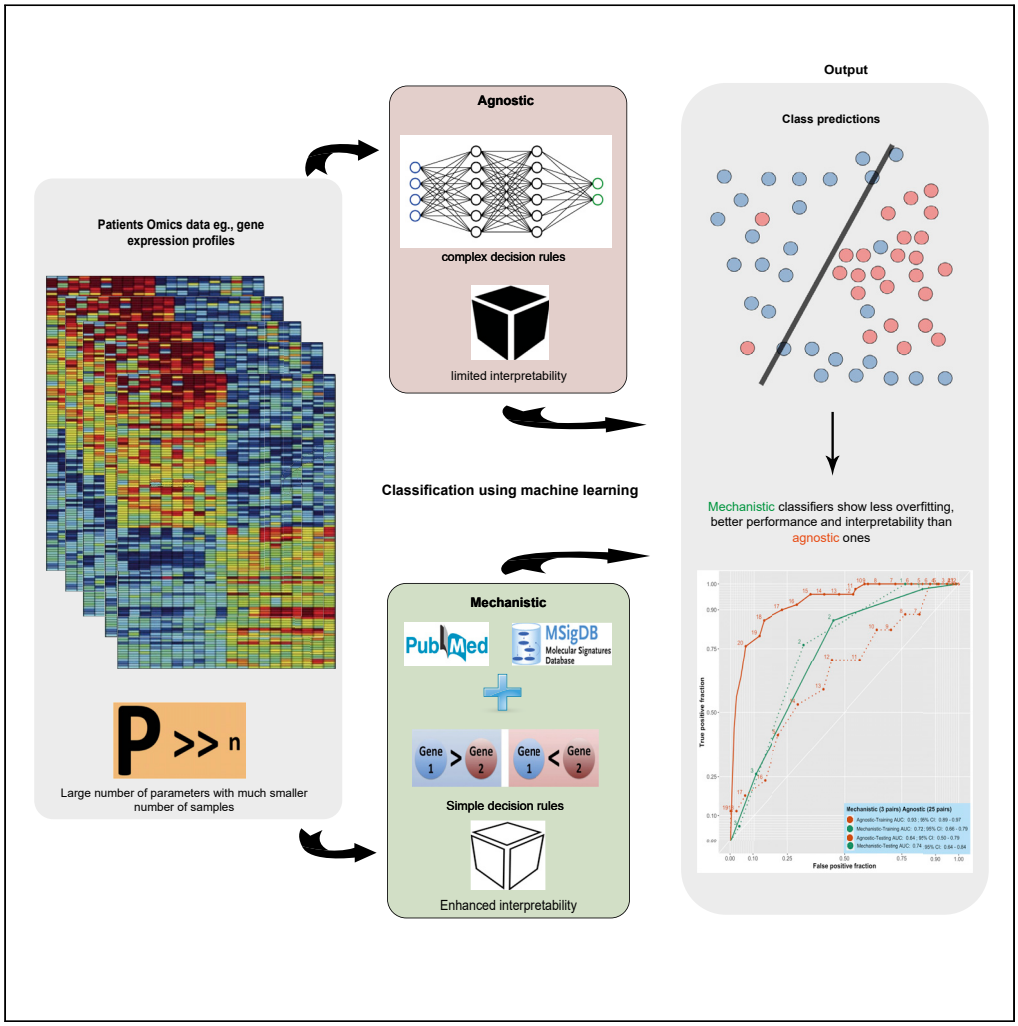
iScience

CellPress
OPEN ACCESS

Article

Using biological constraints to improve prediction in precision oncology



Mohamed Omar, Wikum Dinalankara, Lotte Mulder, ..., Laurent Younes, Donald Geman, Luigi Marchionni

lum4003@med.cornell.edu

Highlights

Most gene signatures suffer from overfitting and limited interpretability

Using known biology in the training can yield robust mechanistic classifiers

We compared the performance of mechanistic models to standard agnostic ones

Mechanistic models tend to have robust performance with enhanced interpretability

Omar et al., iScience 26, 106108
March 17, 2023
https://doi.org/10.1016/j.isci.2023.106108

iScience

CellPress
OPEN ACCESS

Article

Using biological constraints to improve prediction in precision oncology

Mohamed Omar,[1] Wikum Dinalankara,[1] Lotte Mulder,[2] Tendai Coady,[1] Claudio Zanettini,[1] Eddie Luidy Imada,[1] Laurent Younes,[3] Donald Geman,[3] and Luigi Marchionni[1,4,*]

SUMMARY

Many gene signatures have been developed by applying machine learning (ML) on *omics* profiles, however, their clinical utility is often hindered by limited interpretability and unstable performance. Here, we show the importance of embedding prior biological knowledge in the decision rules yielded by ML approaches to build robust classifiers. We tested this by applying different ML algorithms on gene expression data to predict three difficult cancer phenotypes: bladder cancer progression to muscle-invasive disease, response to neoadjuvant chemotherapy in triple-negative breast cancer, and prostate cancer metastatic progression. We developed two sets of classifiers: *mechanistic*, by restricting the training to features capturing specific biological mechanisms; and *agnostic*, in which the training did not use any *a priori* biological information. Mechanistic models had a similar or better testing performance than their agnostic counterparts, with enhanced interpretability. Our findings support the use of biological constraints to develop robust gene signatures with high translational potential.

INTRODUCTION

In oncology, machine learning (ML) algorithms are actively used to decipher gene expression data and identify predictive or prognostic gene signatures for specific cancer phenotypes such as tumor 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] The process of gene signatures discovery and validation is hindered by relevant challenges.[3] The most striking one is the unstable performance of the discovered signatures when tested on different data than the ones used in their 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). In these settings, what can easily happen 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 is known as overfitting.[4] There are several approaches to reduce overfitting and increase robustness, the most important of which is increasing the number of samples; however, this is not always feasible in biomedical research due to financial limitations or rarity of the studied disease phenotypes. Other options include using simple algorithms which are less susceptible to overfitting,[5] using regularization with complex ones,[6,7] and reducing dimensionality by filtering out non-informative features or by using feature selection methods.[8]

We hypothesize that embedding biological, mechanistic constraints in the decision rules during the training process will guide the ML algorithm to a set of features important for the phenotype being predicted, which in turn can reduce overfitting, and improve the performance, robustness, and interpretability of the resulting models. These constraints take the form of gene pairs which are derived from existing biological knowledge from the literature or pre-curated databases, and whose relative ordering determines the predicted class.[9–11] Here, we test if this method can yield more interpretable gene signatures with comparable performance to agnostic methods (*i.e.*, not based on prior biological knowledge) by using gene expression data to develop predictive classifiers in three distinct and hard prediction cases: 1. predicting the progression of non-muscle invasive bladder cancer (NMIBC) (stage T1) to muscle-invasive disease (MIBC) (stages T2-T4); 2. the response to neoadjuvant chemotherapy (NACT) in triple-negative breast cancer (TNBC); and 3. prostate cancer (PCa) metastasis from primary tumor samples. In each of these three cases, we use four different ML algorithms: k-Top Scoring Pairs (k-TSPs), Support Vector Machine (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGB).

[1]Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY 10065, USA

[2]Technical University Delft, 2628 CD Delft, the Netherlands

[3]Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218, USA

[4]Lead contact

*Correspondence: lum4003@med.cornell.edu

https://doi.org/10.1016/j.isci.2023.106108

To build mechanistic models, we restrict the training process to a specific biological mechanism relevant to the phenotype under study. For bladder cancer (BLCA) progression, we use feedforward loops (FFLs) which consist of transcription factors (TFs) and microRNAs (miRNAs) target genes. TFs regulate the expression of their target genes through various mechanisms,[12] while miRNAs—a class of small, non-coding RNAs—play an important role in post-transcriptional gene regulation through the modulation of mRNA degradation.[13] Current evidence shows that both TFs and miRNAs regulate the expression of common target genes and the expression of each other through feedback (FBLs) and feedforward loops (FFLs).[14–17] Moreover, other studies have shown that the interaction between miRNAs targets and TFs is involved in the progression of several cancers including bladder cancer.[18–22] Using the same principle for the TNBC case, we restrict the training process to mechanisms involving gene targets downstream to the Notch and MYC pathways, owing to the role these play in mediating cancer chemoresistance. Notch signaling pathway is involved in promoting cancer angiogenesis and epithelial-mesenchymal transition (EMT).[23] It also promotes chemoresistance in several cancers including breast cancer by inhibiting apoptosis and mediating cancer stem cells (CSC) self-renewal capacity.[24] Similarly, *MYC* promotes chemoresistance by mediating CSC self-renewal and proliferation,[25,26] and also by dysregulating the expression of some ATP-binding cassette (ABC) transporters necessary for cellular drug transport.[27] Finally, since metastatic progression is mediated by several known biological processes, including loss of cell-cell adhesion and hypoxia in the tumor microenvironment (TME),[28,29] we designed a set of mechanistic pairs capturing such processes for predicting PCa metastasis.

In summary, here we embed prior knowledge of cancer biology directly into the algorithmic process to identify robust decision rules. We show that such mechanistic models, even with a relatively small number of features, have a similar or even superior performance and robustness, and an enhanced interpretability compared to agnostic models based on hundreds of genes solely selected based on statistical significance.

RESULTS

Building mechanistic classifiers by embedding prior knowledge in the predictive decision rules

We hypothesized that integrating existing biological knowledge in the training process can yield robust and interpretable models the translational potential of which can surpass that of agnostic methods. For each classification task, we identified several biological processes related to the phenotype under study and used these to build corresponding biological mechanisms. To simplify the decision rules, we designed each mechanism as a list of gene pairs, each consisting of a gene associated with bad prognosis (*e.g.*, progression or chemo-resistance) and another associated with good prognosis. Specifically, for predicting BLCA progression, we built a mechanism based on feedforward loops (FFLs) consisting of a TF which inhibits a downstream miRNA target gene (Figure S1). For the TNBC task, we based the mechanistic constraints on NOTCH and MYC signaling based on their involvement in mediating chemoresistance.[23–26] Specifically, the mechanistic constraints were built by pairing the genes up-regulated with those down-regulated by NOTCH or MYC. Finally, for predicting PCa metastasis, we restricted the training process to gene pairs orchestrating cellular adhesion and O_2 response.

We used such mechanistic pairs to train biologically constrained, rank-based models,[9,11] and then compared their performance to agnostic ones trained without biological constraints, starting from differentially expressed genes or their pairwise combinations (Figure 1). Furthermore, we performed such a comparison using two distinct designs: training bootstrap and cross-study validation. In the bootstrap design, we proceeded as follows: we a) divided the data into training and testing sets; b) bootstrapped the training set 1000 times with a sample size equal to that of the training set (resampling with replacement); c) trained agnostic and mechanistic models on each training resample; and d) evaluated their performance on the untouched testing set. In the cross-study validation design, we used all but one study (*n-1*) for training, while the left-out study was used for testing, this process being repeated *n* times so that each study was used for testing.

Mechanistic models based on feedforward loops outperform agnostic ones in predicting bladder cancer progression

In the BLCA task, we used gene expression profiles from 350 patients with NMIBC (stage T1) to train predictive models for cancer progression to MIBC (stage T2 or higher). We trained mechanistic classifiers using FFLs (37 unique mechanistic pairs) and compared their performance and robustness to agnostic models

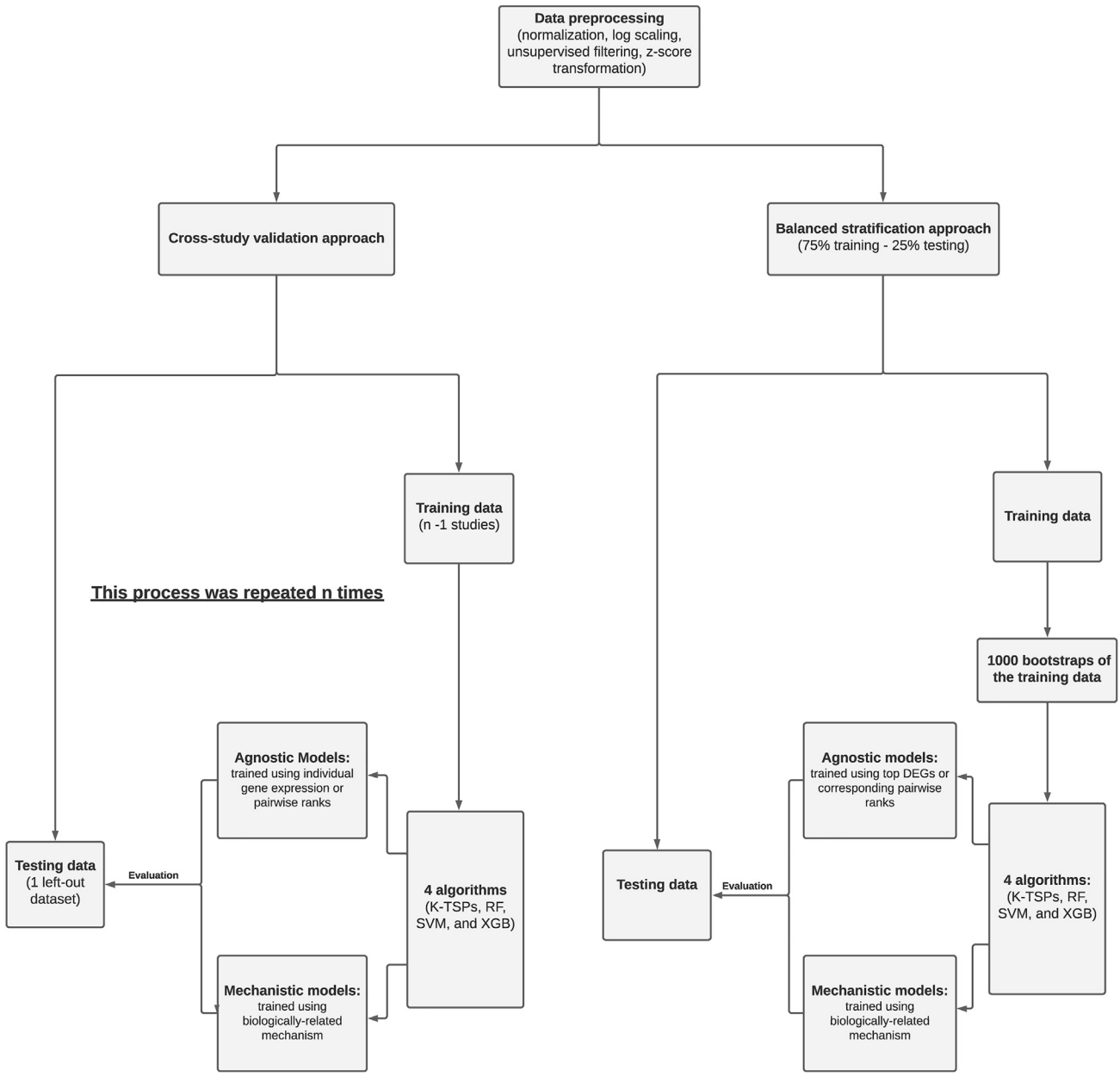


Figure 1. Building mechanistic classifiers by embedding prior knowledge in the predictive decision rules

