Predicting clinical drug response from model systems by non-linear subspace-based transfer learning

Soufiane Mourragui[1,2], Marco Loog[2,3], Daniel J. Vis[1], Kat Moore[1], Anna G. Manjon[4], Mark A. van de Wiel[5,6], Marcel J.T. Reinders[2,7,8], Lodewyk F.A. Wessels[1,2,8]

[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]MRC Biostatistics Unit, Cambridge University, Cambridge, United Kingdom.

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

[8]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. While powerful, these models do not fully recapitulate the complexity of human tumors. Consequently, translating biomarkers of drug response from pre-clinical models to human tumors has been particularly challenging. To explicitly take these differences into account and enable an efficient exploitation of the vast pre-clinical drug response resources, we developed TRANSACT, a novel computational framework for clinical drug response prediction. First, TRANSACT employs non-linear manifold learning to capture biological processes active in pre-clinical models and human tumors. Then, TRANSACT builds predictors on cell line response only and transfers these to Patient-Derived Xenografts (PDXs) and human tumors. TRANSACT outperforms four competing approaches, including Deep Learning approaches, for a set of 15 drugs on PDXs, TCGA cohorts and 226 metastatic tumors from the Hartwig Medical Foundation data. For only four drugs Deep Learning outperforms TRANSACT. We further derived an algorithmic approach to interpret TRANSACT and used it to validate the approach by identifying known biomarkers to targeted therapies and we 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 that correct for these differences are therefore needed to improve the identification of truly predictive biomarkers[12]. In the particular case of cancer, these approaches are divided into two distinct categories. In the 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]. The 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], cell line drug response predictors are inferred[16–18]. The resulting models are then applied to predict the sensitivity of patients to certain drugs[19–21]. Although already

promising, these approaches either do not take into account the fundamental differences between pre-clinical models and human tumors[22], or only model these differences as a technical batch effect[19–21]. Recently, transfer learning and multi-task learning approaches have been developed to explicitly take these differences into account, either partially using tumor responses during training[23], or solely based on pre-clinical labels[24] while only modeling differences linearly.

We present **TRANSACT** (**T**umor **R**esponse **A**ssessment by **N**on-linear **S**ubspace **A**lignment of **C**ell-lines and **T**umors), 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). TRANSACT employs the powerful mathematical framework of Kernel methods[30–35] to capture both linear and non-linear molecular processes expressed in source and target domains. First, we demonstrate that, compared to existing methods[20,21,24], modeling non-linearities using TRANSACT improves drug response prediction in PDXs after training on cell line responses only. We fix the hyperparameter controlling the degree of non-linearity on the PDX data and then employ TRANSACT to transfer predictors of drug response trained on cell lines to two human tumor 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 in TCGA and three drugs in HMF, compared to state-of-the art competing approaches. 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

TRANSACT: Generating non-linear manifold representations to transfer predictors of response from pre-clinical models to tumors

TRANSACT compares genomic signals contained in the source (e.g. pre-clinical models) and target (e.g. human tumors) datasets, and outputs a representation of processes that are present in both datasets. The nature of this representation 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 modeled. For instance, in the case of a Gaussian similarity function, these non-linearities include constant, linear, second and higher-order interaction terms (Methods).

In a first step, TRANSACT computes the similarities between all samples (**Figure 1**A), yielding three matrices: K_s, K_t and K_{st}, containing the similarities between source samples, target samples, 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 the sample similarity space (**Supp**

117 **Figure 1**A). We then quantify the geometric differences between all pairs of NLPCs
118 from the source and target sets by computing the cosine similarity matrix (**Figure 1**C).
119 In a subsequent step (**Figure 1**D), we align these two sets of NLPCs by using the
120 notion of *principal vector*s illustrated in **Supp Figure 1**B. These principal vectors (PV)
121 are pairs of vectors – one from the source, one from the target – ranked by decreasing
122 similarity: the top PVs correspond to geometrically similar factors, while bottom PVs
123 are pairs of almost orthogonal vectors. We restrict further analysis to the most similar
124 PVs based on a similarity of at least 0.5 (Supp. Material). We perform, for each
125 selected pair of PVs, an interpolation between the source and the target vectors and
126 select one intermediate representation that best balances the contribution of the
127 source and the target signals (**Figure 1**E, **Supp Figure 1**C). These vectors, called
128 *consensus features*, define the consensus space, into which we project the source
129 (pre-clinical) and target (tumor) samples. This projection yields, for all source and
130 target samples, consensus feature values that can be used as input to any machine
131 learning model to build predictors of drug response (**Figure 1**G, Method). In the case
132 of a linear similarity function, TRANSACT reduces to PRECISE[24] (Subsection Supp.
133 8) and is fundamentally different from approaches such as Canonical Correlation
134 Analysis (CCA)[37] (Subsection Supp. 9).
135

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

138 When it comes to predicting drug response in one model system, it is known that
139 inducing non-linearities can lead to improved performance[35], although linear methods
140 remain competitive[18,38]. We investigated whether the introduction of non-linearities in
141 the computation of sample similarities resulted in improved response prediction of

142 predictors trained on cell lines (source domain) and transferred to PDXs (target
143 domain). Since gene expression is known to have predictive power comparable to
144 other omics datasets combined[16,18,39,40], we restricted our analysis to the expression
145 of 1780 genes known to be related to cancer[41]. Using TRANSACT, we computed
146 consensus features for cell lines (1049 cell lines from 26 different tissues) and all PDXs
147 (399 samples from 5+ different tissues) (Methods). We projected the gene expression
148 data of all cell lines and all PDXs onto these consensus features. We employed Elastic
149 Net to train models of drug response with the projected cell line expression data as
150 input and the drug response data as output, which is quantified as the AUC, i.e. the
151 area under the drug response curve (Methods). We applied this trained predictor on
152 the projected PDX expression data and compared the predicted response to the
153 measured best average response by Spearman Correlation (**Figure 2**A). We made
154 use of the standard Gaussian similarity function (Methods) to vary the level of non-
155 linearity introduced. This similarity function is characterized by a single scaling factor
156 γ, whose size is directly proportional to the proportion of non-linearity introduced
157 (**Figure 2**B). We studied the predictive performance in PDXs for seven different values
158 of γ, ranging from a set of consensus features with an almost purely linear ($\gamma =$
159 1×10^{-5}) to an almost purely non-linear composition ($\gamma = 1 \times 10^{-2}$). We compared
160 the performance of TRANSACT to two approaches that do not perform domain
161 adaptation: *Elastic Net*[22] and *Deep Learning* regression (DL). We further compared it
162 to two state-of-the-art domain adaptation approaches: ComBat batch correction
163 followed by Deep Learning regression (*ComBat + DL*)[21] and *PRECISE*[24], a linear
164 domain adaptation approach. All models were trained to predict response to seven
165 different drugs (Erlotinib, Cetuximab, Gemcitabine, Afatinib, Paclitaxel, Trametinib and
166 Ruxolitinib) for which we had response data available for both PDX models and cell

167 lines (**Figure 2**C-I). For these seven drugs we observe, in general, a clear
168 improvement of approaches that employ domain-adaptation over those that do not,
169 indicating a necessity to correct the input signal when moving from the source to the
170 target domain. When comparing *Elastic Net* to *Deep Learning*, we observe that
171 modeling non-linearities results in a better transfer for five drugs. We further observe
172 an improvement of TRANSACT over *ComBat+DL*, suggesting that the correction
173 required from cell lines to PDXs is more complicated than correcting for a technical
174 batch effect. We observe that the introduction of non-linearity with TRANSACT
175 generally increases the predictive performance on PDXs when compared to a linear
176 similarity function (*PRECISE*). Specifically, we observe for several drugs that the
177 predictive performance increases with the scaling factor until a maximal performance
178 is reached ($\gamma = 10^{-4}$ for Erlotinib, Cetuximab and Afatinib and $\gamma = 10^{-3}$ for
179 Gemcitabine, Paclitaxel and Trametinib), after which the predictive performance
180 drops. We therefore decided to fix the scaling factor to the average of these two values
181 ($\gamma^* = 5 \times 10^{-4}$) and employ the associated consensus space to transfer the predictors
182 of response to the tumor samples. As a further check, we analyzed the properties of
183 the consensus space obtained using γ^*. We observe a concentration of the offset
184 contribution in the top consensus features and an increasing proportion of non-linear
185 term contribution to lower order features (**Supp Figure 6**C). The UMAP[42] projection
186 of the consensus features shows a clear co-clustering of cell lines and PDXs of the
187 same tissue (**Supp Figure 6**D).
188

189 *Consensus features between cell lines (GDSC) and human tumors*

190 *conserve primary tumor information.*

191 With the scaling factor (γ) calibrated on PDX models, we moved to the clinical setting
192 to investigate domain adaptation between cell lines and two different human tumor
193 datasets: primary tumors from TCGA and metastatic lesions from the HMF. We
194 selected 30 consensus features in the GDSC-TCGA analysis (**Supp Figure 8**) and 20
195 in the GDSC-HMF analysis (**Supp Figure 9**). We arrived at these numbers by first
196 selecting NLPCs based on the inflexion point of the cumulative eigenvalues, and
197 subsequently only retaining PVs with a similarity above 0.5. We observe that the
198 consensus features computed between GDSC and TCGA (**Figure 3**A) and between
199 GDSC and HMF (**Figure 3**B) show the same proportion of non-linearity with a
200 concentration of offset and linearities in the top consensus features.

201 In order to visualize the structure retained in the consensus space, we embedded our
202 consensus scores into a 2D space using UMAP[42]. We observed that primary tumors
203 cluster together based on their tissue type (**Figure 3**C). Cell lines, however, show
204 different behaviors – most do cluster with the tumors from a similar tissue of origin,
205 while a group of cell lines cluster together and away from the tumors, regardless of
206 their tissue of origin, as observed in previous studies[43]. To quantify the degree of co-
207 clustering of cell lines and tumors, we compared distances between tumors and cell
208 lines from similar and non-similar tissues, and observed, as expected, a higher
209 similarity between tumors and cell lines from the same tissue (**Supp Figure 10**C).
210 Metastatic lesions show a weaker clustering based on the primary tumor's tissue of
211 origin (**Supp Figure 10**D-E). This is not unexpected, as the expression profiles are
212 derived from biopsy sites distant from the primary tissue. Of particular interest, we
213 observe the existence of a hematopoietic cell-line cluster that co-clusters with

214 metastatic samples from various biopsy sites. Most of these tumor samples (7 out of
215 12 samples) are lymph node metastases and most likely display a hematopoietic
216 expression profile due to blood infiltration in the samples (**Supp Figure 10**B).
217

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

220 To further validate our approach, we transferred response predictors from cell lines to
221 the TCGA and HMF collections of human tumors. First, we projected the GDSC and
222 TCGA expression data onto the GDSC-TCGA consensus features. Then we trained,
223 for each drug, a regression model using solely the cell line response data (measured
224 as AUC). These drug-specific regression models were then used to predict response
225 on the projected TCGA data. Finally, we compared the predicted response to the
226 known categorical clinical responses using a one-sided Mann-Whitney test and
227 computed the corresponding effect size. We trained models for seventeen different
228 drugs (**Table 1**A). We compared the performance of TRANSACT to the performance
229 obtained by four competing approaches (*Elastic Net, DL, ComBat+DL* and *PRECISE*)
230 (**Table 1**A, Methods). For the Deep Learning approaches (*DL* and *ComBat+DL*), we
231 selected the architecture and hyper-parameters for each drug by 5-fold cross-
232 validation on GDSC (Methods). We subsequently trained 50 models with different and
233 independent initializations (Method) and reported the median performance obtained
234 on TCGA. *Elastic Net* and *PRECISE* obtain significant associations for three and six
235 drugs, respectively, but neither approach ever outperforms all other approaches.
236 *ComBat+DL* and *DL* achieve significant associations for five and eight drugs,
237 respectively – however, both approaches outperform all others for only three and one
238 drugs, respectively. By contrast, TRANSACT achieves significant associations for

239 seven drugs and outperforms all other approaches for five drugs. For both deep
240 learning approaches, we observe an important dependency on the network
241 initialization (**Supp Figure 11**A, **Supp Figure 12**A). More importantly, we observe no
242 correlation between the training error achieved on the source domain and the
243 prediction accuracy on the target domain, making it impossible to select a proper
244 initialization solely based on the source domain performance (**Supp Figure 11**B,
245 **Supp Figure 12**B). Results obtained with TRANSACT, on the contrary, do not depend
246 on a random initialization.
247 For the HMF data, we repeated the steps above, while employing the GDSC-HMF
248 consensus features as well as the HMF and GDSC expression and response data.
249 We trained models for Irinotecan, Carboplatin (using Cisplatin GDSC response),
250 Trastuzumab (using Afatinib GDSC response), Gemcitabine, Paclitaxel and 5-
251 Fluorouracil. Across all approaches, we observe a significant association between the
252 predicted AUC and clinical responses for four of the six drugs (Irinotecan, Carboplatin,
253 Trastuzumab and Gemcitabine) (**Table 1**B, **Figure 4**A). *PRECISE* reaches
254 significance for two drugs, whereas *Elastic Net, DL* and *Combat+DL* reach
255 significance for a single drug. *TRANSACT*, by contrast, outperforms *PRECISE* on
256 three of the five drugs, and *ComBat* + *DL*, *Deep Learning* and *Elastic Net* on four
257 drugs. *TRANSACT* does not obtain a significant association for Paclitaxel. We do,
258 however, observe a trend that indicates a positive association between predicted and
259 clinical response. In contrast, all competing approaches fail to achieve any
260 association. **Figure 4**A provides a summary of the performance of TRANSACT and
261 two state-of-the-art approaches, *PRECISE* and *Combat* + *DL,* on both TCGA and HMF
262 cohorts. Finally, for the HMF cohort, we also observe a large variability in predictions
263 of the deep learning approaches as a result of network initialization (**Supp Figure 11**C,

264 **Supp Figure 12**C). As on the TCGA we do not observe any relationship between the
265 training error on the source domain data and the predictions on the target domain data
266 (**Supp Figure 11**D, **Supp Figure 12**D).
267 Two drugs, already standing out from the GDSC-PDX analysis, are of particular
268 interest. We first observe that Gemcitabine is consistently better predicted by
269 *TRANSACT* than by any other approaches (**Figure 4**B, **Table 1**A). When it comes to
270 Paclitaxel (**Figure 4**C, **Table 1**A), *TRANSACT* shows a clear improvement over
271 competing approaches in TCGA. *TRANSACT* is also the only method to recover a
272 positive association on HMF, although the association is not significant. Finally, for
273 Carboplatin (**Figure 4**C, **Table 1**A), we observe that *TRANSACT* outperforms all
274 approaches on TCGA, and on HMF. When taking the TCGA and HMF results together
275 – and noting that there is overlap in the drugs employed in these two cohorts – it is
276 interesting to observe that the best competing approach is *Deep Learning*. *Deep*
277 *Learning* achieves significant associations in 9 of the 23 comparisons while
278 *ComBat+DL* reaches significance in 6 trials. Based on this metric, *TRANSACT*
279 performs best, reaching significance in 11 of 23 trials. Furthermore, *Deep Learning*
280 outperforms all other approaches in only 3 comparisons, while *TRANSACT* delivers
281 increased performance in 8 comparisons. When also taking the performance on the
282 PDX models into account, we can safely conclude that *TRANSACT* is more accurate
283 in transferring drug response predictors.
284

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

288 We finally made use of the interpretability of our approach to mechanistically validate
289 our predictors (Methods). We first validated targeted therapies with documented
290 modes of action. We started with the TRANSACT predictor of response for Afatinib, a
291 small molecule inhibitor of the EGFR family, which includes HER2 (**Figure 5**A). We
292 performed a gene set enrichment analysis of the linear terms that constitutes to 80%
293 of the predictor (**Supp. Table 1**). Most enriched gene sets are related to breast cancer
294 subtypes as defined by *Charafe and colleagues*[44] where, contrary to the definition
295 based on the intrinsic breast cancer subtypes, the Luminal subtype contains both ER+
296 and HER2+ tumors. The top ranked gene set amongst the genes associated with
297 sensitivity (genes with a negative coefficient in the predictor) are genes associated
298 with the "Luminal" subtypes (FDR < 0.001). Conversely, genes associated with
299 resistance (genes with a positive coefficient in the predictor) show enrichment for the
300 "Mesenchymal" molecular signatures, shared by basal and mesenchymal subtypes.
301 This corresponds with HER2 negative samples, which is in line with our expectation
302 as absence of the drug target would indicate lack of response. Similarly, in the
303 TRANSACT response predictor for Gefitinib (EGFR inhibitor) the genes constituting
304 the linear portion and associated with sensitivity (negative predictor coefficients) show
305 an enrichment for genes downregulated in Gefitinib resistant tumors (**Figure 5**B,
306 **Supp. Table 2**).

