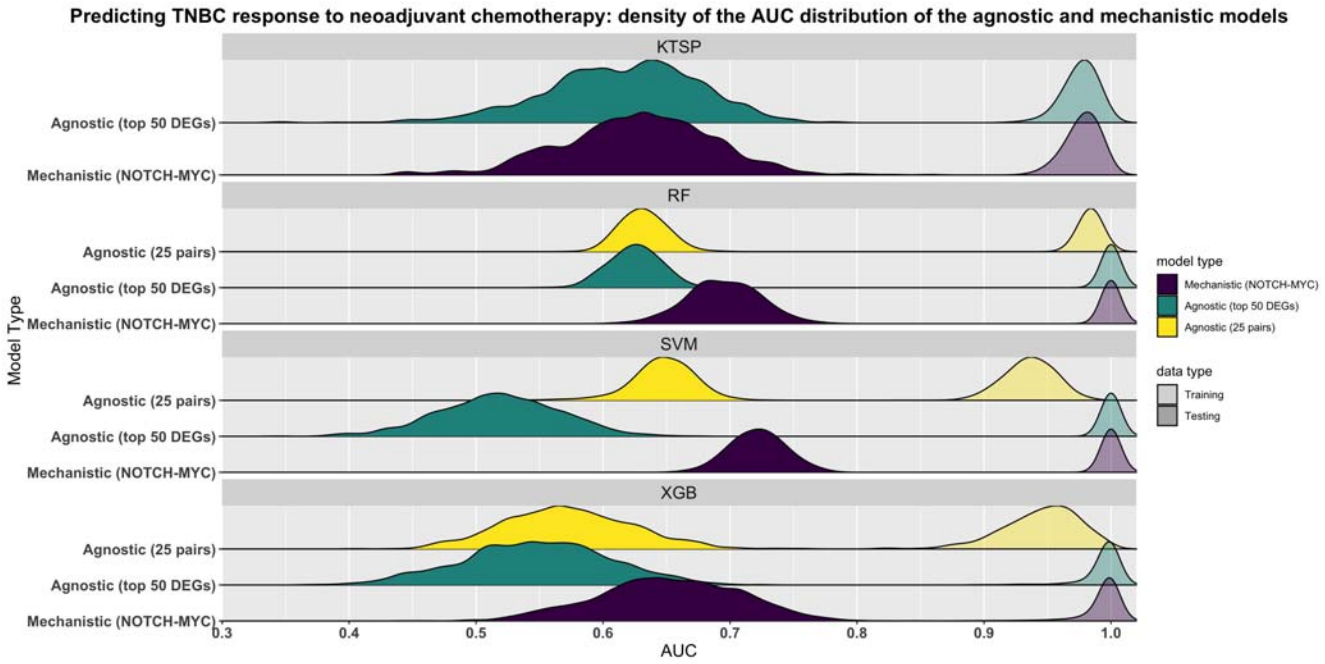

Figure 2. Performance of the agnostic and mechanistic models at predicting the response to neoadjuvant chemotherapy in triple-negative breast cancer using the bootstrap approach. Mechanistic (dark violet) models were trained using the NOTCH-MYC mechanism. Agnostic models were trained either using the top differentially expressed genes (green) or the corresponding pairwise ranks (yellow). The performance of each model was evaluated in the testing data (solid colors) using the Area Under the ROC Curve (AUC). Shown are the smoothed density distributions of the AUC values with each panel corresponding to one of the four algorithms used. KTSP: K-top scoring pairs; RF: random forest; SVM: support vector machine; XGB: extreme gradient boosting; DEGs: differentially expressed genes.

Discussion

Overfitting remains one of the most difficult problems in machine learning especially in transcriptomics owing to the very large number of features and the smaller number of samples[3,4]. The inconsistency of performance of genomic predictive models is one of the reasons why their clinical usage has not been widely implemented and represents a major obstacle to the process of personalizing health care. While some approaches like increasing the sample size may help improve the performance and stability of such models, these may not always be feasible in clinical research owing to financial limitations or the unavailability of samples in cases of rare or lethal cancer types.