Three different cancer cases were considered: predicting bladder cancer progression, predicting the response to neoadjuvant chemotherapy in patients with triple-negative breast cancer, and predicting prostate cancer metastatic progression. We adopted two different experimental designs: the balanced stratification (training bootstrap) and cross-study validation. In the balanced stratification design, all datasets were pooled together after normalization and preprocessing then split into training and testing sets. The training set was bootstrapped 1000 times and on each resample, we trained agnostic and mechanistic models and then evaluated their performance on the testing set. In the cross-study validation, the analysis included n iterations where *n* corresponds to the number of studies. In each iteration, we used all, but one study for training agnostic and mechanistic models and then evaluated their performance on the left-out study. k-TSPs: K-top scoring pairs, RF: random forest, SVM: support vector machine, XGB: extreme gradient boosting, DEGs: differentially expressed genes.

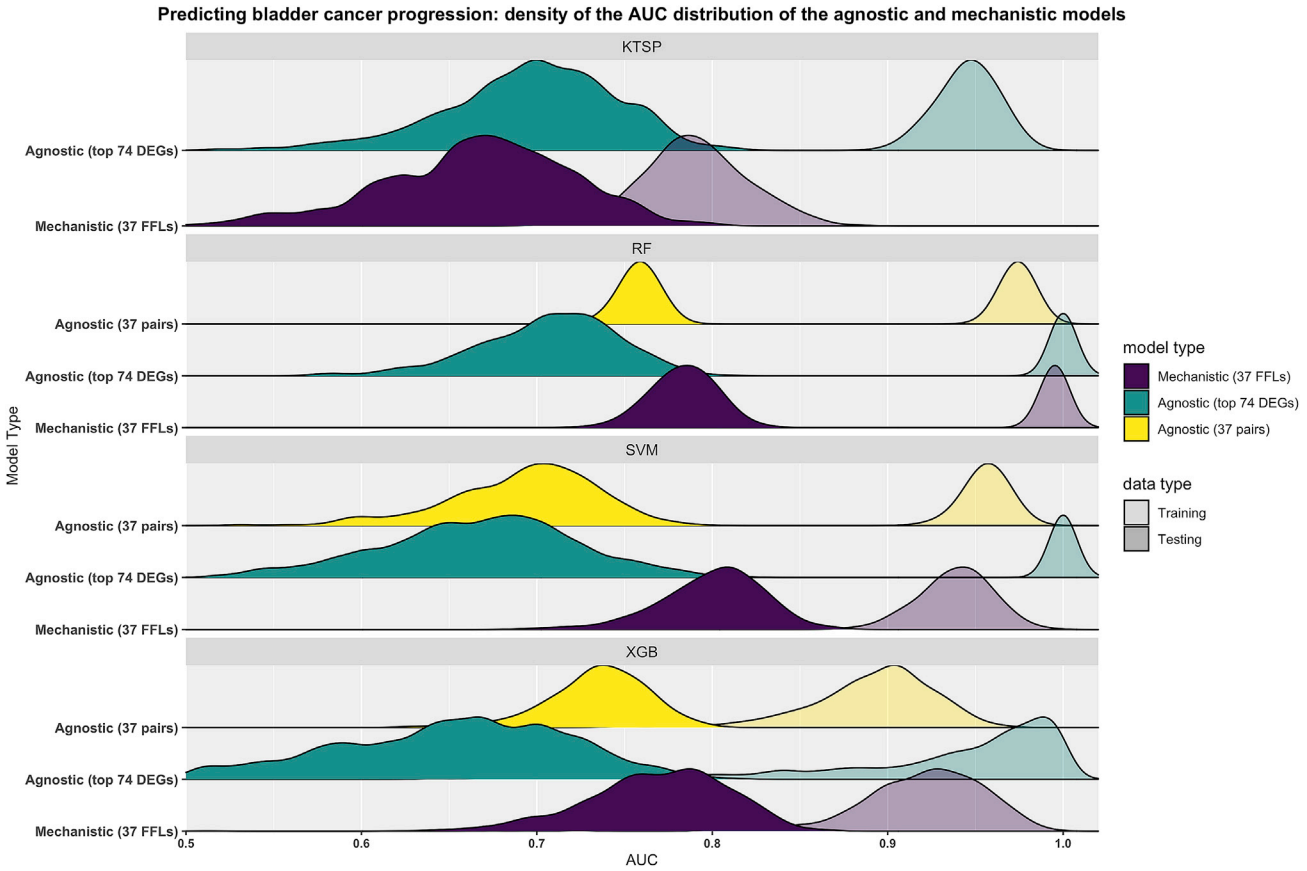


Figure 2. Mechanistic models based on FFLs outperform agnostic ones in predicting bladder cancer progression

The figure depicts the performance of the agnostic and mechanistic models as obtained using the described bootstrap design. Briefly, all models were trained on 1000 bootstraps of the training data (transparent colors), then evaluated on untouched testing data (solid colors) using the Area Under the ROC Curve (AUC) as performance metric. Mechanistic models were based on the feedforward loops (37 pairs) (purple) and agnostic models were trained either using the top differentially expressed genes (74 genes) (green) or the corresponding pairwise comparisons (37 pairs) (yellow). Curves represent the smoothed density distributions 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: feedforward loops; DEGs: differentially expressed genes.

trained on the top differentially expressed genes, or the pairwise combinations of their ranks, without any prior biological consideration. Since the number of features used in the training process can be a major contributor to overfitting, we restricted the initialization of the agnostic models to the top 74 DEGs, matching the starting number of features used for training in the mechanistic case. Additionally, we also combined such top DEGs into 37 rank-based pairs to examine if this could improve the performance. In the bootstrap design, using the Area Under the ROC Curve (AUC) as evaluation metric, the agnostic and mechanistic k-TSPs models had a similar performance at predicting bladder cancer progression in the independent testing set (Figure 2). However, the mechanistic k-TSPs models were more parsimonious yielding on average five gene pairs compared to 16 pairs used by the agnostic ones. Importantly, the testing performance of the mechanistic models was more comparable with what was observed in the training. For the other three algorithms (RF, SVM, and XGB), the mechanistic models showed, on average, a higher testing performance than agnostic ones trained using the top DEGs (Figure 2). Interestingly, using pairwise comparisons derived from the top DEGs—instead of using their individual expression values—improved the performance of the agnostic models, slightly reducing the gap between training and testing. In secondary analyses, we also tested whether increasing the number of starting features in the agnostic case could improve their performance by training additional models using either the top 100, 200, and 500 DEGs or their pairwise comparisons (50, 100, and 250 pairs). Our results show that mechanistic classifiers still had a comparable or superior performance and robustness to agnostic ones, even when increasing the number of features (Figure S2). In addition, we also confirmed that mechanistic models held a clear advantage over agnostic ones trained using randomly selected genes (see Figure S3).

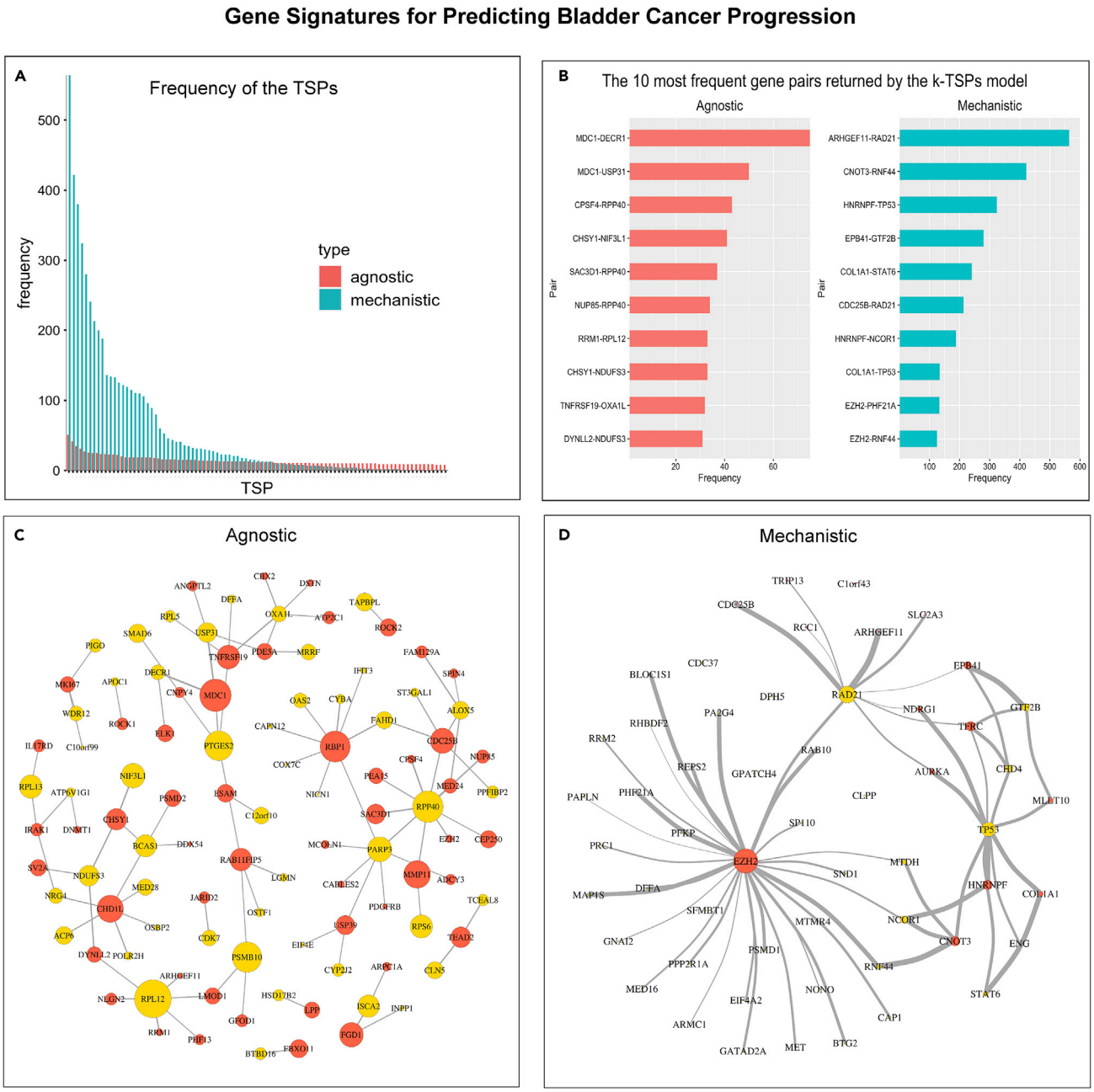


Figure 3. Mechanistic and agnostic k-TSPs signatures for predicting bladder cancer progression from non-muscle- to muscle-invasive stages

(A) Bar plot showing the frequency of the agnostic (red) and mechanistic (blue) scoring pairs across the different bootstraps. The bar plot includes all the mechanistic pairs (n = 93) and the most frequent agnostic pairs (n = 93), both sorted by their frequency across the different bootstraps.

(B) The top 10 most frequent agnostic (red) and mechanistic (blue) pairs sort by their frequency.

(C and D) Networks of the top 93 agnostic (C) and mechanistic (D) top-scoring pairs. Each pair consists of a gene voting for BLCA progression (red) and no progression (yellow). The vertex size corresponds to 2*log2 of the individual gene frequency across unique pairs while the edge thickness corresponds to the log2 of the top scoring pair frequency.