307 Cytotoxic drugs such as Gemcitabine or Paclitaxel have complex modes of actions
308 involving different pathways, crosstalk between which remains challenging to
309 understand. Since the predictions of these two drugs showed a significant association

310 in both PDXs and patients, we set out to interpret the mechanisms of sensitivity or
311 resistance inferred by our predictor. In Gemcitabine (**Supp. Table 3**), we observe that
312 over-expression of the CDC42 pathway is a significant marker of resistance (FDR =
313 0.012, **Figure 5**C.1.) together with pathways linked to microtubule formation and cell
314 migration (**Supp Figure 13**), both known to be promoted by CDC42[45]. Together, these
315 enriched pathways highlight CDC42 over-expression as a potential mechanism of
316 Gemcitabine resistance, which suggests the use of CDC42 inhibitors[46,47] for
317 Gemcitabine-resistant tumors. Another interesting finding is the significant enrichment
318 of TNFα signaling in the sensitive portion (FDR=0.046) (**Figure 5**C.1.). A clinical trial
319 has shown that co-administration of TNF with gemcitabine improves patient survival
320 and further inhibits tumor growth[48], lending additional credence to this finding. Last,
321 we observe a concentration of sensitive interactions involving BLK, a pro-apoptotic
322 Src-proto oncogene involved in B-cell signaling and differentiation (**Figure 5**C.2.).
323 Since hematopoietic cell lines respond better to Gemcitabine, these interactions can
324 either act as a tissue-type marker or could potentially represent a sensitive pathway.
325 Finally, we looked for enriched pathways in Paclitaxel predictor (**Figure 5**D.1., **Supp.**
326 **Table 4**) and observed three potential mechanisms of resistance. We first observe
327 that the resistant linear coefficients are significantly enriched in genes linked to
328 silencing of YBX1[49] (FDR=0.106), a gene associated with proliferation in certain tumor
329 types[50]. In ovarian cancer, YBX1 has been shown to regulate ABCB1 expression
330 levels, a gene related to Paclitaxel resistance[52–56]. Our pan-cancer analysis therefore
331 further supports the role of drug transporters in Paclitaxel resistance. Second, we
332 observe a significant resistant enrichment for PI3K activation (FDR=0.18), which is
333 corroborated by the observed activation of PI3K/AKT/mTOR signaling pathway in
334 Paclitaxel-resistant cancer cells[56,57]. Moreover, a recent investigation suggests that

335 PI3K catalytic subunits can regulate ABCB1 expression[58]. Finally, when it comes to
336 the non-linear part, we observe a concentration of Fibroblast growth factors
337 interactions in the resistant parts of non-linearities, in particular FGF3, FGF20 and
338 FGF8 and FGF4 (**Figure 5**D.2., **Supp. Table 4)**. This behavior, although suggested
339 by previous studies[59,60], is all the more interesting as cell lines do not contain any
340 micro-environment that would elicit such resistance.
341

342 # Discussion

343 We introduced an approach to integrate pre-clinical and clinical data in a fully
344 unsupervised way. Our approach geometrically aligns sample-to-sample similarity
345 matrices and extracts directions of important variations for both datasets, without
346 requiring any sample-level pairing. By performing a geometrical alignment instead of
347 a direct distribution comparison, our approach limits the effect of a potential sample
348 selection bias. This geometrical alignment is implicitly performed in a space induced
349 by our similarity function, which enables the integration of various non-linearities,
350 corresponding to hypothesis made on the system. Although we restricted ourselves to
351 a Gaussian similarity, designing similarity functions that incorporate a wide range of
352 prior knowledge is a potentially promising avenue. Learning the similarity matrix, e.g.
353 using multiple kernel learning[61] or deep learning methods such as variational auto-
354 encoders[62], may also help increase performance.
355 TRANSACT compares directions computed using Kernel PCA, but our approach can
356 be extended to other basis expansion methods by modulating the way the coefficients
357 α^s and α^t are computed. More generally, our method is versatile, generalizable, and
358 can be applied beyond the scope of our study, e.g. to integrate single cell sequencing
359 data.

360 We showed that the consensus features can be used to build translatable predictors
361 of drug response. Although we do not require a strong covariate shift assumption as
362 in a previous study[63], we do assume the functions modeling the response from these
363 consensus features follow the same monotonicity in pre-clinical models and human
364 tumors. This assumption, albeit reasonable, may be questioned.
365 In this study, we limited ourselves to gene expression. Making use of other genomics
366 levels – e.g. mutations, copy number, methylation, chromatin accessibility – may help
367 refine the prediction by providing additional signal. The integration of our approach
368 with multi-omics integration strategies[64,65] may offer a solution to the translation of
369 multi-omics signatures.
370 Finally, we were able to predict response in patients that received a particular
371 treatment, either as monotherapy or in combination with other therapies, even though
372 the cell line predictors were trained on monotherapies only. We envision that large-
373 scale, cell line combination screens that take potential synergistic effects between
374 drugs into account will provide training data for the next generation of cell line
375 predictors to be transferred to human tumors in order to further enhance prediction of
376 patient response. Moreover, in this study, we focus on cytotoxic and targeted
377 therapies, still widely used in the clinic. The recent advent of immuno-therapies calls
378 for methods with the ability to predict the clinical response from model systems. This
379 requires model systems capable of mimicking the action of the immune system and
380 screening technologies able to measure the response for large panels. We believe
381 that our approach can be extended to such problems once data is made available.

382 # Methods

383 ## *Public data download and pre-processing*

384 ### *GDSC dataset, download and processing.*

385 We made use of the GDSC1000 cell line panel[14], which contains complete molecular
386 profiles for 1,049 cell lines (**Supp Figure 2**). Gene expression is provided in the form of
387 both read counts and FPKM. For both settings, we corrected the dataset for library-
388 size using TMM[66] and log-transformed the corrected read counts[67,68]. Finally, we
389 performed a gene-level mean-centering and standardization.

390 ### *Novartis PDXE dataset, download and processing.*

391 We made use of NIBR PDXE dataset for patient-derived xenografts[15], which contains
392 the gene expression profiles of 399 PDXs (**Supp Figure 3**). Gene expression is
393 provided in the form of FPKM. We corrected for library-size using TMM[66] and log-
394 transformed the corrected read counts[67,68]. Finally, we performed a gene-level mean-
395 centering and standardization.

396 ### *TCGA dataset, download and processing.*

397 We made use of the TCGA dataset for analyzing human biopsies[69], which comprises
398 10,347 human tumors (**Supp Figure 4**). Gene expression is provided in the forms of
399 both read counts and FPKM and we used the same pre-processing pipeline as for
400 GDSC. Response data have been obtained from *Ding et al*[70]*.* Following *Ding et al,* for
401 each drug, we consider the response to patients who were administered a particular
402 drug either as monotherapy or in combination with other drugs.

403

404 *Hartwig Medical Foundation dataset (HMF) download and processing.*

405 We validated our approach on a cohort of 1,049 patients provided by the Hartwig
406 Medical Foundation – referred to as *HMF* (**Supp Figure 5**A). Gene expression was
407 measured for each metastatic sample prior to indicated drug regimen. We used
408 MultiQC for quality control[71], salmon v1.0.0 for alignment to reference transcriptome[72],
409 and finally edgeR for gene-level quantification[73]. Comparison with results obtained
410 using STAR[74] and featureCounts[75] shows high degree of concordance (**Supp Figure**
411 **5**D) and we used this comparison to refine our filtering. Read counts were then
412 processed using the same pipeline as in GDSC and TCGA.

413 Drug response was measured in 802 unique metastatic samples using the RECIST
414 criteria. Response was measured differently for each patient (**Supp Figure 5**B) with
415 most patients having one single measure of response around 10 to 15 weeks after
416 treatment started (**Supp Figure 5**C). Since we are interested in the response of the
417 drug given the molecular characterization measured, we considered for each patient
418 the first response after treatment. Since most drugs are administered in combination,
419 we considered, for each drug, all the patients that received it, whether in combination
420 with other drugs or as monotherapy. For instance, in the case of Gemcitabine, we
421 predicted drug response for all patients that received Gemcitabine as part of their
422 treatment strategies.

423

424 *Measure of drug response*

425 In our different analysis, we rely on drug response measurements, either to train a
426 predictor (GDSC), or to validate it (PDXE, TCGA and HMF). For cell lines (GDSC),
427 drug response is measured as *Area Under the drug response Curve*, referred to as
428 AUC. For PDX, drug response is measured as *Best Average Response*, which

corresponds to the relative variation of tumor volume induced by a certain treatment. For both AUC and Best Average Response, large values are associated with resistance. For TCGA and HMF, clinical responses have been measured using the RECIST criteria[76]. Based on various metrics, patients get assigned to one of the following four groups: *Complete Response* (CR), *Partial Response* (PR), *Stable Disease* (SD) and *Progressive Disease* (PD). Following the division used in previous works[21,23], we divide TCGA patients in two categories: *Responders* (CR and PR) and *Non Responders* (SD and PD). For HMF, we discriminate between each possible couple: PR vs PD, PR vs SD and SD vs PD. Since only a couple of patients showed a complete response, we did not consider these patients.

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[77] 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:

$$\begin{aligned} \widetilde{\mathrm{K}}_{\mathrm{s}} &= \mathrm{C}_{\mathrm{n_s}}\mathrm{K_s}\mathrm{C}_{\mathrm{n_s}} \\ \widetilde{\mathrm{K}}_{\mathrm{t}} &= \mathrm{C}_{\mathrm{n_t}}\mathrm{K_t}\mathrm{C}_{\mathrm{n_t}} \text{ , with } \mathrm{C_n} = \mathrm{I_n} - \frac{1}{\mathrm{n}}1_{\mathrm{n}}^{\mathrm{T}}1_{\mathrm{n}}\ (\mathrm{n} \ge 1). \\ \widetilde{\mathrm{K}}_{\mathrm{st}} &= \mathrm{C}_{\mathrm{n_s}}\mathrm{K_{st}}\mathrm{C}_{\mathrm{n_t}} \end{aligned} \tag{3}$$

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 of 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 NLPCs, yielding *non-linear principal vectors* $s_1, .., s_d$ for the source and $t_1, .., t_d$ for the target domains, with $d = \min(d_s, d_t)$ (Def Supp. 4.1). These principal vectors are pairs of vectors: one linear combination of source NLPCs and one linear combination of target NLPCs, ordered by decreasing similarity with the first pair being the most similar. The computation of these PVs relies on the Singular Value Decomposition[78] 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)

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

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

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

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 kth 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):

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

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)

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

Such an interpolation between PVs balances the effect of source and target datasets. We prove that, in the case of a linear kernel, our interpolation scheme is equivalent to the one from previous approaches[26,79] (Supp Subsection 8).

Within each pair of PVs, we select one intermediate representation where the source and target projections match the most. For the kth PV-pair, we compare the source and target projected data using a Kolmogorov-Smirnov statistic and select the interpolation coefficient τ_k^* where the statistic is minimal. We obtain a set of optimal interpolation coefficients $\tau^* \in [0,1]^d$ when, for each PV, source and target influence are balanced. We call the corresponding vector *consensus features*. These consensus features show the minimal difference between source and target domain, a theoretical necessary condition for domain adaptation[80].

Prediction using Elastic Net

In order to predict drug response, we use Elastic Net regression[81]. Elastic Net is a linear model that imposes two penalties on the coefficients to predict: an $\ell_1$penalty that leads to a 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 compared

to complex models[16,18,38]. Using a linear classifier limits the complexity and therefore makes the transfer more robust in practice.

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[82] (Subsection Supp. 7). For the qth consensus feature, we differentiate three kinds of contributions (Def Supp. 7.4):

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

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[83]. 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 correspond to a Euclidean geometric proportion, we used a squared statistic to compare them.

Comparison to competing approaches

We compare TRANSACT to four different approaches. The two first approaches consist in applying a regression model trained on a source dataset to a target dataset without any correction; we use one linear Elastic Net model (referred to as *Elastic Net*) and a non-linear neural network model (referred to as *Deep Learning*). In both cases, we perform a grid-search 5-fold cross-validation on cell lines to select the model with the best performance: on *ElasticNet* we vary the ℓ_1 ratio and the total regularization ; on *Deep Learning*, following the protocol from *Sakellaropoulos et al*[21], we use an hyperbolic-tangent activation function while varying the global network structure, the ℓ_2 penalty and the input and output dropout levels (**Supp Table** 5).

The other two approaches first correct the signal and then train a regression model. The third approach (referred to as *ComBat* + *Deep Learning*) reproduces the approach from *Sakellaropoulos et al*[21] by first performing a ComBat technical batch effect correction between source and target, and then applying a neural network on the corrected signal, similar to *Deep Learning* (**Supp Table** 6). The last competing approach, referred to as *PRECISE,* consists in using a linear similarity function followed by an Elastic Net model, which is equivalent to PRECISE (**Supp Material**).

For the two deep learning approaches, we first performed cross-validation on the source dataset (with or without correction) to select the hyper-parameters and the network structures with the largest predictive performance. We then re-initialize the network and train it on the complete GDSC dataset.

589 *Availability of data and materials*

590 TRANSACT is available as a Python 3.6 module (https://github.com/NKI-
591 CCB/TRANSACT). All our experiments are reproducible and use state-of-the-art
592 libraries[84–89] (https://github.com/NKI-CCB/TRANSACT_manuscript). The dataset(s)
593 supporting the conclusions of this article are available in the "transact_manuscript"
594 repository of the aforementioned GitHub page, except for the HMF data, which can
595 only be obtained upon request to the Hartwig Medical Foundation.