Several studies have shown that using prior knowledge can help to choose the input data or the correct algorithm for the mission at hand[25,26]. In this study, we employed a similar concept to train stable and interpretable predictive models by adding biological constraints to the algorithm decision rules. We examined this approach in two clinically important classification cases: predicting bladder cancer progression from non-muscle invasive to muscle invasive stages and predicting the response to neoadjuvant chemotherapy in TNBC. In each, we used different algorithms to predict the phenotype of interest and compared the performance of biologically constrained rank-based models to agnostic ones trained using the top DEGs or their derived pairwise ranks. In the bladder cancer case, we used FFLs to construct the mechanistic model based on the evidence supporting their involvement in cancer progression and invasion[15–19]. In the TNBC case, we used a mechanism based on the NOTCH and c-MYC signaling pathways based on their role in mediating CSC self-renewal and chemo-resistance[21–23].

When evaluated on the testing data, the mechanistic rank-based models yielded a similar or even better performance than agnostic models trained using the individual expression values of the top DEGs or their corresponding pairwise ranks. Moreover, in the bladder cancer case, mechanistic models, especially K-TSPs, had a more consistent performance between the training and testing data compared to their agnostic counterparts. This was evident in the bootstrap approach in which models were trained on 1000 training bootstraps and tested independently indicating that our results are not confounded by certain specifications in the training data. This was also confirmed in the cross-study validation approach in which the mechanistic models had a higher average testing performance using different metrics. Even in situations where there was a small number of samples

available for training like the TNBC case, the mechanistic models still outperformed their agnostic counterparts in the testing and their performance was matched by the agnostic models only when a large number of features (> 200) was used.

Overall, these results show that models using a small number of biologically important features can have a comparable performance to those using hundreds and thousands of genes. Furthermore, these mechanistic models use simple decision rules that depend entirely on the relative ordering of features in each sample rather than their definite expression values. These simple rank-based decision rules greatly enhance the interpretability of the resulting gene signatures which is crucial to their clinical usage. Also, these models are robust to preprocessing techniques and can be feasibly implemented by another technology as long as the ranking is preserved. For example, signatures derived from microarrays or RNA-Seq studies can be transformed to a more clinically feasible technology like RT-PCR which further increases their translational value.

Notably, our study has some inherent limitations. First, the biological mechanisms used in the mechanistic models are in the form of contrasting gene pairs but this pairwise relationship may not completely capture the underlying biology compared to other formats like gene networks for example. However, this lack of sophistication in the design of the biological mechanism was deemed necessary in our analysis for several reasons related to the format requirement of the used algorithms and the future interpretability of the resulting signatures. For example, the K-TSPs algorithm is naturally dependent on the relative ordering of gene pairs and restricting the training process can either be implemented at the gene or pair-level. At the gene-level, the selection of gene pairs can be restricted to certain features like the top DEGs. At the pair-level, the selection can be restricted to certain pre-specified set of pairs whose rank is expected, based on biological judgment, to differ significantly between the two classes. Additionally, these pairwise comparisons are extremely simple and can be used as predictors for other more complex algorithms and the resulting signatures will still retain a higher level of interpretability compared to models trained using the individual gene expression values or using a network-based biological mechanism. Another limitation resides in the lack of external validation of the results. Although we tested the performance of both the agnostic and mechanistic models on independent testing datasets, there is still the need to validate these results in external patient cohorts to confirm the conclusions.

Despite these limitations, our results show that simple rank-based mechanistic models trained on relevant biological mechanisms have a similar or even better performance than conventional agnostic models. These results are evident in two different clinical cases, using different methodological approaches and different ML algorithms. The simplicity of these rank-based gene signatures allows for their feasible cross-platform implementation and this, combined with their high performance and interpretability, increase their translational potential. Improving the interpretability and stability of gene signatures can eventually increase their integration in clinical practice on a wider scale and bring routine personalized medicine one step closer to reality.