To examine the consistency of gene signatures across the 1000 bootstraps, we ranked the gene pairs returned by the agnostic and mechanistic k-TSPs classifiers by their frequency. Interestingly, the mechanistic models tended to return more frequent pairs across the different training data re-samples compared to the agnostic models, in which the selected pairs differed significantly with each training iteration (Figure 3A). For example, the most frequent gene pairs selected by the mechanistic k-TSPs models were

ARHGEF11-RAD21 (n = 564), *CNOT3-RNF44* (n = 422), *CHD4-DFFA* (n = 380), *HNRNPF-TP53* (n = 324), and *EPB41-GTF2B* (n = 280) (Figure 3B). On the other hand, the five most frequently selected pairs by the agnostic k-TSPs were *MDC1-DECR1* (n = 75), *MDC1-USP31* (n = 50), *CPSF4-RPP40* (n = 43), *CHSY1-NIF3L1* (n = 41), and *SAC3D1-RPP40* (n = 37) (Figure 3B). Each TSP gives a vote for a particular class (BLCA progression vs no-progression) if gene1 is more expressed than gene2. This means that the first gene in each pair can be interpreted as a gene associated with cancer progression, while the second would be expected to have an opposite role. With this notion in mind, we obtained—and then compared—the networks representing the frequency of each TSP and that of each individual gene (across unique pairs), for both the agnostic and mechanistic k-TSPs approaches. The agnostic network was dominated by several genes associated with progression (mainly *MDC1*, *RBP1*, *MMP11*, *CHD1L*, and *CDC25B*) and no-progression (*RPL12*, *RPP40*, *PSMB10*, and *PTGES2*) (Figure 3C). In contrast, the mechanistic network was found to be more coherent, being dominated mainly by *EZH2* serving as progression-associated gene (gene1) and *RAD21*, followed by *TP53*, serving as non-progression associated genes (gene2) (Figure 3D).

To examine the functional states of genes associated with BLCA progression compared to those associated with no-progression, we performed gene set enrichment analysis (GSEA) on all the genes positioned as either gene1 (*i.e.*, *the "bad"* genes) or gene 2 (*i.e.*, *the "good"* genes) in the TSPs, separately for mechanistic and agnostic classifiers. Interestingly, genes in the mechanistic signatures (55 unique genes) were significantly enriched in 62 biological processes, many of which are related to invasion and progression including EMT, cell proliferation, and cell cycle transition (Table S1). On the other hand, genes in the agnostic signatures were significantly enriched in only one biological process despite their larger number (74 unique genes) (Table S1).

Finally, in the cross-study validation design, we found that both the mechanistic and agnostic k-TSPs models had a similar average AUC in the testing data, while the mechanistic one outperformed its agnostic counterpart on multiple metric (balanced accuracy, sensitivity, and MCC,[30] see Table S2). In agreement with the bootstrap results, the testing performance of the mechanistic k-TSPs was also highly comparable with that of the training, suggesting improved generalizability. Similar results were also seen with the other three ML algorithms (RF, SVM, and XGB, see Table S2).

NOTCH-MYC-based models outperform their agnostic counterparts in predicting response to neoadjuvant chemotherapy in triple-negative breast cancer

For predicting the response to NACT in patients with TNBC, we used the pre-treatment gene expression profiles of 369 patients with TNBC. Mechanistic models were trained using the NOTCH-MYC mechanism (241 unique pairs), while agnostic models were trained using the top 500 DEGs, or their corresponding pairwise ranks (250 pairs). Performance was compared with the same two designs described for the bladder cancer case.

In the bootstrap approach, our results show that the mechanistic k-TSPs models have a similar testing performance to the agnostic ones, however, the mechanistic RF, SVM, and XGB models still slightly outperformed their agnostic counterparts trained on the top DEGs (Figure 4). Notably, agnostic models trained on pairwise ranks instead of individual gene expression values had a similar (using RF and SVM) or even better (using XGB) testing performance than mechanistic ones trained on the NOTCH-MYC mechanism (Figure 4). Furthermore, changing the number of training features only improved the testing performance of the agnostic models based on pairwise ranks, while still falling short of the mechanistic ones (Figure S4). Finally, mechanistic models clearly outperformed those trained on random genes (Figure S5).

Ranking out pairs by their frequency across the different iterations showed that a similar and more robust set of mechanistic pairs get selected much more often than agnostic ones (Figure 5A). For instance, the five most frequently returned pairs by the mechanistic k-TSPs models included *DDIT3-DDX18* (n = 531), *TSC2-PLK4* (n = 499), *COL5A1-ITGA6* (n = 442), *GARS-PDCD10* (n = 385), and *NDE1-EZR* (n = 347) (Figure 5B). However, the five most frequent pairs returned by the agnostic k-TSPs models included *SLC43A1-ABT1* (n = 204), *ITGA5-EPHB3* (n = 119), *METRN-MCM5* (n = 116), *PARM1-MAPK9* (n = 97), and *SLC22A5-DCAF7* (n = 96) (Figure 5B). Here, the decision rules follow the same pattern discussed in the BLCA case, with each pair voting for either residual disease (RD) or pathological complete response (pCR) based on the expression of the two genes. In agnostic models, the most frequent gene associated with RD (gene1 in the TSP) was *MAST3* while *RPL39L* was the most frequent gene associated with pCR (gene2 in the TSP)

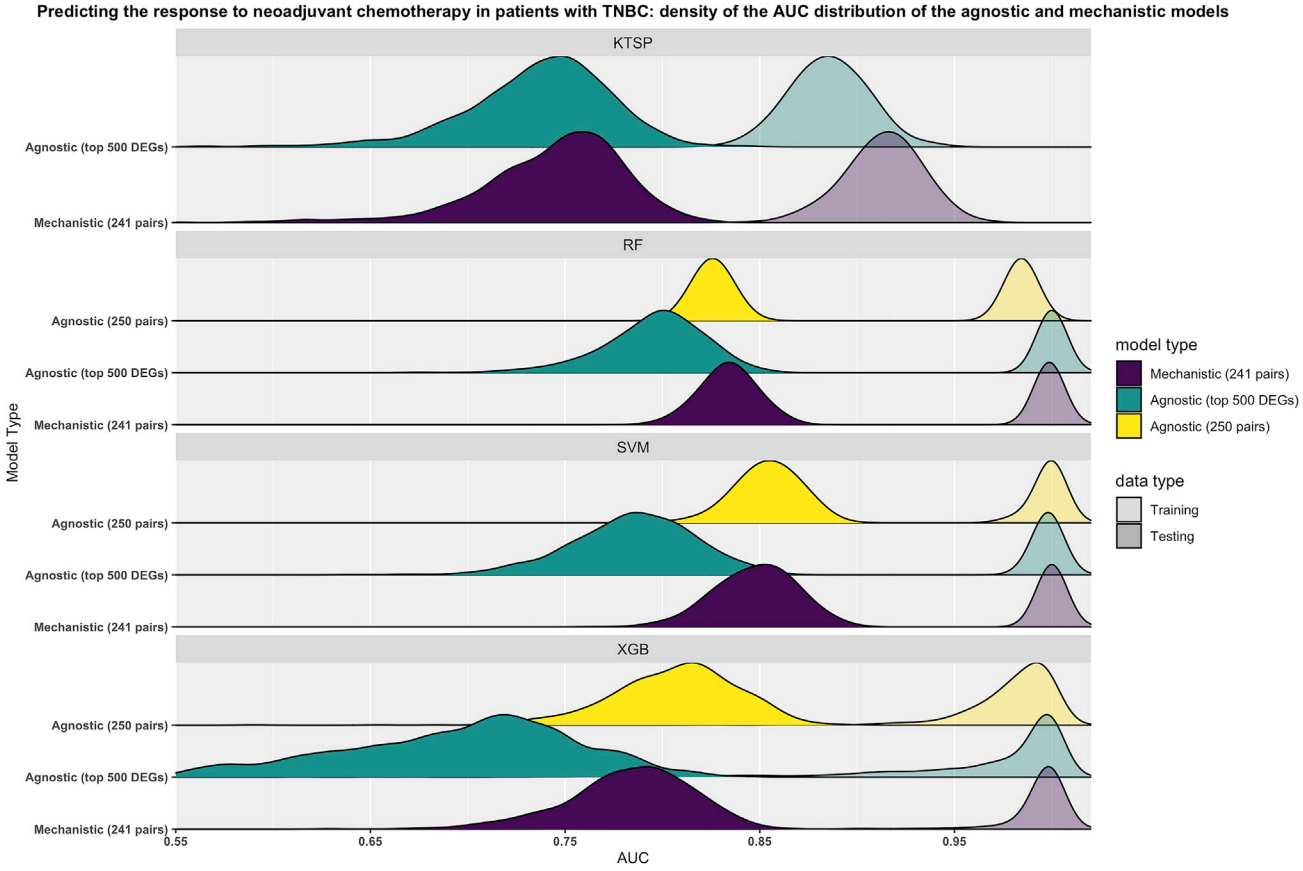


Figure 4. NOTCH-MYC-based models outperform their agnostic counterparts in predicting response to neoadjuvant chemotherapy in patients with triple-negative breast cancer

Models were trained on 1000 bootstraps of the training data (transparent colors) and evaluated on the untouched testing data (solid colors) using the Area Under the ROC curve (AUC) as metric. Mechanistic models were based on the NOTCH-MYC mechanism (241 pairs) (purple) while agnostic models were trained either using the top differentially expressed genes (500 genes) (green) or the corresponding pairwise comparisons (250 pairs) (yellow). 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; TNBC: triple-negative breast cancer.

(Figure 5C). For the mechanistic models, the oncogene *CCND1* was the most frequent gene overexpressed in samples from patients with RD (gene1 in the TSP) while the WNT antagonist *SFRP1*[31] was the most frequently overexpressed in samples from patients who had pCR (gene2 in the TSP) (Figure 5D).

The genes from the mechanistic classifiers (656 unique genes) were significantly enriched in 780 different pathways and processes including those used to build the priori mechanism (NOTCH and MYC signaling) (Table S3). They were also enriched in other cancer-related pathways such as the regulation of apoptosis, beta-catenin-TCF complex assembly, TGF-B signaling, T-cells activation, and differentiation. While genes from agnostic classifiers were larger in number (1335 unique genes), they were significantly enriched in only 49 gene sets without a strong association with cancer biology. Altogether, these results reflect the signatures selected by the k-TSPs models are consistent and more associated with the biological processes underlying chemotherapeutic resistance compared to those returned by agnostic models (Table S3).

Lastly, also in the cross-study validation case, we did not observe significant differences in performance between the model types, however, the mechanistic classifiers showed slightly more consistency in performance metrics like the AUC between training and testing, especially using the k-TSPs and XGB algorithms (Table S4). While using biological constraints did not offer a clear advantage in terms of performance, it significantly enhanced interpretability, and reduced the computational costs by limiting the training to a few hundred features instead several thousand used for the agnostic models.

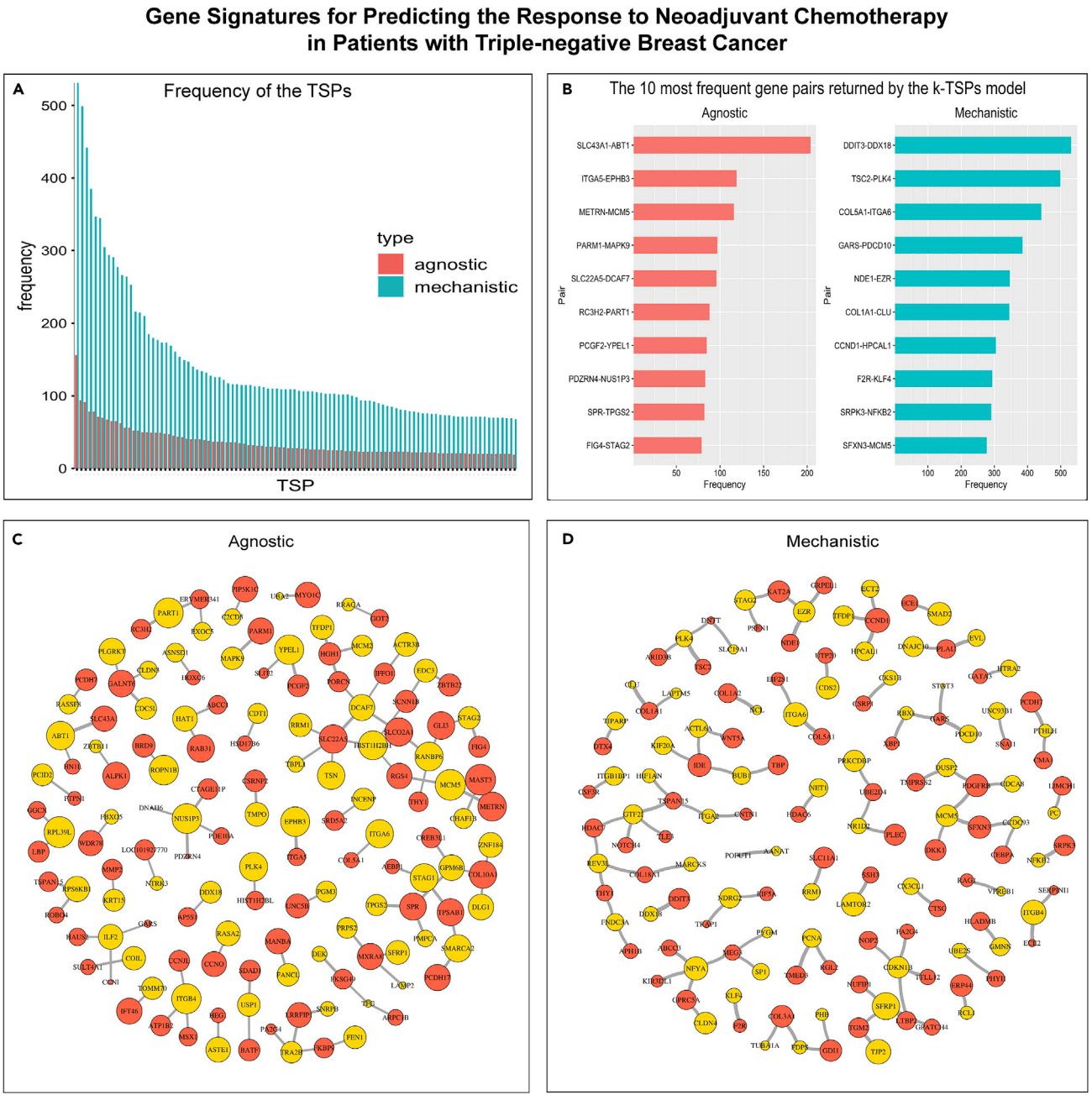


Figure 5. Mechanistic and agnostic k-TSPs signatures for predicting the response to neoadjuvant chemotherapy in patients with triple-negative breast cancer

(A) Bar plot showing the frequency of the top 100 agnostic (red) and mechanistic (blue) scoring pairs across the different bootstraps.

