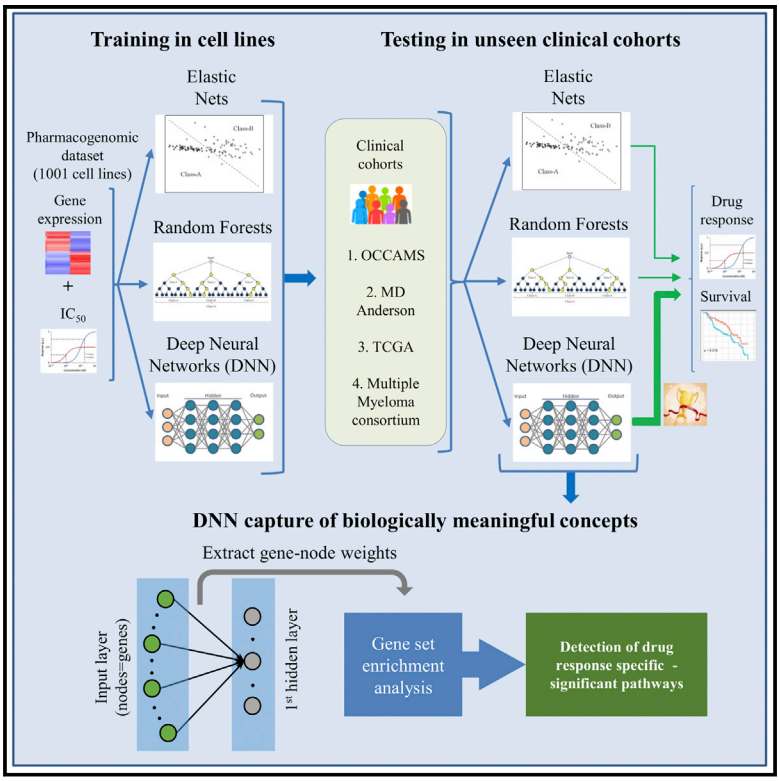
Report

Cell Reports

A Deep Learning Framework for Predicting Response to Therapy in Cancer

Graphical Abstract



Highlights

- A machine learning (ML) workflow is designed to predict drug response in cancer patients
- Deep neural networks (DNNs) surpass current ML algorithms in drug response prediction
- DNNs predict drug response and survival in various large clinical cohorts
- DNNs capture intricate biological interactions linked to specific drug response pathways

Authors

Theodore Sakellaropoulos, Konstantinos Vougas, Sonali Narang, ..., Russell Petty, Aristotelis Tsirigos, Vassilis G. Gorgoulis

Correspondence

kvougas@bioacademy.gr (K.V.), aristotelis.tsirigos@nyulangone.org (A.T.), vgorg@med.uoa.gr (V.G.G.)

In Brief

Sakellaropoulos et al. designed a machine learning workflow to predict drug response and survival of cancer patients. All pipelines are trained on a large panel of cancer cell lines and tested in clinical cohorts. DNN outperforms other machine learning algorithms by capturing pathways that link gene expression with drug response.

Sakellaropoulos et al., 2019, Cell Reports *29*, 3367–3373
December 10, 2019
https://doi.org/10.1016/j.celrep.2019.11.017

CellPress

Cell Reports
Report

A Deep Learning Framework for Predicting Response to Therapy in Cancer

Theodore Sakellaropoulos,[1,2,20] Konstantinos Vougas,[3,4,20,*] Sonali Narang,[1,2] Filippos Koinis,[4] Athanassios Kotsinas,[4] Alexander Polyzos,[5] Tyler J. Moss,[6] Sarina Piha-Paul,[7] Hua Zhou,[8] Eleni Kardala,[4] Eleni Damianidou,[4] Leonidas G. Alexopoulos,[9] Iannis Aifantis,[1,2] Paul A. Townsend,[10] Mihalis I. Panayiotidis,[11,12] Petros Sfikakis,[13,14] Jiri Bartek,[15,16,17] Rebecca C. Fitzgerald,[18] Dimitris Thanos,[3] Kenna R. Mills Shaw,[6] Russell Petty,[19] Aristotelis Tsirigos,[1,2,8,*] and Vassilis G. Gorgoulis[3,4,10,13,21,*]

[1]Department of Pathology, NYU School of Medicine, New York, NY 10016, USA
[2]Laura and Isaac Perlmutter Cancer Center, NYU School of Medicine, New York, NY 10016, USA
[3]Biomedical Research Foundation of the Academy of Athens, 4 Soranou Ephessiou Str., Athens 11527, Greece
[4]Molecular Carcinogenesis Group, Department of Histology and Embryology, School of Medicine, National and Kapodistrian University of Athens, 75 Mikras Asias Str., Athens 11527, Greece
[5]Sanford I. Weill Department of Medicine, Sandra and Edward Meyer Cancer Center, Weill Cornell Medicine, New York, NY 10021, USA
[6]Sheikh Khalifa Bin Zayed al Nahyan Institute for Personalized Cancer Therapy, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd., Houston, TX 77030, USA
[7]Department of Investigational Cancer Therapeutics, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd., Houston, TX 77030, USA
[8]Applied Bioinformatics Laboratories, NYU School of Medicine, New York, NY 10016, USA
[9]School of Mechanical Engineering, National Technical University of Athens, Zografou 15780, Greece
[10]Division of Cancer Sciences, Faculty of Biology, Medicine and Health, Manchester Academic Health Science Centre, Manchester Cancer Research Centre, NIHR Manchester Biomedical Research Centre, University of Manchester, Manchester M20 4GJ, UK
[11]Department of Applied Sciences, Northumbria University, Newcastle upon Tyne NE1 8ST, UK
[12]Department of Electron Microscopy & Molecular Pathology, Cyprus Institute of Neurology & Genetics, Nicosia, 2371, Cyprus
[13]1st Department of Propaedeutic Internal Medicine, Medical School, Laikon Hospital, National and Kapodistrian University of Athens, 75 Mikras Asias Str., Athens 11527, Greece
[14]Center for New Biotechnologies and Precision Medicine, Medical School, National and Kapodistrian University of Athens, 75 Mikras Asias Str., Athens 11527, Greece
[15]Genome Integrity Unit, Danish Cancer Society Research Centre, Strandboulevarden 49, Copenhagen 2100, Denmark
[16]Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University, Hněvotínská, Olomouc 1333/5 779 00, Czech Republic
[17]Science for Life Laboratory, Division of Genome Biology, Department of Medical Biochemistry and Biophysics, Karolinska Institute, Stockholm SE-171 77, Sweden
[18]Medical Research Council Cancer Unit, Hutchison/Medical Research Council Research Centre, University of Cambridge, Cambridge CB2 0XZ, UK
[19]Division of Molecular and Clinical Medicine, Ninewells Hospital and School of Medicine, University of Dundee, Dundee DD1 9SY, UK
[20]These authors contributed equally
[21]Lead Contact
*Correspondence: kvougas@bioacademy.gr (K.V.), aristotelis.tsirigos@nyulangone.org (A.T.), vgorg@med.uoa.gr (V.G.G.)
https://doi.org/10.1016/j.celrep.2019.11.017

SUMMARY

A major challenge in cancer treatment is predicting clinical response to anti-cancer drugs on a personalized basis. Using a pharmacogenomics database of 1,001 cancer cell lines, we trained deep neural networks for prediction of drug response and assessed their performance on multiple clinical cohorts. We demonstrate that deep neural networks outperform the current state in machine learning frameworks. We provide a proof of concept for the use of deep neural network-based frameworks to aid precision oncology strategies.

