*Bioinformatics*, 36(3), 2020, 813–818
doi: 10.1093/bioinformatics/btz645
Advance Access Publication Date: 26 August 2019
Original Paper

Gene expression

# DeepCOP: deep learning-based approach to predict gene regulating effects of small molecules

**Godwin Woo[1], Michael Fernandez [1], Michael Hsing[1], Nathan A. Lack[1,2], Ayse Derya Cavga[3] and Artem Cherkasov[1,*]**

[1]Department of Urologic Sciences, Faculty of Medicine, Vancouver Prostate Centre, University of British Columbia, Vancouver, British Columbia V6H 3Z6, Canada, [2]School of Medicine, Koç University, Rumelifeneri Yolu, Istanbul 34450, Turkey and [3]Chemical and Biological Engineering, Koç University, Rumelifeneri Yolu, Istanbul 34450, Turkey

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on March 4, 2019; revised on June 11, 2019; editorial decision on August 12, 2019; accepted on August 21, 2019

## Abstract

**Motivation:** Recent advances in the areas of bioinformatics and chemogenomics are poised to accelerate the discovery of small molecule regulators of cell development. Combining large genomics and molecular data sources with powerful deep learning techniques has the potential to revolutionize predictive biology. In this study, we present Deep gene COmpound Profiler (DeepCOP), a deep learning based model that can predict gene regulating effects of low-molecular weight compounds. This model can be used for direct identification of a drug candidate causing a desired gene expression response, without utilizing any information on its interactions with protein target(s).

**Results:** In this study, we successfully combined molecular fingerprint descriptors and gene descriptors (derived from gene ontology terms) to train deep neural networks that predict differential gene regulation endpoints collected in LINCS database. We achieved 10-fold cross-validation RAUC scores of and above 0.80, as well as enrichment factors of >5. We validated our models using an external RNA-Seq dataset generated in-house that described the effect of three potent antiandrogens (with different modes of action) on gene expression in LNCaP prostate cancer cell line. The results of this pilot study demonstrate that deep learning models can effectively synergize molecular and genomic descriptors and can be used to screen for novel drug candidates with the desired effect on gene expression. We anticipate that such models can find a broad use in developing novel cancer therapeutics and can facilitate precision oncology efforts.

**Contact:** acherkasov@prostatecentre.com

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Chemogenomics studies chemicals and their effects on cellular states (Stegmaier *et al.*, 2004) and can serve as a very powerful tool for drug discovery in high-throughput screening (Bredel and Jacoby, 2004). Chemical libraries are screened for drug candidates that target-specific cell states (Stegmaier *et al.*, 2004) and can lead to therapies to treat disorders that can be characterized by mis-expressed gene pathways. Mice models have shown that treatments that restore mis-expressed genes are associated with positive physiological effects in tissues (Wagner *et al.*, 2015).

The evolution of chemogenomics has been expedited by the recent emergence of large standardized genomic datasets. Many of them characterize responses of diverse cell lines to various chemicals and are publicly available (Lavertu *et al.*, 2018). These progressively larger datasets are paving the way for a big data revolution in pharmacology (Lavertu *et al.*, 2018). Importantly, the emergence of publicly available pharmacogenomics data, promote the development of novel bioinformatics and cheminformatics tools and approaches for big data mining and machine learning (Lavertu *et al.*, 2018). Machine learning tools such as deep neural networks (DNNs) can identify hidden characteristics of large datasets by generalizing information and storing them as neuron weights (Schmidhuber, 2015). In drug discovery, recent reports have pointed out the superiority of deep learning in the prediction of biological properties of chemical compounds over traditional machine learning methods (Fernandez *et al.*, 2018; Mayr *et al.*, 2016).

In this article, we introduce Deep gene COmpound Profiler (DeepCOP), a deep learning computational tool to predict drug effects on gene expression endpoints from the LINCS L1000 dataset (Subramanian *et al.*, 2017). A DNN in the form of a multilayer perceptron (MLP) was developed for classification and prediction. Initially, we evaluated its efficacy as a prediction model using two binary classification models that determine genes up-regulation and

down-regulation. Consequently, we demonstrated how the developed models can recover up-regulated and down-regulated genes from an in-house RNA-Seq dataset on the LNCaP cell line.

We propose that the DeepCOP approach can be used to screen for drugs that target-specific gene regulations and can be used for applications such as reversing mis-regulated genes in diseases such as cancer.

## 2 Materials and methods

### 2.1 LINCS CMap L1000 cancer genomic dataset

The largest and latest gene perturbation dataset published to date is the LINCS Connectivity Map L1000 dataset (Subramanian *et al.*, 2017). The LINCS project has collected over 1.3 million gene expression profiles with 978 landmark genes against 19 811 compounds on 77 cancer cell lines. Such a large dataset can facilitate research into numerous hypotheses (Lavertu *et al.*, 2018). There has already been a variety of outcome predictions on the previous smaller L1000 dataset such as predicting adverse side effects of drugs (Wang *et al.*, 2016) and predicting cell viability from drug interactions (Szalai *et al.*, 2018). The dataset was split into 2 phases. Phase 1 which contains most of the gene profiles was released in May of 2014, while Phase 2 was released in March of 2016 and continues to be updated every 6 months. They can be found on the https://clue.io website.

