A Deep Learning Framework for Prediction of Clinical Drug Response of Cancer Patients and Identification of Drug Sensitivity Biomarkers using Preclinical Samples

David Earl Hostallero[1], Lixuan Wei[2], Liewei Wang[2], Junmei Cairns[2*], and Amin Emad[1*]

[1] Department of Electrical and Computer Engineering, McGill University, Montreal, QC, Canada

[2] Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic, Rochester, MN, USA

* Corresponding Authors:

Amin Emad

755, McConnell Engineering Building, 3480 University Street, Montreal, Quebec, Canada, H3A 0E9

Email: amin.emad@mcgill.ca

Junmei Cairns

Gonda, 19-418, 1st street SW, Rochester, MN, USA, 55905

Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic

Email: carriehjm@gmail.com

ABSTRACT

Background: Prediction of the response of cancer patients to different treatments and identification of biomarkers of drug sensitivity are two major goals of individualized medicine. In this study, we developed a deep learning framework called TINDL, completely trained on preclinical cancer cell lines, to predict the response of cancer patients to different treatments. TINDL utilizes a tissue-informed normalization to account for the tissue and cancer type of the tumours and to reduce the statistical discrepancies between cell lines and patient tumours. In addition, this model identifies a small set of genes whose mRNA expression are predictive of drug response in the trained model, enabling identification of biomarkers of drug sensitivity.

Results: Using data from two large databases of cancer cell lines and cancer tumours, we showed that this model can distinguish between sensitive and resistant tumours for 10 (out of 14) drugs, outperforming various other machine learning models. In addition, our siRNA knockdown experiments on 10 genes identified by this model for one of the drugs (tamoxifen) confirmed that all of these genes significantly influence the drug sensitivity of the MCF7 cell line to this drug. In addition, genes implicated for multiple drugs pointed to shared mechanism of action among drugs and suggested several important signaling pathways.

Conclusions: In summary, this study provides a powerful deep learning framework for prediction of drug response and for identification of biomarkers of drug sensitivity in cancer.

INTRODUCTION

Cancer is one of the deadliest public health problems worldwide and cases are still rapidly growing. In 2020, it is estimated that around 10 million people have died of cancer [1]. Individualized medicine is a promising concept, which aims to improve the prognosis of patients by adapting the patient's treatment to their unique clinical and molecular characteristics. One of the main goals of individualized medicine is the prediction of the patients' response to different treatments, and identification of biomarkers that enable such predictions. High throughput sequencing technologies along with major initiatives such as The Cancer Genome Atlas (TCGA) [2] have provided a unique opportunity for machine learning (ML) algorithms to address these challenges. However, ML models and particularly deep learning (DL) approaches require a large number of samples with known drug response to train generalizable models. However, data on clinical drug response (CDR) of cancer patients, even in large databases such as TCGA, is usually small for the majority of the drugs and does not lend itself to training of DL models.

On the other hand, large databases of molecular profiles of hundreds of *in-vitro* cancer cell lines (CCLs) and their response to hundreds of drugs [3-5] have enabled development

of various ML algorithms for prediction of drug response [6-8]. Unfortunately, these models, even though accurate in predicting the drug response of held-out CCLs, usually do not generalize well to predicting the CDR of real tumours from cancer patients, and their prediction performance significantly deteriorates due to the major biological and statistical differences between CCLs and tumours [9].

Recognizing these issues, some studies have adopted to utilize tumour samples with known CDR in the training of their models, either by fully training their models on data corresponding to tumour samples [10-12], or by using them in addition to CCLs (e.g., using transfer learning [13]). However, as a result of this strategy, these studies have only been able to develop models on very few drugs due to the small samples sizes of patient cohort data with known drug response. Another strategy is to train ML models completely on preclinical CCLs but use computational approaches to overcome the statistical differences between CCLs and tumours. For example, multiple approaches [9, 14] have used batch removal methods such as ComBat [15] to reduce the discrepancy between the training CCLs and test tumours. One limitation of these methods is that ComBat is used as a preprocessing step such that the gene expression profile of both CCLs (training set) and tumours (test set) are adjusted. As a result, prediction of CDR of new cancer patients requires re-training of the model.

In this study, our goal was to develop a deep learning computational pipeline, fully trained on gene expression profile and drug response of preclinical CCLs, to 1) predict the CDR of cancer patients and 2) to identify biomarkers of drug sensitivity for a variety of cancer drugs. Motivated by Huang et al. [9], who showed that carefully incorporating information on the tissue (or cancer) type of the test samples can improve the predictive power of computational models, we developed a deep learning pipeline with tissue-informed normalization (TINDL) to achieve these goals. Unlike methods mentioned above, TINDL requires normalization of only test samples, and as a result re-training of the model is not necessary for new test samples.

The TINDL pipeline includes two phases. The first phase is responsible for prediction of CDR of cancer patients, while the second phase utilizes these predictions to identify a small number of genes that considerably contribute to the predictive ability of the model. Focusing on drugs shared between the Genomics of Drug Sensitivity in Cancer (GDSC) [3] and TCGA [2], we showed that TINDL can distinguish between the sensitive and resistant patients for 10 (out of 14) drugs, considerably improving the performance of other methods, including our previous work TG-LASSO [9]. TINDL utilizes a simple, yet effective, tissue-informed normalization to reduce the statistical discrepancies between the gene expression profile of the training and test samples. We showed that TINDL outperforms other DL-based models that try to explicitly remove these discrepancies using other techniques such as ComBat or domain adaptation [16, 17].

Focusing on tamoxifen, for which TINDL performed best, we showed that only a small panel of genes identified by TINDL can be used to predict the CDR of cancer patients. Moreover, using siRNA gene knockdown of 10 genes identified by TINDL in a breast cancer cell lines (MCF7), we showed that the knockdown of any of these genes significantly changes the response to tamoxifen. These *in-vitro* experiments further validate the TINDL pipeline and its ability to identify biomarkers of drug sensitivity.

RESULTS

Deep learning prediction of clinical drug response of cancer patients and identification of biomarkers of drug sensitivity using *in vitro* cell line data

We developed a deep learning pipeline with tissue-informed normalization (called TINDL) to 1) predict the clinical drug response (CDR) of cancer patients (test set) and 2) identify predictive biomarkers of drug sensitivity based on models completely trained on preclinical cell line data (training set). The pipeline has two major phases: the modeling phase and the gene identification phase. In the modeling phase (Figure 1a), a neural network is trained using the gene expression (GEx) profiles of cancer cell lines (CCLs) and their response to a drug (e.g., normalized ln(IC50) values in this study). The trained model was then used to predict the drug response of cancer patients based on the carefully normalized GEx profiles of their primary tumours. Details of the DL architecture are provided in Methods.

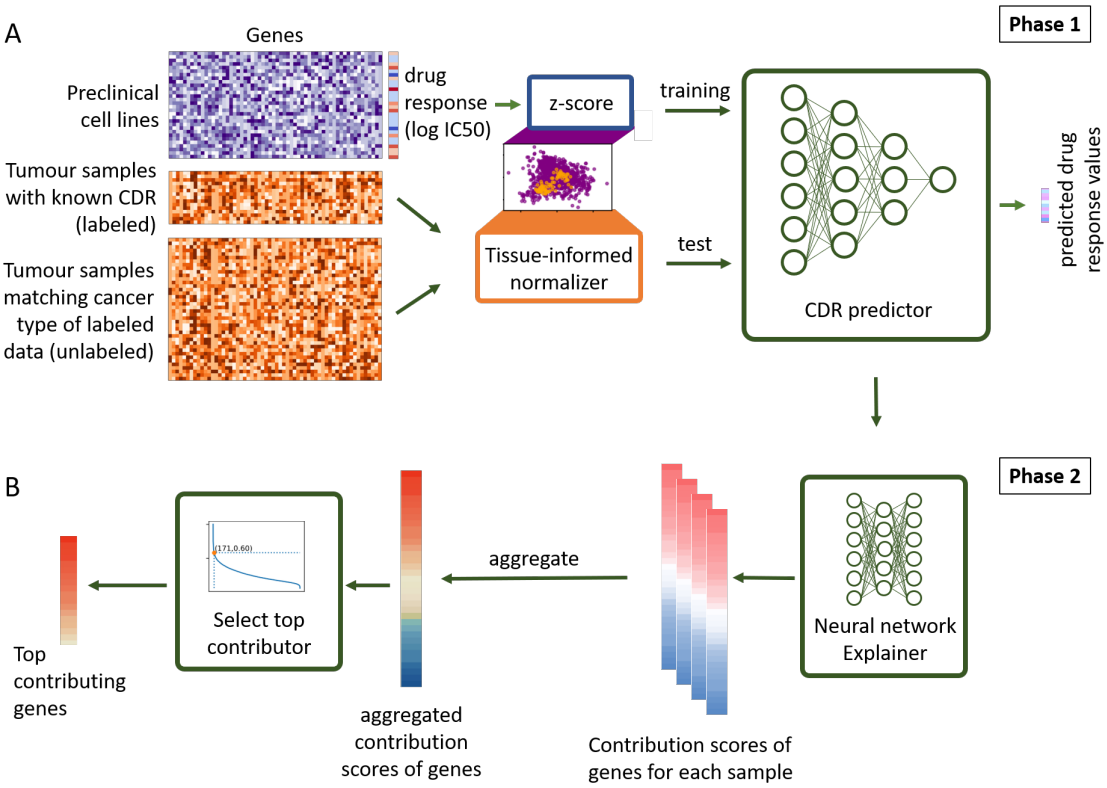

Figure 1. The pipeline used for prediction of drug responses and identification of important genes. In Phase 1, (A) the gene expression data of the cancer cell lines (CCL) and log IC50 are both z-score normalized, while the tumour gene expression (test data) is normalized using the tissue-informed normalizer. We then used the train a cancer drug response (CDR) predictor using the CCL data. After training, we predict the response value for the tumours. In Phase 2 (B), we take the trained CDR predictor and train a neural network explainer using the same training data. We used the trained explainer to give gene contribution scores for each genes of the test samples. We aggregated the scores across samples and then selected the top genes by estimating the point of maximum curvature.

We designed the normalization step of GEx profiles of patient tumours to address two important issues. First, we required this approach to remove the discrepancy between the statistical properties of GEx of CCLs and patient tumours, originating from the technical differences in protocols for measuring the data and the biological differences between pre-clinical CCLs and clinical tumours. Second, we required this approach to incorporate information on the tissue of origin (or cancer type) of tumours in the

prediction task. In a previous study [9], we showed that information on the tissue of origin of samples plays an important role in improving prediction performance, however most commonly used methods for this task are not capable of appropriately incorporating this information. For this purpose and given a drug, we first identified the set of tissues (henceforth referred to as "target tissues") of the clinical samples to whom the drug was administered. Then, we collected additional GEx profile of samples from the same target tissues, independent of what drug was used for their treatment. The GEx profile of each test sample was then normalized against this additional set of "unlabeled" data (see Methods for details).