INTRODUCTION

Predicting the clinical response to therapeutic agents is a major challenge in cancer treatment. To deliver personalized treatment with high efficacy, identifying molecular disease signatures and matching them with the most effective therapeutic interventions are essential. The advent of the ''omics'' era has permitted scientists to dissect the molecular events that are known to drive carcinogenesis (Alexandrov et al., 2013; Negrini et al., 2010). Nonetheless, effective translation of the growing wealth of high-throughput profiling data into clinically meaningful results has been challenging (van't Veer and Bernards, 2008). The latter is primarily hindered by the lack of reliable preclinical models. Although individual cancer cell lines do not reflect the complexity of clinical cancer tissues with fidelity (Weinstein, 2012), when compiled in large panels, they are able to recapitulate the genomic diversity of human cancers (Iorio et al., 2016). These panels can be readily used as platforms upon which expert systems for the prediction of pharmacological response may be developed (reviewed in Vougas et al., 2019). Although, large-scale panels containing pharmacogenomics data have been made available to the public domain, well-validated computational algorithms able

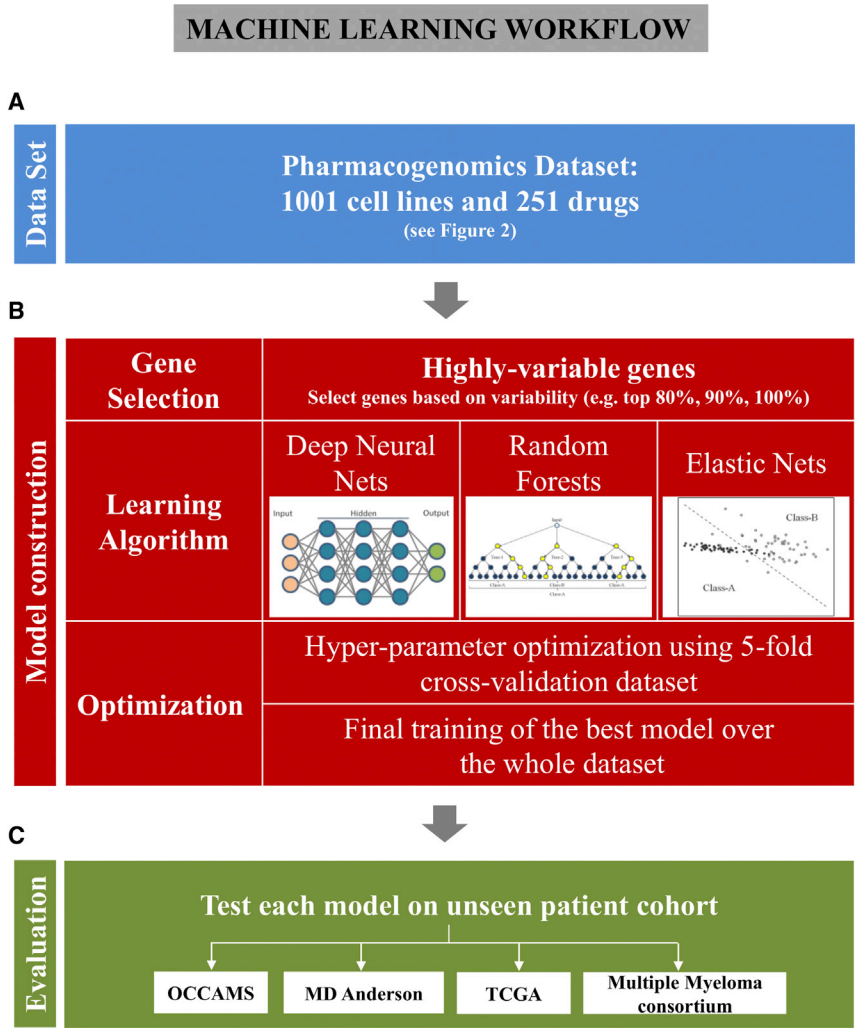


Figure 1. Schematic Representation of the Study Design and Bioinformatics Pipeline

(A) Dataset: the full dataset was compiled using 1,001 cell lines and 251 drugs from the GDSC database.

(B) Model construction: deep neural networks (DNNs) were used to predict patient drug response and compared against two broadly used learning algorithms: random forest (RF) and elastic net (Enet) (right panel).

(C) Evaluation: our models were evaluated in various settings that included the patient drug response dataset obtained from clinical trials and TCGA.

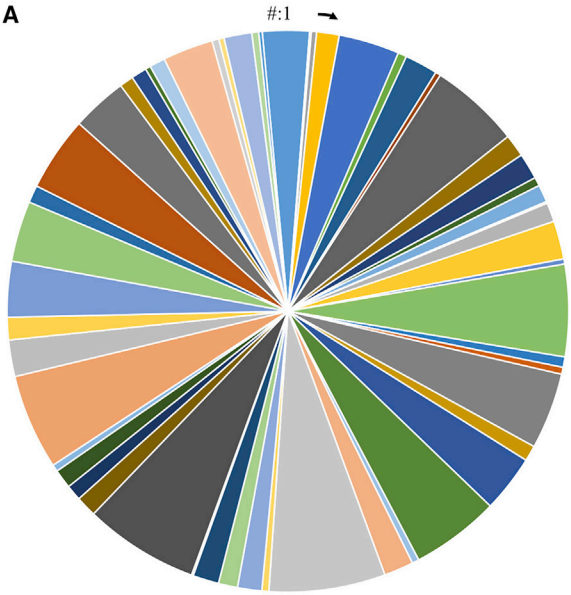
to accurately predict therapeutic response are still lacking. Today's complex omics datasets have appeared too multidimensional to be effectively managed by classical machine learning algorithms (Libbrecht and Noble, 2015). However, deep neural networks (DNNs) have the ability to model biological complexity and have been effectively applied in various fields (e.g., image analysis and text mining) with increased classification accuracy compared with classical computational methods (Schmidhuber, 2015). DNNs are based on the modeling of high-level neural networks in flexible, multilayer systems of connected and interacting neurons, which perform numerous data abstractions and transformations (LeCun et al., 2015). In a recent surge of interest, DNNs have been effectively applied in many fields, such as predicting automated histopathological diagnosis (Coudray et al., 2018). However, the potential of deep neural networks for predicting response to cancer therapy needs to be addressed, and studies in this direction are essential. As shown in Figure S1, its application in clinical settings is almost absent (Chiu et al., 2019). In this study, we address this issue by developing a deep neural network model to predict pharmacological responses using gene expression data.

RESULTS

A DNN-based workflow was designed to predict drug responses in cancer patients using gene expression data, as presented in Figure 1. In brief, the DNN was trained and optimized on a 1,001 cell-line drug response database (Figures 2A–2C), its performance was tested blindly on patient cohorts, and finally it was compared with state-of-the-art models, i.e., random forests (RFs) (Costello et al., 2014) and elastic nets (Enets) (Zou and Hastie, 2005) (Figure 1).