From the LINCS CMap L1000 dataset (Subramanian *et al.*, 2017), we used the Level 5 drug–gene interaction dataset that collects the gene profiles of 978 landmark genes measured by perturbagen experiments on specific cell lines. The direct download link can be accessed here: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE92742. Gene expression was measured using flow cytometry data and was done in replicates of 1–4 each. These values were converted into standardized *z-scores* and were the values we used to train the models. Data from each cell-line was trained separately as its own predictive model. We kept only experiments with an exposure time of 24 h, which represents about 53% of the dataset. In order to standardize doses, we kept only those that had units of measurements in μM, which represents 43% of the experiments. To keep the code simple we did not use experiments measured in units of ng/ml or ng/μl which represent less than 3% of all the experiments measured in μM. Each perturbagen-cell line experiment was mostly performed in four different dose amounts. The number of similar drug–gene experiments with different concentrations were not enough to adequately calculate reliable $IC_{50}$ values, therefore, we chose the concentration with the largest amount of experimental data, which was 10 μM bin and excluded other dosages from the training set. Figure 1A shows the distribution of drug concentrations used in experiments in the dataset that were reported in μM. After culling the dataset, we chose the top six cell lines that had the most drug interactions data, i.e. PC3, MCF7, VCAP, A549, A375 and HT29. Table 1 shows a list of cell lines ordered by the number of tested drug experiments with the top six cell lines in bold.

### 2.2 RNA-Seq data for in-house compounds from the Vancouver Prostate Centre

LNCaP cells ($6 \times 10^6$) were seeded in RPMI 1640 media containing 10% FBS and 1% penicillin/streptomycin. Following O/N incubation, the media was changed to RPMI 1640 containing 5% CSS and the cells were then grown for 48 h at 37°C with 5% $CO_2$. On the day of treatment, CSS media was refreshed and the cells were treated with DHT (10 nM) and either DMSO (Control) or compound MDV3100 (Enzalutamide): 1.5 μM, VPC-17005 (Dalal *et al.*, 2018): 6 μM or VPC-14449 (Li *et al.*, 2014): 6 μM). Concentration of the compound was selected based on the published $IC_{90}$. Cells were treated for 6 h and then total RNA was extracted using Trizol and the Qiagen RNeasy RNA Extraction Kit. Three biological replicas were collected for each condition. The concentrations and the qualities of the RNA samples were checked by Nanodrop and tape station. For sequencing, RNA was subjected to Poly-A selection, and sequenced using the BGISeq500 platform. The raw reads were pre-filtered by removing adaptor sequences, contamination and low-quality reads. The high quality of the filtered reads was confirmed using the FastQC (Andrews, 2014) quality control tool. The reads were mapped and annotated using STAR Aligner (Dobin *et al.*, 2013) (version 2.6.1c) using the latest release reference genome and annotations from the GENCODE (Frankish *et al.*, 2019) project; release 29, GRCh38.p12. The reads were then counted using featureCounts (Liao *et al.*, 2014). Differential gene expression analysis for different treatments was then performed using the DESeq2 method (Love *et al.*, 2014). Note that when comparing predictions with RNA-Seq data using LNCaP cell line, we previously trained our general LINCS-based models with both 3 and 24 h data since they were the only exposure times available that were closest to the 6 h exposures of our RNA-Seq data.

