PRECISE+ predicts drug response in patients by non-linear subspace-based transfer from cell lines and PDX models.

Soufiane Mourragui[1,2], Marco Loog[2,3], Daniel J. Vis[1], Kat Moore[1], Anna G Manjon[4], Mark A. van de Wiel[5], Marcel J.T. Reinders[2,6,7], Lodewyk F.A. Wessels[1,2,7]

[1]*Division of Molecular Carcinogenesis, Oncode Institute, The Netherlands Cancer Institute, Amsterdam, The Netherlands.*

[2]*Department of EEMCS, Delft University of Technology, Delft, The Netherlands.*

[3]*Department of Computer Science, University of Copenhagen, Copenhagen, Denmark.*

[4]*Division of Cell Biology, Oncode Institute, The Netherlands Cancer Institute, Amsterdam, The Netherlands.*

[5]*Epidemiology and Biostatistics, Amsterdam University Medical Center, Amsterdam, The Netherlands.*

[6]*Leiden Computational Biology Center, Leiden University Medical Center, Leiden, The Netherlands.*

[7]*To whom correspondence should be addressed: Lodewyk F.A. Wessels (l.wessels@nki.nl) and Marcel J.T. Reinders (m.j.t.reinders@tudelft.nl).*

Abstract

Pre-clinical models have been the workhorse of cancer research for decades. Albeit powerful, these models do not perfectly recapitulate the complexity of human tumors which has led to a disappointing bench-to-bedside attrition rate. The quest for biomarkers of drug response signatures has been particularly challenging, suffering from poor translatability from pre-clinical models to human tumors. To address this problem, we present a novel computational framework, PRECISE+, that employs non-linear kernel approaches to capture complex biological processes expressed in both pre-clinical models and human tumors. PRECISE+ builds predictors on cell lines that show improved performance over competing approaches for a set of 7 of drugs in Patient-Derived Xenografts, 5 drugs on TCGA cohorts and show significant association with clinical response for 4 drugs in 226 metastatic tumors from the Hartwig Medical Foundation data set. We used the interpretability of PRECISE+ to validate the approach by identifying known biomarkers to targeted therapies and propose novel putative biomarkers of resistance to Paclitaxel and Gemcitabine.

Introduction

The accumulation of somatic alterations on the genome and epigenome transforms healthy cells into malignant tumor cells. Although these alterations are required for tumor growth, they also confer vulnerabilities on tumor cells. Some well-known examples of such genetic vulnerabilities are the amplification of ERBB2 in breast cancer[1], the BRAF^{V600E} mutation in skin melanoma[2] or the BCR/ABL fusion in leukemia[3]. These vulnerabilities have been successfully exploited clinically by directing drugs against them. However, for the vast majority of cancer patients, no clear biomarkers exist. Hence, expanding our arsenal of accurate biomarkers would pave the way for personalized medicine, by identifying, for each patient, the most effective drug[4].

In order to discover such biomarkers, pre-clinical models have been used extensively in the past decades, either in the form of cell lines, patient-derived xenografts (PDX) or organoids. This was partially fueled by the relative ease with which these model systems can be subjected to drug screening. This has led to break-through discoveries with broad clinical impact[5]. However, *Paul Valery*'s statement, "what is simple is always wrong ; what is not, is unusable"[6], also applies to these model systems. Specifically, their simplicity also confers weaknesses: the lack of a micro-environment in cell lines, and the absence of an immune system in cell lines, PDXs and organoids. These shortcomings are further amplified by culture artefacts[7,8] that lead to a reduced clinical significance of these models[9,10] and a high attrition rate in drug development[11].

Computational approaches to correct for these differences are therefore much needed[12]. In the particular case of cancer, these approaches are divided into two distinct categories. In a first category, mechanistic models are developed on pre-clinical models and subsequently "humanized" to focus on the similarities between pre-clinical models and human tumors[13]. A second category approaches the problem in a statistical fashion. Using molecular profiles and drug screens from large-scale panels of pre-clinical models[14,15], drug response in cell lines can be inferred based on molecular profiles[16–18]. The resulting predictive models are then applied to predict the sensitivity of patients to certain drugs[19–21]. Although already promising, these approaches do not take into account the fundamental differences between pre-clinical models and human tumors and show limited transferrability[22]. Recently, transfer learning and multi-task learning approaches have been developed to explicitly take these differences into account, either using partially tumor response[23], or solely based on pre-clinical labels[24].

We present PRECISE+, a general framework for subspace-based transfer learning[25–29] which enables the transfer of drug response predictors trained on a source domain (e.g. cell lines and PDXs) to a target domain (e.g. human tumors). PRECISE+ employs the powerful mathematical framework of Kernel methods[30–35] to capture both linear and non-linear relationships between samples. We show that PRECISE[24], a linear approach, is a special instance of PRECISE+. First, we demonstrate that, compared to linear approaches, PRECISE+ improves drug response prediction in PDXs after training on cell lines. We fix the hyperparameter controlling the degree of non-linearity on the PDX data and then employ PRECISE+ to transfer predictors of drug response trained on cell lines to two human tumors datasets: primary tumors from TCGA and metastatic lesions from the Hartwig Medical Foundation (HMF). Specifically, we show a significant improvement in response prediction for five drugs, including three cytostatic and two targeted therapies, compared to linear and non-adapted methods. Importantly, this performance improvement is attained without any training on data from the human tumors. We finally employ the interpretability of our approach to identify genes and pathways associated with drug response. We provide a full mathematical derivation of our algorithm, a complete reproducible pipeline and a fully open-source software package.

Results

PRECISE+: Generating non-linear subspace representations to transfer predictors of response from pre-clinical models to tumors

PRECISE+ compares genomic signals contained in the source (e.g. pre-clinical models) and target (e.g. human tumors) datasets, and outputs processes that are present in both datasets. The nature of these processes depends on the similarity function K that characterizes the relationships between samples (Methods). Depending on the similarity function employed, various types of non-linear relationships can be modelled. For instance, in the case of a Gaussian similarity function, these non-linearities include constant, linear, second and higher-order interaction terms, all modulated by the exponential of the squared depth (Methods).

In a first step, PRECISE+ computes the similarities between all samples (**Figure 1**A), yielding three matrices: K_s, K_t and K_{st}, containing the similarities between source, target, as well as source and target samples respectively. Using an eigen-decomposition of K_s and K_t, directions of maximal variance in the similarity space are computed using *Kernel PCA*[36] (**Figure 1**B). This decomposition is performed independently on the source and the target spaces and yields two *importance score matrices*: one for the source non-linear principal components (NLPCs) and one for the target NLPCs. These importance scores are *not* the projected sample values — they actually correspond to loadings, in the sense that they represent the geometric directions of the NLPCs in sample similarity space (**Supp Figure 1**A). We then quantify the geometric differences between all pairs of NLPCs from the two different sets by computing the cosine similarity matrix (**Figure 1**C). In a subsequent step (**Figure 1**D), we align these two sets of NLPCs by using the notion of *principal vector*s illustrated in **Supp Figure 1**B. These principal vectors (PV) are pairs of vectors – one from the source, one from the target – ranked by decreasing similarity: the top PVs correspond to geometrically similar factors, while bottom PVs are almost orthogonal. We restrict further analysis to the most similar PVs based on a similarity of at least 0.5 (Supp. Material). We perform, for each selected pair of PVs, an interpolation between the source and the target vectors and select one intermediate representation that best balances the contribution of the source and the target signals (**Figure 1**E, **Supp Figure 1**C). These vectors, called *consensus features*, define the consensus space, into which we project the source (pre-clinical) and target (tumor) samples. This projection yields, for all source and target samples, consensus feature values that can be used as input to any machine learning model to build predictors of drug response (**Figure 1**G, Method). In the case of a linear similarity function, PRECISE+ reduces to PRECISE[24] (Subsection Supp. 8) and is fundamentally different from approaches such as Canonical Correlation Analysis (CCA)[37] (Subsection Supp. 9).

Non-linearities improve response prediction of predictors transferred from cell lines to patient-derived xenografts (PDXs)

When it comes to predicting drug response in one model system, it is known that inducing non-linearities can lead to improved performance[35], although linear methods remain competitive[18,38]. We investigated whether the introduction of non-linearities in the computation of sample similarities resulted in improved response prediction of predictors trained on cell lines (source domain) and transferred to PDXs (target domain). Since gene expression is known to have predictive power comparable to other omics datasets combined[16,18,39,40], we restricted our analysis to the expression of 1 780 genes known to be related to cancer[41]. Using PRECISE+, we computed consensus features for cell lines (1 049 cell lines from 26 different tissues) and all PDXs (399 samples from 5+ different tissues) (Methods). We projected the gene expression data of all cell lines and all PDXs onto these consensus features. We employed Elastic Net to train models of drug response with the projected cell line expression data as input and the measured drug response (AUC) data as output. We applied this trained predictor on the projected PDX expression data and compared the predicted response to the measured best average response by Spearman Correlation (**Figure 2**A). We made use of the standard Gaussian similarity function (Methods) to vary the level of non-linearity introduced. This similarity function is characterized by a single scaling factor γ, whose size is directly proportional to the proportion of non-linearity introduced (**Figure 2**B). We studied the predictive performance in PDXs for seven different values of γ, ranging from a set of consensus features with an almost purely linear ($\gamma = 1 \times 10^{-5}$) to an almost purely non-linear composition ($\gamma = 1 \times 10^{-2}$). We employed a non-domain adapted baseline (Elastic Net regression) and linear PRECISE as references. All models were trained to predict response to seven different drugs (Erlotinib, Cetuximab, Gemcitabine, Afatinib, Paclitaxel, Trametinib and Ruxolitinib) for which we had response data available for both PDX models and cell lines (**Figure 2**C-H). For these seven drugs, we observe, in general, a clear improvement of domain-adaptation over the baseline, indicating a necessity to correct the input signal when moving from the source to the target domain. Except for Cetuximab, we observe that the introduction of non-linearity tends to increase the predictive performance on PDXs. Specifically, we observe for several drugs that the predictive performance increases with the scaling factor until a maximal performance is reached ($\gamma = 10^{-4}$ for Erlotinib, Cetuximab and Afatinib and $\gamma = 10^{-3}$ for Gemcitabine, Paclitaxel and Trametinib), after which the predictive performance drops dramatically. We therefore decided to fix the scaling factor to the average value ($\gamma^* = 5 \times 10^{-4}$) and employ the associated consensus space to transfer the predictors of response to the tumor samples. As a further check, we analyzed the properties of the consensus space obtained using γ^*. We observe a concentration of the offset contribution in the top consensus features and an increasing proportion of non-linear terms contribution to lower order features (**Supp Figure *6***C). The UMAP[42] projection of the consensus features shows a clear co-clustering of cell lines and PDXs of the same tissue (**Supp Figure 6**D).

Consensus features between cell lines (GDSC) and human tumors conserve primary tumor information.

With the scaling factor (γ) calibrated on PDX models, we moved to the clinical setting to investigate domain adaptation between cell lines to two different human tumor datasets: primary tumors from TCGA and metastatic lesions from the HMF. We selected 30 consensus features in the GDSC-TCGA analysis (**Supp Figure 8**) and 20 in the GDSC-HMF analysis (**Supp Figure 9**) after a first selection of NLPCs based on the inflexion point of the cumulative eigenvalues, and a subsequent cut-off of PVs with similarity above 0.5. We observe that the consensus features computed between GDSC and TCGA (**Figure 3**A) and between GDSC and HMF (**Figure 3**B) show the same proportion of non-linearities with a concentration of offset and linearities in the top consensus features.

In order to visualize the structure retained in the consensus space, we embedded our consensus scores into a 2D space using UMAP[42]. We observed that primary tumors cluster together based on their tissue type (**Figure 3**C). Cell lines, however, show different behaviors – most do cluster with the tumors with a similar tissue of origin, while a group of cell lines cluster together and away from the tumors, regardless of their tissue of origin, as observed in previous studies[43]. To quantify the degree of co-clustering of cell lines and tumors, we compared distances between tumors and cell lines from similar and non-similar tissues, and observed, as expected, a higher similarity between tumors and cell lines from the same tissue (**Supp Figure 10**C). Metastatic lesions show a weaker clustering based on the tissue of origin of the primary tumor (**Supp Figure 10**D-E). This is not unexpected, as the expression profiles are derived from biopsy sites distant from the primary tissue. Of particular interest, we observe the existence of a hematopoietic cell-line cluster that co-clusters with metastatic samples from various biopsy sites. Most of these tumor samples (7 out of 12 samples) are lymph nodes metastasis and most likely display a hematopoietic expression profile due to blood infiltration in the samples (**Supp Figure 10**B).