Materials and Methods

Data collection

The bladder cancer case

We used both the NCBI Gene Expression Omnibus (GEO)[27] and ArrayExpress[28] to identify gene expression datasets containing primary tumor samples from non-muscle invasive bladder cancer (NMIBC). We refined the initial results to keep only the datasets with information about the progression status (progression to MIBC versus no progression). Five datasets met our inclusion criteria, four of which are microarray-based (GSE57813[29], GSE13507[30], GSE32894[31] and pmid15930337[32]) and the fifth is RNA-Seq based dataset (E-MTAB-4321[33]).

The breast cancer case

We included the two datasets (GSE25055 and GSE25065) used by Hatzis et al to discover a gene signature associated with the response to neoadjuvant chemotherapy (NACT) in breast cancer[34].

Data preprocessing

The bladder cancer case

In each dataset, we removed non-invasive papillary carcinoma (Ta) and carcinoma in situ (Tis) samples and kept only T1 lesions with information about the progression status. To remove uninformative features in the microarray datasets, we kept genes with raw intensity greater than 100 in at least 50% of the samples. Similarly, in E-MTAB-4321 (RNA-Seq) we kept genes with more than one count per million (CPM) in at least 50% of the

samples. The four microarray datasets were normalized and log2-scaled upon retrieval from GEO. For E-MTAB-4321, the read counts were normalized using trimmed mean of M-values (TMM) and transformed to log2-counts per million (log-CPM). Next, we performed Z-score transformation (by gene) of each normalized dataset separately to ensure that the datasets from both technologies (microarrays and RNA-Seq) are on a similar scale.

To identify a subset of cross-study reproducible genes, we used the integrative correlation coefficient (ICC)[35,36] keeping only genes whose ICC was greater than 0.15 or the 33rd percentile. In summary, the ICC is computed by calculating the Pearson correlation coefficient of the expression values of each pair of genes within and across studies (correlation of correlation). Although the integrative correlation analysis was performed on all data before division into training and testing, this does not violate the validation process since this method only uses the expression data and does not take into account the phenotype information. Of the 5139 genes in common between the five datasets, 3109 genes met the ICC threshold and were used in the downstream analysis.

The breast cancer case

In both datasets, we kept only TNBC samples in which ER, PR, and HER2 were all negative by immunohistochemistry (IHC) and with available information about the response to NACT whether pathological complete response (pCR) or residual disease (RD). This reduced the number of samples to 112 and 58 in the first and second datasets, respectively. Similar to the bladder cancer example, we removed uninformative features using unsupervised filters based on the probe raw intensity. Finally, we mapped each probe ID to the corresponding gene symbol and restricted each expression matrix to the gene symbols in common (4892 genes).

Mechanistic pairs assembly

We took advantage of existing biological knowledge to construct biological mechanisms most relevant to the phenotypes under study. These biological mechanisms are in the form of gene pairs, each consisting of a gene associated with bad prognosis (e.g., progression or chemo-resistance) and another associated with good prognosis.

Feedforward loops

The TF-miRNA mediated gene regulatory loops that we are interested in are the coherent feed-forward loops in which a TF (e.g., MYC) inhibits a target gene (e.g., CD164) directly and indirectly via activation of a hub miRNA (e.g., hsa-miR-346). The TF and target gene have an inverse relationship; over-expression of the TF results in down-regulation of the target gene and vice versa. This inverse relationship makes these pairs suitable for classification. To construct these loops, three different interaction types have to be obtained: the interaction between the TF and target gene (TF-target), the interaction between the TF and miRNA (TF-miRNA), and the interaction between miRNA and target gene (miRNA-target). The TF-target interactions were obtained from Harmonizome[37] using the following databases: ENCODE, ESCAPE, CHEA, JASPAR, MotifMap and TRANSFAC. The TF-miRNA interactions were obtained from the same databases as above together with TransmiR v2.0 database[38]. Finally, the miRNA-target interactions were obtained from TargetScan[39], miRTarBase[40] and miRWalk[41].