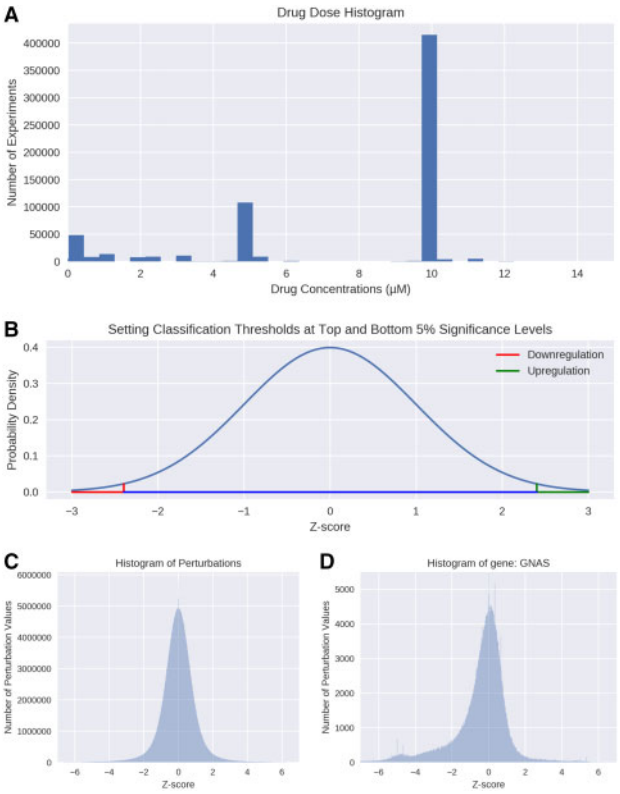


**Fig. 1.** (**A**) Distribution of the number of experiments in the dataset that have reported concentration measures of μM and their drug concentrations. (**B**) Gene perturbation scores (*z*-scores) split into classes at 5% significance levels. (**C**) Histogram of perturbations for the entire L1000 level 5 dataset. (**D**) Histogram of a landmark gene GNAS with the greatest variance

**Table 1.** A list of cell lines sorted by number of compounds they have tested with the top six cell lines used to build the DNN models appear highlighted in bold

| Cell line | Primary site | Subtype | Compounds |
|---|---|---|---|
| **VCAP** | **Prostate** | **Carcinoma** | **6730** |
| **A549** | **Lung** | **Lung cancer carcinoma** | **6410** |
| **A375** | **Skin** | **Malignant melanoma** | **6076** |
| **PC3** | **Prostate** | **Adenocarcinoma** | **5517** |
| **MCF7** | **Breast** | **Adenocarcinoma** | **5508** |
| **HT29** | **Large intestine** | **Colorectal adenocarcinoma** | **5491** |
| HA1E | Kidney | Normal kidney | 4911 |
| HEPG2 | Liver | Hepatocellular carcinoma | 4790 |
| HCC515 | Lung | Carcinoma | 3751 |
| Others | | | <1000 |

### 2.3 Binary classifiers

For the binary classification models, perturbation data were labeled according to a significance level threshold. We calculated *z-score* thresholds that correspond to 5% and 10% significance levels. For the 'up-regulation' model, in each cell line, perturbations above the top threshold were labeled as actively up-regulated, while the rest were labeled as inactive up-regulation representing the genes that were not significantly up-regulated. Similarly, for the 'down-regulation' model, in each cell line, perturbations below the bottom threshold were labeled as actively down-regulated, while the rest were labeled as inactive down-regulation. In the case of multiple experimental replicates, we used a voting system by classifying each replicate, and taking the label with the most replicates. In the case of a tie, we discarded the sample, and it would not be considered actively up-regulated or down-regulated (see Fig. 1B).

### 2.4 Morgan (circular) fingerprints

To correlate the chemical structure to gene perturbation, Morgan descriptors (Morgan, 1965) were calculated using the RDKIT Open-source Python library (Open-source, 2006). This molecular representation continues to be one of the most broadly and intensely used fingerprints in recent studies (Rogers and Hahn, 2010; Wei *et al.*, 2019; Zheng *et al.*, 2018). We calculated the descriptors on the canonical SMILES representation using a radius value of 2 to obtain a one-hot vector of 2048 features for 19 811 compounds. In early tests, we found Morgan descriptors to show better prediction ability compared to other proudly utilized chemical descriptors—Molecular Access System (MACCS) fingerprints.

### 2.5 Gene ontology descriptors

There are currently no broadly-accepted methods for quantifying gene descriptors. However, the gene ontology (GO) consortium (Ashburner *et al.*, 2000) currently has 40 thousand GO terms with over 200 000 qualitative annotations for homo sapiens (The Gene Ontology Consortium, 2018). We devised a novel in-house implementation to quantify this gene information for each of the 978 landmark genes from the L1000 dataset encoded in a one hot vector. These 978 genes were selected by the LINCS consortium using principal component analysis to recover 82% of information in the full transcriptome (Subramanian *et al.*, 2017). From the OntologyX R package (Greene *et al.*, 2017), we took all GO terms that were attributed to at least three or more landmark genes. Each gene was then described in terms of that pool of GO terms. Because GO terms are annotations of a gene's cellular and biological processes and activities, genes that share the same biological pathways or activities will also share the same GO terms. Moreover, we have already previously demonstrated that GO terms can be effectively used gene descriptors in machine learning experiments (Hsing *et al.*, 2008). By using the GO terms as input, a DNN model should be able to differentiate genes by differences in their related biological processes. Our proposed gene quantification method resulted in a one-hot vector of 1107 features per gene. Each feature represents the inclusion or exclusion of each GO term. The concatenation of drug fingerprints one-hot vectors and GO descriptors were used to train the DNN models against the intensity of drug-gene interactions.

### 2.6 Deep neural networks

DNNs have the ability to generalize data iteratively over each successive network layer and as a result exhibits better performance for larger datasets over traditional machine learning algorithms (Mahapatra, 2018). We trained a DNN model with an architecture of a simple back-propagated feed-forward fully connected MLP with four (two hidden) layers. Each of the first three layers had a size of 3155 (2048 + 1107) nodes. Each of the two hidden layers were supplemented with a dropout layer with value of 0.2. The last output layer had two nodes which represented the active and inactive classes. The model was trained using a binary cross-entropy classifier and was optimized with a Scaled Exponential Linear Unit (SELU) input activation function as well as a Rectified Linear Unit (ReLU) activation function in the hidden layers (Fig. 2).