Consensus features increase transfer of response predictors from cell lines to primary tumors and metastatic lesions

To further validate our approach, we transferred response predictors from cell lines to the TCGA and HMF collections of human tumors. First, we projected the GDSC and TCGA expression data onto the GDSC-TCGA consensus features. Then we trained, for each drug, a regression model using solely the cell line response data (AUC). These drug-specific regression models were then used to predict response on the projected TCGA data. Finally, we compared the predicted response to the known categorical clinical responses using a one-sided Mann-Whitney test and computed the effect size using a Cohen's d statistic. We trained models for the seventeen different drugs (**Table 1**A) and observe a significant association for seven drugs : Trastuzumab, Cisplatin, Carboplatin, Etoposide, Gemcitabine, Oxaliplatin and Paclitaxel. We compared the prediction of PRECISE+ to the predictions obtained using PRECISE (linear instantiation of PRECISE+) and a non-adapted baseline (**Figure 4**A). We observe that PRECISE+ out-performs both on five drugs (Afatinib, Carboplatin, Cisplatin, Gemcitabine and Paclitaxel), with the baseline and PRECISE reaching significance in three and five drugs, respectively. For the HMF data, we repeated the steps above, while employing the GDSC-HMF consensus features as well as the HMF and GDSC expression and response data. We trained models for Irinotecan, Carboplatin (using Cisplatin GDSC response), Trastuzumab (using Afatinib GDSC response), Gemcitabine, Paclitaxel and 5-Fluorouracil. We observe a significant association between the predicted AUC and clinical responses for four of the five drugs (Irinotecan, Carboplatin, Trastuzumab and Gemcitabine) (**Table 1**B, **Figure 4**A) – in contrast, the baseline and PRECISE reach significance for zero and two drugs, respectively. PRECISE+ out-performs PRECISE in three of the five drugs (Irinotecan, Cisplatin and Gemcitabine). PRECISE+ does not obtain a significant association for

Paclitaxel, but we do, however, observe a trend that indicates a positive association between predicted and clinical response. In contrast, the non-domain-adapted baseline fails to recapitulate any association.

Two drugs, already standing out from the GDSC-PDX analysis, are of particular interest. We first observe that Gemcitabine is consistently better predicted by PRECISE+ than by PRECISE or the baseline (**Figure 4**B). When it comes to Paclitaxel (**Figure 4**C), PRECISE+ shows a clear improvement over PRECISE and baseline in TCGA. PRECISE+ is also the only method to recover a positive association in HMF, although not significant. Finally, for Carboplatin (**Figure 4**C), we observe that PRECISE+ outperforms PRECISE and the baseline on TCGA, and is the only method to obtain a significant association on HMF. Altogether, these results show that the consensus features can be used to construct non-linear features that, when trained on cell lines, deliver superior performance in predicting drug response in human tumors.

Interpretability of consensus features confirms known mechanisms for targeted therapies and unveils potential biomarkers of sensitivity for cytotoxic drugs

We finally made use of the interpretability of our approach to mechanistically validate our predictors (Methods). We first validated targeted therapies with documented modes of action. We started with the PRECISE+ predictor of response for Afatinib, a small molecule inhibitor of the EGFR family, which includes HER2 (**Figure 5**A). We performed a gene set enrichment analysis of the linear terms that constitutes to 80% of the predictor. Most enriched gene sets are related to breast cancer subtypes as defined by *Charafe and colleagues*[44] where, contrary to the definition based on the intrinsic breast cancer subtypes, the Luminal subtype contains both ER+ and HER2+ tumors. The top ranked gene set amongst the genes associated with sensitivity (genes with a negative coefficient in the predictor) are genes associated with the "Luminal" subtypes (FDR < 0.001). Conversely, genes associated with resistance (genes with a positive coefficient in the predictor) show enrichment for the "Mesenchymal" molecular signatures, shared by basal and mesenchymal subtypes, i.e. HER2 negative samples, which is in line with our expectation as absence of the drug target would indicate lack of response. Similarly, in the PRECISE+ response predictor for Gefitinib (EGFR inhibitor) the genes constituting the linear portion and associated with sensitivity (negative predictor coefficients) show an enrichment for genes downregulated in Gefitinib resistant tumors (**Figure 5**B).

Cytotoxic drugs such as Gemcitabine or Paclitaxel have complex modes of actions involving different pathways which interplays remain challenging to understand. Since the predictions of these two drugs showed a significant association in both PDXs and patients, we set out to interpret the mechanisms of sensitivity or resistance inferred by our predictor. In Gemcitabine, we observe that over-expression of the CDC42 pathway is a significant marker of resistance (FDR = 0.012, **Figure 5**C) together with pathways linked to microtubule formation and cell migration (**Supp Figure 11**), both known to be promoted by CDC42[45]. Together, these enriched pathways highlight CDC42 over-expression as a potential *in-vivo* mechanism of Gemcitabine resistance, which suggests the use of CDC42 inhibitors[46,47] for Gemcitabine-resistant tumors. Another interesting finding is the significant enrichment of TNFα signaling in the sensitive portion (FDR=0.046). A clinical trial has shown that co-administration of TNF with gemcitabine improves patient survival and further inhibits tumor growth[48], lending additional credence to this finding. Last, we observe a concentration of sensitive interactions involving BLK, a pro-apoptotic Src-proto oncogene involved in B-cell signaling and differentiation. Since hematopoietic cell lines respond better to Gemcitabine, these interactions can either act as a tissue-type marker, or could potentially represent a potential sensitive pathway.

Finally, we looked for enriched pathway in Paclitaxel predictor (**Figure 5**D) and observed three potential mechanisms of resistance. We first observe that the resistant linear coefficients are significantly enriched in genes linked to silencing of YBX1[49] (FDR=0.106), a gene associated with proliferation in certain tumor types[50]. In ovarian cancer, YBX1 has been shown to regulate ABCB1 expression levels, a gene related to Paclitaxel resistance[52–56]. Our pan-cancer analysis therefore further supports the role of drug transporters in Paclitaxel resistance. Second, we observe a significant resistant enrichment for PI3K activation (FDR=0.18), which is corroborated by the observed activation of PI3K/AKT/mTOR signaling pathway in Paclitaxel-resistant cancer cells[56,57]. Moreover, a recent investigation suggests that PI3K catalytic subunits can regulate ABCB1 expression[58]. Finally, when it comes to the non-linear part, we observe a concentration of Fibroblast growth factors interactions in the resistant parts of non-linearities, in particular FGF3, FGF20 and FGF8 and FGF4. This behavior, although suggested by previous studies[59,60], is all the most interesting as cell lines do not contain any micro-environment that would elicit such resistance.

Discussion

We introduced an approach to integrate pre-clinical and clinical data in a fully unsupervised way. Our approach geometrically aligns sample-to-sample similarity matrices and extract directions of important variations for both datasets, without requiring any sample-level pairing. By performing a geometrical alignment instead of a direct distribution comparison, our approach limits the effect of any sample selection bias. This geometrical alignment is implicitly performed in a space induced by our similarity function, which enable the integration of various non-linearities, corresponding to hypothesis made on the system. Although we restricted ourselves to the Gaussian similarity, designing similarity functions that incorporate a wide range of prior knowledge is a potentially promising avenue. Learning the similarity matrix, e.g. using multiple kernel learning or deep learning methods, could also help increase performance. Our method is versatile and general and can be applied beyond the scope of our study, e.g. to integrate single cell sequencing data.

We showed that the consensus features can be used to build translatable predictors of drug response. Although we do not require a strong covariate shift assumption as in a previous study[61], we do assume that the functions modelling the response from these consensus features follow the same monotonicity in pre-clinical models and human tumors. This assumption, albeit reasonable, may be questioned.

In this study, we limited ourselves to gene expression. Making use of other genomics levels – e.g. mutations, copy number, methylation, chromatin accessibility – may help refine the prediction by providing additional signal. The integration of our approach with multi-omics integration strategies[62,63] may offer a solution to the translation of multi-omics signatures.

Finally, we focused in our study on cytotoxic and targeted therapies, still widely used in the clinic. The recent advent of immuno-therapies calls for methods able to predict the clinical response from model systems. This requires model systems able to mimic the action of the immune system and screening technologies able to measure the response for large panels. We believe that our approach can be extended to such problems once data is made available.

Methods

Public data download and pre-processing

GDSC dataset, download and processing.

We made use of the GDSC1000 cell line panel[14], which contains complete molecular profiles for 1,049 cell lines (**Supp Figure 2**). Gene expression is provided in the forms of both read counts and FPKM. For both settings, we corrected the dataset for library-size using TMM[64] and log-transformed the corrected read counts[65,66]. Finally, we performed a gene-level mean-centering and standardization. Response to 397 drugs is provided in the form of Area Under the Curve (AUC).

Novartis PDXE dataset, download and processing.

We made use of NIBR PDXE dataset for patient-derived xenografts[15], which contains the gene expression profiles of 399 PDXs (**Supp Figure 3**). Gene expression is provided in the form of FPKM. We corrected for library-size using TMM[64] and log-transformed the corrected read counts[65,66]. Finally, we performed a gene-level mean-centering and standardization. Response to drugs overlapping with GDSC is provided in the form of Best Average Response.

TCGA dataset, download and processing.

We made use of the TCGA dataset for analyzing human biopsies[67], which comprises 10,347 human tumors (**Supp Figure 4**). Gene expression is provided in the forms of both read counts and FPKM and we used the same pre-processing pipeline as for GDSC.

Hartwig Medical Foundation dataset (HMF) download and processing.

We validated our approach on a cohort of 1,049 patients provided by the Hartwig Medical Foundation – referred to as *HMF* (**Supp Figure 5**A). Gene expression was measured for each metastatic sample prior to indicated drug regimen. We used MultiQC for quality control[68], salmon v1.0.0 for alignment to reference transcriptome[69], and finally edgeR for gene-level quantification[70]. Comparison with results obtained using STAR[71] and featureCounts[72] shows high degree of concordance (**Supp Figure**

5D) and we used this comparison to refine our filtering. Read counts were then processed using the same pipeline as in GDSC and TCGA.

Drug response was measured in 802 unique metastatic samples using the RECIST criteria. Response was measured differently for each patient (**Supp Figure 5**B) with most patients having one single measure of response around 2.5 weeks after treatment start (**Supp Figure 5**C). Since we are interested in the response of the drug given the molecular characterization measured, we considered for each patient the first response after treatment.

Mathematical settings

We denote by p the number of genes. We consider one source dataset $\mathcal{X}_s = \{x_1^s, \dots, x_{n_s}^s\} \subset \mathbb{R}^p$and one target dataset $\mathcal{X}_t = \{x_1^t, \dots, x_{n_t}^t\} \subset \mathbb{R}^p$ with corresponding source and target data matrices $X_s \in \mathbb{R}^{n_s \times p}$ and $X_t \in \mathbb{R}^{n_t \times p}$.

We consider a similarity function K -- also called *kernel function* -- that assigns to two samples a scalar value that is large for similar samples and small for dissimilar samples. In this work, we assume the kernel to be positive semi-definite (p.s.d.), which implies[73] that there exists a Hilbert space $\mathcal{H}$ and a mapping $\varphi: \mathbb{R}^p \mapsto \mathcal{H}$ such that

$$\forall x, y \in \mathbb{R}^p, \qquad K(x,y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}. \tag{1}$$

In particular, we use two kernels:

- Linear kernel: $K^{linear}(x,y) = x^T y$.
- Radial Basis Function, also referred to as “Gaussian”: $K_\gamma^{rbf}(x,y) = \exp(-\gamma \|x-y\|^2)$, with $\gamma > 0$.

We denote by K_s the matrix of similarity between source samples, K_t between target samples and K_{st} the matrix of similarity between source and target, formally:

$$\begin{aligned} K_s &= \left(K\left(x_i^s, x_j^s\right)\right)_{1 \le i,j \le n_s} \\ K_t &= \left(K\left(x_i^t, x_j^t\right)\right)_{1 \le i,j \le n_t}. \\ K_{st} &= \left(K\left(x_i^s, x_j^t\right)\right)_{\substack{1 \le i \le n_s \\ 1 \le j \le n_t}} \end{aligned} \tag{2}$$

Implicit mean centering

In a linear-setting, it is standard to perform a gene-level centering prior to any statistical analysis. In the case of non-linear analysis, we perform a feature-level centering in the embedding space $\mathcal{H}$ implicitly using Equation (1). We perform this implicit centering independently in source and target datasets. As shown in (Subsection Supp. 2), the corresponding kernel matrices are:

$$\begin{aligned}\widetilde{\mathrm{K}}_{\mathrm{s}} &= \mathrm{C}_{\mathrm{n_s}}\mathrm{K}_{\mathrm{s}}\mathrm{C}_{\mathrm{n_s}}\\ \widetilde{\mathrm{K}}_{\mathrm{t}} &= \mathrm{C}_{\mathrm{n_t}}\mathrm{K}_{\mathrm{t}}\mathrm{C}_{\mathrm{n_t}}\text{ , with } \mathrm{C}_{\mathrm{n}} = \mathrm{I}_{\mathrm{n}} - \frac{1}{\mathrm{n}}1_{\mathrm{n}}^{\mathrm{T}}1_{\mathrm{n}}\ (\mathrm{n} \geq 1)\\ \widetilde{\mathrm{K}}_{\mathrm{st}} &= \mathrm{C}_{\mathrm{n_s}}\mathrm{K}_{\mathrm{st}}\mathrm{C}_{\mathrm{n_t}}\end{aligned} \tag{3}$$

Kernel PCA by eigen-decomposition of centered kernel matrix for capturing directions of principal variance

Using the embedding introduced in Equation (1), the similarity matrices from Equation (3) can be seen as sample-covariance matrices and therefore decomposed to compute principal components inside the embedded space $\mathcal{H}$, a procedure known as Kernel PCA[36]. We perform Kernel PCA on source and target data independently to compute d_s and d_t principal components respectively. Kernel PCA on the source dataset consists in an eigen-decomposition of the matrix $\widetilde{K}_s$, yielding $\alpha^s \in \mathbb{R}^{d_s \times n_s}$, while Kernel PCA on the target dataset decomposes $\widetilde{K}_t$, yielding $\alpha^t \in \mathbb{R}^{d_t \times n_t}$ (Def Supp 3.1).