596

597 **Acknowledgment**

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

601

602 We thank Mirrelijn van Nee (VUmc), Sander Canisius (NKI), Osman Kayhan (TU
603 Delft), Tycho Bismeijer (NKI), Stavros Makrodimitris (TU Delft), Tesa Severson (NKI),
604 Joe Siefert (NKI), Joris van de Haar (NKI), Ahmed Mahfouz (LUMC), Tamim Abdelaal
605 (TU Delft), David Tax (TU Delft), Bram Thijssen (NKI) and Ziqi Wang (TU Delft) for
606 useful discussions. We thank Mohammed Charrout (TU Delft), Marie Corradi (NKI),
607 Lisa Koob (NKI) and Wouter Kouw (TU Eindhoven) for careful reading of the
608 manuscript.
609 This work was supported by ZonMw TOP grant COMPUTE CANCER [40-00812-98-
610 16012].

611

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.

Competing interests

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

623 1. Slamon, D. J. *et al.* Numb Er 11 Use of Chemotherapy Plus a Monoclonal
624 Antibody Against Her2. *N. Engl. J. Med.* **344**, 783–792 (2001).
625 2. Davies, H. *et al.* Mutations of the BRAF gene in human cancer. *Nature* **417**,
626 949–954 (2002).
627 3. Hughes, T. P. *et al.* Frequency of major molecular responses to imatinib or
628 interferon alfa plus cytarabine in newly diagnosed chronic myeloid leukemia.
629 *N. Engl. J. Med.* **349**, 1423–1432 (2003).
630 4. Kalamara, A., Tobalina, L. & Saez-Rodriguez, J. How to find the right drug for
631 each patient? Advances and challenges in pharmacogenomics. *Curr. Opin.*
632 *Syst. Biol.* **10**, 53–62 (2018).
633 5. Prahallad, A. *et al.* Unresponsiveness of colon cancer to BRAF(V600E)
634 inhibition through feedback activation of EGFR. *Nature* **483**, 100–104 (2012).
635 6. Valery, P. *Mauvaises pensées et autres*. (1941).
636 7. Ben-David, U. *et al.* Patient-derived xenografts undergo mouse-specific tumor
637 evolution. *Nat. Genet.* **49**, 1567–1575 (2017).
638 8. Ben-David, U. *et al.* Genetic and transcriptional evolution alters cancer cell line
639 drug response. *Nature* **560**, 325–330 (2018).
640 9. Gillet, J. P. *et al.* Redefining the relevance of established cancer cell lines to
641 the study of mechanisms of clinical anti-cancer drug resistance. *Proc. Natl.*
642 *Acad. Sci. U. S. A.* **108**, 18708–18713 (2011).
643 10. Gillet, J. P., Varma, S. & Gottesman, M. M. The clinical relevance of cancer
644 cell lines. *J. Natl. Cancer Inst.* **105**, 452–458 (2013).
645 11. Mak, I. W. Y., Evaniew, N. & Ghert, M. Lost in translation: Animal models and
646 clinical trials in cancer treatment. *Am. J. Transl. Res.* **6**, 114–118 (2014).
647 12. Brubaker, D. K. & Lauffenburger, D. A. Translating preclinical models to

648 humans. *Science (80-.).* **367**, 742–743 (2020).
649 13. Webber, J. T., Kaushik, S. & Bandyopadhyay, S. Integration of Tumor
650 Genomic Data with Cell Lines Using Multi-dimensional Network Modules
651 Improves Cancer Pharmacogenomics. *Cell Syst.* **7**, 526-536.e6 (2018).
652 14. Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*
653 (2016). doi:10.1016/j.cell.2016.06.017
654 15. Gao, H. *et al.* High-throughput screening using patient-derived tumor
655 xenografts to predict clinical trial drug response. *Nat. Med.* **21**, 1318–1325
656 (2015).
657 16. Costello, J. C. *et al.* A community effort to assess and improve drug sensitivity
658 prediction algorithms. *Nat. Biotechnol.* **32**, 1202–1212 (2014).
659 17. Ali, M. & Aittokallio, T. Machine learning and feature selection for drug
660 response prediction in precision oncology applications. *Biophys. Rev.* **11**, 31–
661 39 (2019).
662 18. Jang, I. S., Neto, E. C., Guinney, J., Friend, S. H. & Margolin, A. A. Systematic
663 assessment of analytical methods for drug sensitivity prediction from cancer
664 cell line data. *Pacific Symp. Biocomput.* **23**, 1–7 (2013).
665 19. Geeleher, P., Cox, N. J. & Huang, R. S. Clinical drug response can be
666 predicted using baseline gene expression levels and in vitro drug sensitivity in
667 cell lines. *Genome Biol.* **15**, 1–12 (2014).
668 20. Geeleher, P. *et al.* Discovering novel pharmacogenomic biomarkers by
669 imputing drug response in cancer patients from large genomics studies.
670 *Genome Res.* **27**, 1743–1751 (2017).
671 21. Sakellaropoulos, T. *et al.* A Deep Learning Framework for Predicting
672 Response to Therapy in Cancer. *Cell Rep.* **29**, 3367-3373.e4 (2019).

673 22. Kurilov, R., Haibe-Kains, B. & Brors, B. Assessment of modelling strategies for
674 drug response prediction in cell lines and xenografts. *Sci. Rep.* **10**, 2849
675 (2020).
676 23. Noghabi, H. S., Peng, S., Zolotareva, O., Collins, C. C. & Ester, M. AITL:
677 Adversarial Inductive Transfer Learning with input and output space adaptation
678 for pharmacogenomics. *bioRxiv* 2020.01.24.918953 (2020).
679 doi:10.1101/2020.01.24.918953
680 24. Mourragui, S., Loog, M., van de Wiel, M. A., Reinders, M. J. T. & Wessels, L.
681 F. A. PRECISE: a domain adaptation approach to transfer predictors of drug
682 response from pre-clinical models to tumors. *Bioinformatics* **35**, i510–i519
683 (2019).
684 25. Pan, S. J., Kwok, J. T. & Yang, Q. Transfer Learning via Dimensionality
685 Reduction.
686 26. Gong, B., Shi, Y., Sha, F. & Grauman, K. Geodesic flow kernel for
687 unsupervised domain adaptation. *Proc. IEEE Comput. Soc. Conf. Comput. Vis.*
688 *Pattern Recognit.* 2066–2073 (2012). doi:10.1109/CVPR.2012.6247911
689 27. Gopalan, R., Li, R. & Chellappa, R. Domain adaptation for object recognition:
690 An unsupervised approach. *Proc. IEEE Int. Conf. Comput. Vis.* 999–1006
691 (2011). doi:10.1109/ICCV.2011.6126344
692 28. Fernando, B., Habrard, A., Sebban, M. & Tuytelaars, T. Unsupervised visual
693 domain adaptation using subspace alignment. *Proc. IEEE Int. Conf. Comput.*
694 *Vis.* 2960–2967 (2013). doi:10.1109/ICCV.2013.368
695 29. Kouw, W. & Loog, M. A review of domain adaptation without target labels.
696 *IEEE Trans. Pattern Anal. Mach. Intell.* **PP**, 1–1 (2019).
697 30. Hofmann, T., Schölkopf, B. & Smola, A. J. Kernel methods in machine

698 learning. *Ann. Stat.* **36**, 1171–1220 (2008).

699 31. Vert, J.-P., Koji Tsuda & Schölkopf, B. *Kernel methods in computational*
700 *biology*. (MIT Press, 2004).

701 32. He, X., Folkman, L. & Borgwardt, K. Kernelized rank learning for personalized
702 drug recommendation. *Bioinformatics* **34**, 2808–2816 (2018).

703 33. Ammad-Ud-Din, M. *et al.* Drug response prediction by inferring pathway-
704 response associations with kernelized Bayesian matrix factorization.
705 *Bioinformatics* **32**, i455–i463 (2016).

706 34. Li, Y., Wu, F. X. & Ngom, A. A review on machine learning principles for multi-
707 view biological data integration. *Brief. Bioinform.* **19**, 325–340 (2018).

708 35. Paltun, B. G., Mamitsuka, H. & Kaski, S. Improving drug response prediction
709 by integrating multiple data sources : matrix factorization , kernel and network-
710 based approaches. **00**, 1–14 (2019).

711 36. Schölkopf, B., Smola, A. J. & Müller, K.-R. Nonlinear Component Analysis as a
712 Kernel Eigenvalue Problem. *Neural Comput.* **10**, 1299–1319 (1998).

713 37. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-
714 1902.e21 (2019).

715 38. Smith, A. M. *et al.* Standard machine learning approaches outperform deep
716 representation learning on phenotype prediction from transcriptomics data.
717 *BMC Bioinformatics* **21**, 1–18 (2020).

718 39. Aben, N., Vis, D. J., Michaut, M. & Wessels, L. F. A. TANDEM: A two-stage
719 approach to maximize interpretability of drug response models based on
720 multiple molecular data types. *Bioinformatics* **32**, i413–i420 (2016).

721 40. Aben, N. *et al.* ITOP: Inferring the topology of omics data. *Bioinformatics* **34**,
722 i988–i996 (2018).

723 41. Hoogstraat, M. *et al.* Genomic and transcriptomic plasticity in treatment-naïve
724 ovarian cancer. *Genome Res.* **24**, 200–211 (2014).
725 42. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation
726 and Projection for Dimension Reduction. (2018).
727 43. Warren, A. *et al.* Global computational alignment of tumor and cell line
728 transcriptional profiles. *bioRxiv* 2020.03.25.008342 (2020).
729 doi:10.1101/2020.03.25.008342
730 44. Charafe-Jauffret, E. *et al.* Gene expression profiling of breast cell lines
731 identifies potential new basal markers. *Oncogene* 2273–2284 (2006).
732 doi:10.1038/sj.onc.1209254
733 45. Cau, J. & Hall, A. Cdc42 controls the polarity of the actin and microtubule
734 cytoskeletons through two distinct signal transduction pathways. *J. Cell Sci.*
735 **118**, 2579–2587 (2005).
736 46. Guo, Y. *et al.* R-ketorolac targets Cdc42 and Rac1 and alters ovarian cancer
737 cell behaviors critical for invasion and metastasis. *Mol. Cancer Ther.* **14**, 2215–
738 2227 (2015).
739 47. Del Mar Maldonado, M. & Dharmawardhane, S. Targeting rac and Cdc42 GT
740 pases in cancer. *Cancer Res.* **78**, 3101–3111 (2018).
741 48. Murugesan, S. R. *et al.* Combination of human tumor necrosis factor-alpha
742 (hTNF-α) gene delivery with gemcitabine is effective in models of pancreatic
743 cancer. *Cancer Gene Ther.* **16**, 841–847 (2009).
744 49. Basaki, Y. *et al.* Akt-dependent nuclear localization of Y-box-binding protein 1
745 in acquisition of malignant characteristics by human ovarian cancer cells.
746 *Oncogene* **26**, 2736–2746 (2007).
747 50. Frye, B. C. *et al.* Y-box protein-1 is actively secreted through a non-classical

748 pathway and acts as an extracellular mitogen. *EMBO Rep.* **10**, 783–789
749 (2009).
750 51. Housman, G. *et al.* Drug resistance in cancer: An overview. *Cancers (Basel).*
751 **6**, 1769–1792 (2014).
752 52. Goldstein, L. J. MDR1 gene expression in solid tumours. *Eur. J. Cancer* **32**,
753 1039–1050 (1996).
754 53. Vaidyanathan, A. *et al.* ABCB1 (MDR1) induction defines a common
755 resistance mechanism in paclitaxel- and olaparib-resistant ovarian cancer
756 cells. *Br. J. Cancer* **115**, 431–441 (2016).
757 54. Christie, E. L. *et al.* Multiple ABCB1 transcriptional fusions in drug resistant
758 high-grade serous ovarian and breast cancer. *Nat. Commun.* **10**, 5–14 (2019).
759 55. Patch, A. M. *et al.* Whole-genome characterization of chemoresistant ovarian
760 cancer. *Nature* **521**, 489–494 (2015).
761 56. Chen, D. *et al.* Dual PI3K/mTOR inhibitor BEZ235 as a promising therapeutic
762 strategy against paclitaxel-resistant gastric cancer via targeting
763 PI3K/Akt/mTOR pathway article. *Cell Death Dis.* **9**, (2018).
764 57. Hu, L., Hofmann, J., Lu, Y., Mills, G. B. & Jaffe, R. B. Inhibition of
765 phosphatidylinositol 3′-kinase increases efficacy of paclitaxel in in vitro and in
766 vivo ovarian cancer models. *Cancer Res.* **62**, 1087–1092 (2002).
767 58. Zhang, L. *et al.* The PI3K subunits, P110α and P110β are potential targets for
768 overcoming P-gp and BCRP-mediated MDR in cancer. *Mol. Cancer* **19**, 1–18
769 (2020).
770 59. Gan, Y., Wientjes, M. G. & Au, J. L. S. Expression of basic fibroblast growth
771 factor correlates with resistance to paclitaxel in human patient tumors. *Pharm.*
772 *Res.* **23**, 1324–1331 (2006).

773 60. Kim, S. H. *et al.* BGJ398, a pan-FGFR inhibitor, overcomes paclitaxel
774 resistance in urothelial carcinoma with FGFR1 overexpression. *Int. J. Mol. Sci.*
775 **19**, (2018).
776 61. Manica, M., Cadow, J., Mathis, R. & Rodríguez Martínez, M. PIMKL: Pathway-
777 Induced Multiple Kernel Learning. *npj Syst. Biol. Appl.* **5**, (2019).
778 62. Lin, X., Duan, Y., Dong, Q., Lu, J. & Zhou, J. Deep variational metric learning.
779 in *ECCV* 714–729 (2018). doi:10.1007/978-3-030-01267-0_42
780 63. Rampášek, L. Latent variable models for drug response prediction and genetic
781 testing. *PhD Thesis* (2019). doi:.1037//0033-2909.I26.1.78
782 64. Wang, B. *et al.* Similarity network fusion for aggregating data types on a
783 genomic scale. *Nat. Methods* **11**, 333–337 (2014).
784 65. Sharifi-Noghabi, H., Zolotareva, O., Collins, C. C. & Ester, M. MOLI: Multi-
785 omics late integration with deep neural networks for drug response prediction.
786 *Bioinformatics* **35**, i501–i509 (2019).
787 66. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential
788 expression analysis of RNA-seq data. *Genome Biol.* **11**, (2010).
789 67. Dillies, M. A. *et al.* A comprehensive evaluation of normalization methods for
790 Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.* **14**,
791 671–683 (2013).
792 68. Zwiener, I., Frisch, B. & Binder, H. Transforming RNA-Seq data to improve the
793 performance of prognostic gene signatures. *PLoS One* **9**, 1–13 (2014).
794 69. Weinstein, J. N. *et al.* The cancer genome atlas pan-cancer analysis project.
795 *Nat. Genet.* **45**, 1113–1120 (2013).
796 70. Ding, Z., Zu, S. & Gu, J. Evaluating the molecule-based prediction of clinical
797 drug responses in cancer. *Bioinformatics* **32**, 2891–2895 (2016).