We trained our DNN models to predict drug response from gene expression using data from the Genomics of Drug Sensitivity in Cancer (GDSC) database (Garnett et al., 2012). GDSC comprises gene expression from 1,001 cancer-cell lines and drug response data in the form of IC_{50} values for 251 therapeutic compounds (Figures 1A, 2B, and 2C; Data S1a, "Drugs"). In particular, we chose to use the open-source DNN framework provided by H2O.ai (http://www.h2o.ai/), a cluster-read framework, which allows straightforward deployment of our pipeline to a high-performance computing environment. Then, we evaluated and compared DNN to two frequently used learning algorithms: random forest (RF) (Costello et al., 2014) and elastic net (Enet) (Zou and Hastie, 2005) (Figures 1B and 1C). The performance of these models was evaluated with data obtained from clinical cohorts, including clinical trials, in which gene expression data were available before treatment with a drug present in our pharmacogenomics database (Data S1a, "Drugs"). We searched the public domain for patient datasets comprising both gene expression and drug response information. Results from a previous study, describing four trials (three with a single arm and one with multiple ones) suggested that it is indeed possible to predict clinical drug response using baseline gene expression levels (Geeleher et al., 2014). However, the

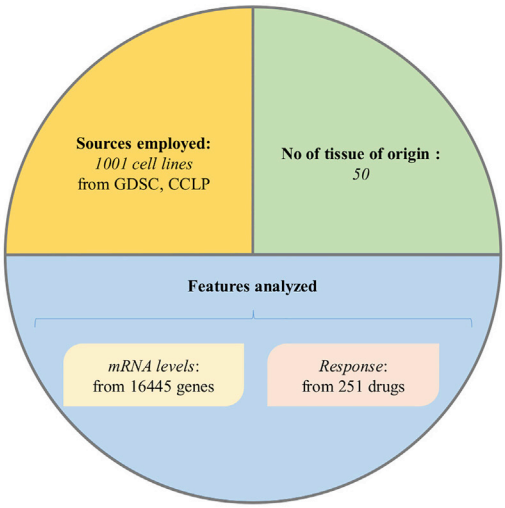
A



#	Type	No	%
1	Acute myeloid leukemia	27	2.69
2	Adrenal gland	1	0.09
3	Anaplastic large cell lymphoma	3	0.29
4	B cell leukemia	13	1.29
5	B cell lymphoma	35	3.49
6	Biliary tract	5	0.49
7	Bladder	19	1.89
8	Bone other	3	0.29
9	Breast	52	5.19
10	Burkitt lymphoma	13	1.29
11	Cervix	15	1.49
12	Chondrosarcoma	5	0.49
13	Chronic myeloid leukemia	10	0.99
14	Digestive system other	1	0.09
15	Endometrium	11	1.09
16	Ewings sarcoma	22	2.19
17	Fibrosarcoma	3	0.29
18	Glioma	53	5.29
19	Haematopoietic neoplasm other	6	0.59
20	Hairy cell leukaemia	4	0.39
21	Head and neck	44	4.39
22	Hodgkin lymphoma	9	0.89
23	Kidney	33	3.29
24	Large intestine	51	5.09
25	Leukemia	4	0.39
26	Liver	17	1.69
27	Lung NSCLC adenocarcinoma	67	6.69
28	Lung NSCLC carcinoid	4	0.39

#	Type	No	%
29	Lung NSCLC large cell	14	1.39
30	Lung NSCLC not specified	11	1.09
31	Lung NSCLC squamous cell carcinoma	15	1.49
32	Lung other	1	0.09
33	Lung small cell carcinoma	66	6.59
34	Lymphoblastic leukemia	12	1.19
35	Lymphoblastic T cell leukemia	9	0.89
36	Lymphoid neoplasm other	11	1.09
37	Medulloblastoma	4	0.39
38	Melanoma	55	5.49
39	Mesothelioma	21	2.09
40	Myeloma	13	1.29
41	Neuroblastoma	32	3.19
42	Oesophagus	35	3.49
43	Osteosarcoma	10	0.99
44	Ovary	43	4.29
45	Pancreas	32	3.19
46	Prostate	8	0.79
47	Rhabdomyosarcoma	9	0.89
48	Skin other	3	0.29
49	Soft tissue other	9	0.89
50	Stomach	29	2.89
51	T cell leukemia	4	0.39
52	Testis	3	0.29
53	Thyroid	16	1.59
54	Urogenital system other	4	0.39
55	Uterus	2	0.19

B



C

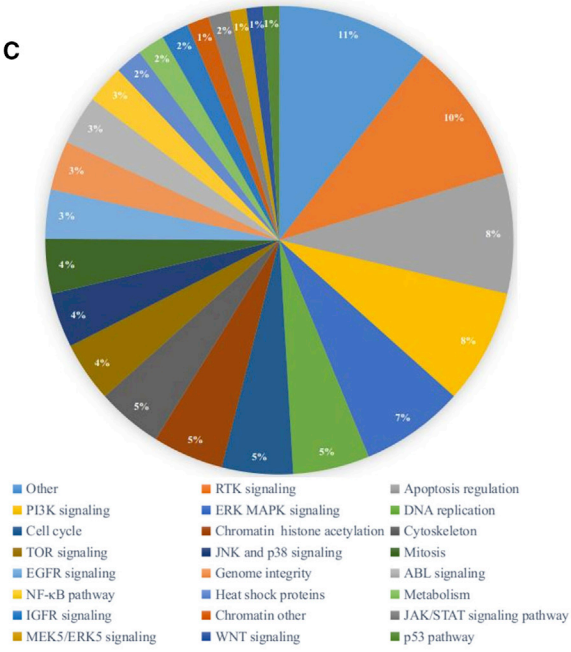


Figure 2. Description of the Dataset Used for Training and Building of the DNN Prediction Framework

(A) Tissue of origin composition of the 1,001 cell lines used as training dataset and platform for generation of our DNN models.

(B) Summary of sources (GDSC: the 1,001 cell lines repository tested on 251 drugs [see C]; CCLP: the same 1,001 cell lines repository examined for expression levels of 16,445 genes), number of tissue of origin and features analyzed by the machine learning algorithms.