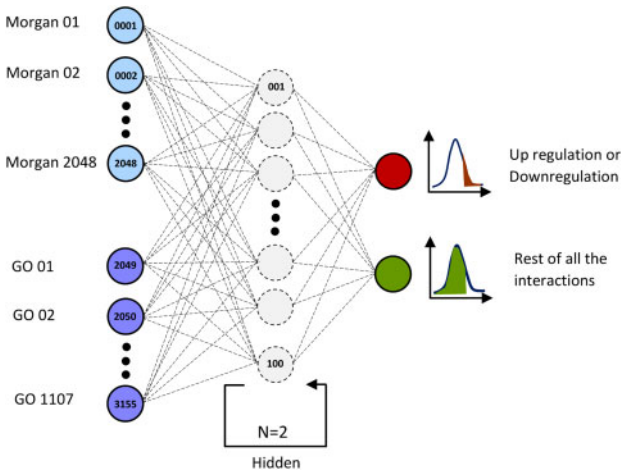


**Fig. 2.** The MLP architecture used in the binary classification model

Different regularization techniques did not improve validation accuracy significantly. Adaptive moment estimation (ADAM) was used as the optimizer. We hand tuned these hyperparameters using grid search to find values that were generic enough to be used on all cell lines. We were able to obtain more specific dropout and regularization values using Bayesian optimization, however, they did not improve validation accuracy significantly. We also compared our results with the popular random forest classifier which was only able to achieve area under the receiver-operating curves (AUC) scores up to 70% (see Supplementary Table S8).

### 2.7 Performance evaluation

Performance was measured using AUC validated against an internal validation split from the LINCS dataset. For each prediction score, we also calculated an associated modified version of enrichment factor (Bender and Glen, 2005) using precision divided by the random probability of finding an active sample ($E_f$).

#### 2.7.1 LINCS training and internal validation

First, we trained our models on LINCS data and evaluated its prediction performance using 10-fold cross-validation to calculate AUC and $E_f$.

#### 2.7.2 External validation on RNA-Seq predictions

We used the trained models to obtain up-regulation and down-regulation predictions for Enzalutamide and in-house compounds VPC-17005, and VPC-14449 on the LNCaP cell line. We then obtained true perturbation values from in-house RNA-Seq experiments by using a *P*-adjusted value with significance level of 5% to determine true active perturbation class. Having both predicted values and the true RNA-Seq values, we calculated and reported AUC and $E_f$.

### 2.8 Source code and data availability

Data processing, training and validation for the LINCS dataset were implemented in Python 3.6. R programming was used to calculate gene_descriptors described in Section 2.5. All Python and R source code is available at https://github.com/godwinwoo/DeepCOP. The RNA-Seq data is available at NCBI NIH Gene Expression Omnibus (GEO accession ID GSE127816).

## 3 Results

### 3.1 Statistical analysis

Initially, we performed basic exploratory statistical analysis of Phase 1 of the L1000 dataset from where we only used the top six cell lines highlighted in bold in Table 1, corresponding to 978 landmark genes. Statistics are listed in Table 2, while Figure 1C shows a

histogram of the perturbation *z-scores* for the entire level five dataset. In order to explore the distribution of the perturbation scores for a typical gene, we measured the variance of each landmark gene and obtained the histogram for gene GNAS which has the highest *z-score* variance depicted in Figure 1D. It can be seen, from the left-skewed distribution, that this particular gene has a tendency to be down-regulated by the small molecules tested in the training set. Table 1 lists all the cell lines in the dataset that have documented experiments with more than 1000 compounds. Figure 1A illustrates the distribution of experiments with their drug concentrations.

### 3.2 Deep learning binary classifiers

We designed binary DNNs to identify 'up-regulation' and 'down-regulation' activity by applying right-tailed and left-tailed significance levels to the *z-score* distributions in Figure 1B as described in the Section 2.3. Gene perturbations were labeled into corresponding classes depending on 5% significance level thresholds. After categorizing the drug–gene interactions according to the specific thresholds, we used 10-fold cross-validation to optimize each DNN model. For each binary classifier, the DNN used an output layer of two nodes with softmax activation and 'binary cross-entropy' loss cost function (Schmidhuber, 2015) to identify up-regulation and down-regulation interactions. The accuracy and AUC measures are reported in Table 3. The AUC scores for the 5% significance level are greater than 80% which is significantly better than random (AUC = 0.5) and is consistent with the scores of other recent DL papers in quantitative biology (Mullane *et al.*, 2019; Sureyya Rifaioglu *et al.*, 2019). Enrichment factors varied between 6–8 for the 5% significance level models and 3–4 for the 10% significance level models.