798 71. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: Summarize
799 analysis results for multiple tools and samples in a single report. *Bioinformatics*
800 **32**, 3047–3048 (2016).
801 72. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon
802 provides fast and bias-aware quantification of transcript expression. *Nat.*
803 *Methods* **14**, 417–419 (2017).
804 73. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor
805 package for differential expression analysis of digital gene expression data.
806 *Bioinformatics* **26**, 139–140 (2009).
807 74. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**,
808 15–21 (2013).
809 75. Liao, Y., Smyth, G. K. & Shi, W. FeatureCounts: An efficient general purpose
810 program for assigning sequence reads to genomic features. *Bioinformatics* **30**,
811 923–930 (2014).
812 76. Therasse, P. tumors. European Organization for Research and Treatment of
813 Cancer, National Cancer Institute of the United States, National Cancer
814 Institute of Canada. *J natl Cancer inst* **92**, 205–16 (2000).
815 77. Aronszajn, N. Theory of Reproducing Kernels. *Trans. Am. Math. Soc.* **68**, 337
816 (1950).
817 78. Golub, G. H. & Van Loan, C. F. *Matrix Computations*. (2013).
818 79. Gopalan, R., Li, R. & Chellappa, R. by Generating Intermediate Data
819 Representations. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**, 2288–2302
820 (2014).
821 80. Ben-David, S. *et al.* A theory of learning from different domains. *Mach. Learn.*
822 **79**, 151–175 (2010).

81. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net Hui. *J. Stat. Soc. Ser. B* **67**, 301–320 (2005).

82. 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).

83. 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).

84. 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).

85. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

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

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

88. Tietz, M., Fan, T. J., Nouri, D., Bossan, B. & skorch Developers. skorch: A scikit-learn compatible neural network library that wraps PyTorch. (2017).

89. Behdenna, A., Haziza, J., Azencott, C.-A. & Nordor, A. pyComBat, a Python tool for batch effects correction in high-throughput molecular data using empirical Bayes methods. *bioRxiv* (2020). doi:10.1101/2020.03.17.995431

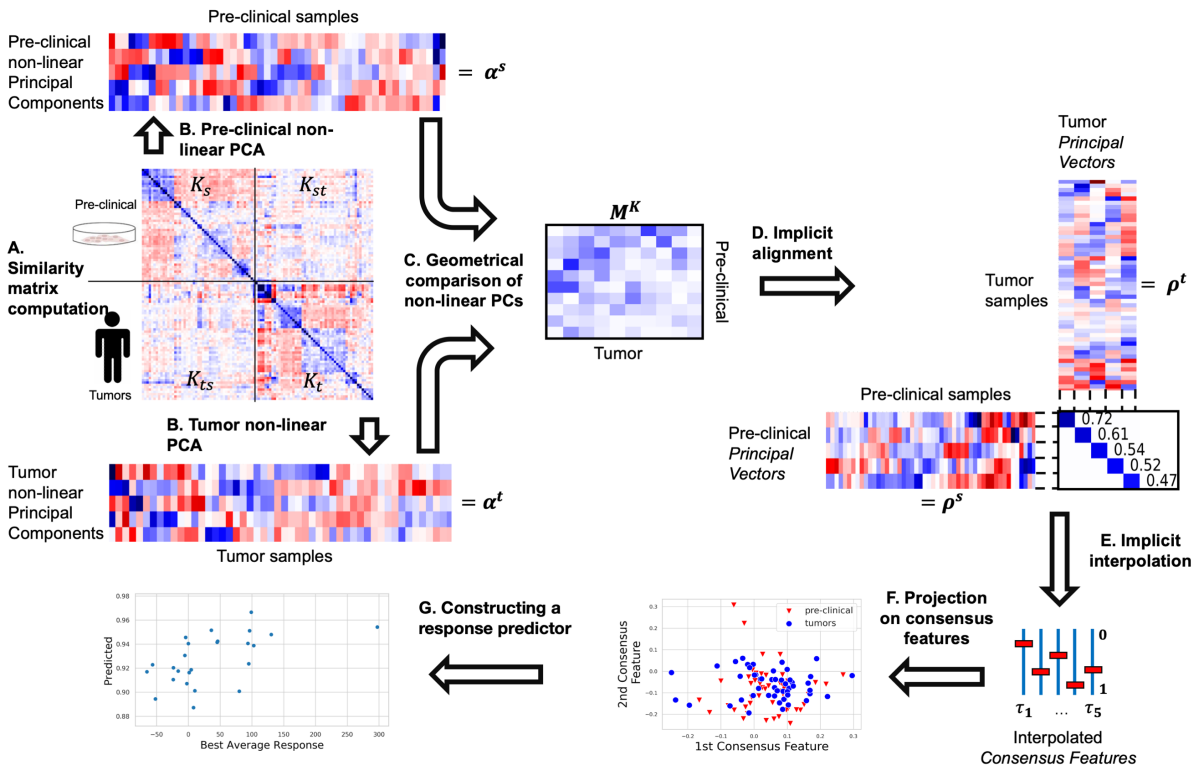


Figure 1 <u>TRANSACT: Generating non-linear manifold representations to transfer predictors of response from pre-clinical models to human tumors.</u> (**A**) Samples are compared using a **similarity function** yielding similarity matrices between pre-clinical models (source, K_s), between tumors samples (target, 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 into **non-linear principal components** (NLPCs) geometrically represented by "sample importance scores" (**Supp Figure 1A**) that represent the importance of each sample in each NLPC (α^s and α^t, for source and target space, respectively). (**C**) Geometrical comparison of pre-clinical and tumor NLPCs results in a **non-linear cosine similarity matrix** $\boldsymbol{M}^K$. (**D**) Alignment of NLPCs using the notion of **principal vectors** (**Supp Figure 1B**). (**E**) Interpolation within each pair of vectors to select one vector per PV-pair that balances the effect of pre-clinical and tumor signals: the **consensus features** (**Supp Figure 1C**). (**F**) Projection of each tumor and pre-clinical sample on the consensus features

861 to obtain **consensus scores**: scores that correspond to the activity of processes
862 conserved between tumors and pre-clinical models. (**G**) Finally, these scores can be
863 used as input to any predictive model, for instance to predict drug response based on
864 these consensus scores.

865

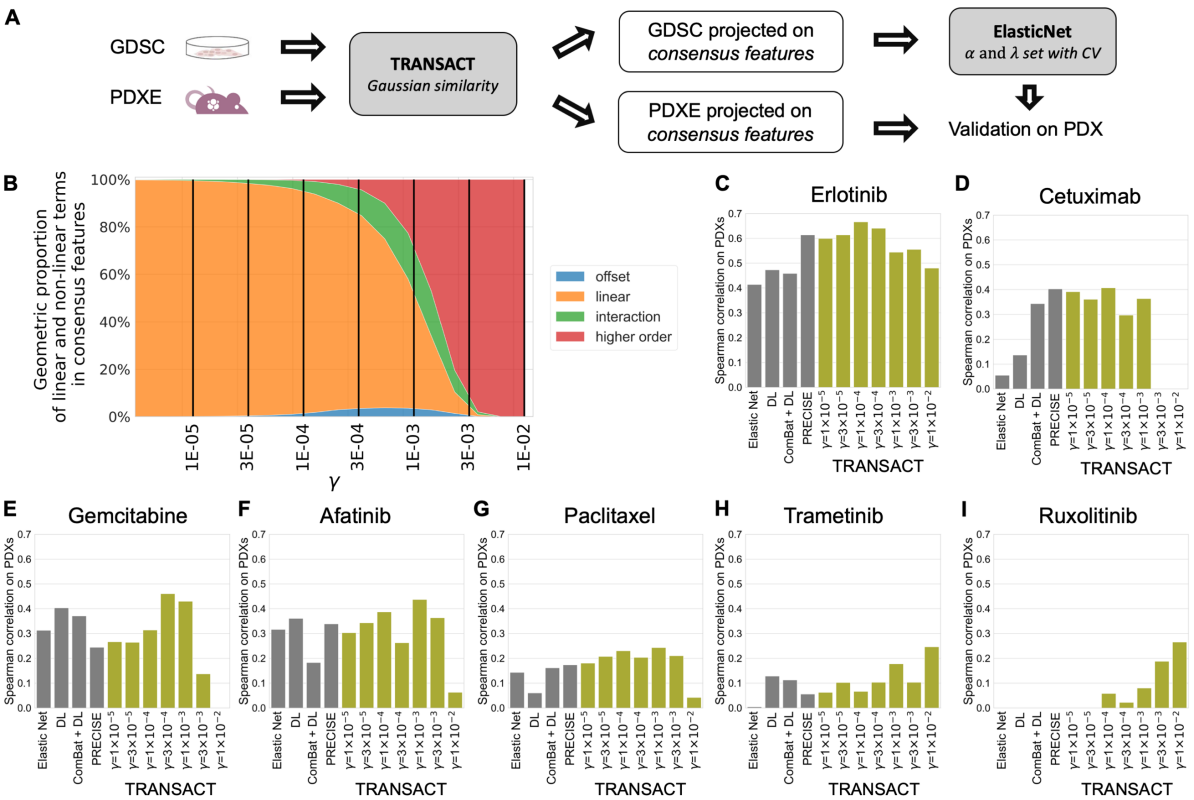
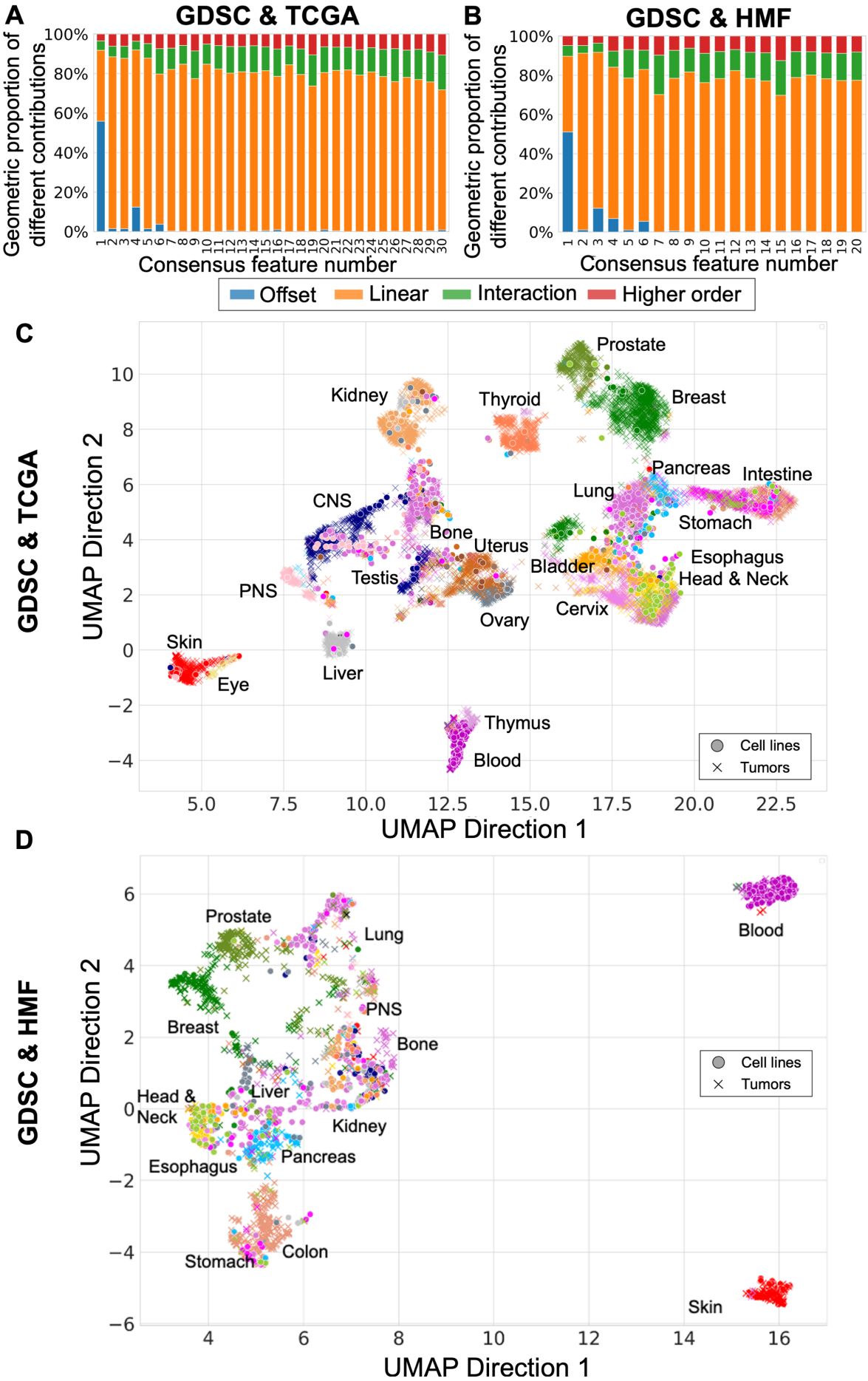


Figure 2 Impact of modeling non-linearities for drug response prediction transfer from cell lines to PDXs. (**A**) Main workflow of the prediction on PDXs. Using cell lines and PDX gene expression, we compute consensus features and project each dataset onto these. We then trained a regression model (Elastic Net) on cell lines projected scores to predict response, measured as Area Under the drug response Curve (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 computed the average contribution over all consensus features of offset, linear, interaction and higher order features (Methods). 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 response prediction on the PDX models for different values of γ, and for four competing approaches: two methods (*ElasticNet* and *Deep Learning*) that do not model the transfer, and two transfer

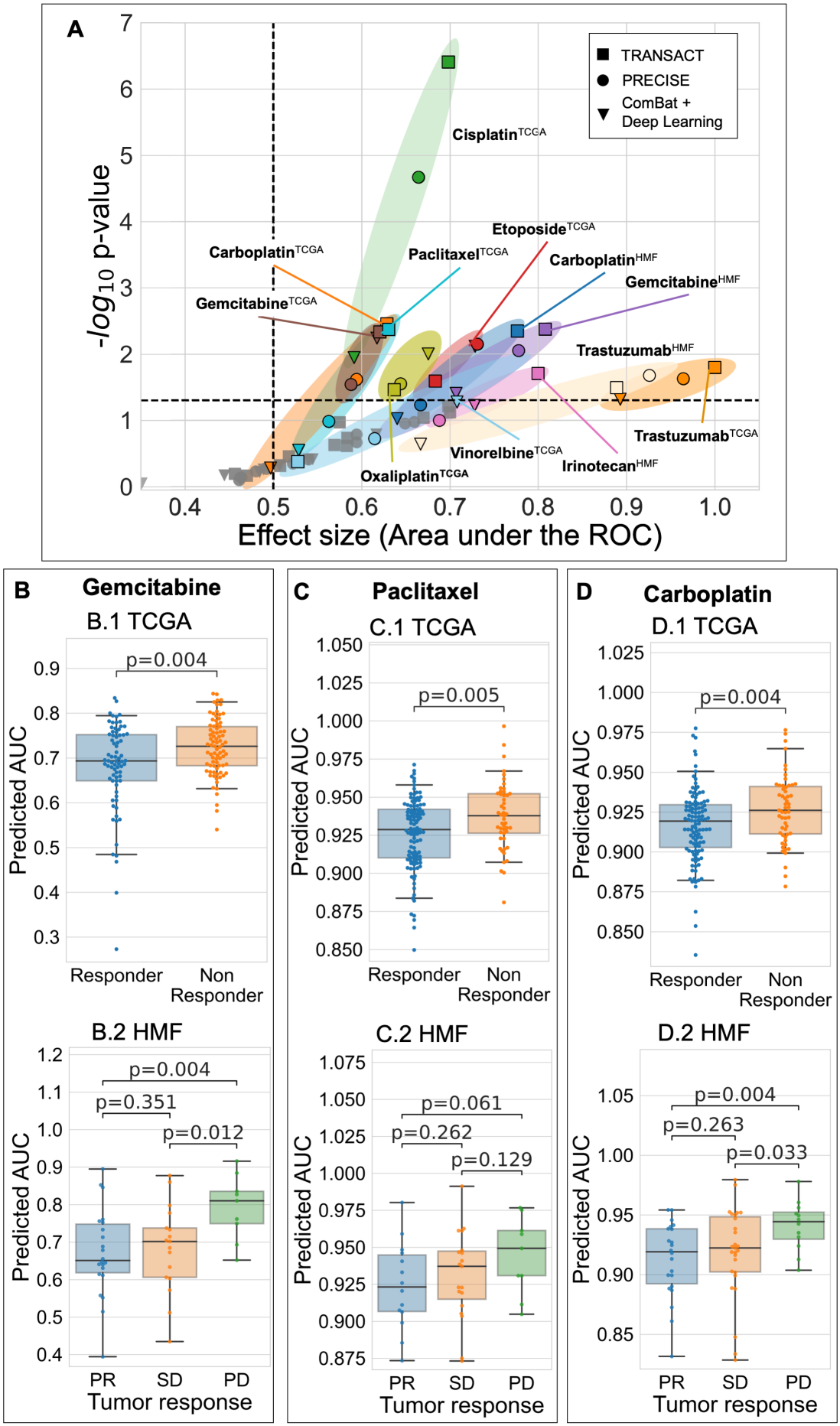
881 learning methods (ComBat followed by Deep Learning, and *PRECISE*). We report
882 results for Erlotinib (**C**), Cetuximab (**D**), Gemcitabine (**E**), Afatinib (**F**), Paclitaxel (**G**),
883 Trametinib (**H**) and Ruxolitinib (**I**). Concordance on PDX is measured as the Spearman
884 correlation between predicted AUC and Best Average Response.