(B) The top 10 most frequent agnostic (red) and mechanistic (blue) pairs sort by their frequency.

(C and D) Networks of the top 100 agnostic (C) and mechanistic (D) top-scoring pairs. Each pair consists of a gene voting for RD (red) and pCR (yellow). The vertex size corresponds to 2*log2 of the individual gene frequency across unique pairs while the edge thickness corresponds to the log2 of the top scoring pair frequency.

Mechanistic models based on cellular adhesion and oxygen response have a similar performance to their agnostic counterparts in predicting prostate cancer metastatic progression

For predicting metastatic progression in prostate cancer, we used seven gene expression datasets comprising 1239 primary tumor samples including 399 with metastatic events.

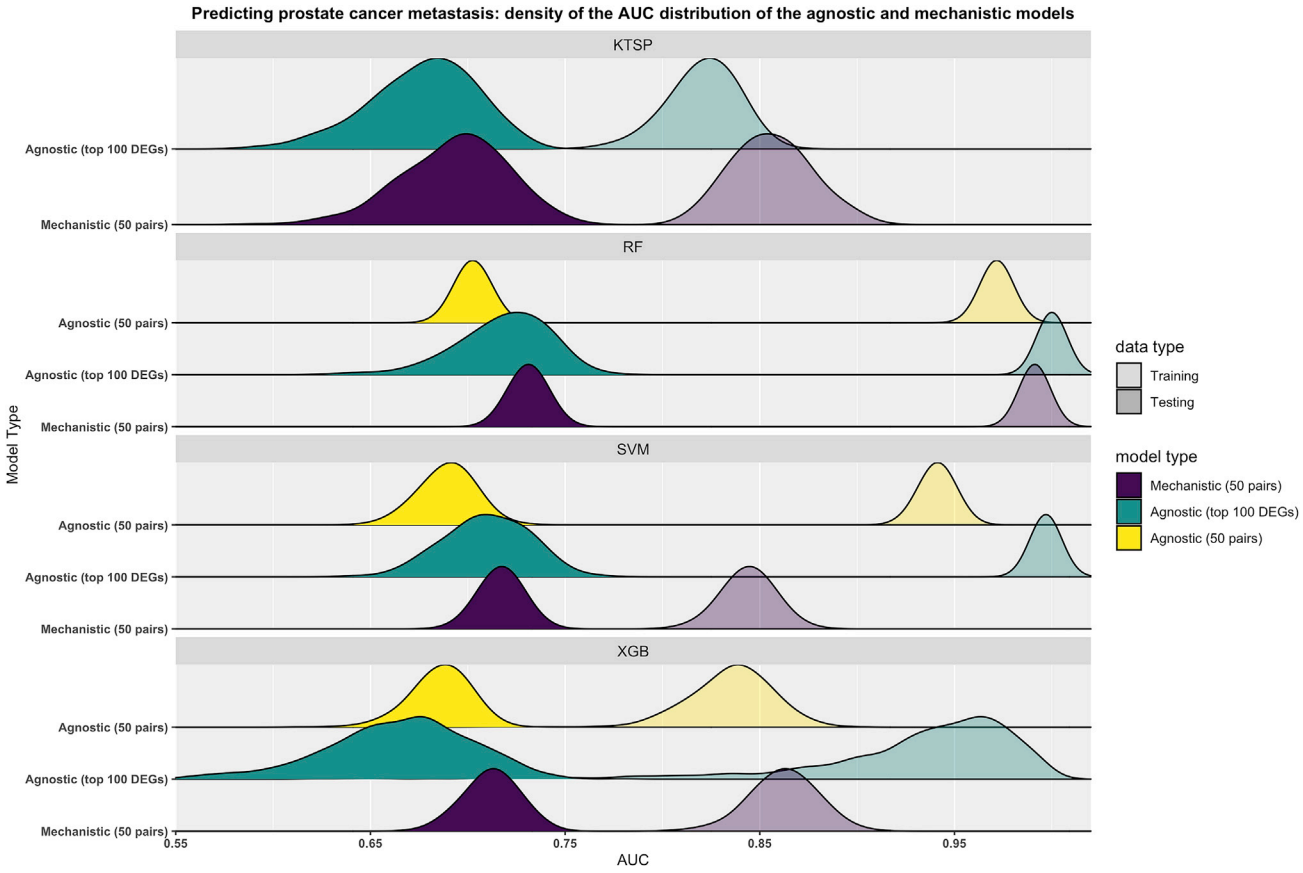


Figure 6. Mechanistic models based on cellular adhesion and oxygen response have similar performance to their agnostic counterparts in predicting prostate cancer metastatic progression

The figure depicts the results from the bootstrap design in which the training set (transparent colors) was resampled 1000 times. On each resample, models were trained to predict metastatic progression in prostate cancer and their performance was evaluated on the untouched testing set (dark colors) using the Area Under the ROC Curve (AUC) as evaluation metric. Mechanistic models were based on the cellular adhesion and O_2 response mechanism (50 pairs) (purple) while agnostic models were trained using either the top differentially expressed genes (100 genes) (green) or the corresponding pairwise comparisons (50 pairs) (yellow). Shown are the smoothed density distributions 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; DEGs: differentially expressed genes.

In the bootstrap approach, mechanistic models using 50 pairs had a similar performance to their agnostic counterparts and incurred less overfitting irrespective of the number of training features (Figure 6). Additionally, these mechanistic classifiers still maintained their performance even when more features were used for training the agnostic models (Figure S6), or when random genes were used in the training process (Figure S7).

Mechanistic TSPs showed high frequency (Figure 7A) with the five most frequent pairs including *S100A10-PTN* (n = 540), *CD74-SATB1* (n = 455), *CBX3-AZGP1* (n = 414), *CXCR4-PCDH18* (n = 237), and *STAT1-DPP4* (n = 214) (Figure 7B). Agnostic models on the other hand frequently returned *CAMK2N1-CDC42EPS* (n = 389), *CXCR4-LPAR3* (n = 230), *EN O 1-CDC42EPS* (n = 217), *RFTN1-DPT* (n = 202), and *GNPTAB-CTBS* (n = 190) (Figures 7B and 7C). Interestingly, in the mechanistic TSPs, genes related to PCa progression and metastases such as *THBS2*,[32] *NRP1*,[33] and *WNT5A*[34] were frequently represented as metastases-voting genes (gene1 in the TSPs) (Figure 7D).

Gene set Enrichment analysis showed that genes derived from the mechanistic signatures (906 unique genes) were significantly enriched in 1920 gene sets, many of which are associated with cell migration, motility, adhesion, and proliferation (Table S5). They were also enriched in other important pathways involved in PCa progression and metastases including the regulation of MAPK and ERK cascades, NF-κB, STAT, TGF-beta, and RAS signaling pathways. Similarly, genes from the agnostic models (1622 unique

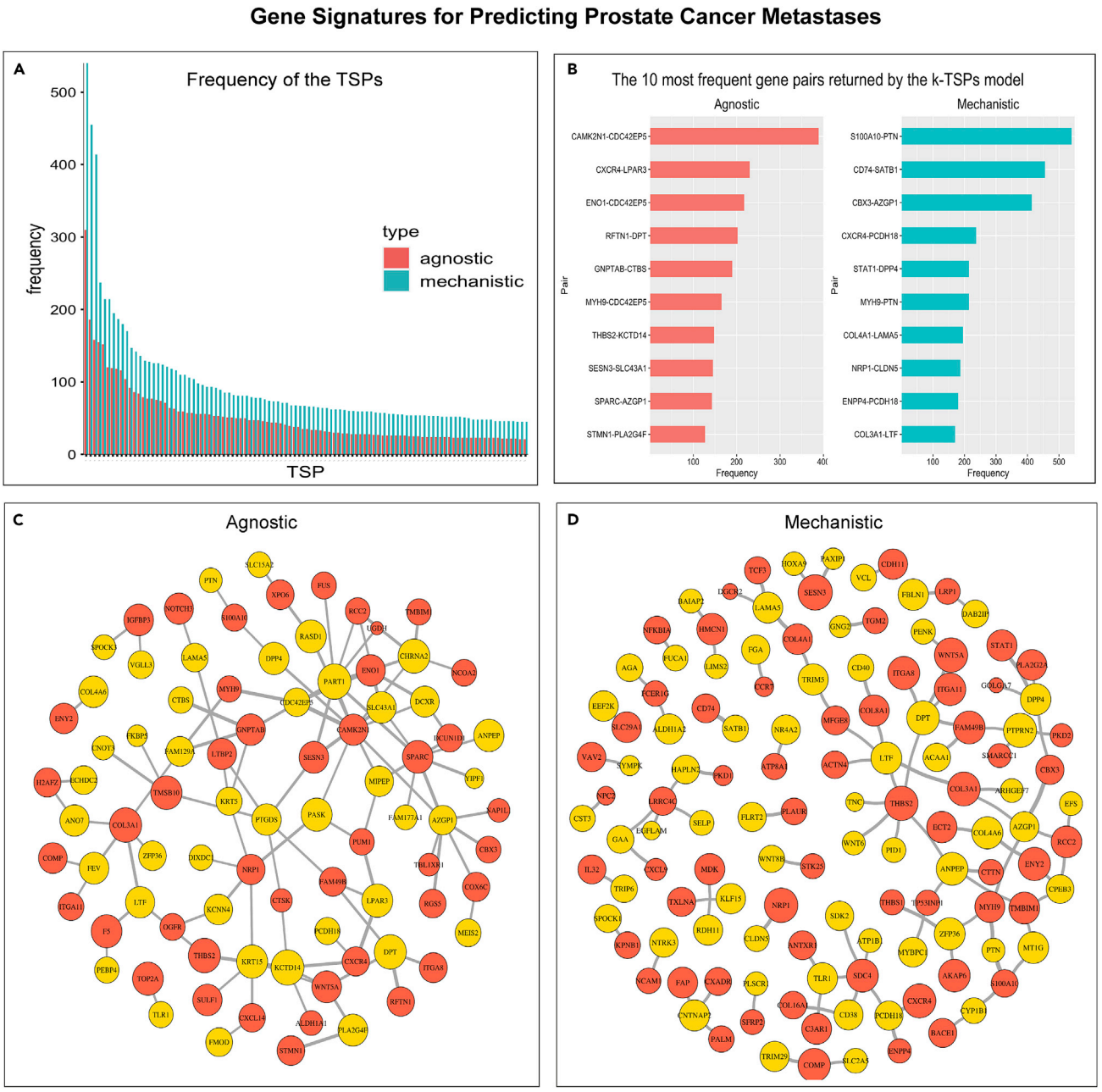


Figure 7. Mechanistic and agnostic k-TSPs signatures for predicting prostate cancer metastases

(A) Bar plot showing the frequency of the top 100 agnostic (red) and mechanistic (blue) scoring pairs across the different bootstraps.

(B) The top 10 most frequent agnostic (red) and mechanistic (blue) pairs sort by their frequency.

(C and D) Networks of the top 100 agnostic (C) and mechanistic (D) top-scoring pairs. Each pair consists of a gene voting for metastasis (red) and no-metastasis (yellow). The vertex size corresponds to the 2*log2 of the individual gene frequency across unique pairs while the edge thickness corresponds to the log2 of the top scoring pair frequency.

genes) were enriched in 96 gene sets, some of which included the regulation of cell migration and motility and PCa-related pathways such as WNT signaling (Table S5).

Finally, these results were also confirmed in the cross-study validation analysis, in which the mechanistic models had similar performance compared to the agnostic ones, but provided superior interpretability

and improved computational efficiency, due to the reduced number of features (tens versus thousands) used in the training process (Table S6).

The right mechanism for the right task: mechanistic constraints should be related to the phenotype under consideration

We have shown that using biological constraints in the training process can improve the performance of the resulting gene signatures, however, we hypothesized that this is contingent on using a mechanism related to the phenotype under study. To investigate this, we assessed the performance of several different mechanisms in each of the three classification tasks we considered. This analysis included the three afore-mentioned cancer-related mechanisms (FFLs, NOTCH-MYC signaling, and cellular adhesion and O_2 response), together with three other, non-cancer related ones, constructed from molecular profiles involved in: a) Alzheimer's disease (266 pairs); b) diabetes (6776 pairs); and c) viral infection (6384 pairs).