The loops were constructed by merging the three different interaction types. It was assumed that the TF always activates the miRNA and always inhibits the target gene. This assumption could be made since loops in which the TF does not activate the miRNA and/or inhibit the target gene, will not be selected as top scoring pairs by the K-TSPs algorithm, as described below. Finally, we chose TF-miRNA and TF-target interactions which were present in at least one of the databases and miRNA-target interactions which were present in at least two databases. This resulted in 985 gene pairs which were used for predicting bladder cancer progression.

The NOTCH-MYC mechanism

We used the Molecular Signature Database (MsigDB)[42] to retrieve gene sets associated with the regulation of the NOTCH signaling pathway or including genes up and downregulated by NOTCH. The NOTCH mechanism was constructed by pairing the genes involved in the positive regulation of NOTCH signaling pathway or genes up-regulated by NOTCH with those involved in the downregulation of the NOTCH signaling pathway or those down-regulated by NOTCH. Similarly, the MYC mechanism was constructed by pairing the genes up-regulated with those down-regulated by c-MYC. Finally, both mechanisms were combined into a single mechanism consisting of 7420 pairs which was further used in the TNBC classification case.

Data splitting

In the bladder cancer analysis, we implemented two data splitting approaches: bootstrap and cross-study validation[43] (see Figure 3). In the bootstrap approach, the five datasets were combined together based on the set of reproducible genes. The data was then divided into 75% training and 25% testing using balanced stratification. This was done to ensure a balanced representation of the parent datasets together with important clinical and pathological variables (age, sex, tumor grade, recurrence status, and intra-vesical therapy) in the training and testing data. Subsequently, models were trained to predict bladder cancer progression on 1000 bootstraps of the training data and their performance was evaluated on the untouched testing data using the Area Under the ROC Curve (AUC) as evaluation metric. In the cross-study validation approach, four out of five datasets were used for model training and the fifth was used for testing. This process was repeated five times so that each dataset was used for testing once.

In the TNBC analysis, we implemented the bootstrap approach using GSE25055 (112 samples) for training and GSE25065 (58 samples) for testing. Models were trained on 1000 bootstraps of the training data to predict the response to NACT (pCR vs RD), and the testing data was used to assess their performance (Figure 3).