885


A
GDSC & TCGA
Geometric proportion of different contributions
100%
80%
60%
40%
20%
0%
Consensus feature number
B
GDSC & HMF
Geometric proportion of different contributions
Consensus feature number
Offset
Linear
Interaction
Higher order
C
GDSC & TCGA
UMAP Direction 2
UMAP Direction 1
Prostate
Kidney
Thyroid
Breast
Pancreas
Intestine
Lung
CNS
Stomach
Bone
Uterus
Esophagus
Bladder
Head & Neck
Testis
PNS
Cervix
Ovary
Skin
Liver
Eye
Thymus
Blood
Cell lines
Tumors
D
GDSC & HMF
UMAP Direction 2
UMAP Direction 1
Prostate
Lung
Blood
PNS
Breast
Bone
Head & Neck
Liver
Kidney
Pancreas
Esophagus
Colon
Stomach
Skin
Cell lines
Tumors

886

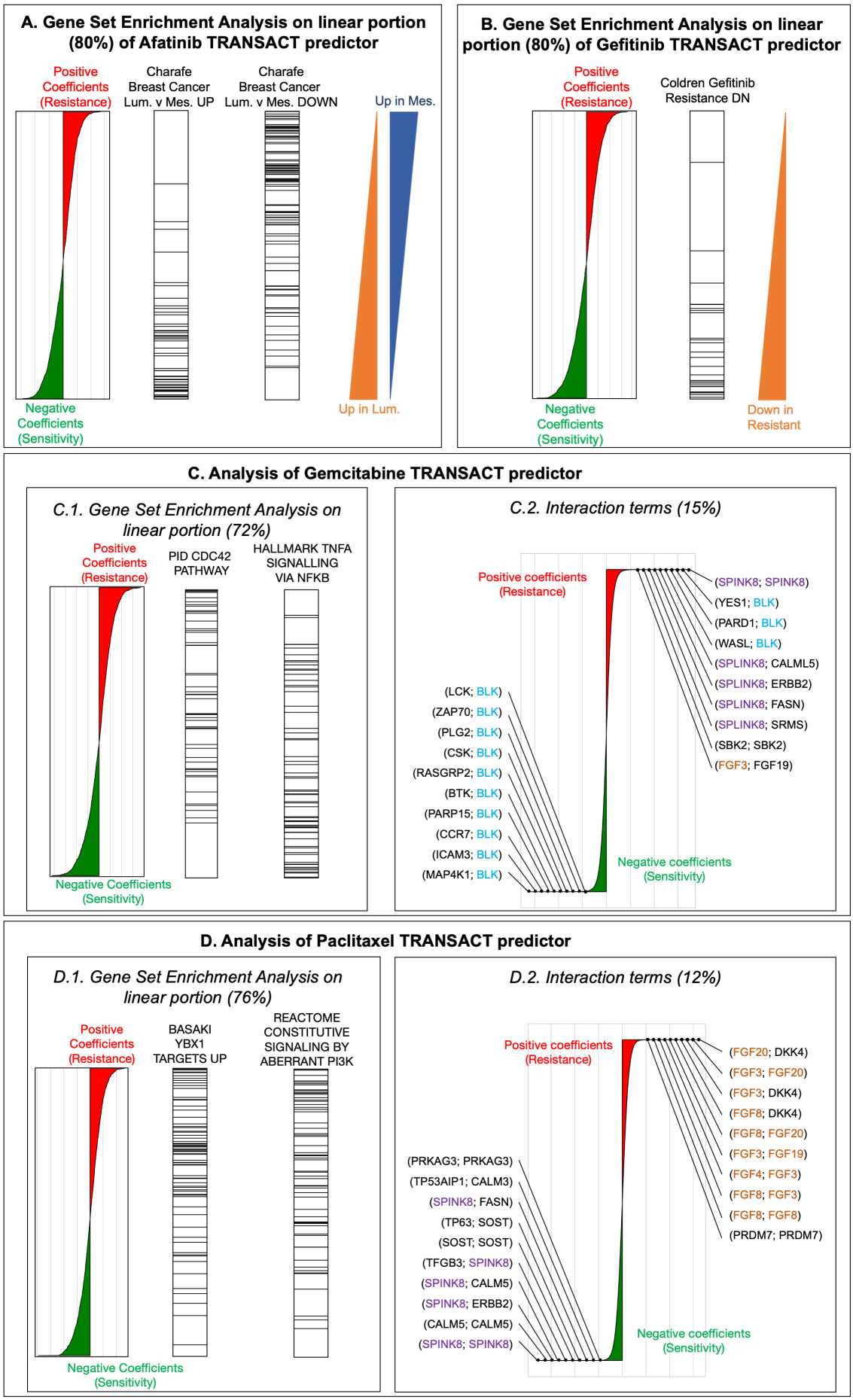
887 **Figure 3** Pan-cancer consensus features between cell lines and tumors conserve
888 tissue type information. We used cell lines (GDSC) as source data and computed
889 two sets of consensus features with two different target datasets: primary tumors
890 (TCGA, A and C) and metastatic lesions (HMF, B and D). (**A**) Proportion of linear
891 and non-linear contributions to each of the 30 GDSC-to-TCGA consensus features.
892 (**B**) Proportion of linear and non-linear contributions to each of the 20 GDSC-to-HMF
893 consensus features. (**C**) UMAP plot of primary tumors (TCGA, 21 tissues) and cell
894 lines (GDSC, 22 tissues) projected on the consensus features, using the same
895 parameters as selected in **Figure 2**. (**D**) UMAP plot of metastatic lesions (HMF) and
896 cell lines, colored by primary tissue for both HMF and GDSC. For both UMAP plots,
897 the full legend can be found in **Supp Figure 10**A.
898
899


A
TRANSACT
PRECISE
ComBat + Deep Learning
Cisplatin TCGA
Etoposide TCGA
Carboplatin TCGA
Paclitaxel TCGA
Carboplatin HMF
Gemcitabine HMF
Gemcitabine TCGA
Trastuzumab HMF
Trastuzumab TCGA
Vinorelbine TCGA
Irinotecan HMF
Oxaliplatin TCGA
-log10 p-value
Effect size (Area under the ROC)
B Gemcitabine
B.1 TCGA
p=0.004
Predicted AUC
Responder
Non Responder
B.2 HMF
p=0.004
p=0.351
p=0.012
PR
SD
PD
Tumor response
C Paclitaxel
C.1 TCGA
p=0.005
C.2 HMF
p=0.061
p=0.262
p=0.129
D Carboplatin
D.1 TCGA
p=0.004
D.2 HMF
p=0.004
p=0.263
p=0.033

900

901 **Figure 4** Consensus features improve response prediction in patients. (**A**) We used
902 consensus features computed between GDSC and each tumor dataset to train a
903 predictor for 17 drugs on TCGA and 5 drugs on HMG, using GDSC response only. We
904 then predicted the Area Under the drug response Curve (AUC) for each drug and
905 compared this predicted value to the observed clinical response using a one-sided
906 Mann-Whitney test (**Table 1**). We performed the same prediction tasks using two
907 state-of-the-art approaches (*ComBat* + *Deep Learning* and *PRECISE*, **Method**) and
908 summarized here the associations for the 3×22 predictors in a Volcano plot, with the
909 corresponding p-values on the y-axis and the effect-sizes on the x-axis. (**B**) Results
910 for the two Gemcitabine predictors. (**B**.1) For the Gemcitabine predictor on TCGA, we
911 compared predicted AUC for each patient to the known clinical response. (**B**.2) Results
912 for Gemcitabine predictor on HMF. (**C**.1) Results for Paclitaxel predictor on TCGA.
913 (**C**.2.) Results for Paclitaxel predictor on HMF. (**D**.1) Results for Carboplatin predictor
914 on TCGA. (**D**.2.) Results for Carboplatin predictor on HMF.

915


A. Gene Set Enrichment Analysis on linear portion (80%) of Afatinib TRANSACT predictor
Positive Coefficients (Resistance)
Charafe Breast Cancer Lum. v Mes. UP
Charafe Breast Cancer Lum. v Mes. DOWN
Up in Mes.
Negative Coefficients (Sensitivity)
Up in Lum.
B. Gene Set Enrichment Analysis on linear portion (80%) of Gefitinib TRANSACT predictor
Positive Coefficients (Resistance)
Coldren Gefitinib Resistance DN
Negative Coefficients (Sensitivity)
Down in Resistant
C. Analysis of Gemcitabine TRANSACT predictor
C.1. Gene Set Enrichment Analysis on linear portion (72%)
Positive Coefficients (Resistance)
PID CDC42 PATHWAY
HALLMARK TNFA SIGNALLING VIA NFKB
Negative Coefficients (Sensitivity)
C.2. Interaction terms (15%)
Positive coefficients (Resistance)
(SPINK8; SPINK8)
(YES1; BLK)
(PARD1; BLK)
(WASL; BLK)
(SPLINK8; CALML5)
(SPLINK8; ERBB2)
(SPLINK8; FASN)
(SPLINK8; SRMS)
(SBK2; SBK2)
(FGF3; FGF19)
(LCK; BLK)
(ZAP70; BLK)
(PLG2; BLK)
(CSK; BLK)
(RASGRP2; BLK)
(BTK; BLK)
(PARP15; BLK)
(CCR7; BLK)
(ICAM3; BLK)
(MAP4K1; BLK)
Negative coefficients (Sensitivity)
D. Analysis of Paclitaxel TRANSACT predictor
D.1. Gene Set Enrichment Analysis on linear portion (76%)
Positive Coefficients (Resistance)
BASAKI YBX1 TARGETS UP
REACTOME CONSTITUTIVE SIGNALING BY ABERRANT PI3K
Negative Coefficients (Sensitivity)
D.2. Interaction terms (12%)
Positive coefficients (Resistance)
(FGF20; DKK4)
(FGF3; FGF20)
(FGF3; DKK4)
(FGF8; DKK4)
(FGF8; FGF20)
(FGF3; FGF19)
(FGF4; FGF3)
(FGF8; FGF3)
(FGF8; FGF8)
(PRDM7; PRDM7)
(PRKAG3; PRKAG3)
(TP53AIP1; CALM3)
(SPINK8; FASN)
(TP63; SOST)
(SOST; SOST)
(TFGB3; SPINK8)
(SPINK8; CALM5)
(SPINK8; ERBB2)
(CALM5; CALM5)
(SPINK8; SPINK8)
Negative coefficients (Sensitivity)

916

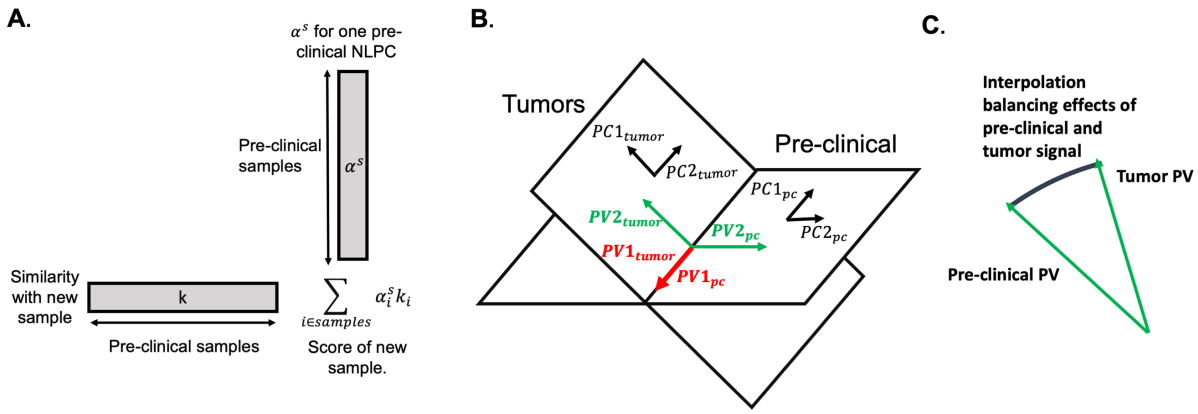
917 **Figure 5** Interpretability of TRANSACT consensus features recapitulates modes of
918 action for Afatinib and Gefitinib, and highlights mechanisms of sensitivity and
919 resistance to Gemcitabine and Paclitaxel. (**A**) For the Afatinib (HER-2 protein kinase
920 inhibitor) response model trained on GDSC samples projected on GDSC-to-TCGA
921 consensus features, we show the contribution of each gene to the predictor (left):
922 positive (negative) weights in the predictor indicate that high (low) expression of the
923 genes leads to resistance (sensitivity) represented by larger (smaller) AUCs. We
924 performed a PreRanked gene set enrichment analysis (GSEA) on these weights and
925 highlight genes associated to two significantly enriched gene sets: "Charafe [...]
926 Luminal vs Mesenchymal UP" (center-left) and "Charafe [...] Luminal vs Mesenchymal
927 DOWN" (center-right). The two triangles on the right illustrate these two gene-set
928 distributions. (**B**) Interpretation of the linear part of the Gefitinib (EGFR inhibitor)
929 model. We display gene weights ordered by contribution (left) and the ranks of genes
930 known to be down regulated in Gefitinib resistant tumors (right). (**C**) Interpretation of
931 Gemcitabine predictor. (*C.1.*) We display the positions of genes associated to two
932 significantly enriched gene sets: "CDC42 pathway", associated to positive weights
933 (resistance), and "TNFα-signalling", associated to negative weights (sensitivity). (*C.2.*)
934 Interaction terms make 15% of Gemcitabine predictor. We show the distribution of
935 their weights annotated with the ten largest weights (resistance) and ten smallest
936 weights (sensitivity). (**D**) Interpretation of Paclitaxel predictor. (*D.1.*) We display the
937 positions of genes associated to two significantly enriched gene sets, both to
938 resistance: "Basaki YBX1 target UP" and "constitutive signaling by PI3K-aberrant
939 signaling". (*D.2.*) Interaction terms for Paclitaxel, that make 12% of the predictor,
940 alongside the top-10 resistant and top-10 sensitive interactions.