Comparing and aligning pre-clinical and tumor non-linear principal components

Similarly to the “cosine similarity matrix” in other related works[24,28], we define the *non-linear cosine similarity matrix* $\boldsymbol{M}^K$ as the matrix that geometrically compares the source NLPCs to the target NLPCs (Def Supp 5.1). This matrix can be computed as follow (Prop Supp 5.2):

$$\boldsymbol{M}^K = \alpha^s \widetilde{K}_{st} \alpha^{t^T} = \alpha^s C_{n_s} K_{st} C_{n_t} \alpha^{t^T}. \tag{4}$$

In a first step of our domain adaptation approach, we use the matrix $\boldsymbol{M}^K$ to align NLPC, yielding *non-linear principal vectors* $s_1,..,s_d$ for source and $t_1,..,t_d$ for target with $d = \min(d_s, d_t)$ (Def Supp. 4.1). These principal vectors are pairs of vectors: one linear combinations of source NLPCs and one linear combinations of target NLPCs, ordered by decreasing similarity with the first pair being the most similar. The computation of these PVs rely on the Singular Value Decomposition[74] of $\boldsymbol{M}^K$, $\boldsymbol{M}^K = \beta^s \Sigma \beta^{t^T}$that helps us define the source and target *sample importance loadings* ρ^s and ρ^t as follows (Def. Supp. 5.4)

$$\rho^s = \beta^{s^T} \alpha^s \text{ and } \rho^t = \beta^{t^T} \alpha^t. \tag{5}$$

We also define the principal angles $\theta_1,..,\theta_d$ as follows (Def. Supp. 5.6)

$$\forall \mathrm{k} \in \{1,..,\mathrm{d}\}, \qquad \cos\, \theta_k = \Sigma_{k,k}\,. \tag{6}$$

Owing to the duality elicited in Equation (1), these principal vectors can be seen both as functions that map samples in cell-view to a one-dimensional vector, or a vector onto which we project the sample embedding (Proposition Supp. 5.8).

Interpolation between kernel principal vectors for balancing effect of source and target

Each pair of principal vectors contains two vectors that are geometrically similar. Projection on them will create two highly correlated covariates that would not be optimal for subsequent statistical treatment. In order to compute one vector out of each pair, we interpolate between the source and the target PV within each pair (Def Supp 6.2). For the k^{th} PV, the interpolation is modulated by a coefficient τ_k that ranges between 0, when the interpolation returns the source PV, and 1, when the interpolation returns the target PV. This interpolation between vectors within each PV pair relies on two functions $\Gamma(\tau) = [\Gamma_1(\tau_1),..,\Gamma_d(\tau_d)]^T$ and $\xi(\tau) = [\xi_1(\tau_1),..,\xi_d(\tau_d)]^T$ defined as (Definition Supp. 6.1):

$$\forall k \in \{1,..,d\}, \quad \Gamma_k(\tau_k) = \frac{\sin(1-\tau_k)\theta_k}{\sin\theta_k} \quad \text{and} \quad \xi_k(\tau_k) = \frac{\sin\tau_k\theta_k}{\sin\theta_k} \tag{7}$$

For a set of d interpolation coefficients $[\tau_1,..,\tau_d]$, we compute the projection of source and target datasets $F(\tau) \in R^{(n_s+n_t)\times d}$ as follows (Theorem Supp. 6.6)

$$F(\tau) = \begin{bmatrix} K^s & K^{st} \\ K^{ts} & K^t \end{bmatrix} \begin{bmatrix} C_{n_s} & 0 \\ 0 & C_{n_t} \end{bmatrix} \begin{bmatrix} \rho^{s^T} & 0 \\ 0 & \rho^{t^T} \end{bmatrix} \begin{bmatrix} \Gamma(\tau) \\ \xi(\tau) \end{bmatrix}. \tag{8}$$

Such an interpolation between PVs balance the effect of source and target datasets. We prove that, in the case of linear kernel, our interpolation scheme is equivalent to the one from previous approaches[26,75] (Supp Subsection 8).

Within each pair of PVs, we select one intermediate representation where the source and target projected match the most. For the k^{th} PV-pair, we compare the source and target projected data using a Kolmogorov-Smirnov statistic and select the interpolation time τ_k^* where the statistics is maximal. We obtain a set of optimal interpolation times $\tau^* \in [0,1]^d$ when for each PV, source and target influence are balanced and we call the corresponding vector *consensus features*. These consensus features show a limited difference between source and target domain, a theoretical necessary condition for domain adaptation[76].

Prediction using Elastic Net

In order to predict drug response, we use Elastic Net regression[77]. Elastic Net is a linear model that imposes two penalties on the coefficients to predict: an $\ell_1$penalty that leads to sparse model and an ℓ_2 penalty that jointly shrinks correlated features.

We chose Elastic Net first because it has repeatedly been shown in the drug response prediction literature to give equivalent, if not better, predictive performance as more complex models[16,18,38]. Second, target error is upper-bounded, among other terms, by the VC dimension of the classifier[76]. Using a linear classifier limits the complexity and therefore makes the transfer theoretically more robust.

Taylor expansion of the similarity function for interpretability of the model

In the case of RBF, we perform a PCA in an infinite-dimensional feature space. Although this space cannot be analytically computed, it can be approximated using a Taylor expansion[78] (Subsection Supp. 7). For the q^{th} consensus feature, we get three kinds of contributions (Def Supp. 7.4):

- *Offset $\mathcal{O}_q$:* Gaussian term that models the squared depth. For each sample $x \in R^p$, it corresponds to $\exp(-\gamma\|x\|^2)$.
- *Linear contributions $\left(\mathcal{L}_{q,j}\right)_{1\leq j\leq p}$:* analog to linear term, i.e. expression of one gene. For each sample $x \in R^p$ and gene $j \in \{1,..,p\}$, it corresponds to $x_j\exp(-\gamma\|x\|^2)$.
- *Interaction terms $\left(\mathcal{I}_{q,j,k}\right)_{1\leq j,k\leq p}$:* analog to interaction term, i.e. product of expressions of two genes. For each sample $x \in R^p$ and gene $j, k \in \{1,..,p\}$, it corresponds to $x_j x_k\exp(-\gamma\|x\|^2)$.

Higher order interactions are also taken into account into the consensus feature, but for computational reasons, we do not look at individual contributions. These contributions can be computed from sample importance loadings of consensus features (Prop. Supp. 7.7). We consider the contributions' sum-of-squares as a geometrical proportion since these sum up to one (Def. Supp. 7.8).

In order to look for enrichment in a particular consensus feature, we look for enrichment of particular gene sets[79]. Specifically, for the linear contribution, we compute the loading of all linear terms (Equation Supp. 48) corresponding each to one gene and we performed a Pre-Ranked gene set enrichment analysis with FDR correction at 20% and 1000 permutations. Since these loadings corresponds to a Euclidean geometric proportion, we used a squared statistic to compare them.

Code and availability.

PRECISE+ is available as a Python 3.6 module (https://github.com/NKI-CCB/PRECISE_plus). All our experiments are reproducible and use state-of-the-art libraries[80–83] (https://github.com/NKI-CCB/precise_plus_manuscript)

Acknowledgment

This publication and the underlying study have been made possible partly on the basis of the data that Hartwig Medical Foundation and the Center of Personalised Cancer Treatment (CPCT) have made available to the study.

We thank Mirrelijn van Nee (VUmc), Sander Canisius (NKI), Tycho Bismeijer (NKI), Stavros Makrodimitris (TU Delft), Tesa Severson (NKI), Joe Siefert (NKI), Joris van de Haar (NKI), Ahmed Mahfouz (LUMC), Tamim Abdelaal (TU Delft), David Tax (TU Delft), Bram Thijssen (NKI) and Ziqi Wang (TU Delft) for useful discussions. We thank Mohammed Charrout (LUMC), Marie Corradi (NKI), Lisa Koob (NKI) and Wouter Kouw (TU Eindhoven) for careful reading of the manuscript.

This work was supported by ZonMw TOP grant COMPUTE CANCER [40-00812-98-16012].

Competing interests

L.F.A.W. received project funding from Genmab BV.

Authors contribution

S.M., L.F.A.W., M.J.T.R., M.L. and M.A.vd.W. designed the study. S.M. performed the experiments. L.F.A.W., M.J.T.R. and M.L. supervised the experiments. S.M., L.F.A.W., M.J.T.R., M.L. and M.A.vd.W. analyzed the results. S.M. and M.L. developed the mathematical framework. S.M. developed the software package. S.M. and D.J.V. aligned the HMF data. K.M. and A.G. interpreted the GSEA results. S.M., L.F.A.W., M.J.T.R. and M.L. wrote the manuscript. All authors read and approved the manuscript.

1. Slamon, D. J. *et al.* Numb Er 11 Use of Chemotherapy Plus a Monoclonal Antibody Against Her2. *N. Engl. J. Med.* **344**, 783–792 (2001).
2. Davies, H. *et al.* Mutations of the BRAF gene in human cancer. *Nature* **417**, 949–954 (2002).
3. Hughes, T. P. *et al.* Frequency of major molecular responses to imatinib or interferon alfa plus cytarabine in newly diagnosed chronic myeloid leukemia. *N. Engl. J. Med.* **349**, 1423–1432 (2003).
4. Kalamara, A., Tobalina, L. & Saez-Rodriguez, J. How to find the right drug for each patient? Advances and challenges in pharmacogenomics. *Curr. Opin. Syst. Biol.* **10**, 53–62 (2018).
5. Prahallad, A. *et al.* Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. *Nature* **483**, 100–104 (2012).
6. Valery, P. *Mauvaises pensées et autres*. (1941).
7. Ben-David, U. *et al.* Patient-derived xenografts undergo mouse-specific tumor evolution. *Nat. Genet.* **49**, 1567–1575 (2017).
8. Ben-David, U. *et al.* Genetic and transcriptional evolution alters cancer cell line drug response. *Nature* **560**, 325–330 (2018).
9. Gillet, J. P. *et al.* Redefining the relevance of established cancer cell lines to the study of mechanisms of clinical anti-cancer drug resistance. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 18708–18713 (2011).
10. Gillet, J. P., Varma, S. & Gottesman, M. M. The clinical relevance of cancer cell lines. *J. Natl. Cancer Inst.* **105**, 452–458 (2013).
11. Mak, I. W. Y., Evaniew, N. & Ghert, M. Lost in translation: Animal models and clinical trials in cancer treatment. *Am. J. Transl. Res.* **6**, 114–118 (2014).
12. Brubaker, D. K. & Lauffenburger, D. A. Translating preclinical models to humans. *Science (80-.).* **367**, 742–743 (2020).
13. Webber, J. T., Kaushik, S. & Bandyopadhyay, S. Integration of Tumor Genomic Data with Cell Lines Using Multi-dimensional Network Modules Improves Cancer Pharmacogenomics. *Cell Syst.* **7**, 526-536.e6 (2018).
14. Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* (2016). doi:10.1016/j.cell.2016.06.017
15. Gao, H. *et al.* High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nat. Med.* **21**, 1318–1325 (2015).

16. Costello, J. C. *et al.* A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.* **32**, 1202–1212 (2014).
17. Ali, M. & Aittokallio, T. Machine learning and feature selection for drug response prediction in precision oncology applications. *Biophys. Rev.* **11**, 31–39 (2019).
18. Jang, I. S., Neto, E. C., Guinney, J., Friend, S. H. & Margolin, A. A. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. *Pacific Symp. Biocomput.* **23**, 1–7 (2013).
19. Geeleher, P., Cox, N. J. & Huang, R. S. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biol.* **15**, 1–12 (2014).
20. Geeleher, P. *et al.* Discovering novel pharmacogenomic biomarkers by imputing drug response in cancer patients from large genomics studies. *Genome Res.* **27**, 1743–1751 (2017).
21. Sakellaropoulos, T. *et al.* A Deep Learning Framework for Predicting Response to Therapy in Cancer. *Cell Rep.* **29**, 3367-3373.e4 (2019).
22. Kurilov, R., Haibe-Kains, B. & Brors, B. Assessment of modelling strategies for drug response prediction in cell lines and xenografts. *Sci. Rep.* **10**, 2849 (2020).
23. Noghabi, H. S., Peng, S., Zolotareva, O., Collins, C. C. & Ester, M. AITL: Adversarial Inductive Transfer Learning with input and output space adaptation for pharmacogenomics. *bioRxiv* 2020.01.24.918953 (2020). doi:10.1101/2020.01.24.918953
24. Mourragui, S., Loog, M., van de Wiel, M. A., Reinders, M. J. T. & Wessels, L. F. A. PRECISE: a domain adaptation approach to transfer predictors of drug response from pre-clinical models to tumors. *Bioinformatics* **35**, i510–i519 (2019).
25. Pan, S. J., Kwok, J. T. & Yang, Q. Transfer Learning via Dimensionality Reduction.
26. Gong, B., Shi, Y., Sha, F. & Grauman, K. Geodesic flow kernel for unsupervised domain adaptation. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 2066–2073 (2012). doi:10.1109/CVPR.2012.6247911
27. Gopalan, R., Li, R. & Chellappa, R. Domain adaptation for object recognition: An unsupervised approach. *Proc. IEEE Int. Conf. Comput. Vis.* 999–1006

(2011). doi:10.1109/ICCV.2011.6126344

28. Fernando, B., Habrard, A., Sebban, M. & Tuytelaars, T. Unsupervised visual domain adaptation using subspace alignment. *Proc. IEEE Int. Conf. Comput. Vis.* 2960–2967 (2013). doi:10.1109/ICCV.2013.368
29. Kouw, W. & Loog, M. A review of domain adaptation without target labels. *IEEE Trans. Pattern Anal. Mach. Intell.* **PP**, 1–1 (2019).
30. Hofmann, T., Schölkopf, B. & Smola, A. J. Kernel methods in machine learning. *Ann. Stat.* **36**, 1171–1220 (2008).
31. Vert, J.-P., Koji Tsuda & Schölkopf, B. *Kernel methods in computational biology*. (MIT Press, 2004).
32. He, X., Folkman, L. & Borgwardt, K. Kernelized rank learning for personalized drug recommendation. *Bioinformatics* **34**, 2808–2816 (2018).
33. Ammad-Ud-Din, M. *et al.* Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. *Bioinformatics* **32**, i455–i463 (2016).
34. Li, Y., Wu, F. X. & Ngom, A. A review on machine learning principles for multi-view biological data integration. *Brief. Bioinform.* **19**, 325–340 (2018).
35. Paltun, B. G., Mamitsuka, H. & Kaski, S. Improving drug response prediction by integrating multiple data sources : matrix factorization , kernel and network-based approaches. **00**, 1–14 (2019).
36. Schölkopf, B., Smola, A. J. & Müller, K.-R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Comput.* **10**, 1299–1319 (1998).
37. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).
38. Smith, A. M. *et al.* Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. *BMC Bioinformatics* **21**, 1–18 (2020).
39. Aben, N., Vis, D. J., Michaut, M. & Wessels, L. F. A. TANDEM: A two-stage approach to maximize interpretability of drug response models based on multiple molecular data types. *Bioinformatics* **32**, i413–i420 (2016).
40. Aben, N. *et al.* ITOP: Inferring the topology of omics data. *Bioinformatics* **34**, i988–i996 (2018).
41. Hoogstraat, M. *et al.* Genomic and transcriptomic plasticity in treatment-naïve ovarian cancer. *Genome Res.* **24**, 200–211 (2014).

42. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. (2018).

43. Warren, A. *et al.* Global computational alignment of tumor and cell line transcriptional profiles. *bioRxiv* 2020.03.25.008342 (2020). doi:10.1101/2020.03.25.008342

44. Charafe-Jauffret, E. *et al.* Gene expression profiling of breast cell lines identifies potential new basal markers. *Oncogene* 2273–2284 (2006). doi:10.1038/sj.onc.1209254

45. Cau, J. & Hall, A. Cdc42 controls the polarity of the actin and microtubule cytoskeletons through two distinct signal transduction pathways. *J. Cell Sci.* **118**, 2579–2587 (2005).

46. Guo, Y. *et al.* R-ketorolac targets Cdc42 and Rac1 and alters ovarian cancer cell behaviors critical for invasion and metastasis. *Mol. Cancer Ther.* **14**, 2215–2227 (2015).

47. Del Mar Maldonado, M. & Dharmawardhane, S. Targeting rac and Cdc42 GT pases in cancer. *Cancer Res.* **78**, 3101–3111 (2018).

48. Murugesan, S. R. *et al.* Combination of human tumor necrosis factor-alpha (hTNF-α) gene delivery with gemcitabine is effective in models of pancreatic cancer. *Cancer Gene Ther.* **16**, 841–847 (2009).

49. Basaki, Y. *et al.* Akt-dependent nuclear localization of Y-box-binding protein 1 in acquisition of malignant characteristics by human ovarian cancer cells. *Oncogene* **26**, 2736–2746 (2007).

50. Frye, B. C. *et al.* Y-box protein-1 is actively secreted through a non-classical pathway and acts as an extracellular mitogen. *EMBO Rep.* **10**, 783–789 (2009).

51. Housman, G. *et al.* Drug resistance in cancer: An overview. *Cancers (Basel).* **6**, 1769–1792 (2014).

52. Goldstein, L. J. MDR1 gene expression in solid tumours. *Eur. J. Cancer* **32**, 1039–1050 (1996).

53. Vaidyanathan, A. *et al.* ABCB1 (MDR1) induction defines a common resistance mechanism in paclitaxel- and olaparib-resistant ovarian cancer cells. *Br. J. Cancer* **115**, 431–441 (2016).

54. Christie, E. L. *et al.* Multiple ABCB1 transcriptional fusions in drug resistant high-grade serous ovarian and breast cancer. *Nat. Commun.* **10**, 5–14 (2019).

55. Patch, A. M. *et al.* Whole-genome characterization of chemoresistant ovarian cancer. *Nature* **521**, 489–494 (2015).

56. Chen, D. *et al.* Dual PI3K/mTOR inhibitor BEZ235 as a promising therapeutic strategy against paclitaxel-resistant gastric cancer via targeting PI3K/Akt/mTOR pathway article. *Cell Death Dis.* **9**, (2018).

57. Hu, L., Hofmann, J., Lu, Y., Mills, G. B. & Jaffe, R. B. Inhibition of phosphatidylinositol 3′-kinase increases efficacy of paclitaxel in in vitro and in vivo ovarian cancer models. *Cancer Res.* **62**, 1087–1092 (2002).

58. Zhang, L. *et al.* The PI3K subunits, P110α and P110β are potential targets for overcoming P-gp and BCRP-mediated MDR in cancer. *Mol. Cancer* **19**, 1–18 (2020).

59. Gan, Y., Wientjes, M. G. & Au, J. L. S. Expression of basic fibroblast growth factor correlates with resistance to paclitaxel in human patient tumors. *Pharm. Res.* **23**, 1324–1331 (2006).

60. Kim, S. H. *et al.* BGJ398, a pan-FGFR inhibitor, overcomes paclitaxel resistance in urothelial carcinoma with FGFR1 overexpression. *Int. J. Mol. Sci.* **19**, (2018).

61. Rampášek, L. Latent variable models for drug response prediction and genetic testing. *PhD Thesis* (2019). doi:.1037//0033-2909.I26.1.78

62. Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **11**, 333–337 (2014).

63. Sharifi-Noghabi, H., Zolotareva, O., Collins, C. C. & Ester, M. MOLI: Multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics* **35**, i501–i509 (2019).

64. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, (2010).

65. Dillies, M. A. *et al.* A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.* **14**, 671–683 (2013).

66. Zwiener, I., Frisch, B. & Binder, H. Transforming RNA-Seq data to improve the performance of prognostic gene signatures. *PLoS One* **9**, 1–13 (2014).

67. Weinstein, J. N. *et al.* The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).

68. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: Summarize

analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).

69. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
70. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2009).
71. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
72. Liao, Y., Smyth, G. K. & Shi, W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
73. Aronszajn, N. Theory of Reproducing Kernels. *Trans. Am. Math. Soc.* **68**, 337 (1950).
74. Golub, G. H. & Van Loan, C. F. *Matrix Computations*. (2013).
75. Gopalan, R., Li, R. & Chellappa, R. by Generating Intermediate Data Representations. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**, 2288–2302 (2014).
76. Ben-David, S. *et al.* A theory of learning from different domains. *Mach. Learn.* **79**, 151–175 (2010).
77. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net Hui. *J. Stat. Soc. Ser. B* **67**, 301–320 (2005).
78. Steinwart, I., Hush, D. & Scovel, C. An Explicit Description of the Reproducing Kernel Hilbert Spaces of Gaussian RBF Kernels. *IEEE Trans. Inf. Theory* **52**, 4635–4643 (2006).
79. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).
80. Van Der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy array: A structure for efficient numerical computation. *Comput. Sci. Eng.* **13**, 22–30 (2011).
81. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

82. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 99–104 (2007).

83. Varoquaux, G. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **19**, 29–33 (2015).

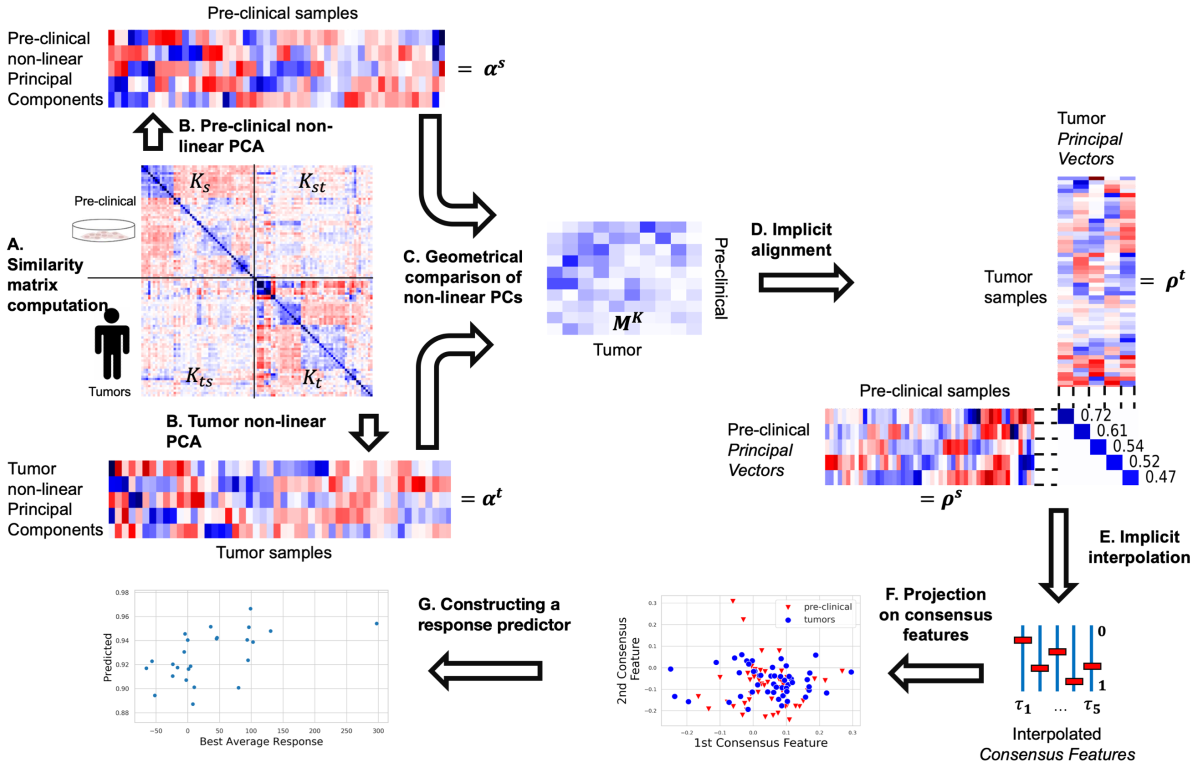


Figure 1 PRECISE+: Generating non-linear subspace representations to transfer predictors of response from pre-clinical models to human tumors. (**A**) Samples are compared using a **similarity function** – also referred to as kernel. This yields a similarity matrix containing similarities between pre-clinical (source) models (K_s), between tumors (target samples) (K_t) and between pre-clinical models and tumors (K_{st}). (**B**) Using non-linear PCA, the pre-clinical and tumor similarity matrices are independently decomposed to compute directions of maximal variance induced by the similarity function – these are referred to as **non-linear principal components** (NLPCs). Each NLPC corresponds to a direction in a very high-dimensional space induced by the similarity function, that is often computationally intractable. Thus, instead of a feature space representation, these NLPCs are geometrically represented by "sample importance scores" (**Supp Figure 1A**) that represent the importance of each sample in each NLPC. These scores are aggregated in the matrices α^s and α^t, for source and target space, respectively. (**C**) These pre-clinical and tumor NLPCs are then geometrically compared in a **non-linear cosine similarity matrix** M^K. Each pre-clinical NLPC (y-axis) is geometrically compared to each tumor NLPC (x-axis) but no clear 1-1 correspondence appears as shown by the large number of off-diagonal elements. (**D**) To find directions of significance for both pre-clinical models and tumors, we align these non-linear principal components using the notion of **principal vectors** (**Supp Figure 1B**). Principal vectors are pairs of vectors (one from pre-clinical NLPCs, one from tumor NLPCs) that are maximally geometrically similar. (**E**) Within each pair of vectors, we perform an interpolation to select one non-linear vector that balances the effect of pre-clinical and tumor signals (**Supp Figure**

1C).This yields a few robust **consensus features** that correspond to non-linear gene combinations of significance for both tumors and pre-clinical models. (**F**) We project each tumor and pre-clinical sample on these consensus features to obtain **consensus scores**. These scores correspond to the activity of processes conserved between tumors and pre-clinical models. (**G**) Finally, these scores can be used as input to any predictive model, for instance to predict drug response based on these consensus scores.