This simple, yet effective, normalization approach used in our pipeline removes the statistical discrepancy between the test and training datasets by mapping the expression of each gene in each dataset to a distribution with unit variance and zero mean. However, since the test samples are normalized while considering the GEx of a much larger unlabeled set of samples, this normalization will not be negatively affected if the size of the test set is small (e.g., if we want to predict the drug response of a single sample). This is different from methods that perform the normalization using only the test samples. In addition, since the normalization is done independently for the training and test sets, one does not need to retrain the DL model every time the drug response of a new test sample is to be predicted (a shortcoming of our previous approach [9]).

153 The second phase of the pipeline seeks to assign a contribution score to each gene based
154 on its contribution to the trained predictive model. In this phase, we first use CXPlain [18]
155 to assign a sample-specific score to each gene. These scores are then averaged over all
156 samples (separately for each gene) and normalized to provide a final contribution score.
157 Additionally, we use the distribution of these scores to systematically identify the critical
158 point that the contribution of the genes diminishes, enabling us to narrow down the top
159 ranked list of genes for follow-up analysis (e.g., pathway enrichment analysis, gene
160 knockdown experiments, etc.). The details of this phase are provided in Methods.
161
162 **TINDL can distinguish between sensitive and resistant patients for the majority of the**
163 **drugs**
164 In order to assess the performance of TINDL in predicting CDR of cancer patients, we
165 obtained GEx profile of primary cancer tumours from the TCGA database [2]. We used
166 data corresponding to RECIST CDR of TCGA patients carefully collected in [10] and
167 identified 14 drugs that satisfy two conditions: 1) there were at least 20 patients with
168 known CDR values for each drug in TCGA database and 2) the ln(IC50) response of these
169 drugs were measured in the GDSC database. Similar to previous studies [9, 14] , we
170 transformed the CDR of these tumours into a Boolean label in which “resistant” referred
171 to patients with CDR of “stable disease” or “progressive disease” and “sensitive” referred
172 to patients with CDR of “complete response” or “partial response”. These CDR values
173 were used to evaluate the predicted drug response values using TINDL and other

algorithms but were not used for training them. The list of these 14 drugs, number of TCGA patients, and their cancer type are provided in Supplementary Table S1. Similarly, we obtained GEx profiles and ln(IC50) drug response of CCLs from different lineages from the GDSC database [3], corresponding to the 14 drugs above (See Supplementary Table S1 for number of training samples for each drug).

Following previous work in this area [9, 14], we used a one-sided Mann-Whitney U test to determine if the predicted ln(IC50) values of resistant patients for a drug are significantly higher than sensitive patients. Table 1 shows the performance of TINDL in prediction of CDR of TCGA samples using preclinical GDSC samples based on this statistical test for different drugs (also see Supplementary Table S2 for the area under the receiver operating characteristic (AUROC) values). TINDL is capable of distinguishing between resistant and sensitive patients for 10 (out of 14) drugs ($p < 0.05$, one-sided Mann-Whitney U test) with a combined p-value of 2.77 E-10 (Fisher's method).

Next, we defined a measure called precision at k^{th} percentile to determine whether patients whose predicted ln(IC50) is within the lower tail of the distribution correspond to sensitive patients (i.e. responders to the drug). For different values of k, tumours with predicted ln(IC50) in the bottom k% were predicted as sensitive and their count was used to calculate precision. Figure 2A and Supplementary Table S3 show precision at k^{th} percentile of TINDL for different values of k. These results suggest that for six drugs

(tamoxifen, etoposide, vinorelbine, cyclophosphamide, bleomycin, and cisplatin) TINDL can identify responders with a precision at kth percentile above 84% for any choice of k. The distribution of predicted CDR values for sensitive and resistant patients for these drugs are shown in Figure 2B.

Table 1: The number of TCGA samples and the performance of TINDL in predicting their CDR for 14 drugs. The first column shows the name of the drug, the second column shows the total number of clinical samples for each drug, and third and fourth columns show the number of sensitive and resistant samples, respectively. The fifth column shows the p-value of a one-sided Mann-Whitney U test to determine if TINDL can distinguish between sensitive and resistant patients. To ensure the results are not biased by the initialization of the model's parameters, TINDL was trained using ten random initializations and the mean aggregate of its prediction were used to calculate the p-values. Drugs are sorted based on their associated p-value.

Drug	**Number of clinical samples**	**Number of sensitive samples**	**Number of resistant samples**	**p-value**
Cisplatin	303	237	66	6.36E-4
Tamoxifen	20	14	6	1.14E-3
Etoposide	84	73	11	4.00E-3
Doxorubicin	100	68	32	1.42E-2
Paclitaxel	158	111	47	2.29E-2
Vinorelbine	30	23	7	2.41E-2
Oxaliplatin	54	33	21	2.41E-2
Temozolomide	95	11	84	2.94E-2
Bleomycin	52	46	6	3.41E-2
Gemcitabine	157	75	82	4.57E-2
Cyclophosphamide	101	96	5	5.60E-2
Pemetrexed	38	18	20	2.86E-1
Irinotecan	23	6	17	3.04E-1
Docetaxel	102	67	35	7.04E-1

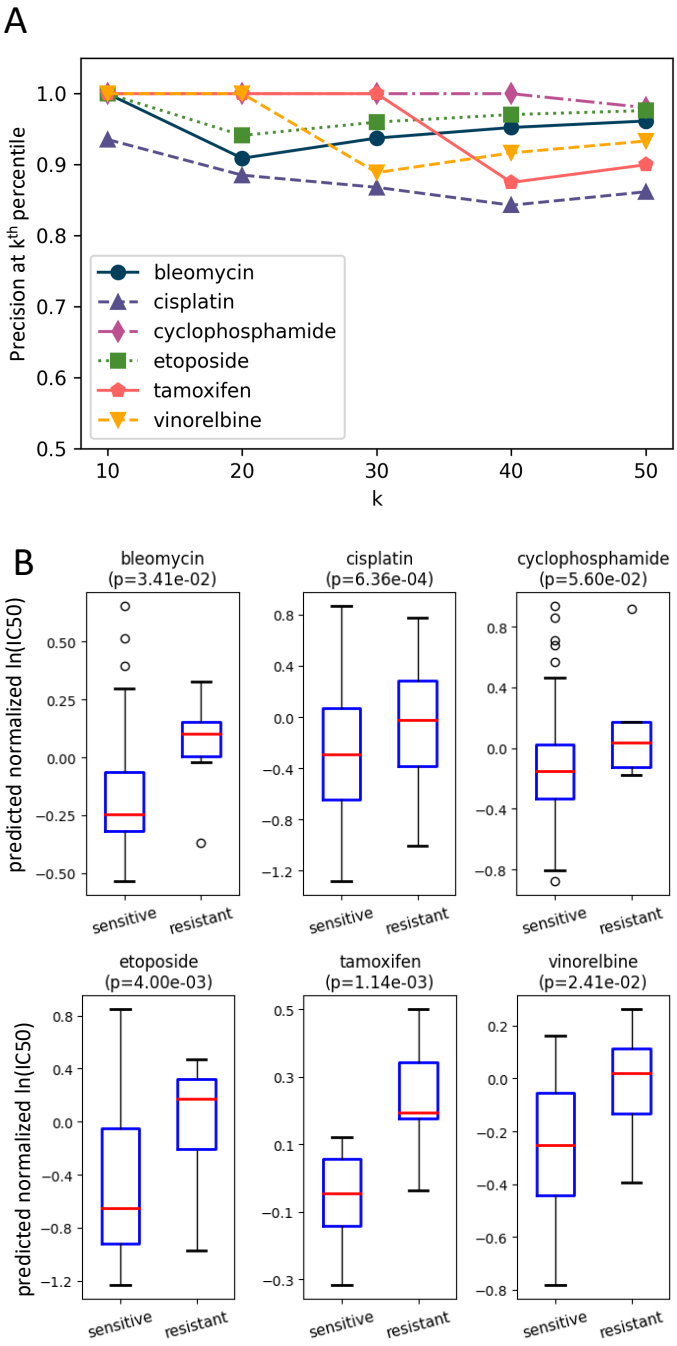

Figure 2: Performance metrics for a subset of the drugs. To prevent the figure from becoming cluttered, the results corresponding to only six drugs are shown (see Supplementary Tables S2 and S3 for performance metrics of all drugs). A) Precision at kth percentile for identification of sensitive patients. B) Distribution of predicted drug response for sensitive and resistant patients. The p-values are calculated using a one-sided Mann Whitney U test.

TINDL outperforms alternative methods in prediction of clinical drug response

Next, we sought to determine how TINDL performs against alternative computational models. For this purpose, we considered multiple traditional and state-of-the-art machine learning models [9, 14] for prediction of CDR of cancer patients from preclinical CCLs. The detailed performance measures for each drug and each model are provided in Supplementary Table S2 and the summary of the results are provided in Table 2. In this table, we used the combined p-value of 14 drugs to summarize the performance of different methods (Fisher's method for combining p-values was used).

Table 2: The performance of different computational models in predicting CDR of TCGA samples using models completely trained on preclinical GDSC CCLs. The first column shows the algorithm. The second column shows the number of drugs for which a one-sided Mann-Whitney U test showed a significant p-value. The third column shows the total number of drugs used for evaluation, and the fourth column shows the combined p-value (combined over all 14 drugs using Fisher's method).