A. GDSC-to-TCGA								
GDSC		**TCGA**		**p-val [effect-size] on TCGA**				
Drug	**Samples**	**Drug**	**Samples**	**Elastic Net**	**Deep Learning**	**DL + ComBat**	**PRECISE**	**TRANS ACT**
Afatinib	800	Trastuzumab	16	**0.089 [0.82]**	**0.034 [0.93]**	**0.048 [0.89]**	**0.026 [0.96]**	**0.016 [1.00]**
Bleomycin	856	Bleomycin	53	0.15 [0.63]	0.19 [0.61]	0.41 [0.53]	0.11 [0.66]	0.091 [0.67]
Cetuximab	868	Cetuximab	19	0.12 [0.67]	0.52 [0.50]	0.39 [0.54]	0.45 [0.52]	0.077 [0.70]
Cisplatin	764	Cisplatin	308	**1.1E-6 [0.69]**	**1.8E-4 [0.64]**	**1.1E-2 [0.59]**	**2.2E-5 [0.66]**	**3.9E-7 [0.70]**
Cisplatin	764	Carboplatin	166	**0.042 [0.58]**	0.10 [0.56]	0.53 [0.50]	**0.024 [0.59]**	**0.0035 [0.63]**
Cyclophosphamide	747	Cyclophosphamide	102	0.57 [0.48]	0.67 [0.45]	0.89 [0.35]	0.11 [0.65]	0.64 [0.46]
Docetaxel	665	Docetaxel	102	0.62 [0.48]	0.27 [0.54]	0.54 [0.49]	0.78 [0.46]	0.11 [0.58]
Doxorubicin	871	Doxorubicin	101	0.81 [0.45]	**0.0055 [0.66]**	0.60 [0.48]	0.99 [0.32]	0.69 [0.47]
Etoposide	880	Etoposide	84	0.21 [0.58]	**0.0032 [0.76]**	**0.0076 [0.73]**	**0.0071 [0.73]**	**0.027 [0.68]**
5-Fluorouracil	801	Fluorouracil	186	0.22 [0.54]	0.72 [0.47]	0.79 [0.46]	0.80 [0.46]	0.35 [0.52]
Gemcitabine	752	Gemcitabine	156	0.25 [0.53]	**0.021 [0.59]**	**0.0057 [0.62]**	**0.029 [0.59]**	**0.0047 [0.62]**
Irinotecan	796	Irinotecan	25	0.81 [0.39]	0.42 [0.53]	0.68 [0.44]	0.54 [0.49]	0.49 [0.51]
Oxaliplatin	724	Oxaliplatin	66	0.38 [0.52]	**0.022 [0.65]**	**0.0099 [0.68]**	**0.028 [0.64]**	**0.035 [0.64]**
Paclitaxel	753	Paclitaxel	160	**0.032 [0.59]**	**0.025 [0.60]**	0.28 [0.53]	0.10 [0.56]	**0.0042 [0.63]**
Pemetrexed	898	Pemetrexed	38	0.16 [0.60]	0.39 [0.53]	0.55 [0.49]	0.16 [0.59]	0.24 [0.57]
Temozolomide	746	Temozolomide	96	0.56 [0.49]	0.29 [0.55]	0.47 [0.51]	0.59 [0.48]	0.19 [0.58]
Vinorelbine	746	Vinorelbine	30	0.31 [0.57]	**0.043 [0.72]**	0.053 [0.71]	0.19 [0.61]	0.42 [0.53]
B. GDSC-to-HMF								
GDSC		**HMF**		**p-val [effect-size] on HMF**				
Drug	**Samples**	**Drug**	**Samples**	**Elastic Net**	**Deep Learning**	**DL + ComBat**	**PRECISE**	**TRANS ACT**
Afatinib	800	Trastuzumab	25	0.18 [0.70]	0.069 [0.81]	0.23 [0.67]	**0.021 [0.93]**	**0.032 [0.89]**
Irinotecan	796	Irinotecan	67	0.060 [0.73]	0.060 [0.73]	0.060 [0.73]	0.10 [0.69]	**0.020 [0.80]**
Cisplatin	764	Carboplatin	64	0.23 [0.58]	0.051 [0.67]	0.094 [0.64]	0.59 [0.67]	**0.0045 [0.78]**
5-Fluorouracil	801	Fluorouracil	65	0.065 [0.68]	0.12 [0.64]	0.16 [0.62]	0.21 [0.59]	0.24 [0.58]
Paclitaxel	753	Paclitaxel	45	0.56 [0.48]	0.32 [0.56]	0.39 [0.54]	0.39 [0.53]	0.061 [0.70]
Gemcitabine	752	Gemcitabine	50	**0.039 [0.71]**	**0.014 [0.76]**	**0.039 [0.71]**	**0.0089 [0.78]**	**0.0042 [0.81]**

Table 1 Results of TRANSACT compared to four competing approaches – *Elastic Net*, *Deep Learning*, *ComBat+DL*, and *PRECISE* – for 17 drugs on TCGA and 5 drugs on HMF. For each drug, we train 5 predictors and compare in each scenario the predicted

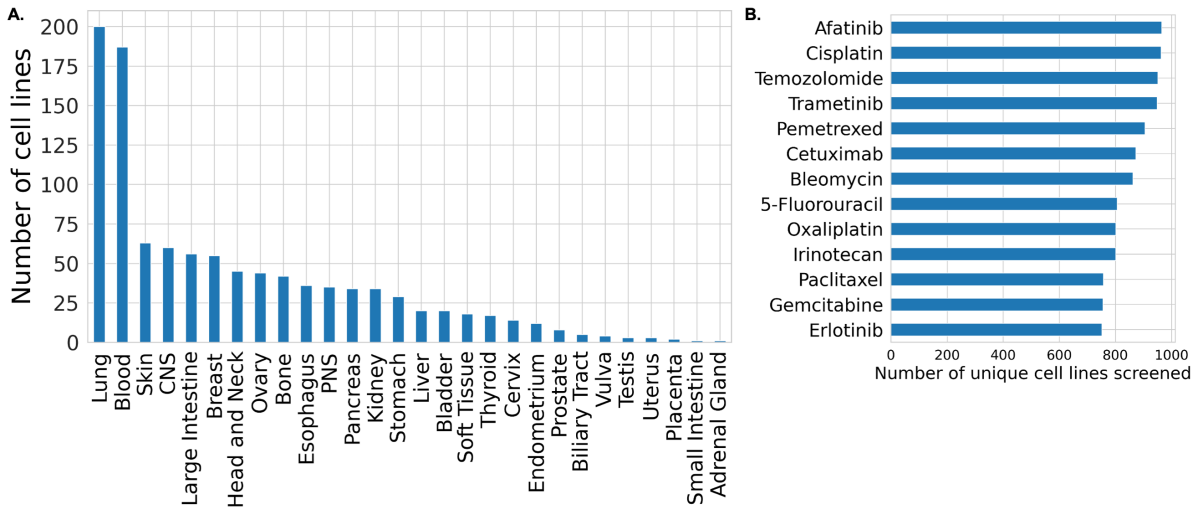
946 AUC to the known clinical response using one-side Mann-Whitney test. The clinical
947 response corresponds to the following division: Responders and Non-Responders for
948 TCGA, PR and PD for HMF (Methods). For each predictor, we report the p-value and
949 the effect size (Area under the ROC, effect size associated with the Mann-Whitney
950 test) under bracket. Bold cells correspond to significant association (pval < 0.05). Red
951 cells correspond to significant associations with the largest effect size across the three
952 methods.

953

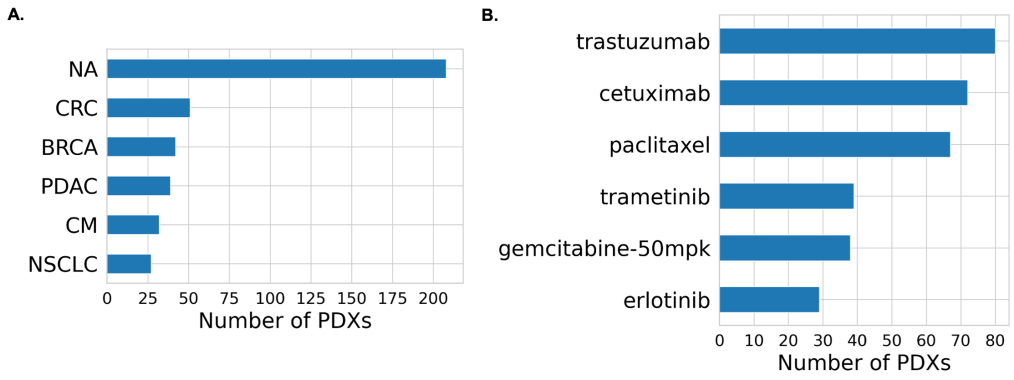


Supp Figure 1 Visual explanation of geometric alignment. **A**: Difference between importance scores (α^s, α^t) and projected scores. Since the space induced by the similarity function K is intractable, we use a dual representation of the NLPC in terms of samples: the importance scores. To project samples 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

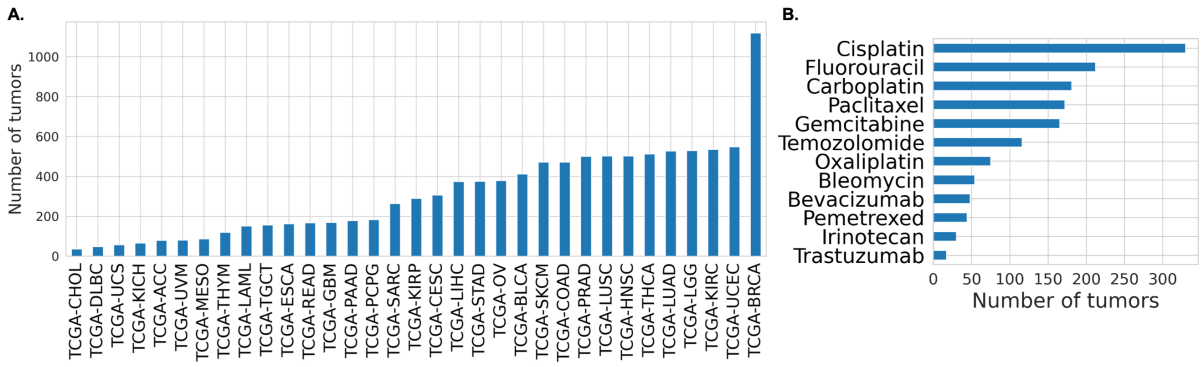
target projected signals are maximally matched. This optimal intermediate vector is called the consensus feature.



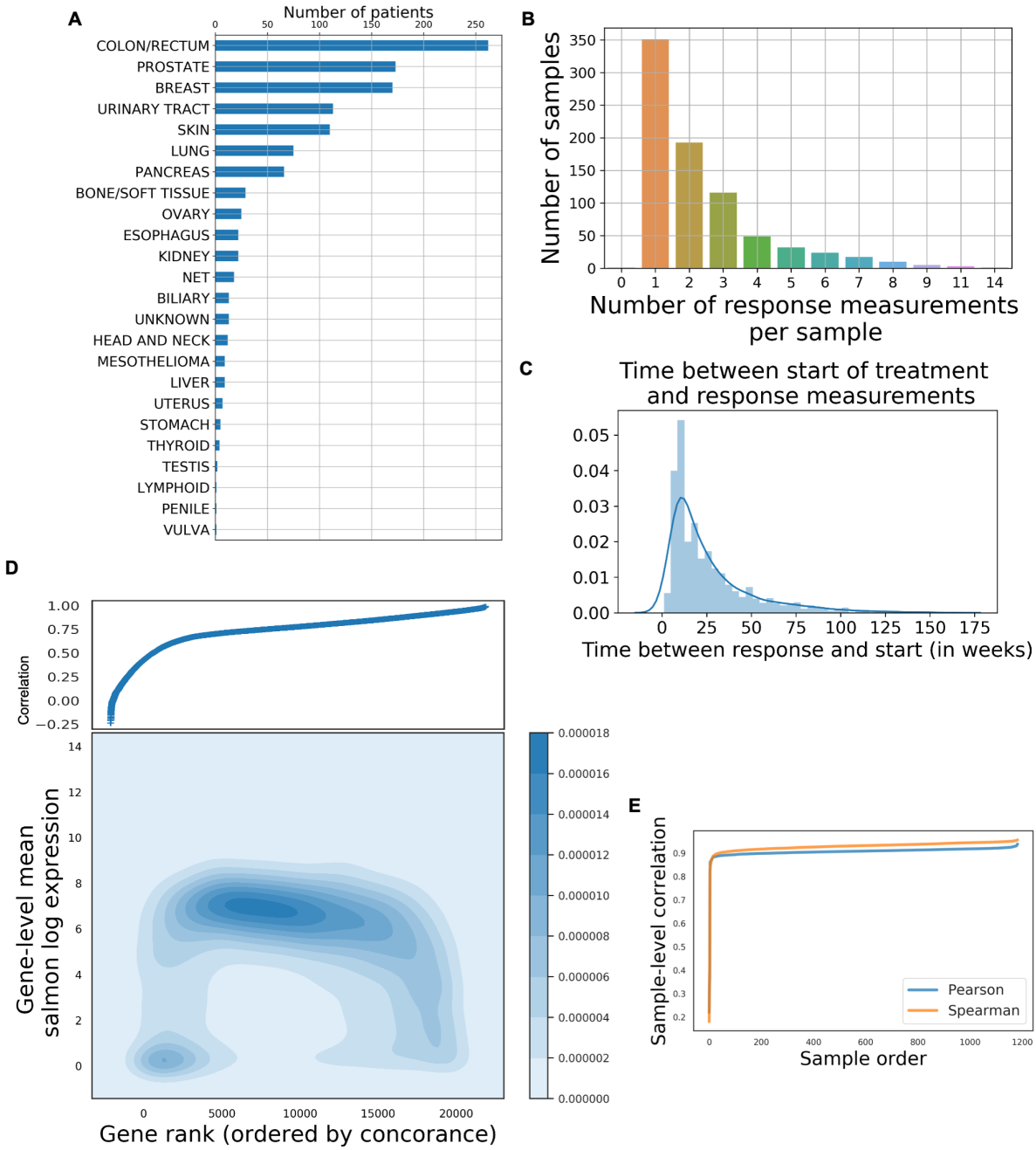
Supp Figure 2 Composition of the GDSC dataset (cell lines). We make use of the GDSC1000 cell line panel[14]. **A**: Number of cell lines per tissue type. **B**: Number of cell lines screened for each drug that we used in our experiments.