For each cancer phenotype, models trained on the relevant mechanism of choice showed the best performance (Table 1). Specifically, classifiers trained using the FFLs had higher testing AUCs for predicting bladder cancer progression compared to those based on the other mechanisms. Similar results were also obtained for NACT response prediction in TNBC using NOTCH-MYC signaling, and for prostate cancer metastatic progression using cellular adhesion and O_2 response.

Notably, the three cancer-related mechanisms maintained a superior performance over the non-cancer related ones in each of the three prediction tasks. For example, while the FFLs-based classifiers had the best performance at predicting bladder cancer progression, the NOTCH-MYC signaling and cellular adhesion/O_2 response mechanisms also achieved a good testing performance. On the other hand, the Alzheimer, diabetes, and viral infection mechanisms performed poorly in the three classification cases (Table 1), highlighting the fact that choosing the right mechanism for the prediction task is essential for achieving an optimal performance.

DISCUSSION

Overfitting and lack of generalization remain among the most difficult problems in machine learning especially in transcriptomics owing to the very large number of features and the much 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 it represents a major obstacle toward personalizing health care. While some approaches such as increasing the sample size may help improve the performance and stability of such models, these may not always be feasible in medical research, owing to financial limitations or the unavailability of samples in cases of rare cancer phenotypes.

Some studies have shown that using prior knowledge can help to choose the input data or the correct algorithm.[35,36] In this study, we employed a similar concept to train robust and interpretable predictive models by adding biological constraints to the decision rules used for classification. We examined this approach in three clinically important classification cases: predicting bladder cancer progression from non-muscle invasive to muscle-invasive stages, predicting the response to neoadjuvant chemotherapy in TNBC, and predicting prostate cancer progression to metastasis disease. In each setting, we employed multiple algorithms and compared the performance of mechanistically constrained models to agnostic ones. In the bladder cancer case, we used FFLs based on the evidence supporting their involvement in cancer progression and invasion.[18–22] In the TNBC case, we focused on NOTCH and c-MYC targets based on their role in mediating cancer stem cells self-renewal and chemo-resistance.[24–26] Finally, for predicting prostate cancer metastasis, we used gene pairs capturing cell-cell adhesion and the cellular response to O_2.[28,29] Since each of these priori mechanisms is already known to be associated with the corresponding phenotype, the features constituting the mechanism are expected to be of high quality for the prediction task undertaken. In this case, the ML algorithm is essentially used to extract the smallest number of features that can serve as a gene signature and has more potential to generalize to other datasets. It is true that agnostic models can also extract a number of features with a similar performance if not better. However, in that case, the search process would be applied to all genes and has more risk to overfit the training data and select non-informative features that can't be generalized. When evaluated on the testing data, the mechanistic models yielded a similar or superior performance than the agnostic ones. Moreover, in both the bladder (using the K-TSPs and SVM algorithms) and prostate (using the SVM and XGB algorithms) cancer cases, mechanistic models demonstrated better generalization from training to testing, even

Table 1. The performance of different mechanisms at predicting each of the three cancer phenotypes assessed by the Area under the ROC curve (AUC)

		k-TSPs		RF		SVM		XGB	
Prediction Task	Mechanism	Train	Test	Train	Test	Train	Test	Train	Test
Bladder cancer progression	FFLs[a]	0.75	0.73	0.98	0.79	0.91	0.82	0.88	0.78
	NOTCH-MYC	0.85	0.72	1.00	0.70	0.98	0.68	1.0	0.67
	Cell adhesion/O_2 response	0.89	0.76	1.00	0.70	1.0	0.64	1.0	0.71
	Alzheimer	0.60	0.49	0.61	0.52	0.54	0.43	0.59	0.52
	Diabetes	0.82	0.65	0.96	0.68	0.86	0.60	0.75	0.52
	Viral infection	0.81	0.63	0.96	0.59	0.90	0.50	0.82	0.57
Response to NACT in TNBC	FFLs	0.74	0.64	0.85	0.73	0.79	0.66	0.75	0.69
	NOTCH-MYC[b]	0.86	0.75	1.00	0.86	1.0	0.87	1.0	0.87
	Cell adhesion/O_2 response	0.88	0.79	1.00	0.85	1.0	0.82	1.0	0.82
	Alzheimer	0.68	0.61	0.77	0.66	0.77	0.59	0.69	0.64
	Diabetes	0.78	0.75	0.94	0.79	0.98	0.74	0.84	0.72
	Viral infection	0.81	0.73	0.96	0.77	0.95	0.69	0.89	0.70
Prostate cancer metastasis	FFLs	0.61	0.63	0.64	0.62	0.62	0.60	0.61	0.60
	NOTCH-MYC	0.73	0.73	0.99	0.69	1.0	0.73	0.83	0.71
	Cell adhesion/O_2 response[c]	0.80	0.75	0.97	0.73	0.97	0.71	0.85	0.73
	Alzheimer	0.58	0.50	0.59	0.53	0.50	0.50	0.60	0.53
	Diabetes	0.69	0.67	0.89	0.69	0.70	0.68	0.71	0.69
	Viral infection	0.68	0.65	0.81	0.64	0.73	0.63	0.72	0.67

Three important cancer phenotypes were considered for prediction: bladder cancer progression from non-muscle invasive to muscle-invasive stages, response to neoadjuvant chemotherapy (NACT) in patients with triple-negative breast cancer (TNBC), and metastatic progression in prostate cancer. For each phenotype, we built *a priori* mechanism capturing the underlying biology and used it for prediction in the main analysis. FeedForward Loops (FFLs) were designed for predicting bladder cancer progression while the NOTCH-MYC signaling and cellular adhesion and O_2 response mechanisms were developed for predicting the response to NACT in TNBC and prostate cancer metastatic progression, respectively. Three cancer unrelated mechanisms were used as negative controls: Alzheimer, diabetes, and viral infection. The performance of the different cancer-related and unrelated mechanisms at predicting each of the three cancer phenotypes was assessed using the Area Under the ROC Curve (AUC). FFLs: feedforward loops, NACT: neoadjuvant chemotherapy, TNBC: triple-negative breast cancer.

[a]The feedforward loops mechanism was used for predicting bladder cancer progression in the main analysis.

[b]The NOTCH-MYC signaling mechanism was used for predicting the response to neoadjuvant chemotherapy in triple-negative breast cancer.

[c]The cell-cell adhesion and O2 response mechanism was used for predicting prostate cancer metastasis.

in situations where there was a small number of samples available for training. Furthermore, these results were not dependent on any specification or characteristics pertaining to the training data, as shown by results from both the bootstrap and the cross-validation analyses.

It is important to note that the mechanism used in the training should be biologically related to the phenotype being predicted. We demonstrated this by assessing the performance of different mechanisms in each classification case and noticed that the mechanisms we designed based on prior knowledge had the best performance at predicting their corresponding phenotype. Interestingly, the three cancer-related mechanisms we have considered achieved optimal performance in each cancer phenotype, while the non-cancer ones performed poorly, which further supports our rationale for using prior related biological knowledge to develop robust classifiers.

Overall, these results show that models using a small number of biologically important features can have a similar or better performance metrics compared to those using hundreds or thousands of genes. Furthermore, mechanistic rank-based decision rules greatly enhance the interpretability of the resulting predictive models, which is crucial to their clinical adaptation. Specifically, the priory mechanism is designed beforehand by pairing genes from pathways positively and negatively associated with the phenotype of interest.

In this case, the mechanism consists of multiple gene pairs belonging to opposing pathways or biological processes. Subsequently, we use the k-top scoring pairs (k-TSPs) algorithm which is rank based and serves by selecting the gene pairs whose expression consistently switch between the two classes of interest. For example, the NOTCH mechanism consists of hundreds of gene pairs up- and down-regulated by NOTCH signaling and we use this mechanism to design a classifier to predict the response to chemotherapy in breast cancer. The resulting classifier consists of a number of gene pairs with each consisting of a gene up- and another down-regulated by NOTCH signaling. Each pair votes for a particular class (sensitive versus resistant) based on the order of expression of the two genes and the final prediction of a patient/sample is determined by the majority of votes. As such, the classifier is interpretable because its decision rules and the function of its genes are known beforehand. Also, such models are more robust to preprocessing techniques, and can be more easily implemented using other technologies already being used for clinical applications. For example, we have shown that such mechanistic signatures derived from microarrays or RNA-Seq studies can be implemented using technologies such as RT-PCR, which further increases their translational value.[37]

Together with their good performance, the mechanistic signatures captured the underlying biology of the associated cancer phenotype as expected. Since we used TF-miRNA targets as *a priori* mechanism to train ML models capable of predicting bladder cancer progression, the resulting signatures were always gene pairs with the first gene being a TF while the second is a miRNA target gene mirroring the priori mechanism. Even with different training data resampling, these signatures were often consistent indicating their predictive value rather than perfectly fitting the training data. For instance, *ARHGEF11-RAD21* was present in more than 50% of the mechanistic BLCA progression signatures voting for progression if *ARHGEF11* is more expressed relative to *RAD21*. While the role *ARHGEF11* in bladder cancer is not well-defined, it was found to be associated with the invasiveness of progression of glioblastoma[38] and hepatocellular carcinoma.[39] Across all mechanistic pairs, the transcriptional repressor *EZH2* was prominent as a progression voting gene being present in 41 unique TSPs in which its expression relative to the second gene determines the vote given by the TSP (progression if *EZH2* is more expressed than gene2). This is consistent with existing evidence linking EZH2 to progression and poor prognosis in several cancers including BLCA.[40,41] Similarly, predicting the response to NACT in patients with TNBC relied on *a priori* mechanism of gene pairs regulated by NOTCH and MYC signaling. Expectedly, the resulting signatures consisted of a small subset of gene pairs up- and down—regulated by either NOTCH or MYC, and among these, genes down-regulated by either TF (*e.g.*, *SFRP1*, *ITGB4*, and *ITGA6*) were predominantly associated with pCR. The same pattern was seen in predicting PCa metastases in which the mechanistic signatures frequently included genes known to be involved in mediating EMT and distant metastases e.g., *THBS2*[32] and *WNT5A*.[34] Gene set enrichment analyses showed that genes from the mechanistic signatures were significantly enriched in many more coherent biological processes and pathways compared to the genes from their agnostic counterpart, despite the latter being in much larger numbers. Moreover, In the three classification tasks, genes from the mechanistic classifiers were enriched in the pathways used to build the priori mechanism (as to be expected), together with other pathways associated with the phenotype under study. Altogether, these results show that mechanistic signatures tend to capture important cancer biology related to their corresponding phenotypes.

Limitations of the study

It is important to note that our study has some inherent limitations. First, the biological mechanisms we used take the form of contrasting gene pairs, but these pairwise relationships may not completely capture the complexity of the underlying biology compared to other formulations such as gene networks. However, this lack of sophistication in the design of the biological mechanism was deemed necessary for the interpretability of the resulting predictive models. Moreover, such pairwise comparisons, which are used by the k-TSPs algorithm, can also be used as input to other more complex approaches (e.g., SVM, RF, and XGB), with the advantage of increasing the level of interpretability of the resulting models. Finally, while our study was focused on cancer, the same conceptual framework can be also applied to other diseases, provided there is available prior knowledge about the underlying pathophysiological mechanisms.

Despite these limitations, our work supports the adoption of mechanistically constrained decision rules for the development of robust prognostic and predictive models. Their high performance and intrinsic interpretability will promote a wider integration into clinical practice, bringing routine personalized medicine one step closer to reality.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Material availability
  - Data and code availability
- METHOD DETAILS
  - Data collection
  - Data preprocessing
  - Mechanistic pairs assembly
  - Data splitting for training and testing
  - Training and evaluating the performance of mechanistic versus agnostic models
  - The k-TSPs classifier
  - Support vector machine
  - Random forest
  - Extreme gradient boosting
  - Gene set enrichment analyses
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.106108.

ACKNOWLEDGMENTS

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 work was supported by the National Institutes of Health-National Cancer Institute [R01CA200859 to LM, WD, DG, and LY].

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 article. LM, DG, and LY supervised the analysis and the article writing. All authors read and approved the final version of the article.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: April 4, 2022
Revised: December 20, 2022
Accepted: January 28, 2023
Published: February 2, 2023

REFERENCES

1. Cardoso, F., van't Veer, L.J., Bogaerts, J., Slaets, L., Viale, G., Delaloge, S., Pierga, J.-Y., Brain, E., Causeret, S., DeLorenzi, M., et al. (2016). 70-Gene signature as an aid to treatment decisions in early-stage breast cancer. N. Engl. J. Med. *375*, 717–729. https://doi.org/10.1056/NEJMoa1602253.

2. Knezevic, D., Goddard, A.D., Natraj, N., Cherbavaz, D.B., Clark-Langone, K.M., Snable, J., Watson, D., Falzarano, S.M., Magi-Galluzzi, C., Klein, E.A., and Quale, C. (2013). Analytical validation of the Oncotype DX prostate cancer assay - a clinical RT-PCR assay optimized for prostate needle biopsies. BMC Genom. *14*, 690. https://doi.org/10.1186/1471-2164-14-690.