Algorithm	Drugs with P<0.05	Drugs	Combined P (Fisher)
TINDL	10	14	2.77 E-10
LASSO	7	14	7.47 E-7
TG-LASSO, Huang et al. [9]	6	14	8.32 E-7
SVR (RBF kernel)	5	14	1.89 E-6
Geeleher, et al. [14]	4	14	5.63 E-3
Random Forests	4	14	3.12 E-3

As shown in table 2, TINDL can distinguish between sensitive and resistant patients for 10 (out of 14) drugs (with a combined p-value of 2.77E-10 for all drugs), while the second-best method in this table can only distinguish between sensitive and resistant patients for 7 drugs. Similar to our previous study [9], we also observed that Lasso and its variation,

238 TG-Lasso, perform reasonably well (when considering all drugs), while Support Vector
239 Regression and Random Forests did not perform as well.
240
241 As discussed earlier, one of the major challenges in predicting the CDR of cancer patients
242 based on ML models trained on preclinical CCLs is the statistical differences between
243 these samples. To assess the performance of TINDL against other DL models that explicitly
244 try to remove these statistical differences, we considered three alternative methods. The
245 first method (referred to as "ComBat-DL") utilizes ComBat [15] as a pre-processing step
246 to remove the statistical discrepancy between CCLs and tumour samples. ComBat [15] is
247 a popular method for removing "batch effects" in gene expression datasets and has been
248 widely used for drug response prediction [9, 14, 19] and other applications [20, 21]. The
249 ComBat-transformed GEx profiles are then used in a DL architecture similar to TINDL for
250 a fair comparison. The second and third methods (called "DANN-DL" and "ADDA-DL"
251 henceforth) are based on DANN [16] and ADDA [17] , two domain adaptation techniques
252 that were originally developed for image processing. Instead of adapting the GEx input
253 features, these methods adjust the latent feature representations learned by the
254 encoder. DANN uses adversarial neural networks to create a shared latent feature space
255 between the datasets. ADDA, on the other hand, is a unidirectional domain adaptation
256 approach that builds over a pre-trained predictor and tries to adapt the first few layers of
257 the neural network such that the target dataset's latent feature representation aligns
258 with that of the source dataset. We trained models of these methods with a similar

architecture to that of TINDL, with the exception of the discriminators, which are specific to ADDA and DANN and are used for domain adaptation. The details of these methods, including their architecture and training procedure are provided in Methods and Supplementary Methods. Table 3 and Supplementary Table S2 show the performance of these DL-based approaches. These results show that in all three cases, only for 7 (out of 14) drugs the predicted normalized ln(IC5) of sensitive patients is significantly smaller than resistant patients.

Table 3: The performance of three deep learning-based methods that explicitly try to remove discrepancy between preclinical training and clinical test datasets. The first column shows the name of the algorithm. The second column shows the number of drugs for which a one-sided Mann-Whitney U test showed a significant p-value. The third column shows the total number of drugs used for evaluation, and the fourth column shows the combined p-value (combined over all 14 drugs using Fisher's method). To ensure a fair comparison, a similar architecture to TINDL was used for all these methods. Additionally, each model was trained using ten random initializations and the mean aggregate of these predictions were used for calculating the p-values.

Algorithm	**Drugs with P<0.05**	**Drugs**	**Combined P (Fisher)**
ComBat-DL	7	14	6.73E-10
ADDA-DL	7	14	2.16E-07
DANN-DL	7	14	1.66E-06

To assess the superior performance of TINDL compared to these DL-based models that use an architecture similar to TINDL, we assessed their ability in removing the discrepancy between preclinical and clinical samples. For this purpose, we assessed the distance of clinical samples and preclinical samples for each method and each drug (see Methods for details of calculating distances). Since methods that use domain adaptation do not modify the input features, but rather seek to remove the domain discrepancies in the latent

space (the output of the encoder), we used the learned representation of each sample in the latent space for all methods. Using a one-sided Wilcoxon signed rank test, we observed that TINDL's learned representations for clinical samples have a significantly smaller average distance to preclinical samples compared to ComBat-DL ($p = 6.10E-5$), ADDA-DL ($p = 4.27E-4$), and DANN-DL ($p = 6.10E-5$), for all drugs (Figure 3A). The effectiveness of tissue-informed normalization of TINDL in removing the statistical discrepancy between the preclinical and clinical embeddings can also be visually observed using principal component analysis and UMAP plots of the representations learned by each method (Figure 3B and Supplementary Figures S1-S4).

Next, we sought to determine whether the latent space representation similarity has an influence on drug response prediction performance of TINDL across different drugs. We observed a negative Spearman's rank correlation ($r = -0.17$, $p = 3.93$ E-2) between the aforementioned distances and the AUROC of prediction for different drugs. In particular, tamoxifen that had the highest AUROC (Supplementary Table S2, AUROC=0.92), also had the smallest average distance between clinical and pre-clinical representations of its samples among all drugs in TINDL. These results further support the conclusion that reducing the discrepancy between the statistical characteristics of clinical and preclinical samples plays an important role in the success of TINDL in the prediction of CDR.

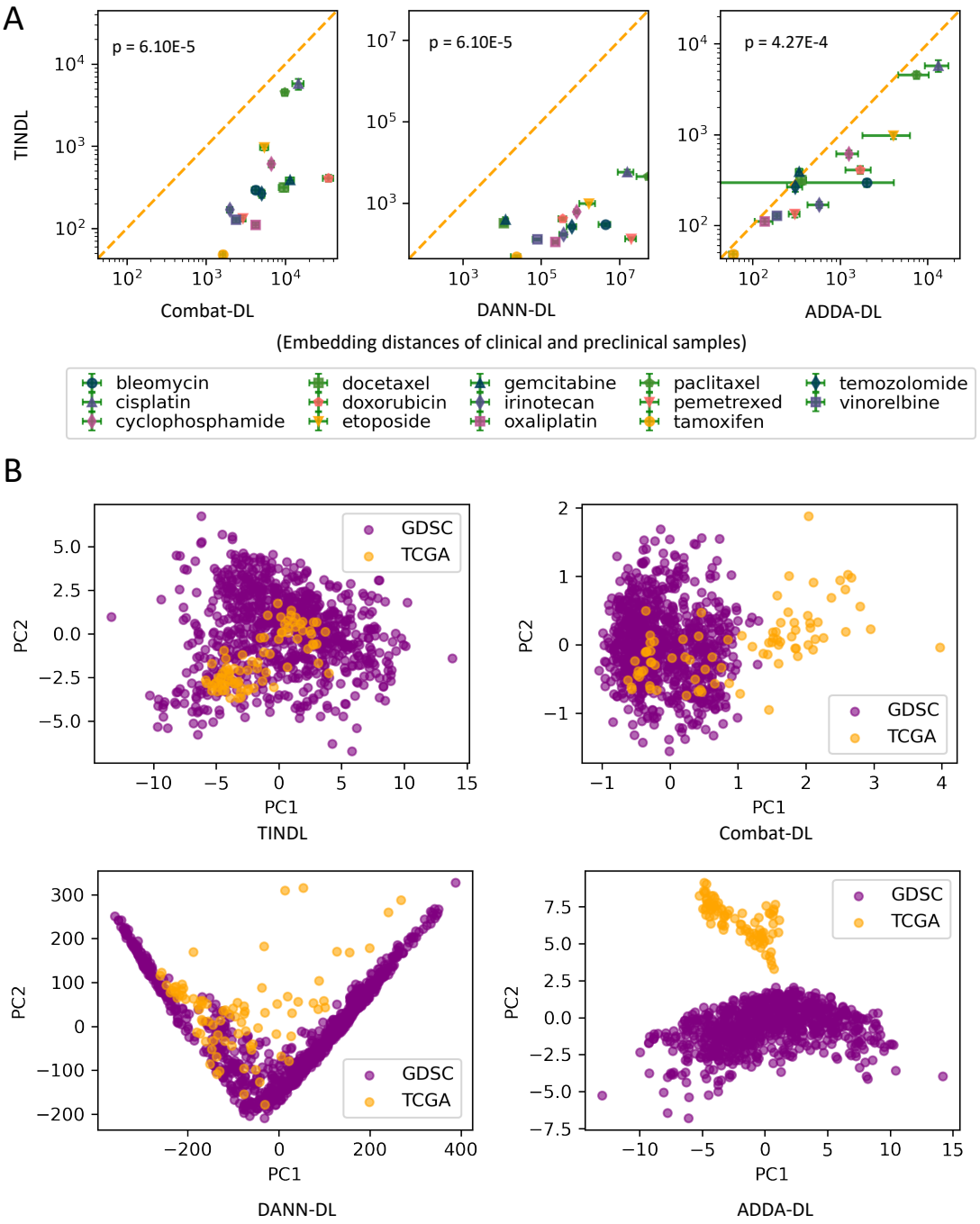

Figure 3: Evaluation of the embeddings used by TINDL and other deep learning methods used for prediction of drug response. A) The scatter plots compare the distance between preclinical samples and clinical samples in the embedding space for each drug. Each point in the scatter plot corresponds to a different drug. The p-values are calculated using a one-sided Wilcoxon signed rank test. The error bars show the 95% confidence intervals and are calculated based on ten runs of each method with random initializations. B) The PCA analysis of the embeddings used by each method to predict the response to etoposide. Visually, the TCGA samples are better mixed (i.e. are not easily separable) with GDSC samples in TINDL compared to other methods.

TINDL identifies biomarkers of drug sensitivity

We used TINDL (Figure 1B) to assign a score to the contribution of each gene in the trained model (see Methods for details). Supplementary Figure S5 shows the distribution of these scores for each drug. To identify the threshold below which the contribution of the genes to the predictive model is small, we used a method called kneedle [22], which systematically determines this threshold for each drug based on the distribution of the scores. This method identified between 64 (for pemetrexed) to 243 (for bleomycin) genes, depending on the drug. The ranked list of genes identified by TINDL that pass this drug-specific threshold are provided in Supplementary Table S4.