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.



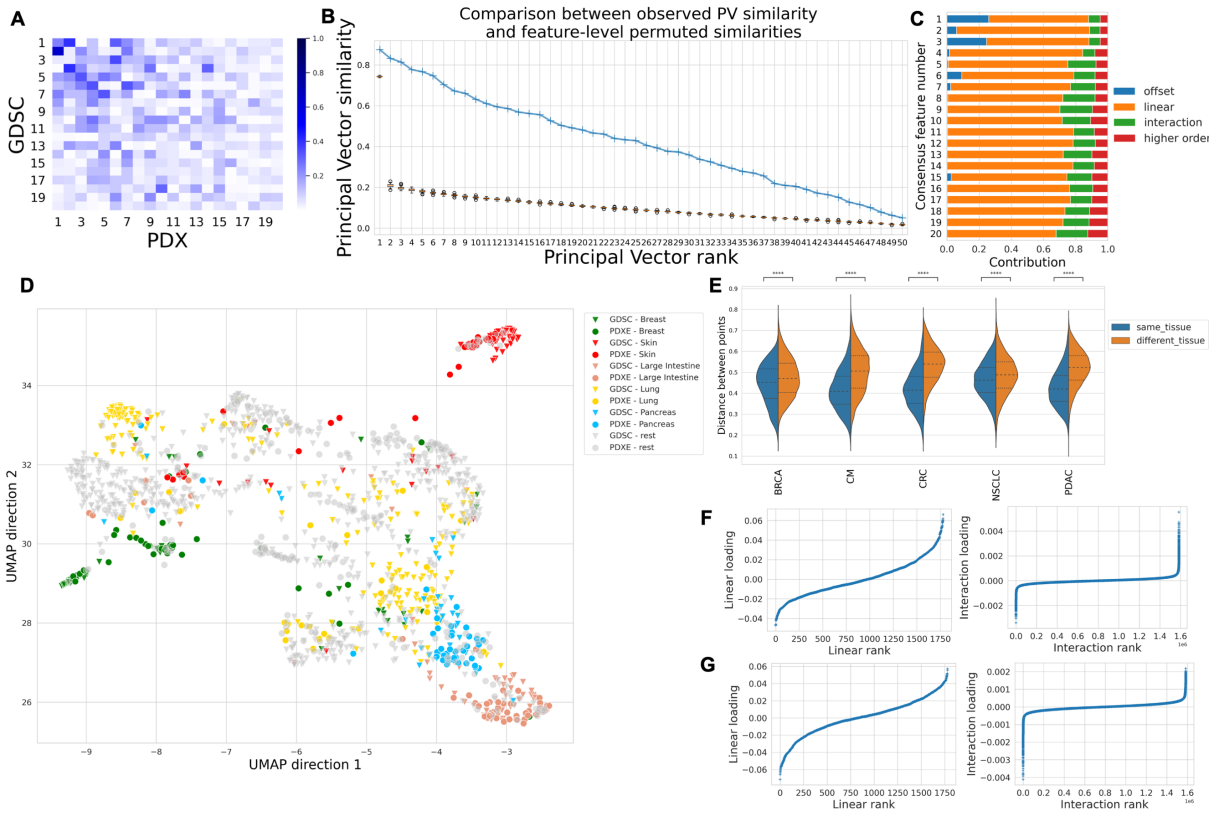
Supp Figure 4 Structure of the TCGA dataset (primary tumors). We make use of the TCGA dataset for primary tumors. **A**: Number of samples per cancer type. **B**: For each drug, number of samples with known response.



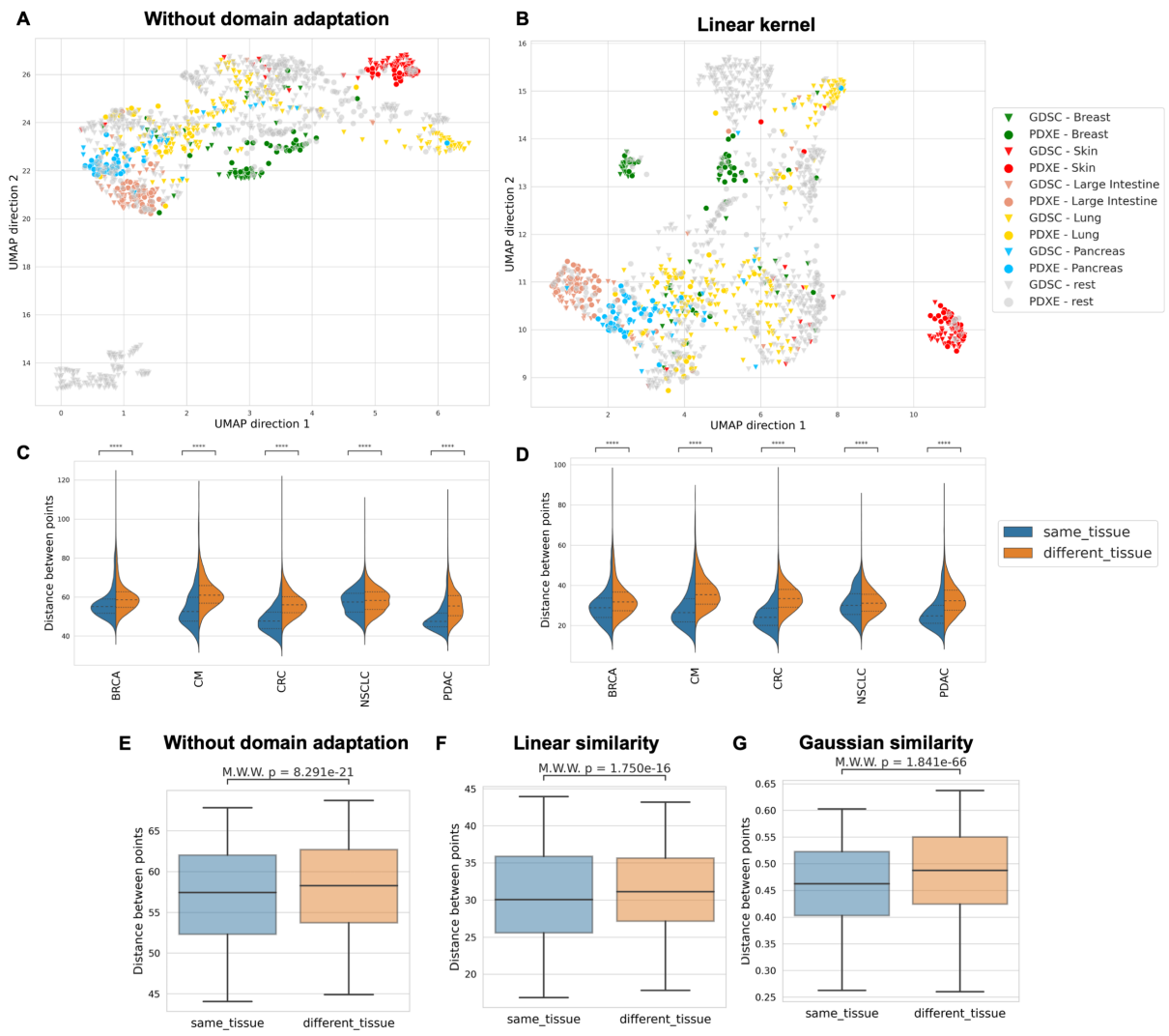
Supp Figure 5 Structure of the HMF dataset (metastatic lesions). We make use of the Hartwig Medical Foundation (HMF) dataset for metastatic lesions. (**A**) Number of samples per cancer type (primary tumor location). (**B**) For each patient, number of response measurements made. For further analysis, we considered the first response measure – i.e. first measure after treatment start. (**C**) Histogram of number of weeks

1002 between treatment start and response measurement. (**D**) For each protein coding
1003 gene, we measure the Spearman correlation between read counts obtained using
1004 Salmon and STAR alignment tools using all samples in the HMF dataset. We then
1005 ranked genes based on the obtained Spearman correlation (x-axis) and plotted it
1006 against the mean-expression of these genes obtained using Salmon (y-axis). Since
1007 lowly concordant genes tend to have low expression, we put a threshold at $corr = 0.5$
1008 and discarded genes below this threshold. (**E**) After the previous selection, we
1009 computed the sample-level Pearson and Spearman correlations between read counts
1010 obtained with STAR and Salmon. All samples but five show a correlation above 0.8 –
1011 these were discarded. We finally further restricted to genes from the mini-cancer
1012 genome.

1013

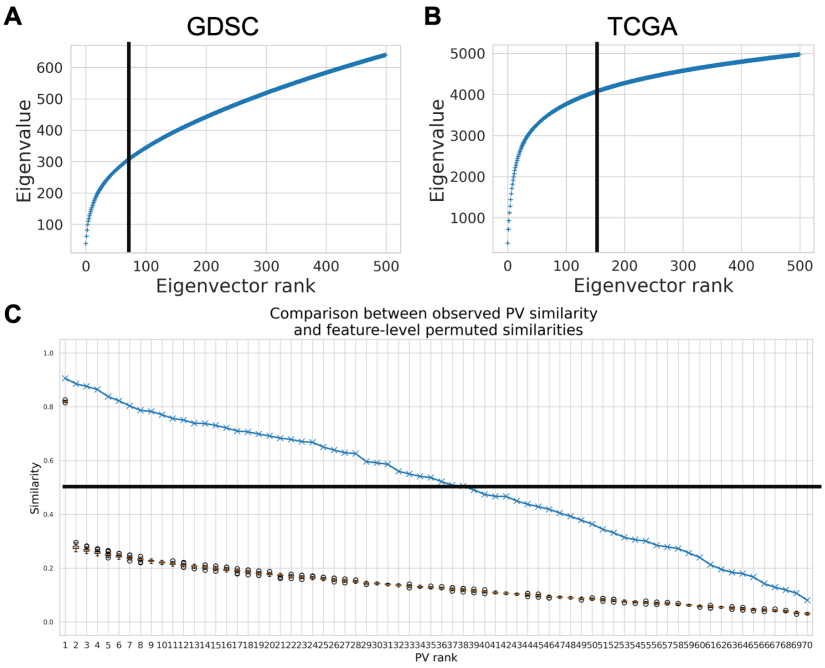


Supp Figure 6 Analysis of consensus features between cell lines (GDSC) and PDXs with $\gamma = 0.0005$. We use a Gaussian similarity matrix with hyper-parameter $\gamma = 0.0005$ and run TRANSACT. (**A**) Cosine similarity between the 20 top source and target NLPCs. (**B**) Similarity between principal vectors (blue line) alongside the similarity obtained after gene-level permutation on GDSC (boxplots). (**C**) For each consensus feature, proportion of offset, linear and interaction term. (**D**) UMAP of data projected on the consensus features, colored by tissue of origin. (**E**) For each tissue type in PDXs, we compare the distances between corresponding PDXs with cell lines from the same tissue of origin (blue), or from another tissue (orange). (**F**) For the first consensus feature, sorted contribution of each linear features (i.e. gene, left) and interaction terms (right). (**G**) For the second consensus feature, sorted contribution of each linear features (i.e. gene, left) and interaction terms (right).

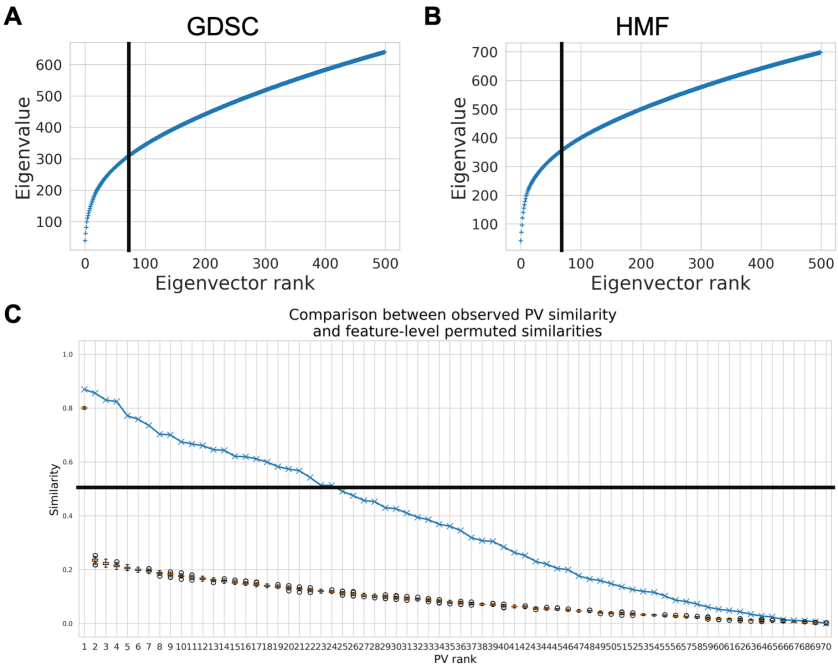


Supp Figure 7 Tissue clustering without domain adaptation and with PRECISE alignment between GDSC and PDXE. (**A**) UMAP plot of cell lines and PDXs colored by tissue type without any domain-adaptation. Data was normalized prior to performing UMAP: cell lines and PDXs were independently mean-centered and scaled to unit variance. (**B**) UMAP plot of cell lines and PDXs colored by tissue type after projection on consensus features obtained with linear PRECISE. (**C**) Comparison of distances between PDXs and cell lines from the same tissue type (blue) or from a different tissue type (orange) without domain adaptation. (**D**) Comparison of distances