(C) Grouping of all available drugs from the GDSC database according to target pathway, out of which 251 models were trained (for details see also Data S1a – "Drugs").

number of patients in all these datasets was very small (maximum of 5 patients in responders and/or non-responders) with the exception of bortezomib, a phase II/III clinical trial in patients with relapsed multiple myeloma (Mulligan et al., 2007). To more precisely address the question and systematically compare previous computational approaches to deep learning, we looked for larger patient datasets with availability of both gene expression and drug response data. We found such datasets on The Cancer Genome Atlas (TCGA) available for two drugs: cisplatin and paclitaxel (Ding et al., 2016). In addition, we obtained unpublished gene expression and patient response data from (1) a clinical trial of a PARP inhibitor, conducted at the MD Anderson Cancer Center, and (2) a cohort of esophageal adenocarcinomas, treated with neo-adjuvant

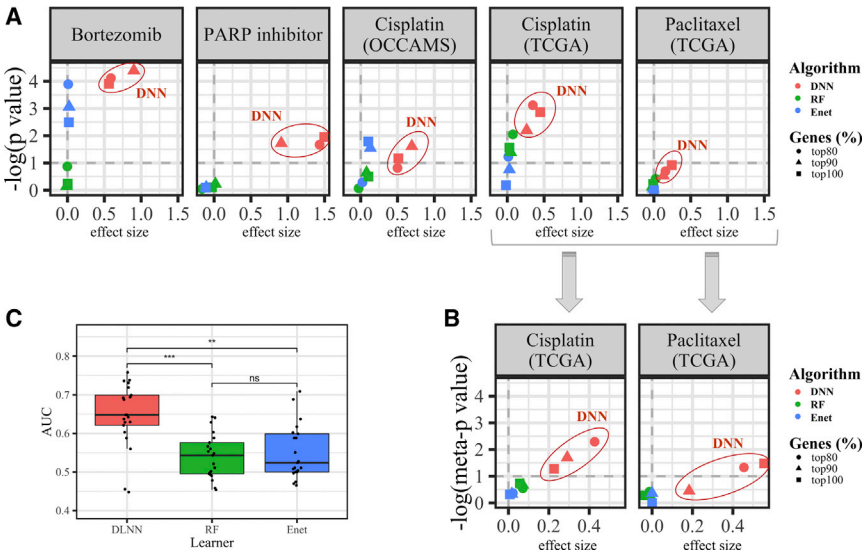


Figure 3. Evaluation of Drug Response Models on Patient Datasets

(A) Volcano plots of effect size (difference of mean IC50 between responders and non-responders) and p values (Wilcox test) of each learning algorithm and percentage of selected genes for each tested algorithm.

(B) Volcano plots of effect size (mean effect size per cancer, weighted by number of samples) and meta-p values (combined using Stouffer's method weighted by number of samples) for cisplatin and paclitaxel after correcting for cancer type.

(C) Boxplot representation and statistical comparison using a paired Wilcoxon test of the AUC values calculated for each learning algorithm, drug, and gene selection percentage (grouped by drug).

chemotherapy as a clinical model for cisplatin-based therapy (Lagergren et al., 2017), under the Esophageal Cancer Clinical and Molecular Stratification (OCCAMS) consortium (Frankell et al., 2019). All datasets are summarized in Table S1, including information about the drug name, the total number of patients, patients in each group (responders versus non-responders), a short description of each cohort and citation (see also Data S1a, "Drugs").

To systematically assess the ability of each algorithm to identify patients that responded or not to each drug, we found the optimal models for each algorithm using the cell-line data alone and then applied those optimal models on each clinical dataset. More specifically, we used 5-fold cross-validation on the cell-line data to automatically select the optimal hyper-parameters for each algorithm (Figure 1B; Table S2). Details on the hyper-parameter space are presented in STAR Methods ("Hyper-parameter Optimization on Cell Line Data"). In addition to the three learning algorithms, we assessed the impact of feature selection by training each algorithm using all genes as features or selecting the most highly variable genes. Gene selection was exclusively performed on the training set to prevent data leak. The optimal models for DNN, RF, and Enet—in combination with each feature selection approach—were selected and then retrained on the entire cell-line dataset (Figure 1B; Data S1h). Finally, these retrained models were applied on the unseen clinical datasets (Figure 1C). We evaluated each model by the effect size—measured as the difference in the mean of predicted IC_{50} values in the responder versus non-responder groups—and the associated p values calculated using the paired Wilcoxon test. The results of our analysis are summarized as volcano plots for each dataset in (Figure 3). Overall, deep neural networks perform consistently better than random forests and elastic nets, both by effect size and by statistical significance, independent of the percentage of highly variable genes used in the corresponding models. In the bortezomib and OCCAMS-cisplatin datasets (Figure 3A), both DNN and Enet predict that patients who respond to the drug have statistically significantly lower (predicted) IC_{50} values, as expected, whereas RF does not yield any significant difference in predicted IC_{50} values between responders and non-responders. Importantly, DNN performs better than Enet in terms of effect size, as measured by the difference of the mean predicted IC_{50} value in the non-responder group compared with the responder group. In the PARP inhibitor dataset (Figure 3A), DNN outperforms the other learning algorithms in both metrics, independent of gene selection. In the TCGA-cisplatin dataset, DNN again outperforms RF and Enet in both metrics (p value and effect size), independent of the percentage of genes selected. RF and Enet can barely pass the significance threshold ($p < 0.1$), but this depends on the percentage of selected genes. No algorithm seems to be close to statistical significance on the paclitaxel dataset with the exception of the DNN model that uses all genes (p value of ~0.1). To further investigate this issue, because the TCGA-cisplatin and paclitaxel datasets include patients with several cancer types, we tested the performance of the algorithms separately for each drug and cancer type. We selected the cancer types with the highest number of patients (at least 40 patients and more than 5 patients in both the responder and the non-responder groups). This resulted in three cancer types for TCGA-cisplatin, bladder cancer (BLDCA), cervical squamous cell carcinoma (CESC), and lung adenocarcinoma (LUAD), and only one cancer type for paclitaxel, breast cancer (BRCA). The greater diversity of the paclitaxel dataset may explain why none of the tested algorithms was able to produce a statistically significant result in Figure 3A. To adjust for the cancer-type diversity in these two datasets, we calculated p values and effect sizes taking into account the cancer type and the number of patients in each type (see STAR Methods for details). The results are shown in Figure 3B. In the TCGA-cisplatin dataset (BLDCA, CESC, and LUAD), we observed an overall reduction in the p values in all learning algorithms (because of the smaller sizes of each cancer-type dataset), and only the DNN models surpassed the statistical significance threshold. In the paclitaxel dataset (BRCA), DNN models improved and exceeded statistical significance in the top

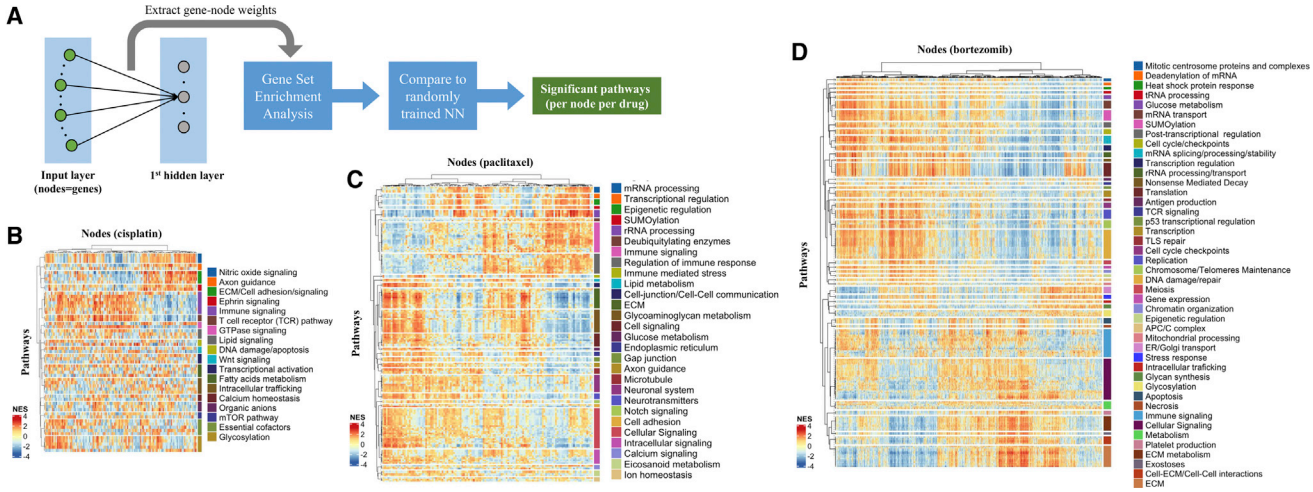


Figure 4. Pathway Enrichment Analysis

(A) Schematic representation of pathway enrichment analysis.