Next, we sought to determine whether the identified genes are drug specific. To this end, we calculated the Jaccard similarity coefficient of drug pairs (Figure 4A). The results revealed a high degree of drug specificity with the average Jaccard similarity coefficient for all drugs equal to only 0.027. However, some genes were implicated for multiple drugs (Figure 4B and Supplementary Table S5). In particular, SLFN11 was implicated for nine drugs and was the top contributor for bleomycin, cisplatin, doxorubicin, etoposide, gemcitabine, and irinotecan, and the top third contributor for oxaliplatin. SLFN11 (Schlafen family member 11) is a putative DNA/RNA helicase that is recruited to the stressed replication fork and inhibits DNA replication. DNA replication is one of the fundamental biological processes in which dysregulation can cause genome instability [23]. This instability is one of the hallmarks of cancer and confers genetic diversity during

329 tumorigenesis [24, 25]. Various studies have shown that the expression of this gene
330 sensitizes cancer cells to many chemotherapeutic agents including cisplatin, oxaliplatin,
331 irinotecan, gemcitabine, doxorubicin, and etoposide [26-30]. Epigenetically mediated
332 suppression of SLFN11 via EZH2 contributes to acquired chemotherapy resistance, one
333 that can be prevented and/or actively remodeled through targeting EZH2 [31]. Several
334 potent and selective EZH2 inhibitors are now in different stages of clinical development
335 with promising safety profile, including phase II (Epizyme) and phase I (Constellation, GSK)
336 trials in multiple solid tumor and hematological indications. Our data supports the notion
337 that the combination of downregulating SLFN11 via EZH2 inhibitor with
338 chemotherapeutic reagents should be considered in multiple cancer types [32, 33].
339
340 To better understand the functional characteristics of genes implicated by TINDL for
341 multiple drugs, we used KnowEnG's gene set characterization pipeline [34] to identify
342 pathways associated with 29 genes identified by TINDL for at least 4 drugs (Figure 4B).
343 This pipeline enables identification of associated pathways while incorporating
344 interactions among genes and their protein products through network-guided analysis.
345 The results (Supplementary Table S5) implicated five pathways, which included
346 "Regulation of toll-like receptor signaling pathway", "Alpha-synuclein signaling", "Arf6
347 trafficking events", "Insulin Pathway", and "RalA downstream regulated genes". Innate
348 immune receptors such as toll-like receptors (TLRs) are responsible for recognizing
349 molecular patterns associated with pathogens and provide critical molecular links

between innate cells and adaptive immune responses. Engagement of TLRs on dendritic cells (DCs) promotes cross-talk between the innate and the adoptive immune system, maturation and migration of DCs into lymph nodes leading to activation, proliferation and survival of tumour antigen-specific naïve CD4+ and CD8+ T cells [35]. Tumour cells themselves do not express molecules which would induce DC maturation, thus application of TLR agonists is an important element of immunotherapy protocols aiming T cell activation [36]. In addition, TLR agonists have been proposed as adjuvants for cancer vaccines [37]. TLR3 agonist as an adjuvant with conventional chemotherapy can break tolerogenic or immunosuppressive effects generated by the tumour and drive T cell responses and tumor rejection [38-41].

Alpha-synuclein (α-syn) is a neuronal protein responsible for regulating synaptic vesicle trafficking. α-syn is frequently expressed in various brain tumours and melanoma [42, 43] and its upregulation has been linked to aggressive phenotypes of meningiomas [44]. Moreover, loss of α-syn results in dysregulation of iron metabolism and suppression of melanoma tumour growth [45]. Oncogenic activation of synuclein contributes to the cancer development by promoting tumor cell survival via activation of JNK/caspase apoptosis pathway and ERK and by providing resistance to certain chemotherapeutic drugs [46, 47], suggesting synuclein as a new therapeutic target for future treatment to overcome resistance to certain chemotherapeutic. ARF6 (ADP-ribosylation factor 6) governs the trafficking of bioactive cargos to tumor-derived microvesicles (TMVs) which

371 comprise a class of extracellular vesicles released from tumor cells that facilitate
372 communication between the tumor and the surrounding microenvironment [48]. Invasive
373 tumor cells shed TMVs containing bioactive cargo and utilize TMVs to degrade
374 extracellular matrix during cell invasion [49]. Indeed, several studies have suggested a
375 correlation between expression of ARF6 and invasion and metastasis of multiple cancers
376 [50-52], suggesting that antagonistic ARF6 signaling can dictate TMV shedding and the
377 overall mode of invasion. Insulin, a signaling molecule that controls systemic metabolic
378 homeostasis, can be seen as enabling tumour development by providing a mechanism for
379 PI3K activation and enhanced glucose uptake [53-58] and plays a role in cytotoxic therapy
380 response [59]. RalA (RAS Like Proto-Oncogene A) is a member of the Ral family, and the
381 RalA pathway contributes to anchorage independent growth, tumorigenicity, migration
382 and metastasis [60-64]. In conclusion, the link between genes implicated for multiple
383 drugs and the pathways above that play different roles in cancer may point to shared
384 mechanisms of action among different anti-cancer drugs. We also performed a similar
385 pathway enrichment analysis for genes implicated for each drug separately and the
386 results are provided in Supplementary Table S6.
387
388

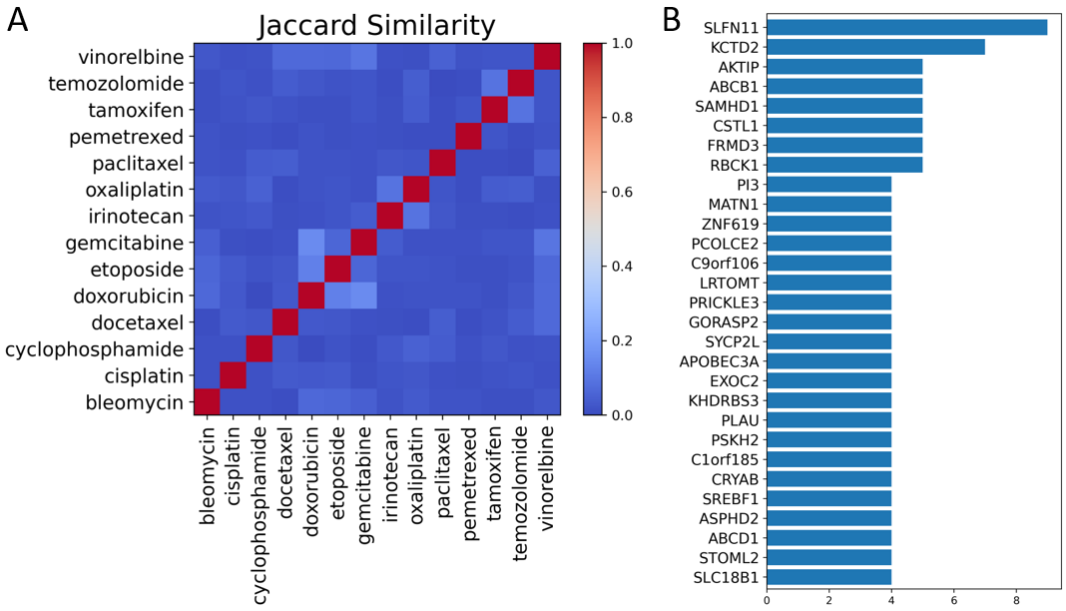

Figure 4: Genes identified by TINDL for different drugs. A) Heatmap of the Jaccard similarity of the selected top genes in the 14 drugs. B) Number of drugs in which the genes were identified as a top gene. Only genes that were present in the top genes of at least four drugs are included.

Functional validations confirm the role of TINDL-identified genes in response to tamoxifen

We sought to evaluate the drug response predictive ability of top identified genes by TINDL, both computationally and experimentally. We focused on tamoxifen due to the good prediction performance of TINDL for this drug (AUROC=0.92, p=1.14E-3 for Mann Whitney U test). First, using only top implicated genes for this drug (n = 136 based on the threshold identified by kneedle), we observed a consistently high value of AUROC and a significant Mann-Whitney U test p-value (Figure 5A, AUROC = 0.89, p =2.32 E-3). Next, we reduced the number of genes for the model to only top twenty and observed that AUROC

403 remains high even with this small number of genes (Figure 5A, AUROC = 0.90, p = 1.65 E-
404 3). This shows that even a small panel of twenty genes can be used to predict the CDR of
405 this drug, suggesting potential clinical applications in precision medicine for these small
406 panels of genes.
407
408 Next, we set out to determine whether genes identified by TINDL as predictive of
409 tamoxifen response could be associated *in-vitro* to significant changes in drug sensitivity.
410 We selected 10 genes identified by TINDL, which included the top 9 ranked genes (RPP25,
411 EMP1, EXTL3, EXOC2, NUP37, RPL13, WBP2NL, RPS6, GBP1) as well as the gene ranked as
412 19 (JAK2), due to its involvement with the type II interferon signaling pathway, an
413 important pathway in cancer [65]. We used estrogen receptor positive breast cancer cell
414 line, MCF7, since tamoxifen has most often been used as the treatment for estrogen
415 receptor positive breast cancer patients in general and 85% of patients in our test dataset
416 for this drug corresponded to breast cancer. We measured the dose-response values of
417 tamoxifen in MCF7 cell line for these ten genes using Cyquant assay which provides an
418 accurate measure of cell number based on DNA content [66-68]. We defined
419 "significance" as a gene knockdown with a significant change in apparent IC50 in
420 comparison with a negative control siRNA. Knockdown of all ten genes with specific
421 siRNAs had a significant effect on tamoxifen sensitivity in MCF7 cell line ($p<0.0001$, extra
422 sum-of-squares F test), validating 100% of tested genes in this cell line (Figure 5B, Table
423 4, Supplementary Figure S6). Taken together, through the functional validation in

estrogen receptor positive breast cancer cells, we found that the expression of ten genes, RPP25, EMP1, EXTL3, EXOC2, NUP37, RPL13, WBP2NL, RPS6, GBP1, and JAK2, were involved in tamoxifen-induced response. The percentage of variation in the IC50 of breast cancer cells that was explained by the variation of these ten genes' expression is provided in Table 4.

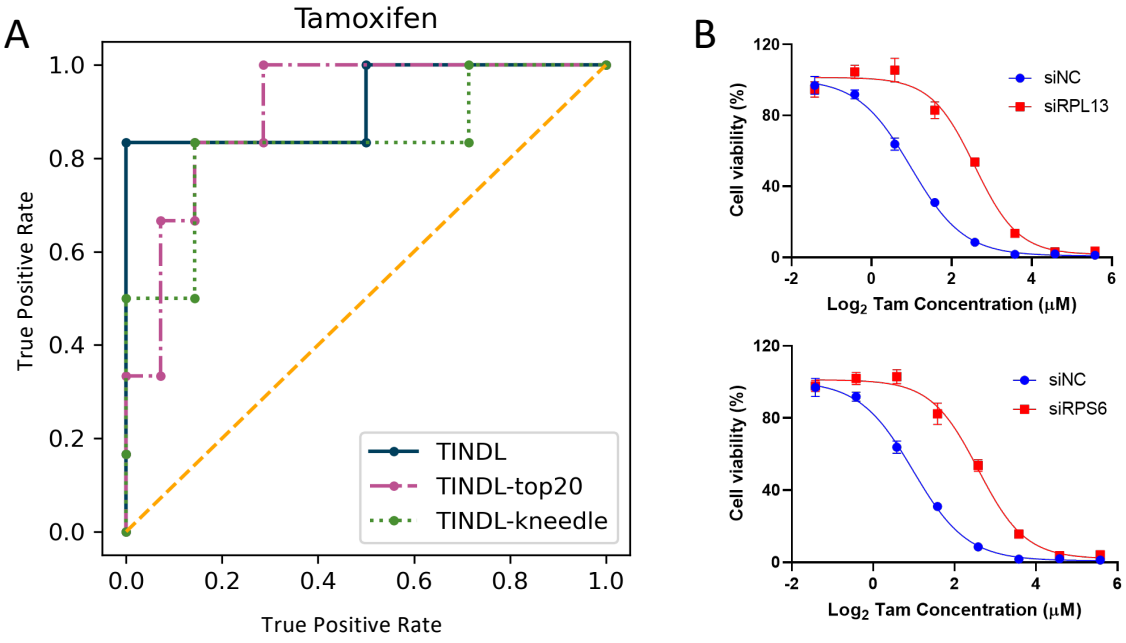

Figure 5: Top genes identified for tamoxifen and their functional validation. A) The ROC curve for tamoxifen, when different number of genes are used for CDR prediction. TINDL utilizes the GEx values of all genes (AUROC = 0.92), while TINDL-top20 (AUROC = 0.90) and TINDL-kneedle (AUROC = 0.83) assign a value of 0 to all genes except for top 20 and top genes identified by kneedle, respectively. B) Tamoxifen dose-response curves corresponding to the siRNA knockdown of RPS6 and RPL13 in MCF7 cells. The dose response curves for all genes are provided in Supplementary Figure S6. The p-values are calculated using an extra sum-of-squares F test.