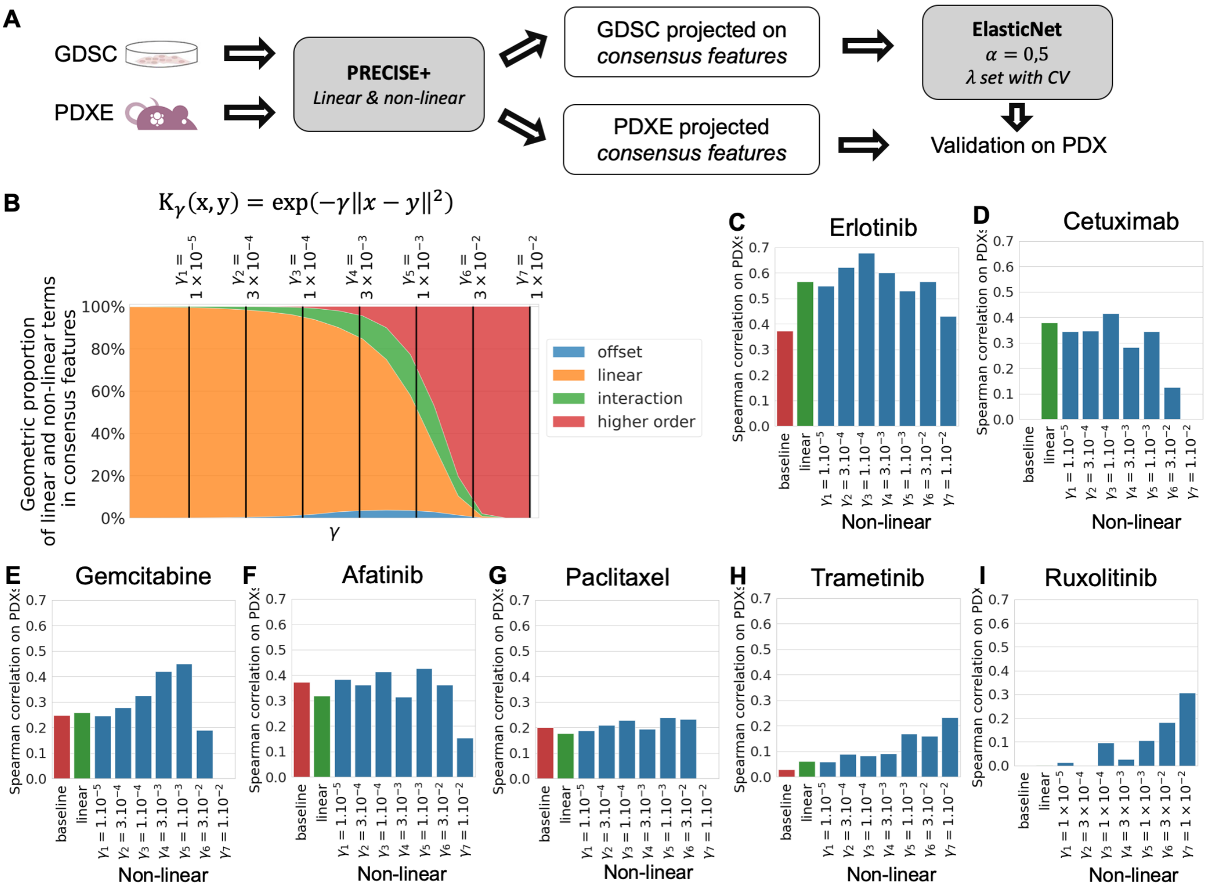


Figure 2 Impact of modelling non-linearities for drug response prediction transfer from cell lines to PDXs. (**A**) Main workflow of the prediction on PDXs. Using cell lines and PDX gene expression, we compute consensus features and project each dataset onto these. We then trained a regression model (Elastic Net) on cell lines projected scores to predict AUC. Finally, we used this regression model to predict drug response in PDXs and correlate the predicted AUC to the known best average response. (**B**) Proportion of non-linearities induced by the Gaussian similarity function as a function of the scaling factor γ. For different values of γ, we compute the average contribution over all consensus features of offset, linear, interaction and higher order features. Offset is here to be understood as the exponential of the squared depth and does not correspond to a constant term. We finally evaluated the PDX prediction for different values of γ, for a linear similarity function, and a non-domain-adapted baseline. We report results for Erlotinib (**C**), Cetuximab (**D**), Gemcitabine (**E**), Afatinib (**F**), Paclitaxel (**G**), Trametinib (**H**) and Ruxolitinib (**I**). Concordance on PDX is measured as the Spearman correlation between predicted AUC and Best Average Response.

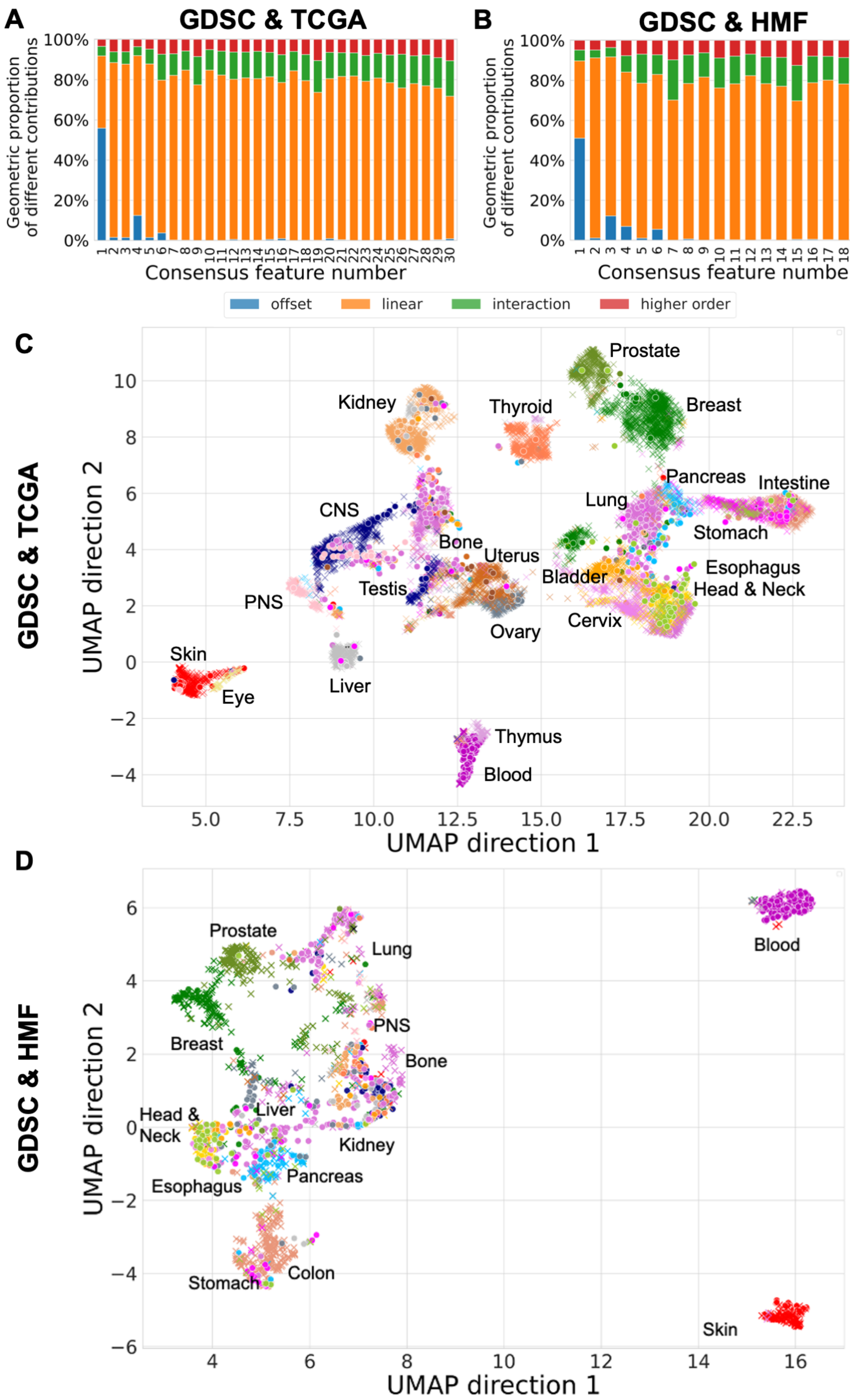

A
GDSC & TCGA
Geometric proportion of different contributions
100%
80%
60%
40%
20%
0%
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Consensus feature number
B
GDSC & HMF
Geometric proportion of different contributions
100%
80%
60%
40%
20%
0%
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Consensus feature number
offset
linear
interaction
higher order
C
GDSC & TCGA
UMAP direction 2
UMAP direction 1
10
8
6
4
2
0
−2
−4
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
Prostate
Kidney
Thyroid
Breast
Pancreas
Intestine
Lung
CNS
Stomach
Bone
Uterus
Esophagus
Bladder
Head & Neck
Testis
PNS
Cervix
Ovary
Skin
Eye
Liver
Thymus
Blood
D
GDSC & HMF
UMAP direction 2
UMAP direction 1
6
4
2
0
−2
−4
−6
4
6
8
10
12
14
16
Prostate
Lung
Blood
PNS
Breast
Bone
Liver
Head & Neck
Kidney
Pancreas
Esophagus
Stomach
Colon
Skin

Figure 3 Pan-cancer consensus features between cell lines and tumors conserve tissue type information. We used cell lines (GDSC) as source data and computed two sets of consensus features with two different target datasets: primary tumors (TCGA, A and C) and metastatic lesions (HMF, B and D). (**A**) Proportion of linear and non-linear contributions to each of the 30 GDSC-to-TCGA consensus features. (**B**) Proportion of linear and non-linear contributions to each of the 20 GDSC-to-HMF consensus features. (**C**) UMAP plot of primary tumors (TCGA, 21 tissues) and cell lines (GDSC, 22 tissues) projected on the consensus features, using the same parameters as selected in **Figure *2***. (**D**) UMAP plot of metastatic lesions (HMF) and cell lines, colored by primary tissue for both HMF and GDSC. For both UMAP plot, the full legend can be found in **Supp Figure *10***A.

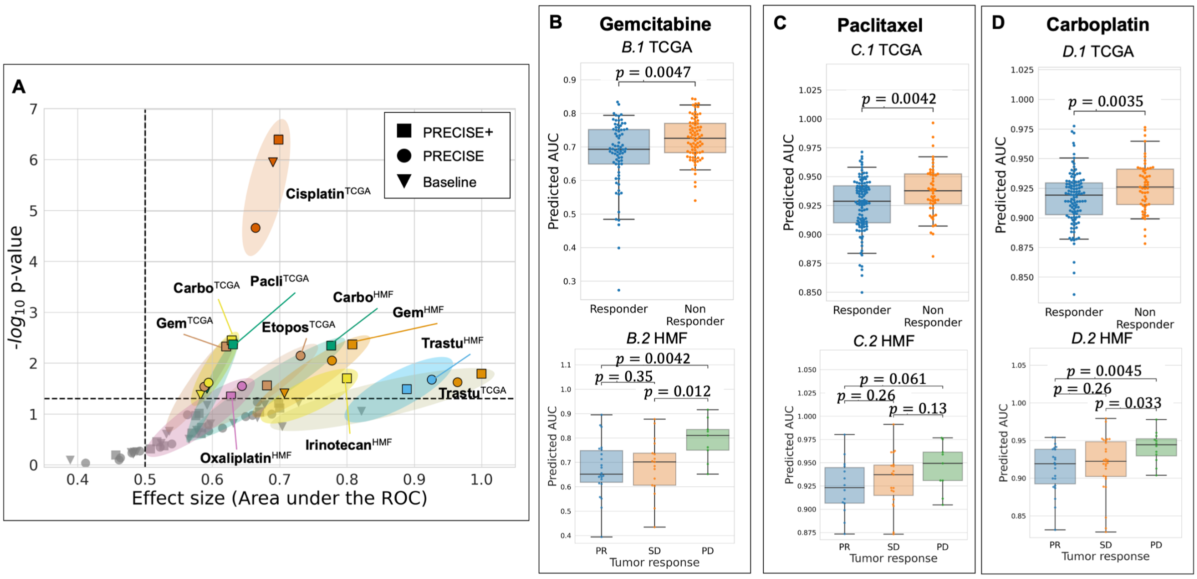


Figure 4 Consensus features improve response prediction in patients. (**A**) We used consensus features computed between GDSC and each tumor dataset to train a predictor for 17 drugs on TCGA and 5 drugs on HMG, using GDSC response only. We then predicted the AUC for each drug and compared this predicted value to the observed clinical response using a one-sided Mann-Whitney test (**Table 1**). We summarized here the associations for the 3×22 predictors in a Volcano plot. (**B**) Results for the two Gemcitabine predictors. (**B**.1) For the Gemcitabine predictor on TCGA, we compared predicted AUC for each patient to the known clinical response (top) and compared this association to the results obtained using a non-domain-adapted baseline and PRECISE (bottom). The black line indicates p = 0.05. (**B**.2) Results for Gemcitabine predictor on HMF. (**C**.1) Results for Paclitaxel predictor on TCGA. (**C**.2.) Results for Paclitaxel predictor on HMF. (**D**.1) Results for Carboplatin predictor on TCGA. (**D**.2.) Results for Carboplatin predictor on HMF.