3. Mirza, B., Wang, W., Wang, J., Choi, H., Chung, N.C., and Ping, P. (2019). Machine learning and integrative analysis of biomedical big data. Genes *10*, 87. https://doi.org/10.3390/genes10020087.

4. Keogh, E., and Mueen, A. (2010). Curse of Dimensionality (Encyclopedia of machine learning), pp. 257–258.

5. Hand, D.J. (2006). Classifier technology and the illusion of progress. Stat. Sci. *21*, 1–14.

6. Neumaier, A. (1998). Solving ill-conditioned and singular linear systems: a tutorial on regularization. SIAM Rev. Soc. Ind. Appl. Math. *40*, 636–666. https://doi.org/10.1137/S0036144597321909.

7. Chicco, D. (2017). Ten quick tips for machine learning in computational biology. BioData Min. *10*, 35. https://doi.org/10.1186/s13040-017-0155-3.

8. Mahendran, N., Durai Raj Vincent, P.M., Srinivasan, K., and Chang, C.-Y. (2020). Machine learning based computational gene selection models: a survey, performance evaluation, open issues, and future research directions. Front. Genet. *11*, 603808. https://doi.org/10.3389/fgene.2020.603808.

9. Geman, D., d'Avignon, C., Naiman, D.Q., and Winslow, R.L. (2004). Classifying gene expression profiles from pairwise mRNA comparisons. Stat. Appl. Genet. Mol. Biol. *3*, Article19. https://doi.org/10.2202/1544-6115.1071.

10. Marchionni, L., Afsari, B., Geman, D., and Leek, J.T. (2013). A simple and reproducible breast cancer prognostic test. BMC Genom. *14*, 336. https://doi.org/10.1186/1471-2164-14-336.

11. Tan, A.C., Naiman, D.Q., Xu, L., Winslow, R.L., and Geman, D. (2005). Simple decision rules for classifying human cancers from gene expression profiles. Bioinformatics *21*, 3896–3904. https://doi.org/10.1093/bioinformatics/bti631.

12. Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and Weirauch, M.T. (2018). The human transcription factors. Cell *172*, 650–665. https://doi.org/10.1016/j.cell.2018.01.029.

13. O'Brien, J., Hayder, H., Zayed, Y., and Peng, C. (2018). Overview of MicroRNA biogenesis, mechanisms of actions, and circulation. Front. Endocrinol. *9*, 402. https://doi.org/10.3389/fendo.2018.00402.

14. Hausser, J., and Zavolan, M. (2014). Identification and consequences of miRNA-target interactions–beyond repression of gene expression. Nat. Rev. Genet. *15*, 599–612. https://doi.org/10.1038/nrg3765.

15. Martinez, N.J., Ow, M.C., Barrasa, M.I., Hammell, M., Sequerra, R., Doucette-Stamm, L., Roth, F.P., Ambros, V.R., and Walhout, A.J.M. (2008). A C. elegans genome-scale microRNA network contains composite feedback motifs with high flux capacity. Genes Dev. *22*, 2535–2549. https://doi.org/10.1101/gad.1678608.

16. Re, A., Corá, D., Taverna, D., and Caselle, M. (2009). Genome-wide survey of microRNA-transcription factor feed-forward regulatory circuits in human. Mol. Biosyst. *5*, 854–867. https://doi.org/10.1039/b900177h.

17. Friard, O., Re, A., Taverna, D., De Bortoli, M., and Corá, D. (2010). CircuitsDB: a database of mixed microRNA/transcription factor feed-forward regulatory circuits in human and mouse. BMC Bioinf. *11*, 435. https://doi.org/10.1186/1471-2105-11-435.

18. Li, Q.-Q., Chen, Z.-Q., Cao, X.-X., Xu, J.-D., Xu, J.-W., Chen, Y.-Y., Wang, W.-J., Chen, Q., Tang, F., Liu, X.-P., and Xu, Z.D. (2011). 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. https://doi.org/10.1038/cdd.2010.103.

19. Guo, Y., Ying, L., Tian, Y., Yang, P., Zhu, Y., Wang, Z., Qiu, F., and Lin, J. (2013). miR-144 downregulation increases bladder cancer cell proliferation by targeting EZH2 and regulating Wnt signaling. FEBS J. *280*, 4531–4538. https://doi.org/10.1111/febs.12417.

20. Liu, J.-J., Lin, X.-J., Yang, X.-J., Zhou, L., He, S., Zhuang, S.-M., and Yang, J. (2014). 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. https://doi.org/10.1093/nar/gku872.

21. Dong, F., Xu, T., Shen, Y., Zhong, S., Chen, S., Ding, Q., and Shen, Z. (2017). Dysregulation of miRNAs in bladder cancer: altered expression with aberrant biogenesis procedure. Oncotarget *8*, 27547–27568. https://doi.org/10.18632/oncotarget.15173.

22. Mullany, L.E., Herrick, J.S., Wolff, R.K., Stevens, J.R., Samowitz, W., and Slattery, M.L. (2018). MicroRNA-transcription factor interactions and their combined effect on target gene expression in colon cancer cases. Genes Chromosomes Cancer *57*, 192–202. https://doi.org/10.1002/gcc.22520.

23. Abdullah, L.N., and Chow, E.K.-H. (2013). Mechanisms of chemoresistance in cancer stem cells. Clin. Transl. Med. *2*, 3. https://doi.org/10.1186/2001-1326-2-3.

24. Ranganathan, P., Weaver, K.L., and Capobianco, A.J. (2011). Notch signalling in solid tumours: a little bit of everything but not all the time. Nat. Rev. Cancer *11*, 338–351. https://doi.org/10.1038/nrc3035.

25. Wang, J., Wang, H., Li, Z., Wu, Q., Lathia, J.D., McLendon, R.E., Hjelmeland, A.B., and Rich, J.N. (2008). c-Myc is required for maintenance of glioma cancer stem cells. PLoS One *3*, e3769. https://doi.org/10.1371/journal.pone.0003769.

26. Zhang, H.-L., Wang, P., Lu, M.-Z., Zhang, S.-D., and Zheng, L. (2019). c-Myc maintains the self-renewal and chemoresistance properties of colon cancer stem cells. Oncol. Lett. *17*, 4487–4493. https://doi.org/10.3892/ol.2019.10081.

27. Porro, A., Haber, M., Diolaiti, D., Iraci, N., Henderson, M., Gherardi, S., Valli, E., Munoz, M.A., Xue, C., Flemming, C., et al. (2010). 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. https://doi.org/10.1074/jbc.M109.078584.

28. Bhandari, V., Hoey, C., Liu, L.Y., Lalonde, E., Ray, J., Livingstone, J., Lesurf, R., Shiah, Y.-J., Vujcic, T., Huang, X., et al. (2019). Molecular landmarks of tumor hypoxia across cancer types. Nat. Genet. *51*, 308–318. https://doi.org/10.1038/s41588-018-0318-2.

29. Oppenheimer, S.B. (2006). Cellular basis of cancer metastasis: a review of fundamentals and new advances. Acta Histochem. *108*, 327–334. https://doi.org/10.1016/j.acthis.2006.03.008.

30. Chicco, D., Tötsch, N., and Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Min. *14*, 13. https://doi.org/10.1186/s13040-021-00244-z.

31. Veeck, J., Niederacher, D., An, H., Klopocki, E., Wiesmann, F., Betz, B., Galm, O., Camara, O., Dürst, M., Kristiansen, G., et al. (2006). Aberrant methylation of the Wnt antagonist SFRP1 in breast cancer is associated with unfavourable prognosis. Oncogene *25*, 3479–3488. https://doi.org/10.1038/sj.onc.1209386.

32. Chen, P.-C., Tang, C.-H., Lin, L.-W., Tsai, C.-H., Chu, C.-Y., Lin, T.-H., and Huang, Y.-L. (2017). Thrombospondin-2 promotes prostate cancer bone metastasis by the up-regulation of matrix metalloproteinase-2 through down-regulating miR-376c expression. J. Hematol. Oncol. *10*, 33. https://doi.org/10.1186/s13045-017-0390-6.

33. Tse, B.W.C., Volpert, M., Ratther, E., Stylianou, N., Nouri, M., McGowan, K., Lehman, M.L., McPherson, S.J., Roshan-Moniri, M., Butler, M.S., et al. (2017). Neuropilin-1 is upregulated in the adaptive response of prostate tumors to androgen-targeted therapies and is prognostic of metastatic progression and patient mortality. Oncogene *36*, 3417–3427. https://doi.org/10.1038/onc.2016.482.

34. Dai, J., Hall, C.L., Escara-Wilke, J., Mizokami, A., Keller, J.M., and Keller, E.T. (2008). Prostate cancer induces bone metastasis through Wnt-induced bone morphogenetic protein-dependent and independent mechanisms. Cancer Res. *68*, 5785–5794. https://doi.org/10.1158/0008-5472.CAN-07-6541.

35. Libbrecht, M.W., and Noble, W.S. (2015). Machine learning applications in genetics and genomics. Nat. Rev. Genet. *16*, 321–332. https://doi.org/10.1038/nrg3920.

36. Yip, K.Y., Cheng, C., Bhardwaj, N., Brown, J.B., Leng, J., Kundaje, A., Rozowsky, J., Birney, E., Bickel, P., Snyder, M., and Gerstein, M. (2012). Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. Genome Biol. *13*, R48. https://doi.org/10.1186/gb-2012-13-9-r48.

37. Ghantous, Y., Omar, M., Broner, E.C., Agrawal, N., Pearson, A.T., Rosenberg, A.J., Mishra, V., Singh, A., El-naaj, I.A., Savage, P.A., et al. (2022). A robust and interpretable gene signature for predicting the lymph node status of primary T1/T2 oral cavity squamous cell carcinoma. Int. J. Cancer *150*, 450–460. https://doi.org/10.1002/ijc.33828.

38. Ding, Z., Dhruv, H., Kwiatkowska-Piwowarczyk, A., Ruggieri, R., Kloss, J., Symons, M., Pirrotte, P., Eschbacher, J.M., Tran, N.L., and Loftus, J.C. (2018). PDZ-RhoGEF is a signaling effector for

TROY-induced glioblastoma cell invasion and survival. Neoplasia *20*, 1045–1058. https://doi.org/10.1016/j.neo.2018.08.008.

39. Du, J., Zhu, Z., Xu, L., Chen, X., Li, X., Lan, T., Li, W., Yuan, K., and Zeng, Y. (2020). ARHGEF11 promotes proliferation and epithelial-mesenchymal transition of hepatocellular carcinoma through activation of β-catenin pathway. Aging (Albany NY) *12*, 20235–20253. https://doi.org/10.18632/aging.103772.

40. Wang, H., Albadine, R., Magheli, A., Guzzo, T.J., Ball, M.W., Hinz, S., Schoenberg, M.P., Netto, G.J., and Gonzalgo, M.L. (2012). Increased EZH2 protein expression is associated with invasive urothelial carcinoma of the bladder. Urol. Oncol. *30*, 428–433. https://doi.org/10.1016/j.urolonc.2010.09.005.

41. Zhou, X., Liu, N., Zhang, J., Ji, H., Liu, Y., Yang, J., and Chen, Z. (2018). Increased expression of EZH2 indicates aggressive potential of urothelial carcinoma of the bladder in a Chinese population. Sci. Rep. *8*, 17792. https://doi.org/10.1038/s41598-018-36164-y.

42. Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., and Edgar, R. (2007). NCBI GEO: mining tens of millions of expression profiles–database and tools update. Nucleic Acids Res. *35*, D760–D765. https://doi.org/10.1093/nar/gkl887.

43. Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y.A., Williams, E., Dylag, M., Kurbatova, N., Brandizi, M., Burdett, T., et al. (2015). ArrayExpress update–simplifying data submissions. Nucleic Acids Res. *43*, D1113–D1116. https://doi.org/10.1093/nar/gku1057.

44. van der Heijden, A.G., Mengual, L., Lozano, J.J., Ingelmo-Torres, M., Ribal, M.J., Fernández, P.L., Oosterwijk, E., Schalken, J.A., Alcaraz, A., and Witjes, J.A. (2016). A five-gene expression signature to predict progression in T1G3 bladder cancer. Eur. J. Cancer *64*, 127–136. https://doi.org/10.1016/j.ejca.2016.06.003.

45. Kim, W.-J., Kim, E.-J., Kim, S.-K., Kim, Y.-J., Ha, Y.-S., Jeong, P., Kim, M.-J., Yun, S.-J., Lee, K.M., Moon, S.-K., et al. (2010). Predictive value of progression-related gene classifier in primary non-muscle invasive bladder cancer. Mol. Cancer *9*, 3. https://doi.org/10.1186/1476-4598-9-3.

46. Sjödahl, G., Lauss, M., Lövgren, K., Chebil, G., Gudjonsson, S., Veerla, S., Patschan, O., Aine, M., Fernö, M., Ringnér, M., et al. (2012). A molecular taxonomy for urothelial carcinoma. Clin. Cancer Res. *18*, 3377–3386. https://doi.org/10.1158/1078-0432.CCR-12-0077-T.

47. Dyrskjøt, L., Zieger, K., Kruhøffer, M., Thykjaer, T., Jensen, J.L., Primdahl, H., Aziz, N., Marcussen, N., Møller, K., and Ørntoft, T.F. (2005). A molecular signature in superficial bladder carcinoma predicts clinical outcome. Clin. Cancer Res. *11*, 4029–4036. https://doi.org/10.1158/1078-0432.CCR-04-2095.