### 3.3 Significance level analysis

Up-regulation and down-regulation signals were defined as: *z-score* <left-tailed significance level and *z-score* >right-tailed threshold, respectively. Figure 3A illustrates how the AUC values of the two binary classifier models for the six cell lines increased from 0.60 to 0.85 as the significance levels decreased in the range from 25% to 5%. The top 5–10% of the gene perturbations were detected with higher accuracies across the six cell lines. An example of the internal validation ROC plots for recognition of gene up-regulation in the PC3 cell-line is depicted in Figure 3B for the 5% significance level. We observed that increasing the up-regulation significance level decreased the DNNs ability to identify specific gene perturbations. Similar behaviors were observed also for the down-regulation classifier across all six cell lines (see Supplementary Table S1 for the data for Fig. 3A).

### 3.4 Gap analysis

The smaller the significance levels were, the better the DNNs were able to predict extreme perturbations. This suggested that we should exclude negative examples that are very close to the positive class significance level threshold. The rationale was that the DNN may have considered these perturbations as ambiguous causing the DNN model to behave randomly around the significance level threshold. This could have been affecting the performance of the binary classifiers causing a decrease in AUC scores. To study this effect, we defined a sample exclusion gap that expanded from the classification significance level

**Table 3.** AUC scores for internal test set predictions for threshold of 5% and 10% significance levels

| Significance level | Cell line | Sample size | Down-regulation | | | Up-regulation | | |
|---|---|---|---|---|---|---|---|---|
| | | | AUC | *F*-score | $E_f$ | AUC | *F*-score | $E_f$ |
| 5% | PC3 | 6 163 650 | 0.84 | 0.36 | 8.44 | 0.84 | 0.36 | 8.56 |
| | MCF7 | 5 876 558 | 0.84 | 0.38 | 8.82 | 0.84 | 0.39 | 8.86 |
| | VCAP | 5 555 089 | 0.84 | 0.36 | 8.28 | 0.84 | 0.36 | 8.27 |
| | A549 | 4 529 363 | 0.81 | 0.34 | 6.14 | 0.81 | 0.34 | 6.03 |
| | A375 | 1 607 343 | 0.82 | 0.40 | 6.23 | 0.83 | 0.40 | 6.28 |
| | HT29 | 1 274 726 | 0.81 | 0.36 | 6.15 | 0.81 | 0.36 | 6.48 |
| 10% | PC3 | 6 163 650 | 0.81 | 0.39 | 4.35 | 0.81 | 0.39 | 4.35 |
| | MCF7 | 5 876 558 | 0.82 | 0.41 | 4.58 | 0.82 | 0.41 | 4.53 |
| | VCAP | 5 555 089 | 0.81 | 0.40 | 4.35 | 0.81 | 0.40 | 4.26 |
| | A549 | 4 529 363 | 0.79 | 0.39 | 3.39 | 0.79 | 0.39 | 3.41 |
| | A375 | 1 607 343 | 0.79 | 0.42 | 3.48 | 0.79 | 0.42 | 3.58 |
| | HT29 | 1 274 726 | 0.77 | 0.39 | 3.55 | 0.77 | 0.39 | 3.48 |

*Note:* AUC was reported separately for both up-regulation and down-regulation classifications as each class was predicted by its own model.

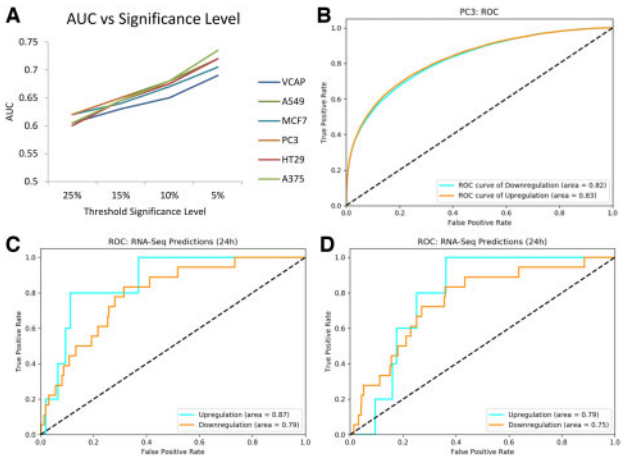


**Fig. 3.** (**A**) AUC values for different up-regulation threshold for six cell types used in this study. (**B**) ROC plots and AUC values for PC3 cell line for threshold of 5% significance level. (**C**) ROC curves for predicting RNA-Seq perturbations using a model trained on 3 h exposure time. (**D**) ROC curves for predicting RNA-Seq perturbations using a model trained on 24 h exposure time