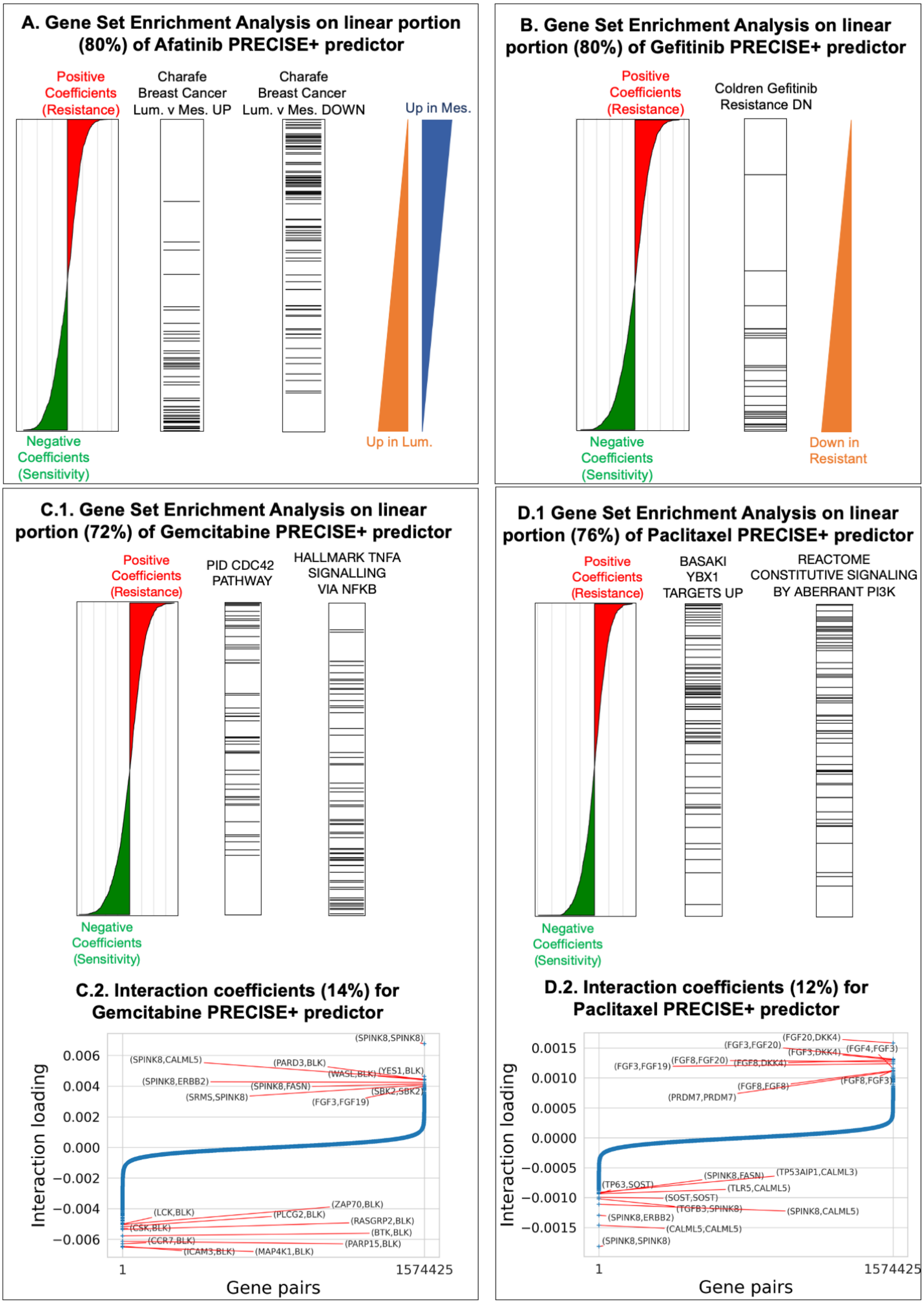
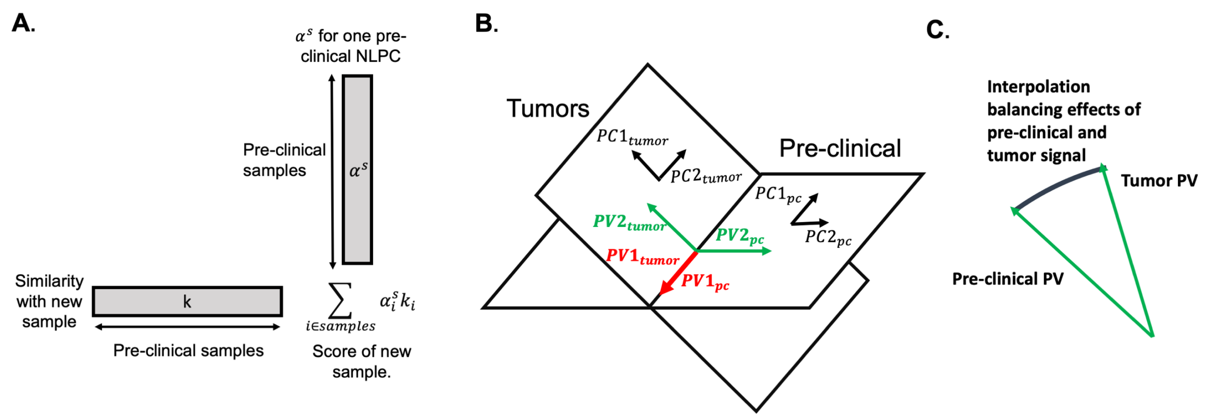


Figure 5 Interpretability of PRECISE+ consensus features highlight mechanisms of sensitivity and resistance to Gemcitabine and Paclitaxel. (**A**) For the Afatinib (HER-2 protein kinase

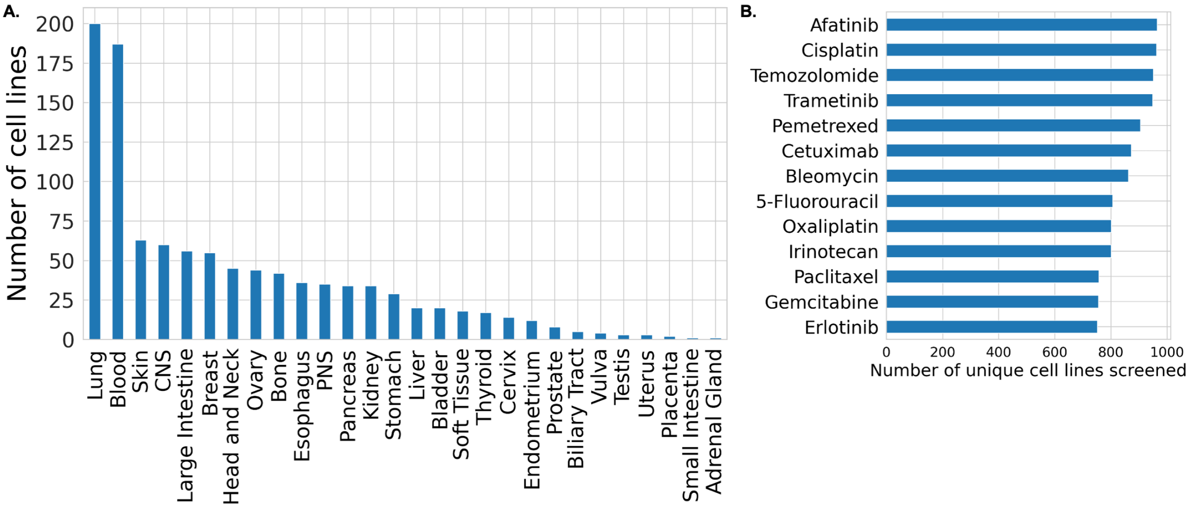
inhibitor) response model trained on GDSC samples projected on GDSC-to-TCGA consensus features we show the proportions of contributions from various terms (left) and the PreRanked gene set enrichment analysis (GSEA) result for genes contributing geometrically to the linear portion of the predictor. Positive (negative) weights in the predictor indicate that high (low) expression of the genes leads to resistance (sensitivity) represented by larger (smaller) AUCs. (**B**) We repeated the same experiment on a Gefitinib (EGFR inhibitor) model and show the significant enrichment for genes known to be downregulated in Gefitinib resistant cell lines. (**C)** Gemcitabine analysis showed enrichment of CDC42 pathway (FDR < 0.001) in resistant coefficients and TNFα-signalling via NF-κB signaling in sensitive coefficients. (**D**) Paclitaxel analysis. Enrichment for genes linked to silencing of YBX1 by shRNA, and genes linked to aberrant PI3K-behavior in resistant coefficients.

A. GDSC-to-TCGA						
GDSC		**TCGA**		**p-val [effect-size] on TCGA**		
Drug	**Samples**	**Drug**	**Samples**	**Baseline**	**PRECISE**	**PRECISE+**
Afatinib	800	Trastuzumab	16	0.089 [0.82]	**0.024** **[0.96]**	**0.016** **[1.00]**
Bleomycin	856	Bleomycin	53	0.153 [0.63]	0.106 [0.67]	0.091 [0.78]
Cetuximab	868	Cetuximab	19	0.118 [0.67]	0.451 [0.52]	0.0765 [0.7]
Cisplatin	764	Cisplatin	308	**1.11E-06** **[0.69]**	**2.15E-05** **[0.66]**	**3.92E-07** **[0.70]**
Cisplatin	764	Carboplatin	166	**0.0419** **[0.58]**	**0.0241** **[0.59]**	**0.0035** **[0.63]**
Cyclophosphamide	747	Cyclophosphamide	102	0.571 [0.48]	0.112 [0.65]	0.642 [0.46]
Docetaxel	665	Docetaxel	102	0.616 [0.48]	0.728 [0.46]	0.107 [0.57]
Doxorubicin	871	Doxorubicin	101	0.0674 [0.59]	0.925 [0.41]	0.0969 [0.58]
Etoposide	880	Etoposide	84	0.209 [0.58]	**0.00708** **[0.73]**	**0.0273** **[0.68]**
5-Fluorouracil	801	Fluorouracil	186	0.219 [0.53]	0.800 [0.46]	0.352 [0.52]
Gemcitabine	752	Gemcitabine	156	0.248 [0.53]	**0.0289** **[0.59]**	**0.00466** **[0.62]**
Irinotecan	796	Irinotecan	25	0.810 [0.39]	0.536 [0.49]	0.488 [0.51]
Oxaliplatin	724	Oxaliplatin	66	0.379 [0.52]	**0.0280** **[0.64]**	**0.0450** **[0.63]**
Paclitaxel	753	Paclitaxel	160	**0.0318** **[0.59]**	0.103 [0.56]	**0.00424** **[0.63]**
Pemetrexed	898	Pemetrexed	38	0.157 [0.59]	0.164 [0.59]	0.2370 [0.57]
Temozolomide	746	Temozolomide	96	0.555 [0.49]	0.587 [0.48]	0.185 [0.58]
Vinorelbine	746	Vinorelbine	30	0.312 [0.56]	0.189 [0.61]	0.422 [0.53]
B. GDSC-to-HMF						
GDSC		**HMF**		**p-val [effect-size] on HMF**		
Drug	**Samples**	**Drug**	**Samples**	**Baseline**	**PRECISE**	**PRECISE+**
Afatinib	800	Trastuzumab	25	0.1776 [0.70]	**0.02098** **[0.93]**	**0.03223** **[0.89]**
Irinotecan	796	Irinotecan	67	0.0596 [0.73]	0.1003 [0.69]	**0.01975** **[0.8]**
Cisplatin	764	Carboplatin	64	0.23 [0.58]	0.05848 [0.59]	**0.004491** **[0.78]**
5-Fluorouracil	801	Fluorouracil	65	0.06548 [0.68]	0.2144 [0.59]	0.2435 [0.58]
Paclitaxel	753	Paclitaxel	45	0.5626 [0.48]	0.3884 [0.53]	0.06137 [0.70]
Gemcitabine	752	Gemcitabine	50	0.03899 [0.71]	**0.008851** **[0.78]**	**0.004233** **[0.81]**

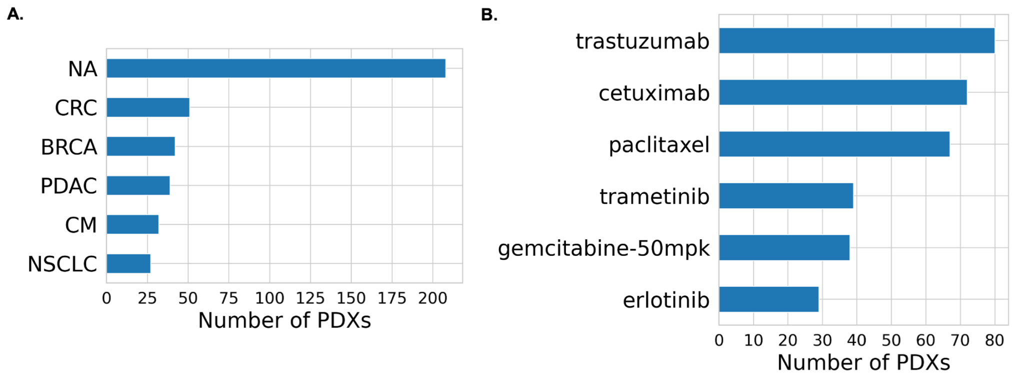
Table 1 Results of PRECISE+ compared to PRECISE (linear) and baseline (ElasticNet without domain adaptation) for 17 drugs on TCGA and 5 drugs on HMF. For each drug, we divide the patients in two categories: *Responders* and *Non Responders*. For TCGA, *Responders* correspond to Partial Responders (PR) and Complete Responders (CR) – for HMF, *Responders* correspond to PR only. For TCGA, *Non Responders* correspond to Stable Disease (SD) and Progressive Disease (PD) – for HMF to PD only. For each drug, we train 3 predictors – baseline, PRECISE and PRECISE+ – and compare in each scenario the predicted AUC to the known clinical response using one-sided Mann-Whitney test. For each predictor, we report the p-value and the effect size (Area under the ROC, effect size associated to Mann-Whitney test) under bracket. Bold cells correspond to significant association (pval < 0.05). Red cells correspond to significant associations with the largest effect size across the 3 methods.