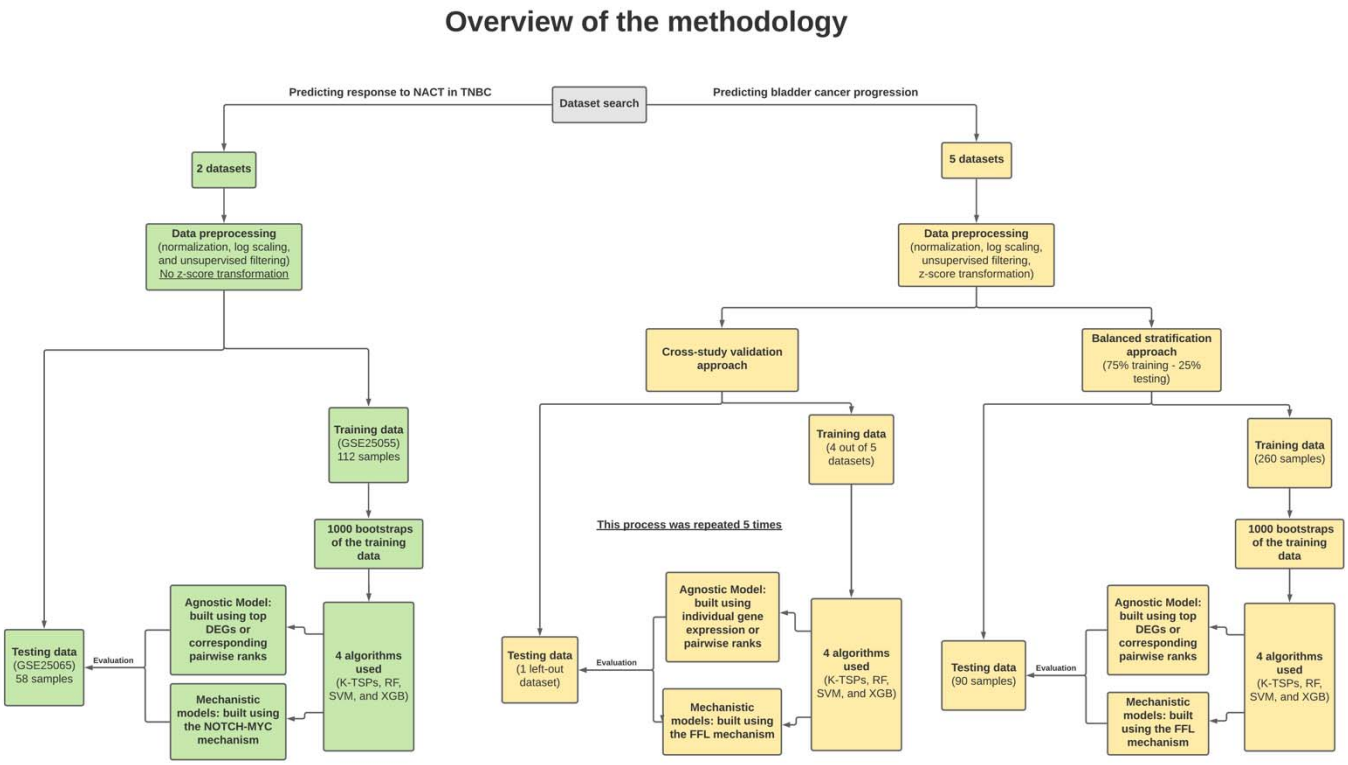

Figure 3. An overview of the data collection, preprocessing and training and testing approaches. Two different classification cases

were considered: predicting bladder cancer progression (yellow section) and predicting the response to NACT in TNBC patients (green section). NACT: neoadjuvant chemotherapy, TNBC: triple-negative breast cancer, K-TSPs: K-top scoring pairs, RF: random forest, SVM: support vector machine, XGB: extreme gradient boosting, DEGs: differentially expressed genes.

Evaluating the performance of mechanistic versus agnostic models

In each classification case, we used four different algorithms: K-TSPs, RF, SVM, and XGB. Each algorithm was trained using two different model types: 1. mechanistic: using a manually curated biological mechanism in the form of pairwise comparisons (see below), and 2. agnostic: using the top differentially expressed genes (DEGs) (agnostic-genes) or the corresponding pairwise comparisons (agnostic-pairs). Importantly, a pairwise comparison is based on the relative ordering of the expression of two genes. For example, in a particular sample, a given gene pair consisting of *gX* and *gY* would be assigned a value of '1' if *gX* is more expressed than *gY* in that sample, and a value of '0' if the opposite is true. Such pairwise comparisons were then used as features in the training process of mechanistic and agnostic-pairs models.

Importantly, both agnostic and mechanistic models were trained and tested on the corresponding training and testing data, respectively. In the bootstrap approach, the AUC of each model was computed in both the training and testing data. The distribution of the AUC values of the mechanistic was plotted against that of the agnostic models to compare their average performance. In the cross-study validation, the average performance across the five iterations of training and testing was computed. Different metrics were used including: the AUC, accuracy, balanced accuracy, sensitivity, specificity, and Matthews correlation coefficient (MCC).

The K-TSPs classifier

The K-TSPs is a rank-based classification method that selects gene pairs (K) whose expression levels switch their ranking between the two classes of interest[44,45]. More specifically, in the training process, if gene X is consistently more expressed relative to gene Y in samples of a particular class compared to the other, it will be selected as a top scoring pair (TSP) and used for classification. In this sense, the output of this algorithm is a number of gene pairs with each voting for a particular class based on the relative ordering of expression values and the final class prediction is determined by the sum of votes. This rank assessment process used by the