**Table 2.** Exploratory statistics of the L1000 dataset including significance levels and corresponding *z-score* values

| Statistics | PC3 | MCF7 | HT29 | A375 | VCAP | A549 | Overall (76) |
|---|---|---|---|---|---|---|---|
| Minimum | −10.0 | −10.0 | −10.0 | −10.0 | −10.0 | −10.0 | −10.0 |
| Maximum | −10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 |
| Mean | −0.013 | −0.012 | −0.003 | −0.001 | −0.005 | −0.017 | −0.005 |
| Median | 0.001 | 0 | 0.005 | 0.001 | 0.002 | 0 | 0.002 |
| Standard deviation | 1.052 | 1.083 | 1.048 | 1.118 | 1.059 | 1.098 | 1.078 |
| *z-scores* at 5% | −1.538 | −1.562 | −1.487 | −1.578 | −1.493 | −1.575 | −1.534 |
| *z-scores* at 10% | −1.061 | −1.08 | −1.027 | −1.089 | −1.044 | −1.092 | −1.065 |
| *z-scores* at 25% | −0.504 | −0.516 | −0.491 | −0.52 | −0.504 | −0.524 | −0.511 |
| *z-scores* at 50% | 0.001 | 0 | 0.005 | 0.001 | 0.002 | 0 | 0.002 |
| *z-scores* at 75% | 0.504 | 0.515 | 0.501 | 0.529 | 0.509 | 0.522 | 0.517 |
| *z-scores* at 90% | 1.036 | 1.058 | 1.028 | 1.096 | 1.033 | 1.067 | 1.058 |
| *z-scores* at 95% | 1.463 | 1.498 | 1.461 | 1.568 | 1.453 | 1.5 | 1.496 |

threshold towards the distribution mean. The negative class thresholds were set using a gap factor as a percentage of the distance from the 10% to 90% significance level thresholds toward the mean (see Supplementary Fig. S2). We trained classifier models for gap factor increments of 10% and evaluated AUC, precision and recall depicted in Figure 4. In general, increasing the gap did increase each model's AUC scores. However, the cost of removing gap molecules would be a decrease in trained chemical diversity (see Supplementary Table S2 data for Fig. 4 along with other gap factor measures).

### 3.5 Predicting RNA-Seq perturbation values

We demonstrated the applicability our LINCS trained model (Phase 2) to identify differential gene perturbation with an external dataset. Specifically, we used RNA-Seq of LNCaP cells treated with a clinically relevant concentration of previously published androgen receptor antagonists. Importantly, these developmental compounds were not included in the LINCS training set and represent an unbiased test set. We used the DNN model with optimal hyperparameters as described in the methods section and obtained the prediction statistics listed on the top half of Table 4. We obtained Morgan fingerprints for Enzalutamide and in-house compounds VPC-17005 and VPC-14449 as well as the GO descriptors for the 978 landmark genes as described in the methods section. After combining the features, we obtained predictions from our models. Predictions from our models identified significant gene regulations in the RNA-Seq data as listed on the bottom half of Table 4. Figure 3C and 3D shows plots of the corresponding ROC curves.

The results show that AUC values hover around 80% when our models were used to predict RNA-Seq values. *F-score* values were somewhat on the lower end ($P < 0.002$) and is likely due to the extreme imbalance between actives and inactive in the RNA-Seq dataset. However, when looking at enrichment scores, they are lower but consistent with the enrichment values that were obtained from 10-fold cross-validation using LINCS data. AUC scores that hover around 80% show that our models were fair to good predictors of RNA-Seq values and has enrichment values from 4 to 11 (see Supplementary Table S3 for *P*-value calculations).

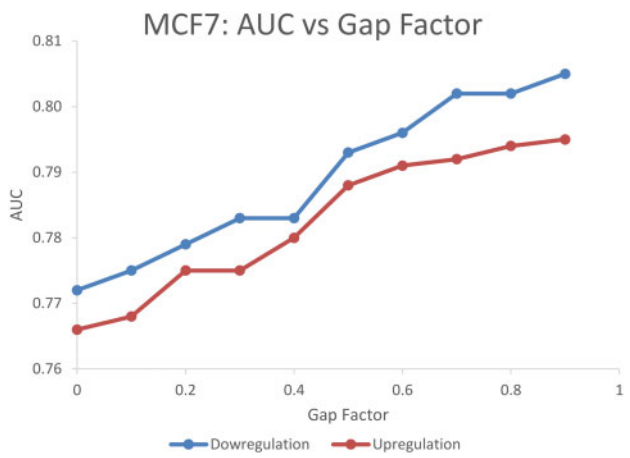


**Fig. 4.** Effect of the gap factor on AUC of up-regulation and down-regulation predictions of the internal validation set