Table 4: The result of siRNA gene knockdown experiments in MCF7 cell line for 10 genes identified by TINDL for tamoxifen. The p-values are calculated using an extra sum-of-squares F test. Genes are sorted based on their rank by TINDL.

Gene	Rank by TINDL	MCF7	
		p-value	Change in IC50
RPP25	1	<0.0001	146%
EMP1	2	<0.0001	69%
EXTL3	3	<0.0001	101%
EXOC2	4	<0.0001	89%
NUP37	5	<0.0001	83%
RPL13	6	<0.0001	201%
WBP2NL	7	<0.0001	113%
RPS6	8	<0.0001	202%
GBP1	9	<0.0001	113%
JAK2	19	<0.0001	134%

DISCUSSION:

Predicting the response of an individual to cancer treatments and identification of predictive biomarkers of drug sensitivity are two major goals of individualized medicine. Computational models that can achieve these goals based on preclinical *in-vitro* data can make a significant impact, due to the significant ease of preclinical data generation and data collection compared to clinical samples. This is particularly important for newly developed or newly approved drugs, for which clinical samples may be very limited or non-existent. However, the biological and statistical differences between cancer cell lines and patient tumours, make this task challenging. In a recent study [9], we assessed the ability of a wide range of machine learning models trained on preclinical CCLs, including those that incorporate auxiliary information such as gene interaction networks, in

454 predicting the CDR of cancer patients. Our analysis confirmed the difficulty of this task
455 and emphasized the importance of carefully designing advanced computational
456 techniques.
457
458 In this study, we developed TINDL, and showed substantial improvement compared to
459 state-of-the-art machine learning models (based on both traditional and deep learning
460 techniques) (Figure 1). Our results showed the importance of removing the statistical
461 discrepancies between preclinical and clinical samples, as well as incorporating the cancer
462 type and tissue of origin of the tumour samples. TINDL is not simply a drug response
463 predictor, but rather allows identification of most predictive biomarkers for each drug.
464 The biomarkers identified by multiple drugs (Figure 4B) suggested important genes and
465 signaling pathways that may play important roles in the mechanism of action of different
466 drugs in cancer. Many genes identified during our study have been reported to have
467 altered levels of expression in response to a given drug, especially SLFN11 for multiple
468 chemotherapies [26-30], SALL4 for cisplatin [69], ABCB1 for taxane and doxorubicin [70,
469 71], PIGB for gemcitabine [72], and BAX to oxaliplatin [73]. These results suggest that our
470 preclinical-to-clinical model could generate biologically relevant candidate genes and
471 pathways for understanding mechanisms underlying drug resistance, and may offer
472 additional combinational therapeutic strategies to overcome certain drug resistance.
473

Focusing on tamoxifen, we were able to show that only a small panel of 20 genes can preserve the predictive performance of TINDL for this drug (Figure 5A). Moreover, functional validation of 10 of these genes identified by TINDL using siRNA knockdown performed with MCF-7 estrogen receptor positive breast cancer cells, confirmed the direct role of these genes in response to tamoxifen (Figure 5B and Supplementary Figure S6). These results suggest that, like many complex traits, response to tamoxifen also involves multiple genes in different pathways. In addition, these results provided us with new insights into novel mechanisms in tamoxifen response. For example, among these ten genes, RPS6 is the canonical substrate of S6 kinase (S6K), which is activated by integrin engagement and inactivated by detachment. Abnormal expression of RPS6 has been indicated as a critical trigger for detachment-induced keratinization related to breast cancer development [74]. Indeed, the prognostic value of RPS6 was assessed by Kaplan-Maier Plotter analysis of gene expression data from estrogen receptor positive/HER2 negative breast tumor samples of 686 patients. High expression of RPS6 was associated with better relapse-free survival (RFS) in this cohort of patients (Supplementary Figure S7A). Decreased phosphorylation of RPS6 was previously observed in tamoxifen resistant breast cancer cells compared to parental cells [75]. However, to the best of our knowledge, no previous study has linked RPS6 to tamoxifen sensitivity. The fact that we found that RPS6 expression can predict tamoxifen sensitivity and that knockdown of RPS6 desensitized breast cancer cells to tamoxifen exposure by two folds suggests a potential role for RPS6 in the estrogen response pathway, in addition to its role as a protein

synthesis regulator. In addition to its prognostic value, further analysis revealed that high mRNA expression of RPS6 was also remarkably associated with prolonged RFS in tamoxifen treated patients (Supplementary Figure S6B). This hypothesis will need to be tested further in future experiments. The second gene that influenced tamoxifen response the most was RPL13, also known as "Ribosomal Protein L13". RPL13 is a component of the 60S ribosomal subunit that expressed at significantly higher levels in benign breast lesions than in breast carcinomas [76], however, to the best of our knowledge, no previous study has linked RPL13 to estrogen signaling or tamoxifen response. Kaplan-Meier analysis revealed that patients with high expression of RPL13 had a significantly longer RFS than those with low RPL13 expression (Supplementary Figure S7C). Our observations here suggest an important role of RPL13 expression level in predicting tamoxifen sensitivity, and could help identify additional drug targets or treatment options to overcome tamoxifen resistance.

Our analysis suggested that TINDL performs better than DL-based domain adaptation techniques in removing the discrepancies between the preclinical and clinical samples. However, these domain adaptation techniques were originally developed for the task of analyzing images. We posit that novel domain adaptation techniques may be able to overcome the shortcoming of current techniques and improve the results. However, such methods need to be carefully designed for the analysis of gene expression data and must take into account biological factors that influence the response of cancer patients to

different drugs. In addition, including information on the cancer type or even subtypes of each cancer may be necessary to achieve better results.

Another important consideration is that due to the limitation of CCLs in mimicking patient tumours (e.g., their growth in 2D environment, being more homogenous than tumours, and not being able to capture the effect of tumour microenvironment, etc.), computational models trained on CCLs are limited in their ability to predict CDR of cancer patients, even if they remove the statistical discrepancies of the training and test sets. As a result, availability of large datasets, pertaining to better models of cancer, such as patient-derived organoids or xenografts play an important role in improving the predictive ability of computational models.

In this study, our focus was on models trained only on gene expression profiles of samples. However, a multi-omics approach that incorporates different molecular characteristics of samples may provide a more complete understanding of the mechanisms of drug response in cancer. Such models, however, need to be carefully designed to avoid over-fitting due to the additional number of features. Another limitation of this study was that all the computational models were trained on CCLs and their response to single drugs. However, some of the patients in the TCGA dataset have received multiple drugs in the course of their treatment, which we had to include in the analysis due to the small number of samples with known CDR. In such cases, any computational model trained on single

drugs can only provide an approximation. To improve the prediction performance in such cases, a computational model must also consider the synergistic and antagonistic effects of the drugs. Recent large publicly available datasets such as DrugComb [77] and DrugCombDB [78] that contain response of different cell lines to pairs of drugs provide an opportunity for developing such methods, a direction that we will pursue in the future.

METHODS:

Datasets:

We used the publicly available data from GDSC and TCGA for training and testing, respectively. For training data, we used the RMA-normalized gene expression data in GDSC, which contains 15650 genes and 958 unique cell lines. For the test data, we used RNAseq (in FPKM) from primary tumors in TCGA, which we transformed using log(FPKM+0.1). We z-score normalized the gene expression data as well as the ln(IC50) values. We obtained clinical drug response of cancer patients from the supplementary file of Ding et al. [10]. Since the number of samples with known drug response in TCGA is relatively small, in our analysis we also included samples that have received multiple drugs in their course of treatment. We only focused on drugs which are common to both datasets and have at least 20 samples with known CDR in TCGA. We used a tissue-informed normalization, which is detailed below. Furthermore, we re-categorized the clinical drug responses to sensitive (corresponding to complete and partial response) and

resistant (corresponding to stable disease and clinically progressive disease). Details on sample counts and tissue types per drug are in Supplementary Table S1.

Tissue-informed Normalization

TINDL trains a separate model for each drug. Each model performs a separate normalization on the GEx profiles of test samples from TCGA to account for the cancer type and tissue of origin of the samples. First, for each drug D the set of tissues/cancer types to which this drug is administered in the TCGA samples is identified (referred to as T_D). All samples corresponding to T_D (excluding those used in the test set) are collected from TCGA, forming the unlabeled dataset. Then, the gene-wise mean (μ_{T_D}) and standard deviation (σ_{T_D}) of these unlabeled samples are calculated and used to normalize labeled test samples corresponding to drug D. More specifically, for a gene i of an arbitrary sample in the test set, the normalized value x_i would be:

$$x_i = \frac{\tilde{x}_i - \mu_{i,T_D}}{\sigma_{i,T_D}},$$

where $\tilde{x}_i$ is the log-transformed expression for gene i of that sample. The test samples are then used as input to the trained model to predict the normalized ln(IC50)s, which were compared to the actual CDR categories for evaluation.

TINDL Architecture, hyperparameter selection and training

We used grid-search and 5-fold cross validation to select the number of epochs, batch size and learning rate of all our DL-based models (including TINDL). Specific hyperparameters chosen using this procedure for TINDL are provided in the Supplementary table S7. We only used the training data corresponding to CCLs (from GDSC) to perform the hyperparameter search. In addition to the input layer (which contained one node for each gene), we used three hidden layers with dense connections, each with 512, 256, and 128 hidden nodes, in the order of their distance to the input layer. We used a rectified linear units (ReLU) activation function and added a dropout layer with 0.2 probability of dropping out prior to the output layer.