algorithm during the training process can be applied to all genes or can be restricted to the top DEGs (agnostic) or to certain predetermined pairs chosen based on prior knowledge (mechanistic). The agnostic K-TSPs models were trained on the top DEGs by Wilcoxon rank sum test using different number of top features: the top 74, 100, 200, and 500 DEGs in the bladder and the top 25, 50, 100, 200, and 500 DEGs in the TNBC case. The training of the mechanistic K-TSPs models was restricted to the FFLs and the NOTCH-MYC mechanisms in the bladder and TNBC cases, respectively. In both, we restricted the number of output pairs (the final signature) to a range between 3 and 25 pairs. Finally, pairwise comparisons derived from the maximum number of unique pairs returned by the mechanistic K-TSPs models (37 and 84 in the bladder and TNBC, respectively) was used to train the mechanistic versions of RF, SVM, and XGB algorithms.

Support Vector Machine

SVM is an algorithm that aims at identifying a hyper-plane separating data points distinctively[46]. We trained the agnostic and mechanistic SVM models using polynomial kernel and used a repeated 10-fold cross-validation (CV) of the training data to identify the best parameter (degree, scale, and cost) values for each model. The final models were trained on the entire training data using the best parameters resulting from the repeated CV process.

Random Forest

RF is an ensemble ML algorithm that consists of a large number of decision trees[47]. Each tree in the forest votes for a specific class and the final predicted class is the one with the majority of votes. To determine the best number of variables randomly selected by the algorithm at each split (mtry), each model was tuned by the *tuneRF* function using the following parameters: *mtryStart* = 1, *ntreeTry* = 500, *stepFactor* = 1, and *improve* = 0.05. To deal with class imbalance, the final model was instructed to draw an equal number of samples from both classes for each tree. This number was set to be equal to the number of samples in the minority class of each of the training data re-samples (in the bootstrap approach) or the training data as a whole (in the cross-study validation approach).

Extreme gradient boosting

Similar to RF, XGB is another ensemble ML algorithm but unlike RF in which each tree is built on a random subset of predictors, XGB sub-models (sub-trees) sequentially add weight or more focus on instances with high error rates[48]. We divided the training data itself into 70% "actual training" and 30% "internal validation". We set the number of iterations to 500 with an early stopping threshold of 50 meaning that the training process will stop if the AUC in the internal validation set did not improve over 50 iterations. This step is necessary to minimize overfitting.

Software and packages

All steps of this analysis were performed using R programming language[49]. The integrative correlation analysis was performed using the *MergeMaid* package[50]. The K-TSPs models were constructed using *SwitchBox* package[51]. The SVM models were constructed using both the Caret[52] and the Kernlab[53] packages. The RF models were constructed using the *RandomForest* package[54] and the XGB models using *xgboost* package[48].

Availability of data and materials

All datasets used in this study are publicly available from the Gene Expression Omnibus (GEO) and ArrayExpress under the corresponding accession number. The code for this analysis can be accessed at the following GitHub repository: https://github.com/marchionniLab/BiologicalConstraints

Acknowledgements

We thank Dr. Soren Vang (Department of Molecular Medicine, Aarhus University Hospital, Denmark) for providing the raw RNA-Seq counts of the E-MTAB-4321 dataset. This publication was made possible through support from the NIH-NCI grant R01CA200859.

Author contributions

LM and DG conceived the research question. MO, LM, TC, CZ, ELI, and WD collected the datasets and gene sets. MO performed the analysis and wrote the manuscript. LM, DG, and LY supervised the analysis and the manuscript writing. All authors read and approved the final version of the manuscript.

Competing Interests Statement

The authors declare no competing interests.