The limitations of the RNA-Seq data comparison arise from the fact that the LINCS training data for LNCaP cell line was limited in size with only 58 compounds tested. Further comparisons with RNA-Seq experiments on cell-lines such as VCAP and A549 which have a significantly more diverse set of drugs would expect to yield higher AUC, *F*-scores and enrichment factors. Another limitation comes from the exposure time differences. The LINCS data for LNCaP only included experiments with 3 and 24 h exposure times whereas the RNA-Seq experiments were done with an exposure time of 6 h. These differences would result in lower than expected prediction scores due to differences in gene perturbation over time. We also see that the 6 h RNA-Seq experiment exposure times is closer to the 3 h conditions of the LINCS training data than the 24 h conditions as can be seen by the increased performance. There were also slight differences in dose amounts being 1.5 μM and 6 μM as described in Section 2.2 compared to the 10 μM dosage from the training set. We would also like to note that the DESeq2 protocol uses a strict *P*-value adjustment that leads us to have highly imbalanced data contributing to lower *F*-scores. Additionally, we tried to improve our *F*-scores by using a version of conformal predictions that have been shown to improve the predictions of existing models (Ahmed *et al.*, 2018; Svensson *et al.*, 2017). Yet, results did not improve on any of the measures in our case.

### 3.6 Applicability domain

Prediction machine learning models are always accompanied by a question of reliability (Jaworska *et al.*, 2005). In Quantitative Structure Activity Relationship (QSAR) models, application domain (AD) can be used to define the scope of molecules that were used in the training set. When an AD is computed, we can determine how molecules used for prediction, fit in the scope of the AD. For a similarity based approach, a molecule that has high similarity scores with the compounds of a training set will be considered to have a more reliable prediction than a molecule that has low similarity scores. There are multiple approaches to calculate the similarity scores between the feature representations of molecules, each with their own strengths and weaknesses. Since the features we used for training were binary, we used the Jaccard index approach.

Figure 5 shows the calculated scores between the molecules that were used for training to predict the RNA-Seq values. The similarity scores of the molecules predicted are also marked in the diagram. The plot demonstrates that the compounds in the RNA-Seq experiments fall well within the applicability domain and have an average similarity score to the chemicals from the LINCS training set in the case of Enzalutamide and better than average similarity score in the case of our two in-house compounds.

## 4 Conclusions

In this article, we report the first study of the efficacy of deep learning in the direct prediction of drug–gene interactions. The internal

**Table 4.** The AUC scores for the 10-fold cross validation training on the top half and performance on predicting RNA-Seq values using the trained models on the bottom half

| Measurement | Train data experiment conditions | Sample size | Down-regulation | | | Up-regulation | | |
|---|---|---|---|---|---|---|---|---|
| | | | AUC | *F*-score | $E_f$ | AUC | *F*-score | $E_f$ |
| LINCS internal 10-fold CV | 10 μM 3 h | 56 723 | 0.73 | 0.34 | 9.42 | 0.72 | 0.35 | 9.33 |
| | 10 μM 24 h | 55 745 | 0.82 | 0.49 | 4.77 | 0.81 | 0.48 | 4.79 |
| RNA-Seq predictions | 10 μM 3 h | 2895 | 0.68 | 0.08 | 4.35 | 0.85 | 0.14 | 10.88 |
| | 10 μM 24 h | 2895 | 0.73 | 0.05 | 2.56 | 0.67 | 0.03 | 2.30 |

*Note:* The maximum *F-score* is considerably lower for RNA-Seq predictions (*P*-values <0.002) likely due to imbalanced data between active and inactive regulations in the RNA-Seq data, however, enrichment values were slightly lower but remained consistent with 10-fold cross-validation values.

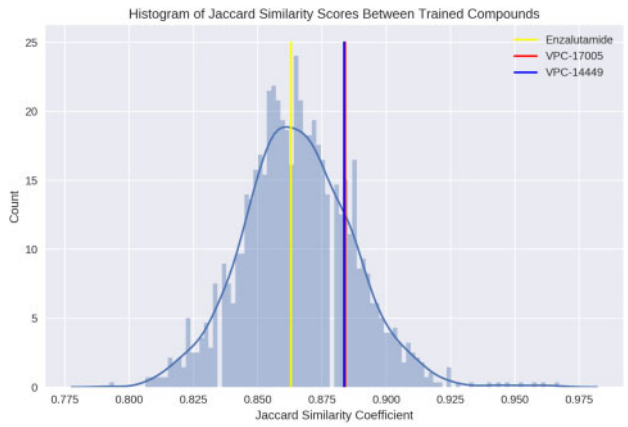


**Fig. 5.** Jaccard index between the molecules used in training to predict RNA-Seq compounds

model validation resulted in good prediction statistics with the corresponding cross-validated AUC parameters of ~0.80. When comparing the prediction for our in house RNA-Seq data for three different antiandrogens we see overall decreased accuracy compared to the training statistics, but there is still utility in the form of enrichment. Thus, DeepCOP did not show ideal results when applied to an external set but it provided a valid proof of principle starting point, termed for future improvement. Such drug-to-gene deep learning models can be significantly improved in the future when more abundant external data becomes available and when more standardized and controlled experimental conditions are enforced. The proposed deep learning tools can be also be further improved by the use of more sophisticated gene and molecular descriptors; when optimized, DeepCOP will be used to screen an ever growing massive library of available chemicals for drug candidates with desired gene regulating properties and can manipulate cellular pathways. We believe, that such direct prediction of the gene regulating effect of small molecules can pave the way for precision oncology studies where therapeutic candidates are tailored specifically for each patient's unique gene expression profile.