Models were trained using mean squared error (MSE) as the loss function, and the normalized ln(IC50) as the labels. During hyperparameter tuning, models were allowed to train up to a maximum of 1000 epochs, but early stopping was applied when the model's loss did not decrease after 30 epochs. After hyperparameter tuning, we retrained a final model using all the labeled CCL samples. We used 10 different random initializations (i.e., seeds) to ensure robustness of the results. A similar technique was used for ADDA-DL, DANN-DL, and ComBat-DL.

Calculating contribution scores of genes

In the second phase of TINDL (Figure 1B) we used CXPlain [18] as the explainer to assign a contribution score to each gene in each sample. CXPlain is a method that attempts to

provide causal explanations of a trained model's predictions. This is achieved by training a separate model (called "explainer") using the outputs of the trained model (called "predictor"). This method utilizes Granger's causality [79] to evaluate the contribution of a single feature (gene in our case) by zeroing out features one by one and calculating the normalized difference of the predictor's original error and its error when the feature is zeroed out. In our case, we define error as $\varepsilon_X = (y_X - \hat{y}_X)^2$, where y_X is the true value and $\hat{y}_X$ is the output of the predictor for sample $X = \{x_1, \dots, x_p\}$, p being the number of features. Prior to training the explainer, the real contribution vectors, $\Omega_X = \{\omega_1(X), \dots, \omega_p(X)\}$, are calculated for each training sample as follows:

$$\omega_i(X) = \frac{\Delta\varepsilon_{X,i}}{\sum_{j=1\dots p} \Delta\varepsilon_{X,j}},$$

where $\Delta\varepsilon_{X,i} = \varepsilon_{X\setminus\{i\}} - \varepsilon_X$. Here, $\varepsilon_{X\setminus\{i\}}$ denotes the predictor's error when given X but with feature i zeroed out. The explainer has an architecture such that the dimensions of the input vector X and the output vector $\widehat{\Omega}_X = \{\widehat{\omega}_1(X), \dots, \widehat{\omega}_p(X)\}$ are the same. Each of the outputs correspond to the predicted contribution for the corresponding feature. The explainer is trained by minimizing the KL divergence $KL(\Omega_X, \widehat{\Omega}_X)$ of the real contributions Ω_X and predicted contributions $\widehat{\Omega}_X$ of the training set.

We used a neural network with two layers and 512 hidden units for the explainer, and used the ensemble mode, which trains 10 independent explainers and reports their median as the final contribution values. We modified the CXPlain library's code to fit our

application, which we also included in our published code. Once trained, we predicted the contribution values of each genes in each of the samples in the testing set. To obtain drug-specific gene contribution scores, we calculated the mean contribution score of each gene across all the labeled test samples for that drug and normalized it such that the largest contribution score of a drug equals 1.

Identifying genes with highest contribution scores

After obtaining contribution scores to each gene for a drug, we sought to identify the top genes that substantially affect our model's predictions. We sorted the genes according to their final test contribution scores and plotted a curve (Supplementary Figure S5), where the x-axis is the rank of the gene i and the y-axis is gene i's drug-specific contribution score $\overline{\omega}_i$. We used the kneedle algorithm [22] to identify the point of maximum curvature, called "knee", which we then treated as the cutoff for the top genes. Kneedle relies on the idea that if one forms a line l from (1, $\overline{\omega}_{max}$) to ($n, \overline{\omega}_{min}$) and rotate the curve around the point ($n, \overline{\omega}_{min}$), the "knee" can be approximated by the set of points in the local maxima. Among these points, the point that is farthest from the line l is then identified as the knee.

Knowledge-guided Pathway Enrichment Analysis

We identified pathways associated with the top identified genes using KnowEng's gene set characterization (GSC) pipeline [34]. We used the network-guided mode, which

incorporates knowledge in the form of gene-gene interactions to augment the analysis. For the knowledge network, we selected the STRING Experimental PPI [80], which contains experimentally verified protein-protein interactions. We then proceeded with the default 50% network smoothing parameter and used the "Enrichr" pathway collection. This pipeline does not provide a p-value, but rather uses a score called "Difference Score" to implicate top pathways. Any pathway above the 0.5 threshold is considered associated with the input query set. A value above this threshold shows that the pathway has a high relevance score to the input query set (using a random walk with restarts algorithm), compared to the background [34].

Precision at kth percentile

For each drug, we used TINDL's predictions of ln(IC50) of the tumour samples, and identified the kth percentiles of the distribution (k ≤ 50), which we denote as t_k. We stratified the predictions such that all predictions below t_k is predicted as positives (i.e. sensitive). We then calculated the precision at kth percentile as $Precision_k = \frac{TP_k}{TP_k+FP_k}$, where TP_k and FP_k are the true positives and false positives at kth percentile, respectively.

Baseline models

SVR, Random Forest, and Lasso Regression were all implemented using the Scikit-learn. Geeleher's method [14] was reimplemented using Scikit-Learn and PyComBat, a python

659 implementation of ComBat [15]. We used the available implementation of TG-Lasso [9].
660 All hyperparameters were tuned as described in the previous subsections except for TG-
661 Lasso, which has its built-in hyperparameter tuning.
662
663 To ensure a fair comparison, all DL-based baseline models used a similar architecture to
664 TINDL. Additionally, the hyperparameter tuning and training procedure was also similar
665 to the one described above for TINDL. Below, we describe model-specific considerations.
666 For ComBat-DL we used ComBat [15] for removing the discrepancy between TCGA and
667 GDSC datasets. Similar to TINDL, we used both labeled and unlabeled samples of TCGA
668 for this purpose.
669
670 ADDA-DL utilizes adversarial discriminative domain adaptation (ADDA) [17], to remove
671 the discrepancy between TCGA and GDSC datasets. ADDA is a unidirectional domain
672 adaptation technique, which takes a pre-trained neural network and attempts to adapt
673 the network to the target dataset by forcing the latent feature space of the target dataset
674 (TCGA) to be similar to that of the source dataset's (GDSC). We used the TINDL model as
675 the pre-trained network which we adapt through ADDA's adversarial losses. We used the
676 unlabeled tumour samples from the drugs target tissues during training. Details are
677 provided in the Supplementary Methods.
678

DANN-DL utilizes domain adaptive neural network (DANN) [16] to remove the discrepancies between TCGA and GDSC datasets. DANN utilizes the shared latent feature space to allow the model to be used on the target dataset despite only being trained using the source dataset's labels. This is done by incorporating a gradient-reversed discriminative loss function such that a discriminator cannot tell whether the given embedding came from the source (GDSC) or target (TCGA) datasets. Similar to ADDA-DL, we used the unlabeled tumours from the drug's target tissues for training of the discriminator.

Measuring distance of clinical and preclinical samples in the latent space of DL-based models

To assess the ability of each DL-based model in removing discrepancy between preclinical and clinical samples, we used pairwise Euclidean distance of samples based on their representation learned by the encoder of the DL models. Since these representations are used by the decoder to make predictions, comparing these latent representations is more meaningful than comparing input feature representations. We used Ward's method [81] to assess the distance of preclinical samples and clinical samples, which is one of the most popular methods in assessing the distance of two groups of samples. This method, that is widely used in hierarchical clustering, has the advantage that not only analyzes the Euclidean distances of the data points, but also incorporates their variance in determining the distance of two groups of samples.

Chemicals and reagents

Dulbecco's minimum essential medium (DMEM) medium was purchased from Life Technologies, Inc. (Carlsbad, CA, USA). Fetal bovine serum (FBS) and charcoal-stripped FBS were from Invitrogen (Carlsbad, CA, USA). Ontarget-plus SMARTpool small interfering RNAs (siRNA) targeting RPP25, EMP1, EXTL3, EXOC2, NUP37, RPL13, WBP2NL, RPS6, GBP1, and JAK2 as well as negative control siRNA were purchased from Dharmacon (Thermo Scientific Dharmacon, Inc.). Reagents and primers for real time PCR were purchased from Qiagen (Valencia, CA, USA). 17β-estradiol (E2) and 4-hydroxytamoxifen (OH-TAM) were purchased from Sigma Aldrich (Saint Louis, MO USA).

Cell lines

MCF-7 cell line wase obtained from American Type Culture Collection (ATCC) (Manassus, VA) in 2014 and the identities of all cell lines were confirmed by the medical genome facility at Mayo Clinic Center (Rochester MN) using short tandem repeat profiling upon receipt. MCF-7 cells were cultured in DMEM containing 10% fetal bovine serum (FBS).

Transfection and gene silencing

Specific short interfering RNAs (siRNAs) that targeted RPP25, EMP1, EXTL3, EXOC2, NUP37, RPL13, WBP2NL, RPS6, GBP1, JAK2, and negative siRNA controls (Dharmacon, Lafayette, CO) were transfected into MCF-7 cells in 96-well plates using Lipofectamine

721 RNAiMAX Reagent (Thermo Fisher Scientific, Waltham, MA) according to the vendor's
722 protocol [67, 68]. Total RNA was extracted 48 hours after transfection for RNA
723 quantification. Specific siGENOME siRNA SMARTpool reagents against a given gene as
724 well as a negative control, siGENOME Non-Targeting siRNA, were purchased from
725 Dharmacon Inc. (Lafayette, CO, USA). For the purpose of drug tamoxifen response assay,
726 cells were plated in base medium supplemented with 5% charcoal stripped FBS for 24
727 hours, and then cultured in FBS-free DMEM media for another 24 hours before
728 transfection. Different treatments were started 24 hours after transfection.
729
730 **Tamoxifen sensitivity assay**
731 Drugs were dissolved in DMSO, and aliquots of stock solutions were frozen at −80°C.
732 Cytotoxicity assays were performed in triplicate at each drug concentration. Specifically,
733 4000 breast cancer cells were seeded in 96-well plates and were cultured in base media
734 containing 5% (vol/vol) charcoal-stripped FBS for 24 hours and were subsequently
735 cultured in FBS-free base media for another 24 hours. Cells were then transfected with
736 either control siRNA or siRNA targeting specific gene. Twenty-four hours after
737 transfection the media was replaced with fresh FBS-free base media and the cells were
738 treated with 10 µL of tamoxifen at final concentrations of 0, 0.1875, 0.375, 0.75, 1.5, 3, 6,
739 12, 24, and 48 µM [82]. After incubation for an additional 72 hours, cytotoxicity was
740 determined by quantification of DNA content using CYQUANT assay (#C35012, Invitrogen)
741 following the manufacturer's instructions [66, 83, 84]. 100µL of CyQUANT assay solution

was added, and plates were incubated at 37°C for one hour, and then read in a Safire2 plate reader with filters appropriate for 480 nm excitation and 520 nm emission.