References

1. Cardoso, F. *et al.* 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. *New England Journal of Medicine* **375**, 717–729 (2016).
2. Knezevic, D. *et al.* Analytical validation of the Oncotype DX prostate cancer assay - a clinical RT-PCR assay optimized for prostate needle biopsies. *BMC genomics* **14**, 690–690 (2013).
3. Mirza, B. *et al.* Machine Learning and Integrative Analysis of Biomedical Big Data. *Genes* **10**, 87 (2019).
4. Keogh, E. & Mueen, A. Curse of dimensionality. *Encyclopedia of machine learning* 257–258 (2010).
5. Hand, D. J. Classifier Technology and the Illusion of Progress. *Statistical Science* **21**, 1–14 (2006).
6. Neumaier, Arnold. Solving Ill-Conditioned and Singular Linear Systems: A Tutorial on Regularization. *SIAM Review* **40**, 636–666 (1998).
7. Chicco, D. Ten quick tips for machine learning in computational biology. *BioData mining* **10**, 35–35 (2017).
8. Mahendran, N., Durai Raj Vincent, P. M., Srinivasan, K. & Chang, C.-Y. Machine Learning Based Computational Gene Selection Models: A Survey, Performance Evaluation, Open Issues, and Future Research Directions. *Frontiers in genetics* **11**, 603808–603808 (2020).
9. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).
10. O'Brien, J., Hayder, H., Zayed, Y. & Peng, C. Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. *Frontiers in Endocrinology* **9**, 402 (2018).
11. Hausser, J. & Zavolan, M. Identification and consequences of miRNA-target interactions–beyond repression of gene expression. *Nat Rev Genet* **15**, 599–612 (2014).
12. Martinez, N. J. *et al.* A C. elegans genome-scale microRNA network contains composite feedback motifs with high flux capacity. *Genes Dev* **22**, 2535–2549 (2008).
13. Re, A., Cora, D., Taverna, D. & Caselle, M. Genome-wide survey of microRNA-transcription factor feed-forward regulatory circuits in human. *Mol Biosyst* **5**, 854–867 (2009).
14. Friard, O., Re, A., Taverna, D., De Bortoli, M. & Cora, D. CircuitsDB: a database of mixed microRNA/transcription factor feed-forward regulatory circuits in human and mouse. *BMC Bioinformatics* **11**, 435 (2010).
15. Li, Q.-Q. *et al.* Involvement of NF-κB/miR-448 regulatory feedback loop in chemotherapy-induced epithelial-mesenchymal transition of breast cancer cells. *Cell Death Differ* **18**, 16–25 (2011).

16. Guo, Y. *et al.* miR-144 downregulation increases bladder cancer cell proliferation by targeting EZH2 and regulating Wnt signaling. *FEBS J* **280**, 4531–4538 (2013).

17. Liu, J.-J. *et al.* A novel AP-1/miR-101 regulatory feedback loop and its implication in the migration and invasion of hepatoma cells. *Nucleic Acids Res* **42**, 12041–12051 (2014).

18. Dong, F. *et al.* Dysregulation of miRNAs in bladder cancer: altered expression with aberrant biogenesis procedure. *Oncotarget* **8**, 27547–27568 (2017).

19. Mullany, L. E. *et al.* MicroRNA-transcription factor interactions and their combined effect on target gene expression in colon cancer cases. *Genes, chromosomes & cancer* **57**, 192–202 (2018).

20. Abdullah, L. N. & Chow, E. K.-H. Mechanisms of chemoresistance in cancer stem cells. *Clin Transl Med* **2**, 3 (2013).

21. Ranganathan, P., Weaver, K. L. & Capobianco, A. J. Notch signalling in solid tumours: a little bit of everything but not all the time. *Nat Rev Cancer* **11**, 338–351 (2011).

22. Wang, J. *et al.* c-Myc Is Required for Maintenance of Glioma Cancer Stem Cells. *PLoS One* **3**, (2008).

23. Zhang, H.-L., Wang, P., Lu, M.-Z., Zhang, S.-D. & Zheng, L. c-Myc maintains the self-renewal and chemoresistance properties of colon cancer stem cells. *Oncol Lett* **17**, 4487–4493 (2019).

24. Porro, A. *et al.* Direct and coordinate regulation of ATP-binding cassette transporter genes by Myc factors generates specific transcription signatures that significantly affect the chemoresistance phenotype of cancer cells. *J Biol Chem* **285**, 19532–19543 (2010).

25. Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. *Nature reviews. Genetics* **16**, 321–332 (2015).