## Acknowledgements

G.W. would like to thank the members of the Cherkasov laboratory for guidance and suggestions that helped prepare this article.

## Funding

This work has been supported by the Canadian Institutes of Health Research (CIHR) Project (#156094) and Operating (#390757).

*Conflict of Interest*: none declared.

## References

Ahmed,L. *et al.* (2018) Efficient iterative virtual screening with Apache Spark and conformal prediction. *J. Cheminform.*, **10**, 8.

Andrews,S. (2014) FastQC: a quality control tool for high throughput sequence data. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Bender,A. and Glen,R.C. (2005) A discussion of measures of enrichment in virtual screening: comparing the information content of descriptors with increasing levels of sophistication. *J. Chem. Inf. Model.*, **45**, 1369–1375.

Bredel,M. and Jacoby,E. (2004) Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.*, **5**, 262.

Dalal,K. *et al.* (2018) Selectively targeting the dimerization interface of human androgen receptor with small-molecules to treat castration-resistant prostate cancer. *Cancer Lett.*, **437**, 35–43.

Dobin,A. *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

Fernandez,M. *et al.* (2018) Toxic colors: the use of deep learning for predicting toxicity of compounds merely from their graphic images. *J. Chem. Inf. Model.*, **58**, 1533–1543.

Frankish,A. *et al.* (2019) GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.*, **47**, D766–D773.

Greene,D. *et al.* (2017) ontologyX: a suite of R packages for working with ontological data. *Bioinformatics*, **33**, 1104–1106.

Hsing,M. *et al.* (2008) The use of Gene Ontology terms for predicting highly-connected 'hub' nodes in protein-protein interaction networks. *BMC Syst. Biol.*, **2**, 80.

Jaworska,J. *et al.* (2005) QSAR applicabilty domain estimation by projection of the training set descriptor space: a review. *Altern. Lab. Anim.*, **33**, 445–459.

Lavertu,A. *et al.* (2018) Pharmacogenomics and big genomic data: from lab to clinic and back again. *Hum. Mol. Genet.*, **27**, R72–R78.

Li,H. *et al.* (2014) Discovery of small-molecule inhibitors selectively targeting the DNA-binding domain of the human androgen receptor. *J. Med. Chem.*, **57**, 6458–6467.

Liao,Y. *et al.* (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.

Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

Mahapatra,S. (2018) Why deep learning over traditional machine learning? Available at: https://towardsdatascience.com/why-deep-learning-is-needed-over-traditional-machine-learning-1b6a99177063.

Mayr,A. *et al.* (2016) DeepTox: toxicity prediction using deep learning. *Front. Environ. Sci.*, **3**, 80.

Morgan,H.L. (1965) The generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service. *J. Chem. Doc.*, **5**, 107–113.

Mullane,S. *et al.* (2019) Machine learning for classification of protein helix capping motifs. In: *Systems and Information Engineering Design Symposium (SIEDS)*. doi: 10.1109/SIEDS.2019.8735646.

Open-source (2006) RDKit: Open-Source Cheminformatics Software. http://www.rdkit.org/ (11 April 2013, date last accessed).

Rogers,D. and Hahn,M. (2010) Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, **50**, 742–754.

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

Stegmaier,K. *et al.* (2004) Gene expression–based high-throughput screening (GE-HTS) and application to leukemia differentiation. *Nat. Genet.*, **36**, 257–263.

Subramanian,A. *et al.* (2017) A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, **171**, 1437–1452.e17.

Sureyya Rifaioglu,A. *et al.* (2019) DEEPred: automated protein function prediction with multi-task feed-forward deep neural networks. *Sci. Rep.*, **9**, 7344.

Svensson,F. *et al.* (2017) Improving screening efficiency through iterative screening using docking and conformal prediction. *J. Chem. Inf. Model.*, **57**, 439–444.

Szalai,B. *et al.* (2018) Signatures of cell death and proliferation in perturbation transcriptomics data—from confounding factor to effective prediction. bioRxiv, 454348. doi: 10.1101/454348.

The Gene Ontology Consortium. (2018) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.

Wagner,A. *et al.* (2015) Drugs that reverse disease transcriptomic signatures are more effective in a mouse model of dyslipidemia. *Mol. Syst. Biol.*, **11**, 791.

Wang,Z. *et al.* (2016) Drug-induced adverse events prediction with the LINCS L1000 data. *Bioinformatics*, **32**, 2338–2345.

Wei,J.N. *et al.* (2019) Rapid prediction of electron-ionization mass spectrometry using neural networks. *ACS Cent. Sci.*, **5**, 700–708.

Zheng,S. *et al.* (2018) e-Bitter: bitterant prediction by the consensus voting from the machine-learning methods. *Front. Chem.*, **6**, 82.