DECLARATIONS:

Ethics approval and consent to participate:

Not applicable.

Consent for publication:

Not applicable.

Availability of data and materials:

An implementation of TINDL in python, with appropriate documentation, is available at: https://github.com/ddhostallero/tindl. Data generated in this study are provided as supplementary files.

Competing Interests:

The authors declare that they have no competing interests.

Funding:

This work was supported by the Government of Canada's New Frontiers in Research Fund (NFRF) [NFRFE-2019-01290] (AE and JC), by Natural Sciences and Engineering Research

763 Council of Canada (NSERC) grant RGPIN-2019-04460 (AE), and by McGill Initiative in
764 Computational Medicine (MiCM) (AE). This work was also funded by Génome Québec, the
765 Ministère de l'Économie et de l'Innovation du Québec, IVADO, the Canada First Research
766 Excellence Fund and Oncopole, which receives funding from Merck Canada Inc. and the
767 Fonds de Recherche du Québec – Santé (AE).

768

769 **Authors' contributions:**

770 AE and JC conceived the study and designed the project. AE led the computational aspects
771 of the study. DEH designed the algorithms, implemented the pipeline and performed the
772 statistical analyses of the results. JC led the experimental validation of the results. LW
773 performed the gene knockdown experiments. All authors contributed to the drafting of
774 the manuscript and critical discussion of the results. All authors read and approved the
775 final manuscript.

776

777 **Additional Files:**

778 **Additional File 1:** This file contains Supplementary Methods, all supplementary figures
779 and their captions, as well as captions of all supplementary tables. Some tables are
780 provided as separate files.

781

782 **REFERENCES:**

783 1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F:
784 **Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and**

Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* 2021, **71:**209-249.

2. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM: **The Cancer Genome Atlas Pan-Cancer analysis project.** *Nat Genet* 2013, **45:**1113-1120.
3. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, Bindal N, Beare D, Smith JA, Thompson IR, others: **Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells.** *Nucleic acids research* 2012, **41:**D955-D961.
4. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D, et al: **The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.** *Nature* 2012, **483:**603-607.
5. Basu A, Bodycombe NE, Cheah JH, Price EV, Liu K, Schaefer GI, Ebright RY, Stewart ML, Ito D, Wang S, et al: **An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules.** *Cell* 2013, **154:**1151-1161.
6. Costello JC, Heiser LM, Georgii E, Gonen M, Menden MP, Wang NJ, Bansal M, Ammad-ud-din M, Hintsanen P, Khan SA, et al: **A community effort to assess and improve drug sensitivity prediction algorithms.** *Nat Biotechnol* 2014, **32:**1202-1212.
7. Jiang P, Sellers WR, Liu XS: **Big Data Approaches for Modeling Response and Resistance to Cancer Drugs.** *Annu Rev Biomed Data Sci* 2018, **1:**1-27.
8. Yang J, Li A, Li Y, Guo X, Wang M: **A novel approach for drug response prediction in cancer cell lines via network representation learning.** *Bioinformatics* 2019, **35:**1527-1535.
9. Huang EW, Bhope A, Lim J, Sinha S, Emad A: **Tissue-guided LASSO for prediction of clinical drug response using preclinical samples.** *PLoS computational biology* 2020, **16:**e1007607.
10. Ding Z, Zu S, Gu J: **Evaluating the molecule-based prediction of clinical drug responses in cancer.** *Bioinformatics* 2016, **32:**2891-2895.
11. Wang Z, Li R, Wang M, Li A: **GPDBN: deep bilinear network integrating both genomic data and pathological images for breast cancer prognosis prediction.** *Bioinformatics* 2021.
12. Malik V, Kalakoti Y, Sundar D: **Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer.** *BMC Genomics* 2021, **22:**214.
13. Sharifi-Noghabi H, Peng S, Zolotareva O, Collins CC, Ester M: **AITL: Adversarial Inductive Transfer Learning with input and output space adaptation for pharmacogenomics.** *Bioinformatics* 2020, **36:**i380-i388.
14. Geeleher P, Cox NJ, Huang RS: **Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines.** *Genome Biol* 2014, **15:**R47.

15. Johnson WE, Li C, Rabinovic A: **Adjusting batch effects in microarray expression data using empirical Bayes methods.** *Biostatistics* 2007, **8:**118-127.
16. Ganin Y, Lempitsky V: **Unsupervised domain adaptation by backpropagation.** In *International conference on machine learning*. 2015: 1180-1189.
17. Tzeng E, Hoffman J, Saenko K, Darrell T: **Adversarial discriminative domain adaptation.** In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017: 7167-7176.
18. Schwab P, Karlen W: **CXPlain: Causal Explanations for Model Interpretation under Uncertainty.** In *Advances in Neural Information Processing Systems (NeurIPS)*. 2019
19. Dong Z, Zhang N, Li C, Wang H, Fang Y, Wang J, Zheng X: **Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection.** *BMC Cancer* 2015, **15:**489.
20. Li B, Severson E, Pignon JC, Zhao H, Li T, Novak J, Jiang P, Shen H, Aster JC, Rodig S, et al: **Comprehensive analyses of tumor immunity: implications for cancer immunotherapy.** *Genome Biol* 2016, **17:**174.
21. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R: **Integrating single-cell transcriptomic data across different conditions, technologies, and species.** *Nat Biotechnol* 2018, **36:**411-420.
22. Satopaa V, Albrecht J, Irwin D, Raghavan B: **Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior.** In *2011 31st International Conference on Distributed Computing Systems Workshops*. 2011: 166-171.
23. Burrell RA, McClelland SE, Endesfelder D, Groth P, Weller MC, Shaikh N, Domingo E, Kanu N, Dewhurst SM, Gronroos E, et al: **Replication stress links structural and numerical cancer chromosomal instability.** *Nature* 2013, **494:**492-496.
24. Hanahan D, Weinberg RA: **The hallmarks of cancer.** *Cell* 2000, **100:**57-70.
25. Hanahan D, Weinberg RA: **Hallmarks of cancer: the next generation.** *Cell* 2011, **144:**646-674.
26. Murai J, Thomas A, Miettinen M, Pommier Y: **Schlafen 11 (SLFN11), a restriction factor for replicative stress induced by DNA-targeting anti-cancer therapies.** *Pharmacol Ther* 2019, **201:**94-102.
27. Deng Y, Cai Y, Huang Y, Yang Z, Bai Y, Liu Y, Deng X, Wang J: **High SLFN11 expression predicts better survival for patients with KRAS exon 2 wild type colorectal cancer after treated with adjuvant oxaliplatin-based treatment.** *BMC Cancer* 2015, **15:**833.
28. Coleman N, Zhang B, Byers LA, Yap TA: **The role of Schlafen 11 (SLFN11) as a predictive biomarker for targeting the DNA damage response.** *Br J Cancer* 2021, **124:**857-859.

29. Luan J, Gao X, Hu F, Zhang Y, Gou X: **SLFN11 is a general target for enhancing the sensitivity of cancer to chemotherapy (DNA-damaging agents).** *J Drug Target* 2020, **28:**33-40.
30. Winkler C, Armenia J, Jones GN, Tobalina L, Sale MJ, Petreus T, Baird T, Serra V, Wang AT, Lau A, et al: **SLFN11 informs on standard of care and novel treatments in a wide range of cancer models.** *Br J Cancer* 2021, **124:**951-962.
31. Gardner EE, Lok BH, Schneeberger VE, Desmeules P, Miles LA, Arnold PK, Ni A, Khodos I, de Stanchina E, Nguyen T, et al: **Chemosensitive Relapse in Small Cell Lung Cancer Proceeds through an EZH2-SLFN11 Axis.** *Cancer Cell* 2017, **31:**286-299.
32. Fillmore CM, Xu C, Desai PT, Berry JM, Rowbotham SP, Lin YJ, Zhang H, Marquez VE, Hammerman PS, Wong KK, Kim CF: **EZH2 inhibition sensitizes BRG1 and EGFR mutant lung tumours to TopoII inhibitors.** *Nature* 2015, **520:**239-242.
33. Munakata W, Shirasugi Y, Tobinai K, Onizuka M, Makita S, Suzuki R, Maruyama D, Kawai H, Izutsu K, Nakanishi T, et al: **Phase 1 study of tazemetostat in Japanese patients with relapsed or refractory B-cell lymphoma.** *Cancer Sci* 2021, **112:**1123-1131.
34. Blatti C, 3rd, Emad A, Berry MJ, Gatzke L, Epstein M, Lanier D, Rizal P, Ge J, Liao X, Sobh O, et al: **Knowledge-guided analysis of "omics" data using the KnowEnG cloud platform.** *PLoS Biol* 2020, **18:**e3000583.
35. Gelman AE, Zhang J, Choi Y, Turka LA: **Toll-like receptor ligands directly promote activated CD4+ T cell survival.** *J Immunol* 2004, **172:**6065-6073.
36. Alexopoulou L, Holt AC, Medzhitov R, Flavell RA: **Recognition of double-stranded RNA and activation of NF-kappaB by Toll-like receptor 3.** *Nature* 2001, **413:**732-738.
37. Li JK, Balic JJ, Yu L, Jenkins B: **TLR Agonists as Adjuvants for Cancer Vaccines.** *Adv Exp Med Biol* 2017, **1024:**195-212.
38. Nowak AK, Robinson BW, Lake RA: **Synergy between chemotherapy and immunotherapy in the treatment of established murine solid tumors.** *Cancer Res* 2003, **63:**4490-4496.
39. Rakoff-Nahoum S, Medzhitov R: **Toll-like receptors and cancer.** *Nat Rev Cancer* 2009, **9:**57-63.
40. Jego G, Bataille R, Geffroy-Luseau A, Descamps G, Pellat-Deceunynck C: **Pathogen-associated molecular patterns are growth and survival factors for human myeloma cells through Toll-like receptors.** *Leukemia* 2006, **20:**1130-1137.
41. Bohnhorst J, Rasmussen T, Moen SH, Flottum M, Knudsen L, Borset M, Espevik T, Sundan A: **Toll-like receptors mediate proliferation and survival of multiple myeloma cells.** *Leukemia* 2006, **20:**1138-1144.