26. Yip, K. Y. *et al.* Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol* **13**, R48 (2012).

27. Barrett, T. *et al.* NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res* **35**, D760-5 (2007).

28. Kolesnikov, N. *et al.* ArrayExpress update–simplifying data submissions. *Nucleic acids research* **43**, D1113—6 (2015).

29. van der Heijden, A. G. *et al.* A five-gene expression signature to predict progression in T1G3 bladder cancer. *Eur J Cancer* **64**, 127–136 (2016).

30. Kim, W.-J. *et al.* Predictive value of progression-related gene classifier in primary non-muscle invasive bladder cancer. *Mol Cancer* **9**, 3 (2010).

31. Sjodahl, G. *et al.* A molecular taxonomy for urothelial carcinoma. *Clin Cancer Res* **18**, 3377–3386 (2012).

32. Dyrskjøt, L. *et al.* A Molecular Signature in Superficial Bladder Carcinoma Predicts Clinical Outcome. *Clinical Cancer Research* **11**, 4029 (2005).

33. Hedegaard, J. *et al.* Comprehensive Transcriptional Analysis of Early-Stage Urothelial Carcinoma. *Cancer Cell* **30**, 27–42 (2016).

34. Hatzis, C. *et al.* A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *JAMA* **305**, 1873–1881 (2011).

35. Parmigiani, G., Garrett-Mayer, E. S., Anbazhagan, R. & Gabrielson, E. A cross-study comparison of gene expression studies for the molecular classification of lung cancer. *Clin Cancer Res* **10**, 2922–2927 (2004).

36. Cope, L., Naiman, D. Q. & Parmigiani, G. Integrative correlation: Properties and relation to canonical correlations. *Journal of multivariate analysis* **123**, 270–280 (2014).

37. Rouillard, A. D. *et al.* The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* **2016**, (2016).

38. Tong, Z., Cui, Q., Wang, J. & Zhou, Y. TransmiR v2.0: an updated transcription factor-microRNA regulation database. *Nucleic Acids Res* **47**, D253–D258 (2019).

39. Agarwal, V., Bell, G. W., Nam, J.-W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* **4**, (2015).

40. Chou, C.-H. *et al.* miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Research* **46**, D296–D302 (2021).

41. Sticht, C., De La Torre, C., Parveen, A. & Gretz, N. miRWalk: An online resource for prediction of microRNA binding sites. *PLOS ONE* **13**, e0206239- (2018).

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

43. Xu, Y. & Goodacre, R. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *Journal of Analysis and Testing* **2**, 249–262 (2018).

44. Tan, A. C., Naiman, D. Q., Xu, L., Winslow, R. L. & Geman, D. Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics* **21**, 3896–3904 (2005).

45. Geman, D., d'Avignon, C., Naiman, D. Q. & Winslow, R. L. Classifying Gene Expression Profiles from Pairwise mRNA Comparisons. *Stat Appl Genet Mol Biol* **3**, Article19 (2004).

46. Noble, W. S. What is a support vector machine? *Nature Biotechnology* **24**, 1565–1567 (2006).

47. Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).

48. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016). doi:10.1145/2939672.2939785.

49. Ihaka, R. & Gentleman, R. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics* **5**, 299–314 (1996).

50. Cope, L., Zhong, X., Garrett, E. & Parmigiani, G. MergeMaid: R tools for merging and cross-study validation of gene expression data. *Stat Appl Genet Mol Biol* **3**, Article29 (2004).

51. Afsari, B., Fertig, E. J., Geman, D. & Marchionni, L. switchBox: an R package for k–Top Scoring Pairs classifier development. *Bioinformatics* **31**, 273–274 (2015).

52. Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software, Articles* **28**, 1–26 (2008).

53. Karatzoglou, A., Smola, A., Hornik, K. & Zeileis, A. kernlab – An S4 Package for Kernel Methods in R. *Journal of Statistical Software* **11**, 1–20 (2004).

54. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *undefined* /paper/Classification-and-Regression-by-randomForest-Liaw-Wiener/6e633b41d93051375ef9135102d54fa097dc8cf8 (2007).