48. Hedegaard, J., Lamy, P., Nordentoft, I., Algaba, F., Høyer, S., Ulhøi, B.P., Vang, S., Reinert, T., Hermann, G.G., Mogensen, K., et al. (2016). Comprehensive transcriptional analysis of early-stage urothelial carcinoma. Cancer Cell *30*, 27–42. https://doi.org/10.1016/j.ccell.2016.05.004.

49. Hatzis, C., Pusztai, L., Valero, V., Booser, D.J., Esserman, L., Lluch, A., Vidaurre, T., Holmes, F., Souchon, E., Wang, H., et al. (2011). A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. JAMA *305*, 1873–1881. https://doi.org/10.1001/jama.2011.593.

50. Edlund, K., Madjar, K., Lebrecht, A., Aktas, B., Pilch, H., Hoffmann, G., Hofmann, M., Kolberg, H.-C., Boehm, D., Battista, M., et al. (2021). Gene expression–based prediction of neoadjuvant chemotherapy response in early breast cancer: results of the prospective multicenter EXPRESSION trial. Clin. Cancer Res. *27*, 2148–2158. https://doi.org/10.1158/1078-0432.CCR-20-2662.

51. Birkbak, N.J., Li, Y., Pathania, S., Greene-Colozzi, A., Dreze, M., Bowman-Colin, C., Sztupinszki, Z., Krzystanek, M., Diossy, M., Tung, N., et al. (2018). Overexpression of BLM promotes DNA damage and increased sensitivity to platinum salts in triple-negative breast and serous ovarian cancers. Ann. Oncol. *29*, 903–909. https://doi.org/10.1093/annonc/mdy049.

52. Popovici, V., Chen, W., Gallas, B.G., Hatzis, C., Shi, W., Samuelson, F.W., Nikolsky, Y., Tsyganova, M., Ishkin, A., Nikolskaya, T., et al. (2010). Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. Breast Cancer Res. *12*, R5. https://doi.org/10.1186/bcr2468.

53. Shi, L., Campbell, G., Jones, W.D., Campagne, F., Wen, Z., Walker, S.J., Su, Z., Chu, T.-M., Goodsaid, F.M., Pusztai, L., et al. (2010). The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat. Biotechnol. *28*, 827–838. https://doi.org/10.1038/nbt.1665.

54. Shen, K., Song, N., Kim, Y., Tian, C., Rice, S.D., Gabrin, M.J., Symmans, W.F., Pusztai, L., and Lee, J.K. (2012). A systematic evaluation of multi-gene predictors for the pathological response of breast cancer patients to chemotherapy. PLoS One *7*, e49529. https://doi.org/10.1371/journal.pone.0049529.

55. Tabchy, A., Valero, V., Vidaurre, T., Lluch, A., Gomez, H., Martin, M., Qi, Y., Barajas-Figueroa, L.J., Souchon, E., Coutant, C., et al. (2010). Evaluation of a 30-gene paclitaxel, fluorouracil, doxorubicin, and cyclophosphamide chemotherapy response predictor in a multicenter randomized trial in breast cancer. Clin. Cancer Res. *16*, 5351–5361. https://doi.org/10.1158/1078-0432.CCR-10-1265.

56. Miyake, T., Nakayama, T., Naoi, Y., Yamamoto, N., Otani, Y., Kim, S.J., Shimazu, K., Shimomura, A., Maruyama, N., Tamaki, Y., and Noguchi, S. (2012). GSTP1 expression predicts poor pathological complete response to neoadjuvant chemotherapy in ER-negative breast cancer. Cancer Sci. *103*, 913–920. https://doi.org/10.1111/j.1349-7006.2012.02231.x.

57. Jain, S., Lyons, C.A., Walker, S.M., McQuaid, S., Hynes, S.O., Mitchell, D.M., Pang, B., Logan, G.E., McCavigan, A.M., O'Rourke, D., et al. (2018). Validation of a Metastatic Assay using biopsies to improve risk stratification in patients with prostate cancer treated with radical radiation therapy. Ann. Oncol. *29*, 215–222. https://doi.org/10.1093/annonc/mdx637.

58. Ragnum, H.B., Vlatkovic, L., Lie, A.K., Axcrona, K., Julin, C.H., Frikstad, K.M., Hole, K.H., Seierstad, T., and Lyng, H. (2015). The tumour hypoxia marker pimonidazole reflects a transcriptional programme associated with aggressive prostate cancer. Br. J. Cancer *112*, 382–390. https://doi.org/10.1038/bjc.2014.604.

59. Ross, A.E., Feng, F.Y., Ghadessi, M., Erho, N., Crisan, A., Buerki, C., Sundi, D., Mitra, A.P., Vergara, I.A., Thompson, D.J.S., et al. (2014). A genomic classifier predicting metastatic disease progression in men with biochemical recurrence after prostatectomy. Prostate Cancer Prostatic Dis. *17*, 64–69. https://doi.org/10.1038/pcan.2013.49.

60. Erho, N., Crisan, A., Vergara, I.A., Mitra, A.P., Ghadessi, M., Buerki, C., Bergstralh, E.J., Kollmeyer, T., Fink, S., Haddad, Z., et al. (2013). Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy. PLoS One *8*, e66855. https://doi.org/10.1371/journal.pone.0066855.

61. Boormans, J.L., Korsten, H., Ziel-van der Made, A.J.C., van Leenders, G.J.L.H., de Vos, C.V., Jenster, G., and Trapman, J. (2013). Identification of TDRD1 as a direct target gene of ERG in primary prostate cancer. Int. J. Cancer *133*, 335–345. https://doi.org/10.1002/ijc.28025.

62. Ross, A.E., Johnson, M.H., Yousefi, K., Davicioni, E., Netto, G.J., Marchionni, L., Fedor, H.L., Glavaris, S., Choeurng, V., Buerki, C., et al. (2016). Tissue-based genomics augments post-prostatectomy risk stratification in a natural history cohort of intermediate- and high-risk men. Eur. Urol. *69*, 157–165. https://doi.org/10.1016/j.eururo.2015.05.042.

63. Ross-Adams, H., Lamb, A.D., Dunning, M.J., Halim, S., Lindberg, J., Massie, C.M., Egevad, L.A., Russell, R., Ramos-Montoya, A., Vowler, S.L., et al. (2015). Integration of copy number and transcriptomics provides risk stratification in prostate cancer: a discovery and validation cohort study. EBioMedicine *2*, 1133–1144. https://doi.org/10.1016/j.ebiom.2015.07.017.

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

65. Cope, L., Naiman, D.Q., and Parmigiani, G. (2014). Integrative correlation: properties and relation to canonical correlations.

J. Multivariate Anal. *123*, 270–280. https://doi.org/10.1016/j.jmva.2013.09.011.

66. Rouillard, A.D., Gundersen, G.W., Fernandez, N.F., Wang, Z., Monteiro, C.D., McDermott, M.G., and Ma'ayan, A. (2016). The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database *2016*, baw100. https://doi.org/10.1093/database/baw100.

67. Tong, Z., Cui, Q., Wang, J., and Zhou, Y. (2019). TransmiR v2.0: an updated transcription factor-microRNA regulation database. Nucleic Acids Res. *47*, D253–D258. https://doi.org/10.1093/nar/gky1023.

68. Agarwal, V., Bell, G.W., Nam, J.-W., and Bartel, D.P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. Elife *4*, e05005. https://doi.org/10.7554/eLife.05005.

69. Chou, C.-H., Shrestha, S., Yang, C.-D., Chang, N.-W., Lin, Y.-L., Liao, K.-W., Huang, W.-C., Sun, T.-H., Tu, S.-J., Lee, W.-H., et al. (2018). miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. Nucleic Acids Res. *46*, D296–D302. https://doi.org/10.1093/nar/gkx1067.

70. Sticht, C., De La Torre, C., Parveen, A., and Gretz, N. (2018). miRWalk: an online resource for prediction of microRNA binding sites. PLoS One *13*, e0206239.

71. Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics *27*, 1739–1740. https://doi.org/10.1093/bioinformatics/btr260.

72. Wu, Z., Guo, H., Chow, N., Sallstrom, J., Bell, R.D., Deane, R., Brooks, A.I., Kanagala, S., Rubio, A., Sagare, A., et al. (2005). Role of the MEOX2 homeobox gene in neurovascular dysfunction in Alzheimer disease. Nat. Med. *11*, 959–965. https://doi.org/10.1038/nm1287.

73. Kaizer, E.C., Glaser, C.L., Chaussabel, D., Banchereau, J., Pascual, V., and White, P.C. (2007). Gene expression in peripheral blood mononuclear cells from children with diabetes. J. Clin. Endocrinol. Metab. *92*, 3705–3711. https://doi.org/10.1210/jc.2007-0979.

74. Akl, H., Badran, B.M., Zein, N.E., Bex, F., Sotiriou, C., Willard-Gallo, K.E., Burny, A., and Martiat, P. (2007). HTLV-I infection of WE17/10 CD4+ cell line leads to progressive alteration of Ca2+ influx that eventually results in loss of CD7 expression and activation of an antiapoptotic pathway involving AKT and BAD which paves the way for malignant transformation. Leukemia *21*, 788–796. https://doi.org/10.1038/sj.leu.2404585.

75. Dorn, A., Zhao, H., Granberg, F., Hösel, M., Webb, D., Svensson, C., Pettersson, U., and Doerfler, W. (2005). Identification of specific cellular genes up-regulated late in adenovirus type 12 infection. J. Virol. *79*, 2404–2412. https://doi.org/10.1128/JVI.79.4.2404-2412.2005.

76. Marshall, D.R., Olivas, E., Andreansky, S., La Gruta, N.L., Neale, G.A., Gutierrez, A., Wichlan, D.G., Wingo, S., Cheng, C., Doherty, P.C., and Turner, S.J. (2005). Effector CD8+ T cells recovered from an influenza pneumonia differentiate to a state of focused gene expression. Proc. Natl. Acad. Sci. USA *102*, 6074–6079. https://doi.org/10.1073/pnas.0501960102.

77. Xu, Y., and Goodacre, R. (2018). On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. J. Anal. Test. *2*, 249–262. https://doi.org/10.1007/s41664-018-0068-2.

78. Noble, W.S. (2006). What is a support vector machine? Nat. Biotechnol. *24*, 1565–1567. https://doi.org/10.1038/nbt1206-1565.

79. Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32. https://doi.org/10.1023/A:1010933404324.

80. Chen, T., and Guestrin, C. (2016). XGBoost: a scalable tree boosting system. In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD '16 (ACM), pp. 785–794. https://doi.org/10.1145/2939672.2939785.

81. Fisher, R.A. (1992). Statistical methods for research workers. In Breakthroughs in Statistics: Methodology and Distribution Springer Series in Statistics, S. Kotz and N.L. Johnson, eds. (Springer), pp. 66–70. https://doi.org/10.1007/978-1-4612-4380-9_6.

82. Fisher, R.A. (1922). On the interpretation of χ 2 from contingency tables, and the calculation of P. J. Roy. Stat. Soc. *85*, 87–94. https://doi.org/10.2307/2340521.

83. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B *57*, 289–300.

84. Cope, L., Zhong, X., Garrett, E., and Parmigiani, G. (2004). MergeMaid: R tools for merging and cross-study validation of gene expression data. Stat. Appl. Genet. Mol. Biol. *3*, Article29. https://doi.org/10.2202/1544-6115.1046.

85. Afsari, B., Fertig, E.J., Geman, D., and Marchionni, L. (2015). switchBox: an R package for k–Top Scoring Pairs classifier development. Bioinformatics *31*, 273–274. https://doi.org/10.1093/bioinformatics/btu622.

86. Kuhn, M. (2008). Building predictive models in R using the caret package. J. Stat. Software *28*, 1–26. https://doi.org/10.18637/jss.v028.i05.

87. Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). Kernlab – an S4 package for kernel methods in R. J. Stat. Software *11*, 1–20.

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

89. Davison, A.C., and Hinkley, D.V. (1997). Bootstrap Methods and Their Application (Cambridge University Press). https://doi.org/10.1017/CBO9780511802843.

90. Csardi, G., and Nepusz, T. (2005). The Igraph Software Package for Complex Network Research (InterJournal *Complex Systems*), p. 1695.