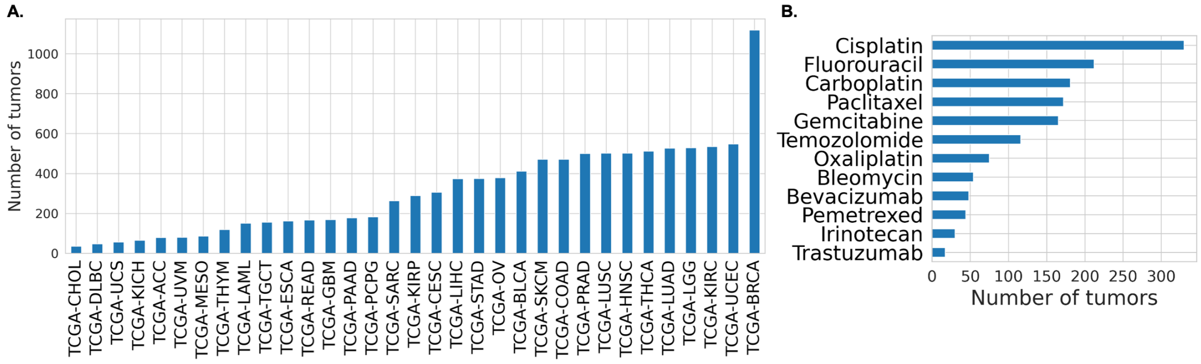
Supp Figure 1 Visual explanation of geometric alignment. **A**: Difference between importance scores (α^s, α^t) and projected scores. Since the space induced by the similarity function K is intractable, we use a dual representation of the NLPC in terms of samples: the importance scores. To project samples on NLPCs, one needs to compute the similarity between this sample and all of the samples used to gauge the NLPC. The projected score is obtained by taking the vector-product between this similarity vector and the importance scores. The same rational holds principal vectors that are represented by $\gamma^s$and γ^t. **B**: Visual example of principal vectors (PV). We here consider 3 genes (features) and 2 NLPCs. The pre-clinical (source) and tumor (target) NLPCs intersect in one direction, which form the pair of closest vectors: the first PV forms the pair of the two red vectors – although these are identical. The second pair of PVs is defined orthogonally to the red pair. This defines the green vectors (with a swap in direction for visual purposes). These pairs reconstruct the original NLPC spaces and are ordered by similarity. **C**: Interpolation between PVs. For one pair of PVs – e.g. the green one in B – source and target vectors are different. In order to generate one robust vector out of these two and avoid redundancy, we draw an arc between these two vectors. We then project source and target datasets onto these interpolated vectors and select one intermediate representation where source and target projected signals are maximally matched. This optimal intermediate vector is called the consensus feature.



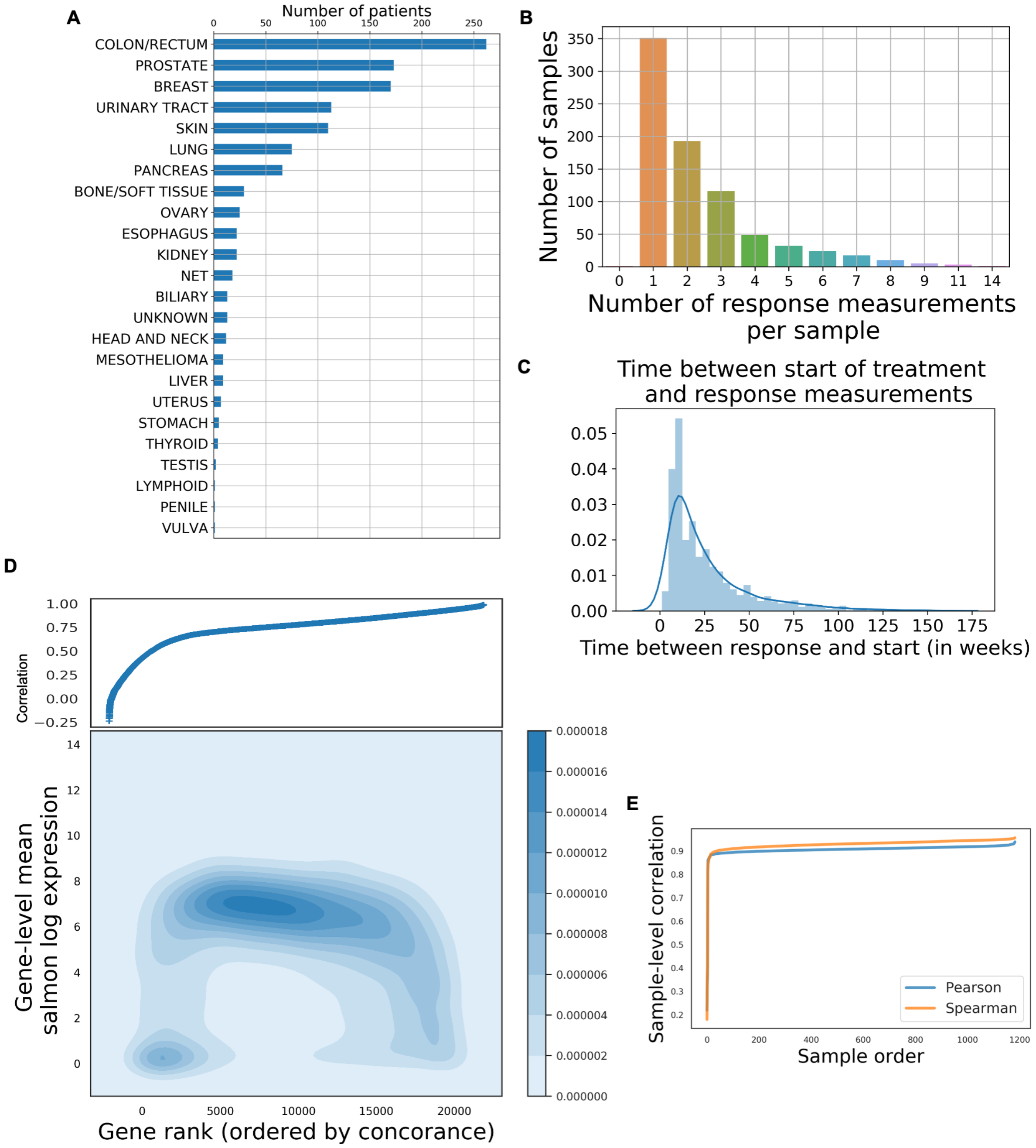
Supp Figure 2 Composition of the GDSC dataset (cell lines). We make use of the GDSC1000 cell line panel[14]. **A**: Number of cell lines per tissue type. **B**: Number of cell lines screened for each drug that we used in our experiments.



Supp Figure 3 Composition of the NIBR PDXE dataset (patient derived xenografts). We make use of the NIBR PDXE patient derived xenograft panel[15]. **A**: Number of PDXs per tissue type. **B**: Number of unique PDXs screened for each drug that we used in our experiments.

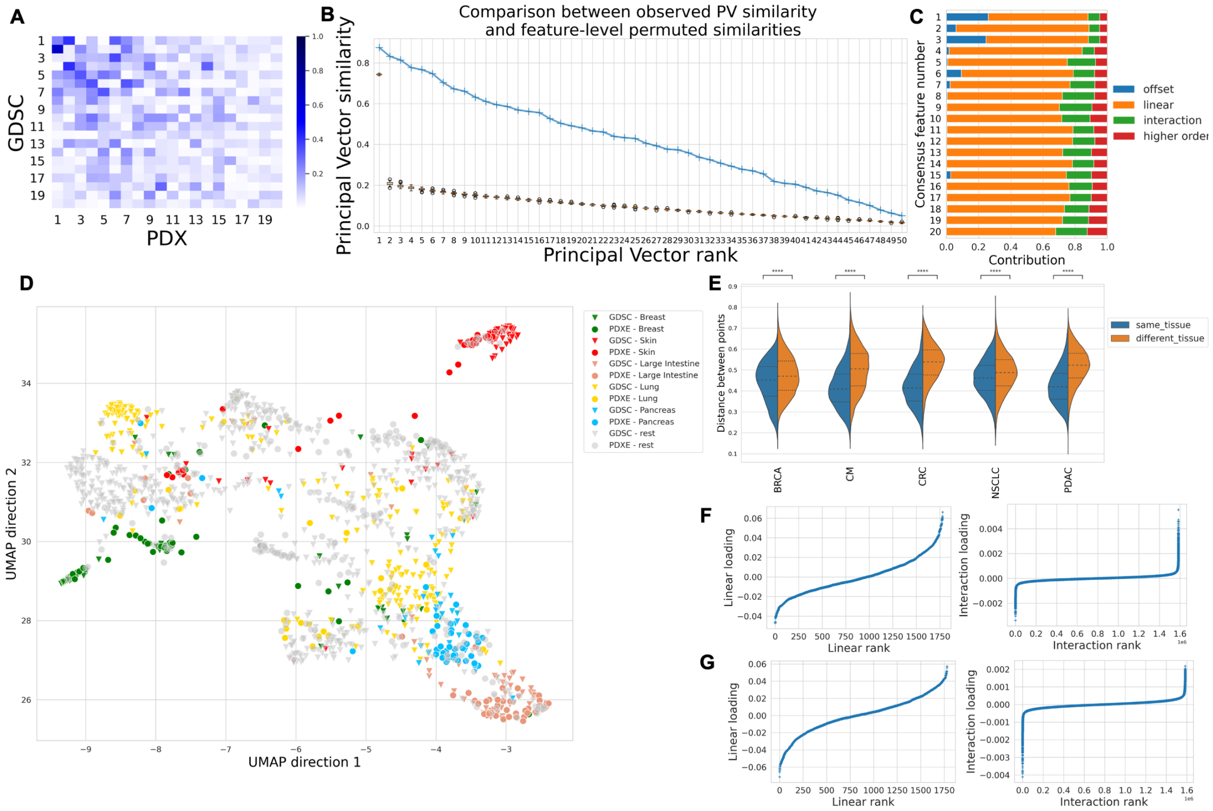


Supp Figure 4 Structure of the TCGA dataset (primary tumors). We make use of the TCGA dataset for primary tumors. **A**: Number of samples per cancer type. **B**: For each drug, number of samples with known response.

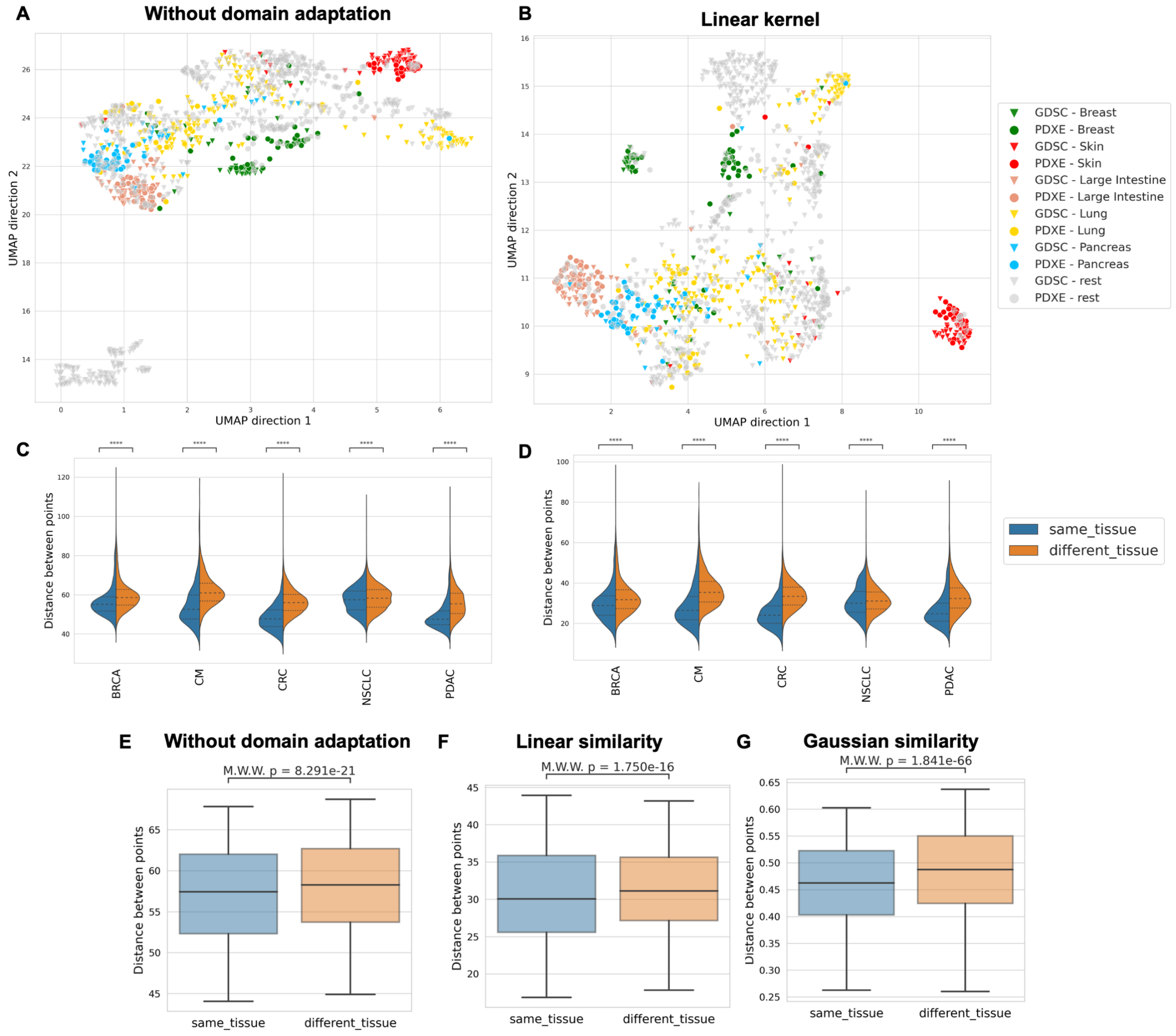


Supp Figure 5 Structure of the HMF dataset (metastatic lesions). We make use of the Hartwig Medical Foundation (HMF) dataset for metastatic lesions. (**A**) Number of samples per cancer type (primary tumor location). (**B**) For each patient, number of response measurements made. For further analysis, we considered the first response measure – i.e. first measure after treatment start. (**C**) Histogram of number of weeks between treatment start and response measurement. (**D**) For each protein coding gene, we measure the Spearman correlation between read counts obtained using Salmon and STAR alignment tools using all samples in the HMF dataset. We then ranked genes based on the obtained Spearman correlation (x-axis) and plotted it against the mean-expression of these genes obtained using Salmon (y-axis).