(B–D) Heatmap representation of normalized enrichment scores for each significant pathway across the nodes of the first hidden layer of the cisplatin (B), paclitaxel (C), and bortezomib (D) neural network models, respectively: rows correspond to significant pathways (for each drug), and columns correspond to nodes of the first hidden layer of each network.

80% and 100% selected genes, while the other two algorithms failed to produce statistically significant results. Finally, we evaluated the algorithms based on the AUC (area under the curve). In Figure 3C, we show boxplots of the AUC values—grouped by learning algorithm—across drugs and gene selection percentages. We found that DNN models yield statistically significantly higher AUCs compared with both RF and Enet models (Wilcoxon paired test). All metrics (effect sizes, p values, and AUCs with confidence intervals) for each algorithm, gene selection percentage, and patient dataset are reported in Data S1b, "supplementary_clinical."

To corroborate our findings, we performed survival analyses using the predicted IC_{50} values to split the patients into high- and low-sensitivity cohorts for every combination of learner and feature selection scheme (Figure S2). Overall, DNN performed better than Enet and RF models across drugs, selection, and significance level (Figure S3). In the cases of bortezomib and cisplatin (TCGA), it achieved significance (p value less than 0.05) more than the competition, whereas in the case of the PARP inhibitor, it was the only method to achieve significance at any selection level (Figures S2A, S2B, and S2D).

We then asked whether neural networks can learn biologically meaningful concepts, such as regulatory pathways. To answer this question, we extracted the weights connecting the input layer (gene expression) to each of the nodes of the first hidden layer of the optimal neural network architecture determined by the nested cross-validation approach for cisplatin, paclitaxel, and bortezomib. For each drug, we used the weights to perform pathway enrichment analysis, independently for each node of the first hidden layer. The weights of the first node were not driven by the magnitude of the gene expression (Table S3). As a control, we retrained the neural network on randomly permuted IC_{50} values, performed pathway enrichment analysis, and repeated the process 100 times (see STAR Methods for details). Finally, we kept only the statistically significantly enriched pathways (bootstrapped p value = 5%, using the randomly trained networks as control). In Figure 4A, we provide a schematic of the approach. In Figures 4B–4D (cisplatin, paclitaxel, and bortezomib, respectively), we show heatmap representations of the normalized enrichment scores (NESs) for the significant pathways for each node. We observed that nodes cluster into subgroups and each subgroup has its own signature of enriched pathways, suggesting possible connections between certain pathways and drug mechanisms. To find evidence linking these pathways to the action of each of the three drugs, we performed a detailed literature search. Publication matching was strict, scoring positive only when the status of a pathway clearly influenced the drug response. The results of our analysis (Data S1i–S1k) suggest that the neural network framework can recognize biological pathways that dictate the responsiveness of a given drug. For the examined drugs cisplatin, paclitaxel, and bortezomib, the degree of confirmation with prior knowledge (literature) was very high: 96%, 79%, and 68%, respectively. As an overall observation, we noted that although certain pathways were common among the interrogated drugs, there were also distinct ones that reflected their different mode of action. As an example, effectiveness to cisplatin depended on the status of the DNA damage response network, drug transporters, and RAS-like signaling molecules (Damia and Broggini, 2019; Housman et al., 2014; Galluzzi et al., 2014). In the case of paclitaxel, mechanisms or factors implicated in microtubule dynamics determine paclitaxel responsiveness (Orr et al., 2003; Barbuti and Chen, 2015; Marcus et al., 2005). Lastly, various signaling pathways affect response to bortezomib, particularly that of nuclear factor κB (NF-κB), the master regulator of immune response (Reddy and Czuczman, 2010; Kumar and Rajkumar, 2008).

DISCUSSION

Our study presents a clinical validation of cell-line-trained DNN models to predict drug response from gene expression. It appears that DNN captures the intricate biological interactions more effectively than the current state-of-the-art machine learning frameworks. Based on our findings, we believe that in the future, thorough molecular profiling of large cell-line collections followed by drug response assays applied on organotypic cultures (recapitulating tissue architecture) will provide an appealing training platform for delivering DNN-based tools that can eventually become an integral part of broader precision oncology efforts. To successfully pursue this vision, it is clear that a large amount of additional drug response and genomic data will be necessary to train accurate deep learning models, while extensive evaluation should be performed on multiple clinical datasets.

STAR★METHODS

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

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- METHOD DETAILS
  - 1. Gene expression cell line and patient datasets
  - 2. RNA-seq analysis
  - 3. Prediction of drug response
  - 4. Pathway analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND CODE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.celrep.2019.11.017.

ACKNOWLEDGMENTS

Financial support to V.G.G. and his team is from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant 722729 (SYNTRAIN); the Welfare Foundation for Social & Cultural Sciences (KIKPE), Greece; Pentagon Biotechnology, UK; DeepMed IO, UK; and NKUA-SARG grants 70/3/9816 and 70/3/12128. J.B. is financially supported by the Novo Nordisk Foundation grant 16584, the Danish Cancer Society (R204-A12617), and the Swedish Cancerfonden (170176). Financial support to P.A.T. is from the Medical Research Council – MRC (Confidence in Concept to support Dr. Tom Jackson for early pilot data and a subsequent DTP studentship to P.A.T. and V.G.G.) and the infrastructure of Manchester Cancer Research Centre and CRUK Manchester Institute. Support to D.T. is from the KMW offsets program. A.T. and the NYU Applied Bioinformatics Laboratories (ABL) are partially supported by the Cancer Center Support Grant P30CA016087 at the Laura and Isaac Perlmutter Cancer Center. On behalf of the OCCAMS Consortium, we acknowledge that infrastructure was supported by Cancer Research UK (CRUK).

AUTHOR CONTRIBUTIONS

K.V., study conception and design, scripting, bioinformatics analysis, experimental procedures, results interpretation, writing, and manuscript and figure preparation; T.S., scripting, development of deep-drug code, bioinformatics analysis, results interpretation, and figure preparation; A.P. and A.K., RNA sequencing (RNA-seq) analysis, scripting, and data analysis; H.Z., I.A., and L.G.A., scripting, bioinformatics analysis, and data analysis; D.T., P.S., F.K., P.A.T., and V.G.G., results interpretation and guidance; T.J.M., S.P.-P., K.R.M.S., R.F., and R.P., RNA-seq experimental procedures and analysis and clinical data preparation; A.K. and J.B., results interpretation, guidance, figures, and manuscript preparation; A.T., study design, supervision and guidance, results interpretation, figures, and manuscript preparation; and V.G.G., study conception and design, experimental procedures, data analysis and interpretation, guidance, manuscript preparation, and writing. All authors discussed the results and commented on the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 27, 2018
Revised: July 16, 2019
Accepted: November 5, 2019
Published: December 10, 2019

SUPPORTING CITATIONS

The following references appear in the Supplemental Information: Bibault et al. (2018); Chang et al. (2018); Larder et al. (2007); Xia et al. (2018).