42. Kawashima M, Suzuki SO, Doh-ura K, Iwaki T: **alpha-Synuclein is expressed in a variety of brain tumors showing neuronal differentiation.** *Acta Neuropathol* 2000, **99:**154-160.
43. Matsuo Y, Kamitani T: **Parkinson's disease-related protein, alpha-synuclein, in malignant melanoma.** *PLoS One* 2010, **5:**e10481.
44. Ge Y, Xu K: **Alpha-synuclein contributes to malignant progression of human meningioma via the Akt/mTOR pathway.** *Cancer Cell Int* 2016, **16:**86.
45. Shekoohi S, Rajasekaran S, Patel D, Yang S, Liu W, Huang S, Yu X, Witt SN: **Knocking out alpha-synuclein in melanoma cells dysregulates cellular iron metabolism and suppresses tumor growth.** *Sci Rep* 2021, **11:**5267.
46. Tzivion G, Luo Z, Avruch J: **A dimeric 14-3-3 protein is an essential cofactor for Raf kinase activity.** *Nature* 1998, **394:**88-92.
47. Pan ZZ, Bruening W, Giasson BI, Lee VM, Godwin AK: **Gamma-synuclein promotes cancer cell survival and inhibits stress- and chemotherapy drug-induced apoptosis by modulating MAPK pathways.** *J Biol Chem* 2002, **277:**35050-35060.
48. Clancy JW, Zhang Y, Sheehan C, D'Souza-Schorey C: **An ARF6-Exportin-5 axis delivers pre-miRNA cargo to tumour microvesicles.** *Nat Cell Biol* 2019, **21:**856-866.
49. Clancy JW, Tricarico CJ, Marous DR, D'Souza-Schorey C: **Coordinated Regulation of Intracellular Fascin Distribution Governs Tumor Microvesicle Release and Invasive Cell Capacity.** *Mol Cell Biol* 2019, **39**.
50. Li R, Peng C, Zhang X, Wu Y, Pan S, Xiao Y: **Roles of Arf6 in cancer cell invasion, metastasis and proliferation.** *Life Sci* 2017, **182:**80-84.
51. Hu Z, Xu R, Liu J, Zhang Y, Du J, Li W, Zhang W, Li Y, Zhu Y, Gu L: **GEP100 regulates epidermal growth factor-induced MDA-MB-231 breast cancer cell invasion through the activation of Arf6/ERK/uPAR signaling pathway.** *Exp Cell Res* 2013, **319:**1932-1941.
52. Hashimoto S, Onodera Y, Hashimoto A, Tanaka M, Hamaguchi M, Yamada A, Sabe H: **Requirement for Arf6 in breast cancer invasive activities.** *Proc Natl Acad Sci U S A* 2004, **101:**6647-6652.
53. Hopkins BD, Pauli C, Du X, Wang DG, Li X, Wu D, Amadiume SC, Goncalves MD, Hodakoski C, Lundquist MR, et al: **Suppression of insulin feedback enhances the efficacy of PI3K inhibitors.** *Nature* 2018, **560:**499-503.
54. Nencioni A, Caffa I, Cortellino S, Longo VD: **Fasting and cancer: molecular mechanisms and clinical application.** *Nat Rev Cancer* 2018, **18:**707-719.
55. Poloz Y, Stambolic V: **Obesity and cancer, a case for insulin signaling.** *Cell Death Dis* 2015, **6:**e2037.
56. Malaguarnera R, Belfiore A: **The insulin receptor: a new target for cancer therapy.** *Front Endocrinol (Lausanne)* 2011, **2:**93.
57. Belfiore A, Malaguarnera R: **Insulin receptor and cancer.** *Endocr Relat Cancer* 2011, **18:**R125-147.

58. Hua H, Kong Q, Yin J, Zhang J, Jiang Y: **Insulin-like growth factor receptor signaling in tumorigenesis and drug resistance: a challenge for cancer therapy.** *J Hematol Oncol* 2020, **13:**64.
59. Agrawal S, Wozniak M, Luc M, Makuch S, Pielka E, Agrawal AK, Wietrzyk J, Banach J, Gamian A, Pizon M, Ziolkowski P: **Insulin enhancement of the antitumor activity of chemotherapeutic agents in colorectal cancer is linked with downregulating PIK3CA and GRB2.** *Sci Rep* 2019, **9:**16647.
60. Bodemann BO, White MA: **Ral GTPases and cancer: linchpin support of the tumorigenic platform.** *Nat Rev Cancer* 2008, **8:**133-140.
61. Tchevkina E, Agapova L, Dyakova N, Martinjuk A, Komelkov A, Tatosyan A: **The small G-protein RalA stimulates metastasis of transformed cells.** *Oncogene* 2005, **24:**329-335.
62. Moghadam AR, Patrad E, Tafsiri E, Peng W, Fangman B, Pluard TJ, Accurso A, Salacz M, Shah K, Ricke B, et al: **Ral signaling pathway in health and cancer.** *Cancer Med* 2017, **6:**2998-3013.
63. Chien Y, White MA: **RAL GTPases are linchpin modulators of human tumour-cell proliferation and survival.** *EMBO Rep* 2003, **4:**800-806.
64. Neel NF, Martin TD, Stratford JK, Zand TP, Reiner DJ, Der CJ: **The RalGEF-Ral Effector Signaling Network: The Road Less Traveled for Anti-Ras Drug Discovery.** *Genes Cancer* 2011, **2:**275-287.
65. Gocher AM, Workman CJ, Vignali DAA: **Interferon-gamma: teammate or opponent in the tumour microenvironment?** *Nat Rev Immunol* 2021.
66. Cairns J, Ingle JN, Dudenkov TM, Kalari KR, Carlson EE, Na J, Buzdar AU, Robson ME, Ellis MJ, Goss PE, et al: **Pharmacogenomics of aromatase inhibitors in postmenopausal breast cancer and additional mechanisms of anastrozole action.** *JCI Insight* 2020, **5**.
67. Cairns J, Kalari KR, Ingle JN, Shepherd LE, Ellis MJ, Goss PE, Barman P, Carlson EE, Goodnature B, Goetz MP, et al: **Interaction between SNP Genotype and Efficacy of Anastrozole and Exemestane in Early Stage Breast Cancer.** *Clin Pharmacol Ther* 2021.
68. Cairns J, Ly RC, Niu N, Kalari KR, Carlson EE, Wang L: **CDC25B partners with PP2A to induce AMPK activation and tumor suppression in triple negative breast cancer.** *NAR Cancer* 2020, **2:**zcaa039.
69. Li Y, Wang M, Yang M, Xiao Y, Jian Y, Shi D, Chen X, Ouyang Y, Kong L, Huang X, et al: **Nicotine-induced ILF2 facilitates nuclear mRNA export of pluripotency factors to promote stemness and chemoresistance in human esophageal cancer.** *Cancer Res* 2021.
70. Kikuchi H, Maishi N, Annan DA, Alam MT, Dawood RIH, Sato M, Morimoto M, Takeda R, Ishizuka K, Matsumoto R, et al: **Chemotherapy-Induced IL8 Upregulates MDR1/ABCB1 in Tumor Blood Vessels and Results in Unfavorable Outcome.** *Cancer Res* 2020, **80:**2996-3008.
71. Kubiliute R, Januskeviciene I, Urbanaviciute R, Daniunaite K, Drobniene M, Ostapenko V, Daugelavicius R, Jarmalaite S: **Nongenotoxic ABCB1 activator**

tetraphenylphosphonium can contribute to doxorubicin resistance in MX-1 breast cancer cell line. *Sci Rep* 2021, **11:**6556.

72. Li L, Fridley BL, Kalari K, Niu N, Jenkins G, Batzler A, Abo RP, Schaid D, Wang L: **Discovery of genetic biomarkers contributing to variation in drug response of cytidine analogues using human lymphoblastoid cell lines.** *BMC Genomics* 2014, **15:**93.
73. Lindner AU, Concannon CG, Boukes GJ, Cannon MD, Llambi F, Ryan D, Boland K, Kehoe J, McNamara DA, Murray F, et al: **Systems analysis of BCL2 protein family interactions establishes a model to predict responses to chemotherapy.** *Cancer Res* 2013, **73:**519-528.
74. Wang CC, Bajikar SS, Jamal L, Atkins KA, Janes KA: **A time- and matrix-dependent TGFBR3-JUND-KRT5 regulatory circuit in single breast epithelial cells and basal-like premalignancies.** *Nat Cell Biol* 2014, **16:**345-356.
75. Leung E, Kannan N, Krissansen GW, Findlay MP, Baguley BC: **MCF-7 breast cancer cells selected for tamoxifen resistance acquire new phenotypes differing in DNA content, phospho-HER2 and PAX2 expression, and rapamycin sensitivity.** *Cancer Biol Ther* 2010, **9:**717-724.
76. Franco GR, Tanaka M, Simpson AJ, Pena SD: **Characterization of a Schistosoma mansoni homologue of the gene encoding the breast basic conserved protein 1/L13 ribosomal protein.** *Comp Biochem Physiol B Biochem Mol Biol* 1998, **120:**701-708.
77. Zagidullin B, Aldahdooh J, Zheng S, Wang W, Wang Y, Saad J, Malyutina A, Jafari M, Tanoli Z, Pessia A, Tang J: **DrugComb: an integrative cancer drug combination data portal.** *Nucleic Acids Res* 2019, **47:**W43-W51.
78. Liu H, Zhang W, Zou B, Wang J, Deng Y, Deng L: **DrugCombDB: a comprehensive database of drug combinations toward the discovery of combinatorial therapy.** *Nucleic Acids Res* 2020, **48:**D871-D881.
79. Granger CW: **Investigating causal relations by econometric models and cross-spectral methods.** *Econometrica: journal of the Econometric Society* 1969**:**424-438.
80. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, et al: **STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets.** *Nucleic Acids Res* 2019, **47:**D607-D613.
81. Ward Jr JH: **Hierarchical grouping to optimize an objective function.** *Journal of the American statistical association* 1963, **58:**236-244.
82. Cairns J, Ingle JN, Wickerham LD, Weinshilboum R, Liu M, Wang L: **SNPs near the cysteine proteinase cathepsin O gene (CTSO) determine tamoxifen sensitivity in ERalpha-positive breast cancer through regulation of BRCA1.** *PLoS Genet* 2017, **13:**e1007031.

83. Cairns J, Fridley BL, Jenkins GD, Zhuang Y, Yu J, Wang L: **Differential roles of ERRFI1 in EGFR and AKT pathway regulation affect cancer proliferation.** *EMBO Rep* 2018, **19**.

84. Cairns J, Ingle JN, Kalari KR, Shepherd LE, Kubo M, Goetz MP, Weinshilboum RM, Wang L: **The lncRNA MIR2052HG regulates ERalpha levels and aromatase inhibitor resistance through LMTK3 by recruiting EGR1.** *Breast Cancer Res* 2019, **21:**47.