91. Jawaid, W. (2022). enrichR: Provides an R Interface to "Enrichr".

92. Xie, Z., Bailey, A., Kuleshov, M.V., Clarke, D.J.B., Evangelista, J.E., Jenkins, S.L., Lachmann, A., Wojciechowicz, M.L., Kropiwnicki, E., Jagodnik, K.M., et al. (2021). Gene set knowledge discovery with enrichr. Curr. Protoc. *1*, e90. https://doi.org/10.1002/cpz1.90.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
BLCA1 dataset	(van der Heijden et al., 2016)[44]	GEO: GSE57813
BLCA2 dataset	(Kim et al., 2010)[45]	GEO: GSE13507
BLCA3 dataset	(Sjodahl et al., 2012)[46]	GEO: GSE32894
BLCA4 dataset	(Dyrskjøt et al., 2005)[47]	https://doi.org/10.1158/1078-0432.CCR-04-2095
BLCA5 dataset	(Hedegaard et al., 2016)[48]	ArrayExpress: E-MTAB-4321
BRCA1 dataset	(Hatzis et al., 2011)[49]	GEO: GSE25055
BRCA2 dataset	(Hatzis et al., 2011)[49]	GEO: GSE25065
BRCA3 dataset	(Edlund et al., 2021)[50]	GEO: GSE140494
BRCA4 dataset	(Birkbak et al., 2018)[51]	GEO: GSE103668
BRCA5 dataset	(Popovici et al., 2010; Shi et al., 2010)[52,53]	GEO: GSE20194
BRCA6 dataset	(Shen et al., 2012; Tabchy et al., 2010)[54,55]	GEO: GSE20271
BRCA7 dataset	(Miyake et al., 2012)[56]	GEO: GSE32646
PCA1 dataset	(Jain et al., 2018)[57]	GEO: GSE116918
PCA2 dataset	(Ragnum et al., 2015)[58]	GEO: GSE55935
PCA3 dataset	(Ross et al., 2014)[59]	GEO: GSE51066
PCA4 dataset	(Erho et al., 2013)[60]	GEO: GSE46691
PCA5 dataset	(Boormans et al., 2013)[61]	GEO: GSE41408
PCA6 dataset	(Ross et al., 2016)[62]	JHU Natural History cohort
PCA7 dataset	(Ross-Adams et al., 2015)[63]	GEO: GSE70769
TF-target and TF-miRNA interactions	(Rouillard et al., 2016)[66]	Harmonizome
miRNA-target interactions1	(Agarwal et al., 2015)[68]	TargetScan
miRNA-target interactions2	(Chou et al., 2021)[69]	miRTarBase
miRNA-target interactions3	(Sticht et al., 2018)[70]	miRWalk
The feed-forward loops mechanism	This paper	GitHub repository
The NOTCH-MYC signaling mechanism	This paper	GitHub repository
The cell-cell adhesion, activation, and O2 response mechanism	This paper	GitHub repository
The Alzheimer disease mechanism	This paper	GitHub repository
The diabetes mechanism	This paper	GitHub repository
The viral infection mechanism	This paper	GitHub repository
Code for reproducibility	This paper	GitHub repository
Software and algorithms		
R (v4.0.3)	(Ihaka and Gentleman, 1996)	https://www.r-project.org/
MergeMaid (v3.1)	(Cope et al., 2004)[84]	https://bioconductor.riken.jp/packages/3.1/bioc/html/MergeMaid.html
SwitchBox (v1.24.0)	(Afsari et al., 2015)[85]	https://www.bioconductor.org/packages/release/bioc/html/switchBox.html
caret (v6.0-86)	(Kuhn, 2008)[86]	https://topepo.github.io/caret/
limma (v3.44.3)	(Ritchie et al., 2015)	https://bioconductor.org/packages/release/bioc/html/limma.html

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
boot (v1.3-25)	(Davison and Hinkley, 1997)[89]	https://cran.r-project.org/web/packages/boot/index.html
randomForest (v4.6-14)	(Breiman, 2001)[79]	https://cran.r-project.org/web/packages/randomForest/index.html
kernlab (v0.9-29)	(Karatzoglou et al., 2004)[87]	https://cran.r-project.org/web/packages/kernlab/index.html
xgboost (v1.2.0.1)	(Chen and Guestrin, 2016)[80]	https://cran.r-project.org/web/packages/xgboost/index.html
pROC (v1.16.2)	(Robin et al., 2011)	https://cran.r-project.org/package=pROC
MsigDB (v6.1)	(Subramanian et al., 2005)[71]	https://www.gsea-msigdb.org/gsea/msigdb/
igraph (v1.3.0)	(Csardi and Nepusz, 2005)[90]	https://cran.r-project.org/web/packages/igraph/index.html
enrichR (v3.0)	(Jawaid, 2022; Xie et al., 2021)[91,92]	https://cran.r-project.org/web/packages/enrichR/index.html

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to the lead contact, Dr. Luigi Marchionni (lum4003@med.cornell.edu).

Material availability

This study did not generate new materials.

Data and code availability

This paper analyzes existing, publicly available data. The accession numbers for the datasets are listed in the key resources table. All original code together with our curated biological mechanisms have been deposited at https://github.com/MohamedOmar2020/Biological_Constraints and are publicly available. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Data collection

Bladder cancer

We used both the NCBI Gene Expression Omnibus (GEO)[42] and ArrayExpress[43] 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,[44] GSE13507,[45] GSE32894 [46] and pmid15930337 [47]) and the fifth is RNA-Seq based dataset (E-MTAB-4321 [48]).

Breast cancer

We identified seven datasets with pre-treatment gene expression profiles from patients with breast cancer who received neoadjuvant chemotherapy: GSE25055,[49] GSE25065, GSE140494,[50] GSE103668,[51] GSE20194,[52,53] GSE20271,[54,55] and GSE32646.[56]

Prostate cancer

For predicting PCa metastatic progression, we identified seven gene expression datasets containing primary tumor samples from PCa patients: GSE116918,[57] GSE55935,[58] GSE51066,[59] GSE46691,[60] GSE41408,[61] JHU natural history cohort,[62] and GSE70769.[63]

Data preprocessing

Bladder cancer

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.

Breast cancer

The seven breast cancer datasets originally included 1013 samples which we filtered to keep only 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 used in downstream analysis to 369 TNBC samples. All datasets were normalized and log2-scaled when retrieved from GEO, except for GSE20271 in which the expression values were not logged and so was log2-transformed. Finally, we mapped each probe ID to the corresponding gene symbol and filtered the expression matrices to the gene symbols in common (13299 genes).

Prostate cancer

In each of the seven PCa datasets, we kept primary tumor samples with information about metastatic events resulting in 1239 primary tumor samples eligible for downstream analysis. Normalization and preprocessing were performed as described above followed by z-score transformation for each dataset separately. Finally, probe IDs were mapped to their corresponding gene symbols and expression matrices were restricted to the genes in common.

Platform harmonization

For the prediction tasks in bladder and prostate cancers, which included data obtained across different platforms and technologies (microarrays and RNAseq), we harmonized the final datasets by identifying a subset of cross-study reproducible genes using integrative correlation coefficient (ICC)[64,65] 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 consider the phenotype information. Using the ICC threshold mentioned above, 3109 and 4055 genes were identified and used in the bladder and prostate cases, respectively.

Mechanistic pairs assembly

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., has-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 (Figure S1). This inverse relationship makes these pairs suitable for classification. To construct these loops, three different interaction types must 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[66] 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.[67] Finally, the miRNA-target interactions were obtained from TargetScan,[68] miRTarBase[69] and miRWalk.[70]

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 BLCA progression.

The NOTCH-MYC signaling mechanism

We used the Molecular Signature Database (MsigDB)[71] 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 MYC. Finally, both mechanisms were combined into a single mechanism consisting of 78672 pairs which was further used in the TNBC classification case.

The cell-cell adhesion, activation, and oxygen response mechanism

We used three gene ontology (GO) biological processes to build a mechanism that can capture the biology of metastasis. These processes included: *GOBP_CELL_CELL_ADHESION*, *GOBP_CELL_ACTIVATION*, and *GOBP_CELLULAR_RESPONSE_TO_OXYGEN*. Unique genes were paired together resulting in 409965 gene pairs which we used as biological constraints for training classifiers to predict PCa metastatic progression.

Non-cancer related mechanisms

As negative controls, we designed three non-cancer related mechanisms and assessed their performance at predicting each of the three main phenotypes. First, we built a mechanism for Alzheimer disease using two gene sets including genes up- and down-regulated in the brain endothelial cells of patients with Alzheimer disease.[72] Up- and down-regulated genes were paired together to form a mechanism consisting of 266 pairs. Second, we built a diabetes mechanism consisting of 6776 pairs by pairing up- and down-regulated genes in the peripheral blood monocytes from patients with diabetes at the time of the diagnosis versus 1-4 months later.[73] Finally, we designed a mechanism consisting of 6384 pairs representing the changes in the gene expression profiles of immune cells following viral infections.[74–76] It is important to note that these three mechanisms were chosen randomly with the aim of using them as negative controls to test our hypothesis.

Data splitting for training and testing

In each classification case, we implemented two data splitting designs: bootstrap and cross-study validation[77] (see Figure 1). In the bootstrap design, all datasets were combined 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. In the BLCA case, the clinical variables used in the stratification included age, sex, tumor grade, recurrence status, and intra-vesical therapy while in the breast cancer case, we included age, tumor grade, T and N stages. Finally, in the PCa case, we focused on age, Gleason score, tumor stage, and prostate-specific antigen (PSA) levels. Subsequently, models were trained to predict the phenotype of interest on 1000 bootstraps of the training data and their performance was evaluated on the unseen testing data using the Area Under the ROC Curve (AUC) as evaluation metric.

In the cross-study validation design, we used all but one dataset (*n-1*) for models training and the left-out dataset was used for testing. This process was repeated *n* times so that each dataset was used for testing once (see Figure 1).

Training and 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 three different model types: 1. mechanistic: using a manually curated biological mechanism in the form of pairwise comparisons (see below), 2. agnostic: using the top *2*k* differentially expressed genes (DEGs) (agnostic-genes) where k is the number of pairs used in the mechanistic models or their corresponding *k* pairwise comparisons (agnostic-pairs), 3. random genes: using a set of *2*k* randomly selected genes.

It should be noted that 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, all model types 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 those of the agnostic and random genes models to compare their average performance. In the cross-study validation, the average performance across all *n* 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.[9,11] More specifically, in the training process, if gene *X* is consistently more expressed relative to gene *Y* in samples belonging to 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 specific 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 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 both the TNBC and prostate cancer cases. The training of the mechanistic k-TSPs models was restricted to the mechanism of choice. In all cases, we restricted the number of output pairs (the final signature) to a range between 3 and 25 pairs. Finally, for each clinical problem, we obtained a matrix of binary values (*i.e.*, 0,1) representing the relative order of expression between genes pairs across individual samples (*i.e.*, samples by gene pairs), which we subsequently used for training purposes with the other three prediction algorithms besides k-TSP. Such matrices of pairwise gene comparisons were generated to match the maximum number of unique starting pairs used as input to train the mechanistic k-TSPs models (37 in the bladder, 241 in the TNBC, and 50 in the prostate), selecting an appropriate number of differentially expressed genes via Wilcoxon rank sum test for the agnostic case.

Support vector machine

SVM is an algorithm that aims at identifying a hyper-plane separating data points distinctively.[78] 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.[79] 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.[80] 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 was necessary to minimize overfitting. Hyperparameters including *gamma*, *lambda*, *alpha*, and

subsample were tuned using grid search process on the training data while stabilizing the learning rate and maximum depth. In the bootstrap analysis, this process was done on the training data before resampling. In the cross-study validation analysis, the hyperparameters were tuned on each training data partition.

Gene set enrichment analyses

To characterize the functional roles associated with the agnostic and mechanistic signatures, we performed GSEA to compute the overlap between the genes derived from the k-TSPs signatures and gene sets from the gene ontology (GO) biological processes database. In each prediction task, all unique mechanistic TSPs from the bootstrap processes were identified together with an equal number of agnostic TSPs. The GSEA analysis was performed on the genes associated with bad prognosis (positioned as gene1 in the TSPs) and those associated with good prognosis (positioned as gene2 in the TSPs), separately. *P*-values were calculated using Fisher's exact test[81,82] and were corrected using the Benjamini-Hochberg (BH) method for multiple hypotheses testing.[83] Finally, gene sets with an adjusted p-value greater than 0.05 were considered insignificant and removed.

QUANTIFICATION AND STATISTICAL ANALYSIS

All steps of this analysis were performed using R version 4.0.3 (2020-10-10). The integrative correlation analysis was performed using the *MergeMaid* package.[84] The k-TSPs models were trained using the *SwitchBox* R package.[85] The SVM models were trained using both the *Caret*[86] and *Kernlab*[87] packages. The RF and XGB models were trained using the *RandomForest*[88] and *xgboost*[80] packages, respectively. Bootstrapping (resampling with replacement) was performed using the *Boot* package.[89] The values plotted in Figures 2, 4, and 6 (Figures S2 and S7) represent the training and testing AUC values of the models trained on 1000 bootstraps of the training data and tested on the untouched testing data (not resampled). The values reported in Tables S2, S4, and S6 represent the average performance across all iterations of the cross-study validation process. The AUC values were computed using the prediction probabilities (RF, SVM, and XGB) or the votes (k-TSPs) returned by the classifier. The prediction probabilities were converted to predicted class labels using the optimal threshold which was determined from the training data using the ROC curve. The accuracy, balanced accuracy, sensitivity, specificity and MCC were computed by comparing the predicted class labels to the ground truth labels. The k-TSPs pairs returned from each bootstrap were ranked based on their frequency and individual genes were ranked based on their frequency in unique pairs. In the BLCA case, all mechanistic (n = 93) and an equal number of agnostic pairs were used to plot the networks shown in Figures 3C and 3D while in the TNBC and PCa cases, the top 100 pairs were used to plot the networks shown in Figures 5, 7C, and 7D. All networks were built using the *igraph* R software package.[90] The size of the network vertices corresponds to twice the log2 of the gene frequency in unique pairs while the edge thickness corresponds to the log2 of frequency of the gene pair across the 1000 bootstraps. Enrichment analyses were performed using the *enrichR* package[91,92] by computing the overlap between our gene lists and gene sets from the "GO_Biological_Process_2021" database with Benjamini-Hochberg (BH) method for multiple hypotheses testing.[83]