REFERENCES

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.L., et al.; Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; ICGC PedBrain (2013). Signatures of mutational processes in human cancer. Nature *500*, 415–421.

Barbuti, A.M., and Chen, Z.S. (2015). Paclitaxel Through the Ages of Anticancer Therapy: Exploring Its Role in Chemoresistance and Radiation Therapy. Cancers (Basel) *7*, 2360–2371.

Bibault, J.E., Giraud, P., Housset, M., Durdux, C., Taieb, J., Berger, A., Coriat, R., Chaussade, S., Dousset, B., Nordlinger, B., and Burgun, A. (2018). Deep Learning and Radiomics predict complete response after neo-adjuvant chemoradiation for locally advanced rectal cancer. Sci. Rep. *8*, 12611.

Chang, Y., Park, H., Yang, H.J., Lee, S., Lee, K.Y., Kim, T.S., Jung, J., and Shin, J.M. (2018). Cancer Drug Response profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature. Sci. Rep. *8*, 8857.

Chiu, Y.C., Chen, H.H., Zhang, T., Zhang, S., Gorthi, A., Wang, L.J., Huang, Y., and Chen, Y. (2019). Predicting drug response of tumors from integrated genomic profiles by deep neural networks. BMC Med. Genomics *12* (*Suppl 1*), 18.

Costello, J.C., Heiser, L.M., Georgii, E., Gönen, M., Menden, M.P., Wang, N.J., Bansal, M., Ammad-ud-din, M., Hintsanen, P., Khan, S.A., et al.; NCI DREAM Community (2014). A community effort to assess and improve drug sensitivity prediction algorithms. Nat. Biotechnol. *32*, 1202–1212.

Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., and Tsirigos, A. (2018). Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat. Med. *24*, 1559–1567.

Damia, G., and Broggini, M. (2019). Platinum Resistance in Ovarian Cancer: Role of DNA Repair. Cancers (Basel) *11*, E119.

Ding, Z., Zu, S., and Gu, J. (2016). Evaluating the molecule-based prediction of clinical drug responses in cancer. Bioinformatics *32*, 2891–2895.

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., et al. (2018). The Reactome Pathway Knowledgebase. Nucleic Acids Res. *46* (*D1*), D649–D655.

Frankell, A.M., Jammula, S., Li, X., Contino, G., Killcoyne, S., Abbas, S., Perner, J., Bower, L., Devonshire, G., Ococks, E., et al.; Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) Consortium (2019). The

landscape of selection in 551 esophageal adenocarcinomas defines genomic biomarkers for the clinic. Nat. Genet. *51*, 506–516.

Galluzzi, L., Vitale, I., Michels, J., Brenner, C., Szabadkai, G., Harel-Bellan, A., Castedo, M., and Kroemer, G. (2014). Systems biology of cisplatin resistance: past, present and future. Cell Death Dis. *5*, e1257.

Garnett, M.J., Edelman, E.J., Heidorn, S.J., Greenman, C.D., Dastur, A., Lau, K.W., Greninger, P., Thompson, I.R., Luo, X., Soares, J., et al. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature *483*, 570–575.

Gautier, L., Cope, L., Bolstad, B.M., and Irizarry, R.A. (2004). affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics *20*, 307–315.

Geeleher, P., Cox, N.J., and Huang, R.S. (2014). Clinical drug response can be predicted using baseline gene expression levels and *in vitro* drug sensitivity in cell lines. Genome Biol. *15*, R47.

Housman, G., Byler, S., Heerboth, S., Lapinska, K., Longacre, M., Snyder, N., and Sarkar, S. (2014). Drug resistance in cancer: an overview. Cancers (Basel) *6*, 1769–1792.

Iorio, F., Knijnenburg, T.A., Vis, D.J., Bignell, G.R., Menden, M.P., Schubert, M., Aben, N., Gonçalves, E., Barthorpe, S., Lightfoot, H., et al. (2016). A Landscape of Pharmacogenomic Interactions in Cancer. Cell *166*, 740–754.

Kumar, S., and Rajkumar, S.V. (2008). Many facets of bortezomib resistance/susceptibility. Blood *112*, 2177–2178.

Lagergren, J., Smyth, E., Cunningham, D., and Lagergren, P. (2017). Oesophageal cancer. Lancet *390*, 2383–2396.

Larder, B., Wang, D., Revell, A., Montaner, J., Harrigan, R., De Wolf, F., Lange, J., Wegner, S., Ruiz, L., Pérez-Elías, M.J., et al. (2007). The development of artificial neural networks to predict virological response to combination HIV therapy. Antivir. Ther. (Lond.) *12*, 15–24.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444.

Leek, J.T., Johnson, W.E., Parker, H.S., Fertig, E.J., Jaffe, A.E., and Storey, J.D. (2016). sva: Surrogate Variable Analysis.

Libbrecht, M.W., and Noble, W.S. (2015). Machine learning applications in genetics and genomics. Nat. Rev. Genet. *16*, 321–332.

Marcus, A.I., Peters, U., Thomas, S.L., Garrett, S., Zelnak, A., Kapoor, T.M., and Giannakakou, P. (2005). Mitotic kinesin inhibitors induce mitotic arrest and cell death in Taxol-resistant and -sensitive cancer cells. J. Biol. Chem. *280*, 11569–11577.

Mulligan, G., Mitsiades, C., Bryant, B., Zhan, F., Chng, W.J., Roels, S., Koenig, E., Fergus, A., Huang, Y., Richardson, P., et al. (2007). Gene expression profiling and correlation with outcome in clinical trials of the proteasome inhibitor bortezomib. Blood *109*, 3177–3188.

Negrini, S., Gorgoulis, V.G., and Halazonetis, T.D. (2010). Genomic instability—an evolving hallmark of cancer. Nat. Rev. Mol. Cell Biol. *11*, 220–228.

Orr, G.A., Verdier-Pinard, P., McDaid, H., and Horwitz, S.B. (2003). Mechanisms of Taxol resistance related to microtubules. Oncogene *22*, 7280–7295.

R Core Team (2016). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).

Reddy, N., and Czuczman, M.S. (2010). Enhancing activity and overcoming chemoresistance in hematologic malignancies with bortezomib: preclinical mechanistic studies. Ann. Oncol. *21*, 1756–1764.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics *12*, 77.

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. Neural Netw *61*, 85–117.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.

van't Veer, L.J., and Bernards, R. (2008). Enabling personalized cancer medicine through analysis of gene-expression patterns. Nature *452*, 564–570.

Vougas, K., Sakellaropoulos, T., Kotsinas, A., Foukas, G.P., Ntargaras, A., Koinis, F., Polyzos, A., Myrianthopoulos, V., Zhou, H., Narang, S., et al. (2019). Machine learning and data mining frameworks for predicting drug response in cancer: An overview and a novel in silico screening process based on association rule mining. Pharmacol. Ther. *30*, 107395.