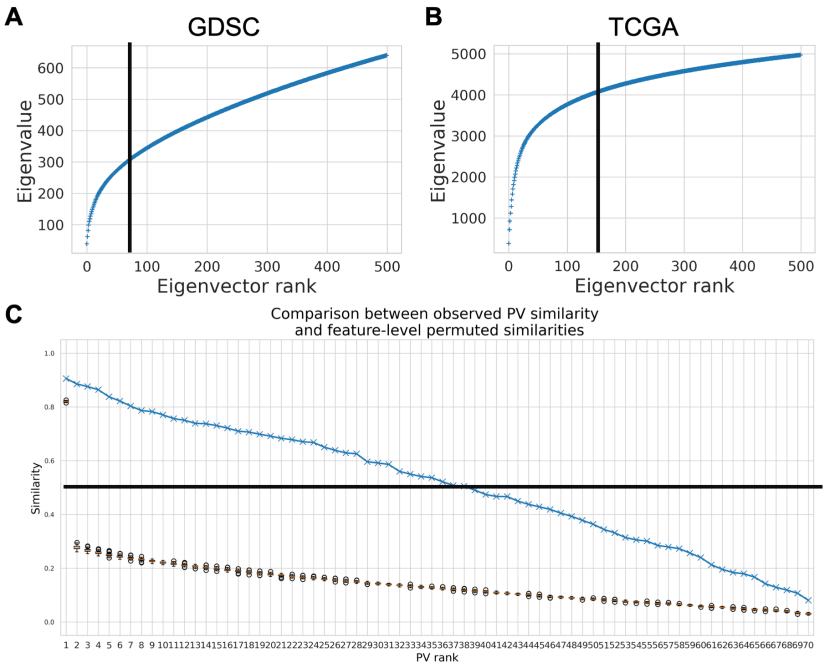
Since lowly concordant genes tend to have low expression, we put a threshold at $corr = 0.5$ and discarded genes below this threshold. (**E**) After the previous selection, we computed the sample-level Pearson and Spearman correlations between read counts obtained with STAR and Salmon. All samples but 5 shows a correlation above 0.8 – these were discarded. We finally further restricted to genes from the mini-cancer genome.



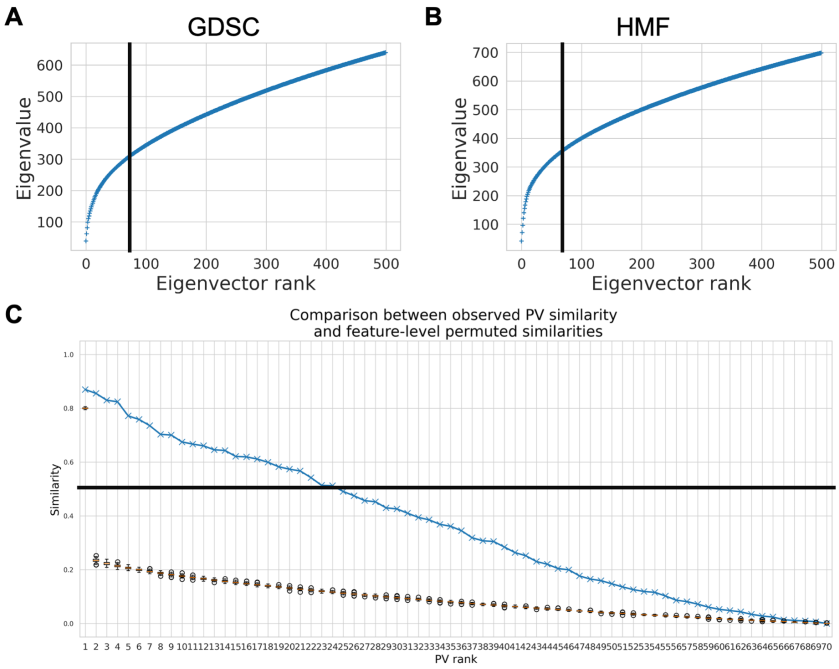
Supp Figure 6 Analysis of consensus features between cell lines (GDSC) and PDXs with $\gamma = 0.0005$.



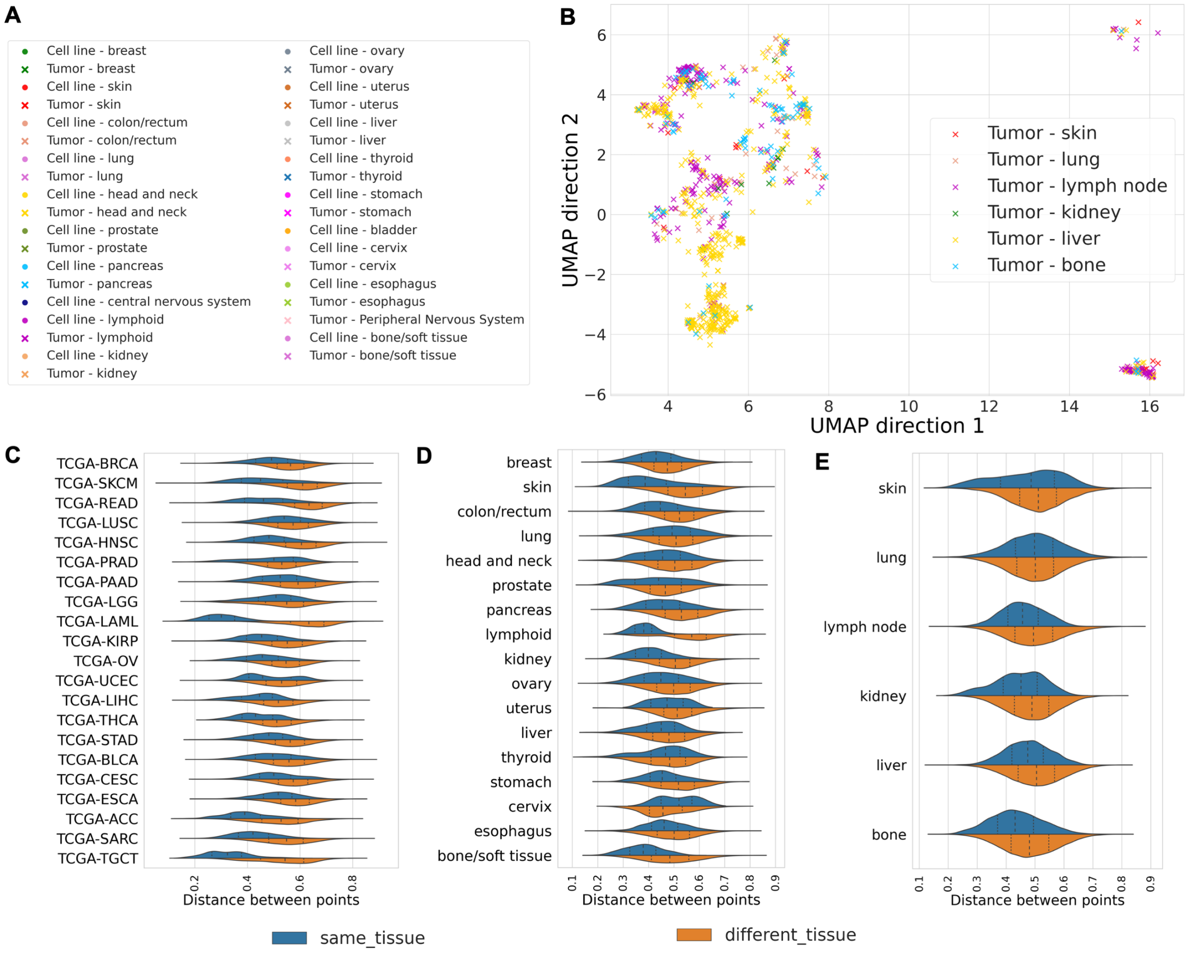
Supp Figure 7 Tissue clustering analysis for baseline and linear similarity.



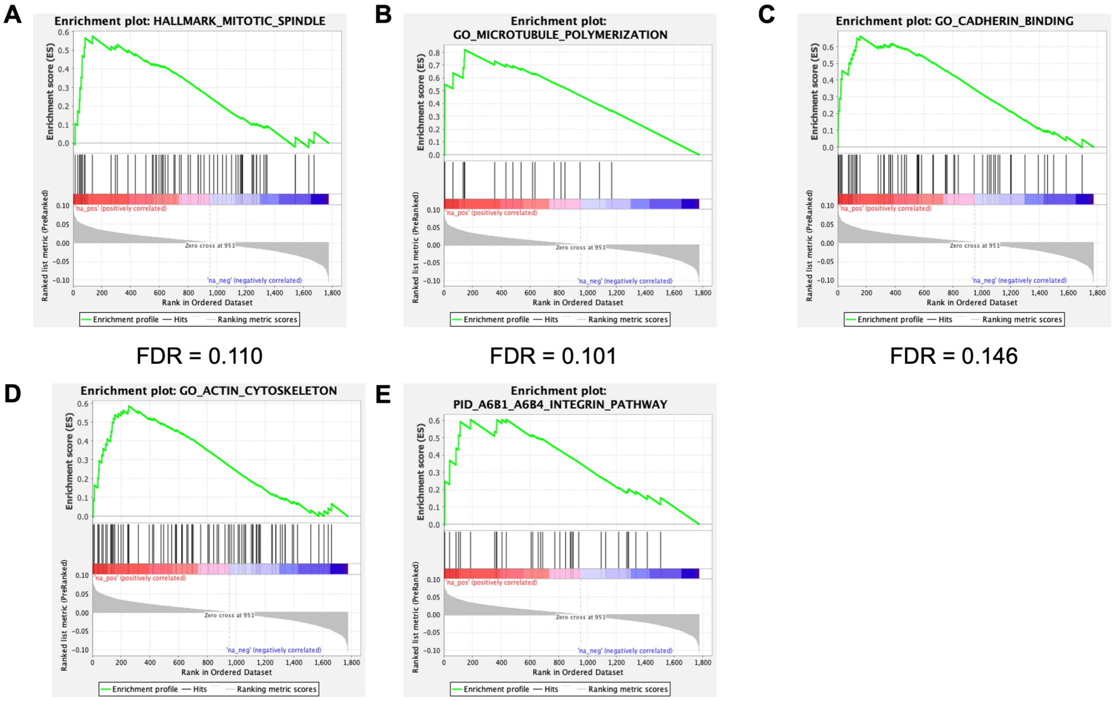
Supp Figure 8 Choice of the number of NLPCs and consensus features between GDSC and TCGA. (**A**) Cumulative sum of eigenvalues of $\widetilde{K_s}$ (GDSC) with $\gamma^* = 5 \times 10^{-4}$. The cumulative sum increases steeply, reaches an inflexion points and then follows an almost-linear behavior. We select all the NLPCs corresponding before this almost-linear zone, corresponding to 75 NLPCs. (**B**) Cumulative sum of eigenvalues of $\widetilde{K_t}$ (TCGA) with $\gamma^* = 5 \times 10^{-4}$. Following a similar thinking as in (**A**), we restrict the study to the first 150 NLPCs. (**C**) Similarity between PV when 75 NLPCs are considered for GDSC and 150 for TCGA. We observe that the 33 first PVs have a similarity above 0.5 (our cut-off) and round the selection to 30 PVs.



Supp Figure 9 Choice of the number of NLPCs and consensus features between GDSC and HMF. (**A**) Cumulative sum of eigenvalues of $\widetilde{K_s}$ (GDSC) with $\gamma^* = 5 \times 10^{-4}$. The cumulative sum increases steeply, reaches an inflexion points and then follows an almost-linear behavior. We select all the NLPCs corresponding before this almost-linear zone, corresponding to 75 NLPCs. (**B**) Cumulative sum of eigenvalues of $\widetilde{K_t}$ (HMF) with $\gamma^* = 5 \times 10^{-4}$. Following a similar thinking as in (**A**), we restrict the study to the first 75 NLPCs. (**C**) Similarity between PV when 75 NLPCs are considered for both GDSC and HMF. We observe that the 21 first PVs have a similarity above 0.5 (our cut-off) and round the selection to 20 PVs.



Supp Figure 10 Pan-cancer consensus features between cell lines and tumors conserve tissue type information (Supplement of Figure 3) **A**: Legend of UMAP plots for **Figure3**C-D. **B**: UMAP plot of HMF metastatic lesions (same as Figure 3D) colored by metastatic site. **C**: In TCGA, for each tumor type, distance between tumors and cell lines from similar (blue) and non-similar (orange) tissue. **D**: In HMF, for each primary tumor type, distance between metastatic sample and cell line from similar and non-similar tissue of origin. **E**: In HMF, for each metastatic site, distance between metastatic sample and cell line from tissue of origin similar (blue) or dissimilar from the metastatic site.



FDR = 0.110

FDR = 0.101

FDR = 0.146

FDR = 0.065

FDR = 0.134

Supp Figure 11 Pathway enriched for resistant linear coefficients in GDSC-to-TCGA Gemcitabine drug response predictor.