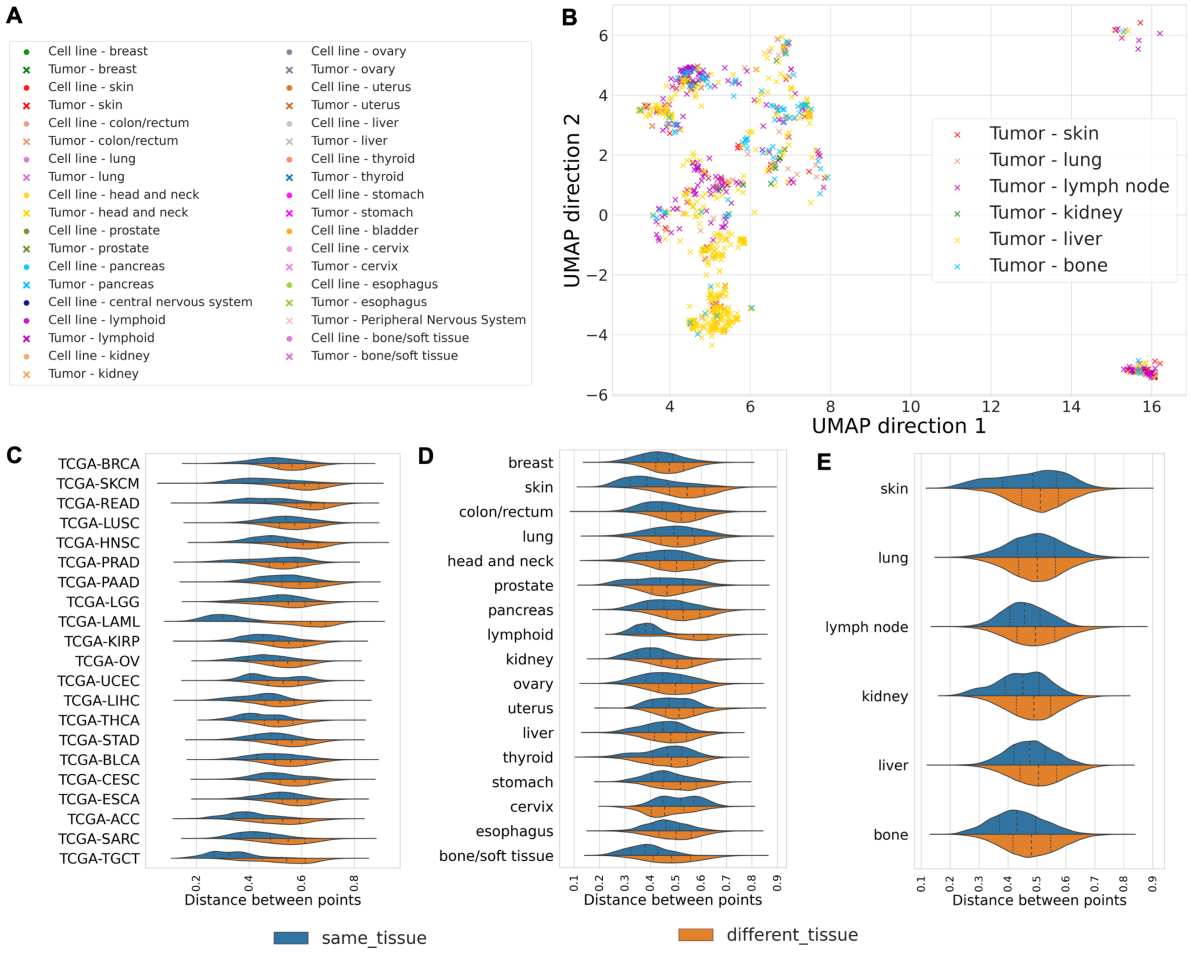
when using linear PRECISE. We zoom in on lung (NSCLC) without domain adaptation (**E**), with linear PRECISE (**F**) or with TRANSACT (**G**) using same setting as in **Supp Figure 6.**



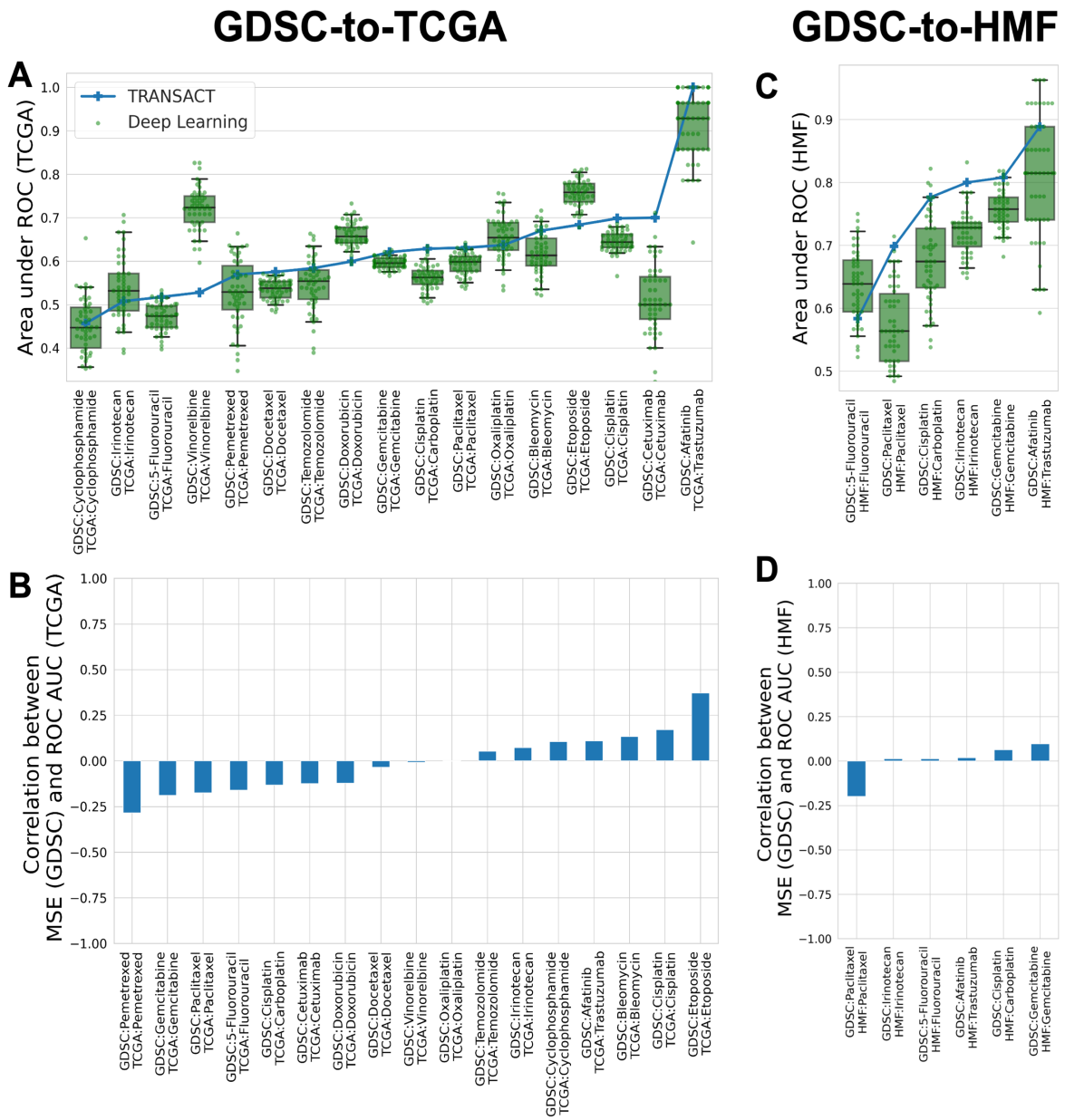
Supp Figure 8 Choice of the number of NLPCs and consensus features between GDSC and TCGA. (**A**) Cumulative sum of eigenvalues of $\widetilde{K_s}$ (GDSC) with $\gamma^* = 5 \times 10^{-4}$. The cumulative sum increases steeply, reaches an inflexion points and then follows an almost-linear behavior. We select all the NLPCs 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 similar reasoning as in (**A**), we restrict the study to the first 150 NLPCs. (**C**) Similarity between PVs 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 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 similar reasoning as in (**A**), we restrict the study to the first 75 NLPCs. (**C**) Similarity between PVs 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.

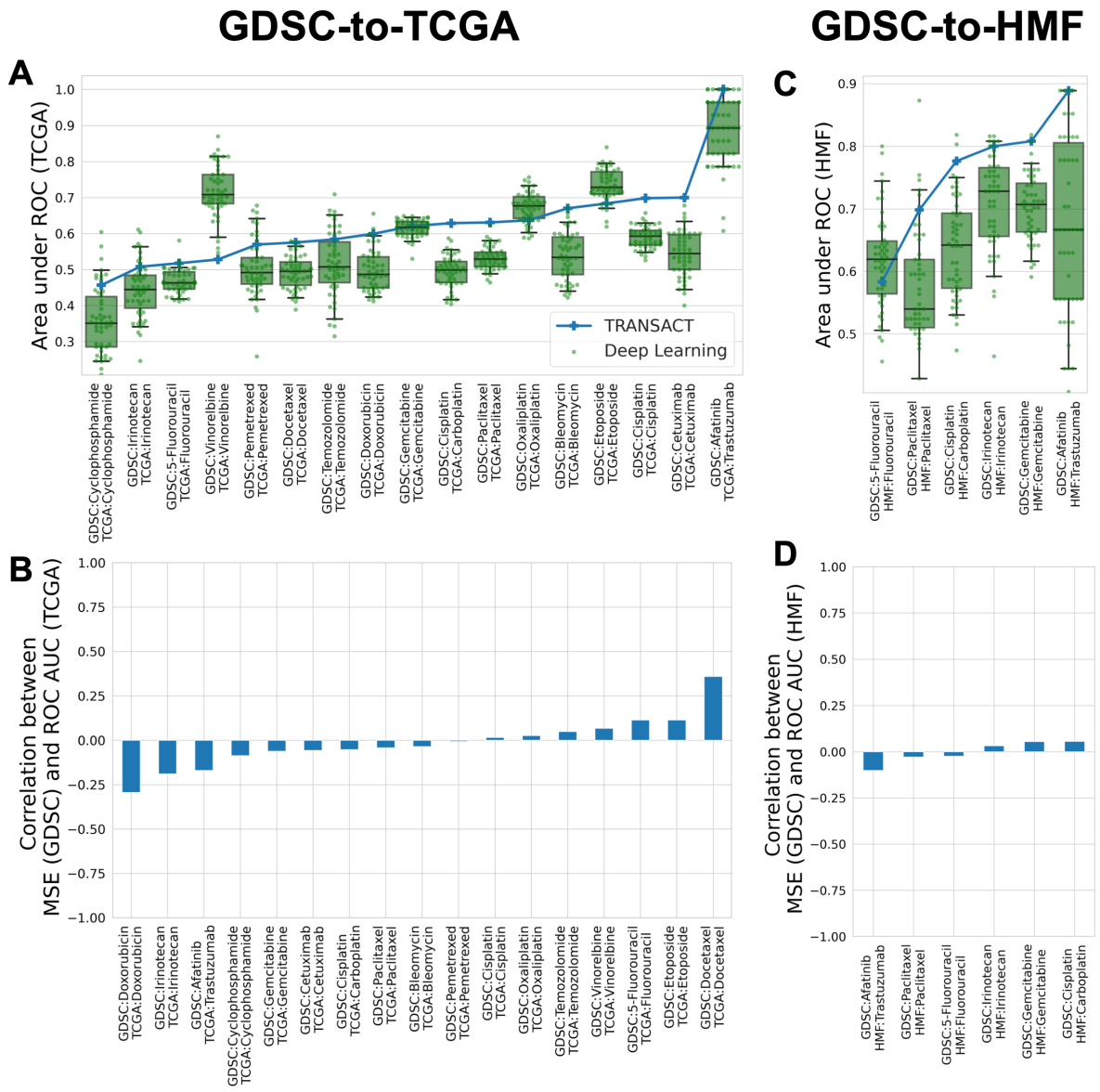


Supp Figure 10 Pan-cancer consensus features between cell lines and tumors conserve tissue type information (Supplement of Figure 3) **A**: Legend of UMAP plots for **Figure 3**C-D**. B**: UMAP plot of HMF metastatic lesions (same as Figure 3D) colored by metastatic site. **C**: In TCGA, for each tumor 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.



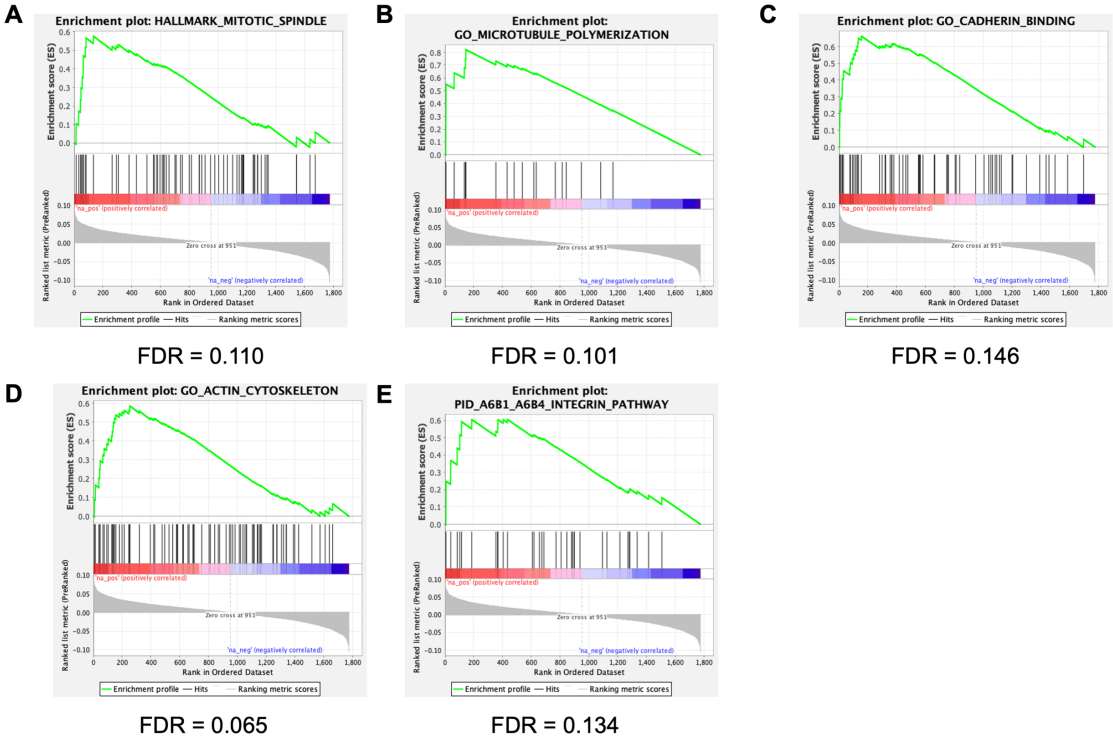
Supp Figure 11 Impact of initialization on results for the *Deep Learning (DL)* approach. For each drug on TCGA and HMF, we considered the architecture and the set of hyper-parameters with the lowest Mean Squared Error on GDSC given an initialization. We then randomly generated 50 independent initializations of the resulting networks and trained them using the GDSC data. Each of these trained networks was then employed to predict the TCGA or HMF response. The resulting prediction accuracies (area under the ROC) are plotted for the different drugs on the TCGA and HMF data. **A**: Deep Learning prediction accuracies for each drug on TCGA

(green boxplots). For comparison, the results from TRANSACT are displayed (blue curve). **B**: Pearson correlation of the Mean Square Error of the predictor on GDSC to the Area under the ROC of the same predictor on TCGA. **C**: Boxplot of performances on each HMF drug for 50 different initializations. **D**: Pearson correlation on HMF between MSE (GDSC) and Area under the ROC (HMF).



Supp Figure 12 Impact of initialization on results for the *ComBat+DL* approach. For each drug on TCGA and HMF, we considered the architecture and the set of hyper-parameters with the lowest Mean Squared Error on GDSC given an initialization. We then randomly generated 50 independent initializations of the resulting networks and trained them using the GDSC data. Each of these trained networks was then employed to predict the TCGA or HMF response. The resulting predictions accuracies (area under the ROC) are plotted for the different drugs on the TCGA and HMF data. **A**: *ComBat+DL* prediction accuracies for each drug on TCGA (green boxplots). For comparison, the results from TRANSACT are displayed (blue curve). **B**: Pearson

correlation of the Mean Square Error of the predictor on GDSC to the Area under the ROC of the same predictor on TCGA. **C**: Boxplot of performances on each HMF drug for 50 different initializations. **D**: Pearson correlation on HMF between MSE (GDSC) and Area under the ROC (HMF).



Supp Figure 13 Pathway enriched for resistant linear coefficients in GDSC-to-TCGA Gemcitabine drug response predictor. Additional pathways significantly enriched in the linear part of the GDSC-to-TCGA predictor.