Weinstein, J.N. (2012). Drug discovery: Cell lines battle cancer. Nature *483*, 544–545.

Xia, F., Shukla, M., Brettin, T., Garcia-Cardona, C., Cohn, J., Allen, J.E., Maslov, S., Holbeck, S.L., Doroshow, J.H., Evrard, Y.A., et al. (2018). Predicting tumor cell line response to drug pairs with deep learning. BMC Bioinformatics *19* (*Suppl 18*), 486.

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

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Files are stored as rds (R binary format). There are 2 files per drug < drug > _cells.rds with the data used for training and < drug > _clinical.rds with the data used for testing. Data are batch normalized. Raw data are stored in the raw directory under the same link.	this paper	https://genome.med.nyu.edu/public/tsirigoslab/deep-drug-response/
Software and Algorithms		
Code for training models	this paper	https://github.com/TeoSakel/deep-drug-response
R version 3.5.1 with the following packages:	R Core Team, 2016	https://www.R-project.org
ComplexHeatmap_2.0.0		
ReactomePA_1.28.0		
broom_0.5.2		
h2o_3.20.0.2		
jsonlite_1.6		
org.Hs.eg.db_3.8.2		
pROC_1.15.3		
survival_2.44-1.1		
survminer_0.4.6		
tidyverse_1.2.1		

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests should be directed to and will be fulfilled by the Lead Contact, Vassilis G Gorgoulis (vgorg@med.uoa.gr). This study did not generate new unique reagents.

METHOD DETAILS

All scripting, data-processing, statistical calculations, except RNA-seq analysis, have been performed with R-language for statistical computing (R Core Team, 2016).

1. Gene expression cell line and patient datasets

Gene Expression of GDSC cell lines

Raw gene expression data for the GDSC cell lines were collected from the Array Express repository (E-MTAB-3610). Drug response data were obtained from: ftp://ftp.sanger.ac.uk/pub4/cancerrxgene/releases/release-5.0/gdsc_manova_input_w5.csv. The raw data was used to train our drug response models after batch correction as described below (“Batch correction between GDSC cell line data and clinical datasets”). The Bioconductor Affy package (Gautier et al., 2004) was used to apply Robust Multiarray Averaging (RMA) normalization to the aforementioned dataset. The normalized gene values can be found in file “raw/cells.rds” available in the provided link (see Data and Code Availability).

Bortezomib

Raw gene expression data for this clinical trial were not available. Only the pre-processed MAS5 normalized data was publicly available at GEO:GSE9782. Because concurrent processing of two data-sets with different normalizations (RMA and MAS5) can be problematic, we proceeded to the analysis as described below (Batch correction between GDSC cell line data and clinical datasets). The actual data file, which also contains the clinical response information, was retrieved from (Geeleher et al., 2014).

Cisplatin clinical trial

The pre-processed RMA-normalized gene expression data also containing the clinical response information was retrieved from (Geeleher et al., 2014).

TCGA cisplatin and paclitaxel datasets

These datasets were obtained from the (Frankell et al., 2019).

PARP inhibitor dataset

Note, the name of the inhibitor used in the dataset from MD Anderson cannot be disclosed as it was applied in the context of a clinical trial.

OCCAMS dataset

Raw gene expression data from esophageal adenocarcinomas treated with three cycles of neoadjuvant cisplatin-based combination chemotherapy followed by esophagectomy and corresponding histological response and tumor regression grade was obtained from the International Cancer Genome Consortium (ICGC) dataset through the OCCAMS consortium (Frankell et al., 2019).

Batch correction between GDSC cell line data and clinical datasets

To correct for the batch effect between the raw GDSC cell line and clinical datasets we used the ComBat function of the SVA Bioconductor package (Leek et al., 2016) with batch, cell lines versus clinical, as the only covariate. DNN, Enet and RF regression models were trained as described in methods above using training with cross-validation only on the cell line gene expression data to determine the optimal models for each learning algorithm. The optimal models were utilized to predict z-score normalized IC_{50} values for each patient in the clinical dataset from the patient's gene expression data. The code for normalization is also available in the GitHub page accompanying this publication.

2. RNA-seq analysis

RNA-seq analysis of the OCCAMS clinical samples has been previously described (Frankell et al., 2019), while the MD Anderson clinical set has been analyzed as follows. RNA was purified and the polyA+ mRNA fraction was used to generate stranded cDNA libraries according to the following: Quantification of Genomic RNA using Picogreen (Invitrogen, Carlsbad, CA, USA) and quality assessment using a 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA) was performed for each sample. RNA from each sample (1 μg) was fragmented and converted into double stranded cDNA and then preceded to library prep using TrueSeq RNA Sample Preparation from Illumina according to the manufacturer protocol. The library prep includes repair ends, A-tailing, Adaptor Indexes ligation followed by PCR amplification (15 cycles). The PCR primers were removed using 1.8x volume of Agencourt AMPure PCR Purification kit (Agencourt Bioscience Corporation). At the end of the library prep, samples were analyzed and quantified on TapeStation (Agilent) using the DNA High Sensitivity kit (Agilent) to verify correct fragment size and to ensure the absence of extra bands. Equimolar amounts of DNA were pooled for capture (8 samples per pool) and verified by TapeStation. The captured libraries were sequenced on a HiSeq 4000 (Illumina Inc., San Diego, CA, USA) on a version 3 TruSeq paired end flowcell according to manufacturer's instructions at a cluster density between 700 – 1000 K clusters/mm2. Sequencing was performed for 2 × 100 paired end reads with a 7 nt read for indexes using Cycle Sequencing v3 reagents (Illumina). The resulting BCL files containing the sequence data were converted into ".fastq.gz" files and individual libraries within the samples were demultiplexed using CASAVA 1.8.2 with no mismatches. All regions were covered by > 20 reads.

Raw reads from the RNASeq samples were processed using TopHat 2 for read alignment, FastQC and RSeQC for read and alignment quality assessment and HTSeq for expression count. The reads were aligned to the hg19 version of the human genome and mapping to the human transcriptome according to UCSC gene annotations.

3. Prediction of drug response

3.1. Gene selection

Gene selection was performed by selecting the top percentage of highly variable genes in the training set. The median absolute deviation (MAD) was used to quantify gene variability. Various percentages of highly variable genes were tested to evaluate the robustness of each algorithm to the number of genes used for modeling drug response.

3.2. Hyper-parameter optimization on cell line data

The hyperparameters for the models were estimated using random grid search and cross-validation on the cell line dataset. The optimal models were then tested on the patient drug response dataset. In particular, for each learning algorithm, 30 sets of model parameters were selected at random from the grid and a 5-fold cross validation was performed for every drug and feature selection strategy. For every split, the most variable genes were extracted and all 30 models were trained and evaluated against the left-out dataset. The parameters with the lowest average mean-square-error (MSE) across all splits were used to train the final model on the entire cell line dataset, which was used to predict the patients' IC_{50} z-score from the clinical gene expression data. The grid used is summarized in Table S2. Deep Neural Networks (DNN) and Random Forests (RF) were constructed using the H2O.ai platform [http://www.h2o.ai/]. Elastic Net models were constructed using the glmnet R package.

3.3. Deep Neural Network Architecture

The basic unit in the model is the neuron, a biologically inspired model of the human neuron. In humans, the varying strengths of the neurons' output signals travel along the synaptic junctions and are then aggregated as input for a connected neuron's activation. In the model, the weighted combination ($\alpha = \Sigma n_{i=1} w_i x_i + b$) of input signals is aggregated, and then an output signal $f(\alpha)$ transmitted by the connected neuron. The function f represents a nonlinear activation function used throughout the network and allows it to model non-linear patterns. Multi-layer, feed-forward neural networks consist of many layers of interconnected neuron units, starting with an input layer to match the feature space, followed by multiple layers of nonlinearity, and ending with a

classification layer to match the output space. The inputs and outputs of the model's units follow the basic logic of the single neuron described above. Bias units are included in each non-output layer of the network. The weights linking neurons and biases with other neurons fully determine the output of the entire network. Learning occurs when these weights are adapted via backpropagation during the training phase to minimize the error on the labeled training data. More specifically, for each training example j, the objective is to minimize the loss function, L(W, B | j). Here, W is the collection $\{W_i\}_{1:N-1}$, where W_i denotes the weight matrix connecting layers i and i + 1 for a network of N layers. Similarly B is the collection $\{b_i\}_{1:N-1}$, where b_i denotes the column vector of biases for layer i + 1. Apart from backpropagation, modern architectures use a plethora of engineering tricks to fit the weights of a network such as dropout and batch normalization. For this application, the hyperbolic tangent (tanh) with dropout was used as neuron activation function and the mean squared error (MSE) as loss function. DNNs have the power to learn feature representations of the sample-space over multiple levels of abstraction. This capability negates the need for feature selection and engineering, offering at the same time superior generalisation potential.

3.4. Nested cross-validation

We also performed nested cross-validation on the cell line data in order to estimate the performance of the models exclusively on the cell line data. In this scenario, the same process and grid were used as described above, but the optimal models (using MSE as a metric) for each learner and gene selection percentage, were evaluated on the left-out data of the outer 5-fold cross validation split of the cell line dataset. For each run, we report the MSE, mean-absolute-error (MAE) and the Pearson correlation coefficient between the predicted and the actual IC_{50} values.

3.5. Evaluation metrics for drug response in clinical datasets

In all the clinical datasets analyzed (Table S1; Data S1) in the current study except the one obtained from MD Anderson, patients whose response was marked as Stable Disease (SD) and Progressive Disease (PD) were treated as non-responders, while the ones with Partial Response (PR) and Complete Response (CR) as responders. For the MD Anderson dataset the clinical Primary Investigator suggested that patients with CR, PR and SD > = 6 months should be treated as responders while the ones with SD < 6months and PD as non-responders. For the OCCAMs cohort response to neo-adjuvant cisplatin-based chemotherapy was assessed histopathologically in the esophagectomy specimen and classified according to Mandard with Tumor Regression Grade (TRG); TRG 1, 2 and 3 considered to be responders and TRG 4 and 5 considered to be non-responders. The effect size was computed as the mean difference in predicted IC_{50} z-scores of responders versus non-responders. The p value was calculated using the Wilcoxon rank sum (Mann-Whitney) test with alternative hypothesis that responders have a higher mean IC_{50}. To adjust for the cancer type diversity in our datasets, we computed the effect size and p value for each cancer type individually and then merged the result by taking the weighted average of the effect sizes and using Stouffer's method (sum of z) for the p values. For this meta-analysis, we only considered cancer types with more than 5 patients in each group and 40 patients in total. We also quantified the ability of each model to classify patients as responders and non-responders using the Area Under the Curve (AUC) calculated on the patients ranked by their predicted IC_{50} values. The AUC values and confidence intervals were computed using the pROC package (Robin et al., 2011). The Wilcoxon signed ranked test was used to access the differences between the pairs of training algorithms.

4. Pathway analysis

For the optimal network of each drug, as determined by nested cross-validation, we extracted the weights of the first hidden layer linking genes from the input layer to the nodes. Using these weights we ran a gene set enrichment analysis (GSEA) (Subramanian et al., 2005) and calculated the normalized enrichment score of every node against every pathway in Reactome (Fabregat et al., 2018). To call a drug-pathway score significant, we averaged the enrichment score of the positively and the negatively enriched nodes for both the original optimal network as well as its bootstrapped versions. The bootstrapped versions were generated as follows: we permuted the IC_{50} values of the drug 100 times and reran the training with all the other hyperparameters fixed (per drug). Then, we calculated a p value for positive/negative enrichment as the number of times the average enrichment score of the original dataset was higher/lower than the respective bootstrapped versions and adjusted using false discovery rate. We called an interaction significant if either the positive or negative enrichment p value was less than 0.05. We created one heatmap per drug: rows represent pathways, columns correspond to nodes of the first layer of the neural network and cells are colored based on the normalized enrichment score. Pearson correlation distance and the Ward's method were used to cluster rows and columns.

QUANTIFICATION AND STATISTICAL ANALYSIS

For every prediction model, we split the patients, based on their predicted IC_{50}, into 3 quantile groups for the TCGA drugs (*Bortezomib*, *Cisplatin*, and *Paclitaxel*) and 2 groups for *Cisplatin* OCCAMS and *PARP inhibitor*, since they had far fewer patients. We then performed Kaplan-Meyer survival analysis to contrast the groups of the lowest and highest IC_{50}. For Paclitaxel, Enet failed to predict any variance for two feature selection schemes so the corresponding facets are empty. Survival models and p values were calculated with R's survival package (v2.44-1.1) and were plotted using survminer (v0.4.6). Models with p value less than 0.05 were considered significant (table in Figure S3). We also draw the cumulative distribution of all the p value (graph in Figure S3).

DATA AND CODE AVAILABILITY

All data files used for training and testing are available via this link: https://genome.med.nyu.edu/public/tsirigoslab/deep-drug-response/. More specifically, the following files are made available:

- bortezomib_cells.rds: batch-normalized GDSC/bortezomib dataset used for training
- bortezomib_clinical.rds: bortezomib clinical response data
- cisplatin-occams_cells.rds: batch-normalized GDSC/cisplatin (OCCAMS) dataset used for training
- cisplatin-occams_clinical.rds: cisplatin (OCCAMS) clinical response data
- cisplatin_cells.rds: batch-normalized GDSC/cisplatin (TCGA) dataset used for training
- cisplatin_clinical.rd: cisplatin (TCGA) clinical response data
- paclitaxel_cells.rds: batch-normalized GDSC/paclitaxel (TCGA) dataset used for training
- paclitaxel_clinical.rds: paclitaxel (TCGA) clinical response data
- parpi_cells.rds: batch-normalized GDSC/PARP inhibitor dataset used for training
- parpi_clinical.rds: PARP inhibitor clinical response data

The training drug response data before batch-normalization can be found in the "raw/" folder under the same link.

The R code used to normalize, train, test and validate the DNN, RF and Enet models is deposited in GitHub: https://github.com/TeoSakel/deep-drug-response.

For script parameters, see Data S2.