Towards Explainable Anticancer Compound Sensitivity Prediction via Multimodal Attention-based Convolutional Encoders

Matteo Manica [*1] **Ali Oskooei** [*1] **Jannis Born** [*123] **Vigneshwari Subramanian** [4] **Julio Sáez-Rodríguez** [45]
María Rodríguez Martínez [1]

Abstract

In line with recent advances in neural drug design and sensitivity prediction, we propose a novel architecture for interpretable prediction of anticancer compound sensitivity using a multimodal attention-based convolutional encoder. Our model is based on the three key pillars of drug sensitivity: compounds' structure in the form of a SMILES sequence, gene expression profiles of tumors and prior knowledge on intracellular interactions from protein-protein interaction networks. We demonstrate that our multiscale convolutional attention-based (MCA) encoder significantly outperforms a baseline model trained on Morgan fingerprints, a selection of encoders based on SMILES as well as previously reported state of the art for multimodal drug sensitivity prediction (R2 = 0.86 and RMSE = 0.89). Moreover, the explainability of our approach is demonstrated by a thorough analysis of the attention weights. We show that the attended genes significantly enrich apoptotic processes and that the drug attention is strongly correlated with a standard chemical structure similarity index. Finally, we report a case study of two receptor tyrosine kinase (RTK) inhibitors acting on a leukemia cell line, showcasing the ability of the model to focus on informative genes and submolecular regions of the two compounds. The demonstrated generalizability and the interpretability of our model testify its potential for in-silico prediction of anticancer compound efficacy on unseen cancer cells, positioning it as a valid solution for the development of personalized therapies as well as for the evaluation of candidate compounds in de novo drug design.

[*]Equal contribution [1]IBM Research Zurich, Switzerland [2]ETH Zurich, Switzerland [3]University of Zurich, Switzerland [4]RWTH Aachen University, Germany [5]Heidelberg University, Germany. Correspondence to: Matteo Manica, Ali Oskooei, Jannis Born <{tte,osk,jab}@zurich.ibm.com>.

Workshop on Computational Biology at the International Conference on Machine Learning (ICML), Long Beach, CA, 2019.

1 Introduction

1.1 Motivation

Discovery of novel compounds with a desired efficacy and improving existing therapies are key bottlenecks in the pharmaceutical industry, which fuel the largest R&D business spending of any industry and account for 19% of the total R&D spending worldwide (Petrova, 2014; Goh et al., 2017). Anticancer compounds, in particular, take the lion's share of drug discovery R&D efforts, with over 34% of all drugs in the global R&D pipeline in 2018 (5,212 of 15,267 drugs) (Lloyd et al., 2017). Despite enormous scientific and technological advances in recent years, serendipity still plays a major role in anticancer drug discovery (Hargrave-Thomas et al., 2012) without a systematic way to accumulate and leverage years of R&D to achieve higher success rates in drug discovery. On the other hand, there is strong evidence that the response to anticancer therapy is highly dependent on the tumor genomic and transcriptomic makeup, resulting in heterogeneity in patient clinical response to anticancer drugs (Geeleher et al., 2016). This varied clinical response has led to the promise of personalized (or precision) medicine in cancer, where molecular biomarkers, e.g., the expression of specific genes, obtained from a patient's tumor profiling may be used to choose a personalized therapy.

These challenges highlight a need across both pharmaceutical and healthcare industries for multimodal quantitative methods that can jointly exploit disparate sources of knowledge with the goal of characterizing the link between the molecular structure of compounds, the genetic and epigenetic alterations of the biological samples and drug response (De Niz et al., 2016). In this work, we present a multimodal attention-based convolutional encoder that enables us to tackle the aforementioned challenges.

1.2 Related work

There have been a plethora of works on the prediction of drug sensitivity in cancer cells (Garnett et al., 2012; Yang et al., 2012; Costello et al., 2014; Ali & Aittokallio, 2018; Kalamara et al., 2018). While the majority of them have focused on the analysis of unimodal datasets (genomics or transcriptomics, e.g., De Niz et al. (2016); Tan (2016); Turki & Wei (2017); Tan et al. (2018)), a handful of pre-

vious works have integrated omics and chemical descriptors to predict cell line-drug sensitivity using a variety of methods including but not limited to: simple neural networks (one hidden layer) and random forests (Menden et al., 2013), kernelized Bayesian matrix factorization (Ammad-Ud-Din et al., 2014), Pearson correlation-based similarity networks (Zhang et al., 2015), a Kronecker product kernel in conjunction with SVMs (Wang et al., 2016), autoencoders in combination with elastic net and SVMs (Ding et al., 2018), matrix factorization (Wang et al., 2017), trace norm regularization (Yuan et al., 2016), link predictions (Stanfield et al., 2017) and collaborative filtering (Liu et al., 2018; Zhang et al., 2018b). In addition to genomic and chemical features, previous studies have demonstrated the value of complementing drug sensitivity prediction models with prior knowledge in the form of protein-protein interactions (PPI) networks (Oskooei et al., 2018b). For example, in a network-based per-drug approach integrating these data sources, Zhang et al. (2018a) surpassed various earlier models and reported a performance drop of 3.6% when excluding PPI information.

However, all previous attempts at incorporating chemical information in drug sensitivity prediction rely on molecular fingerprints as chemical descriptors. Traditionally, fingerprints were applied extensively for drug discovery, virtual screening and compound similarity search (Cereto-Massagué et al., 2015), but it has recently been argued that the usage of engineered features constraints the learning ability of machine learning algorithms (Goh et al., 2017). Furthermore, for many applications, molecular fingerprints may not be relevant, informative or even available.

With the rise of deep learning methods and their proven ability to learn the most informative features from raw data, machine learning methods used in molecular design and drug discovery have also experienced a shift (Chen et al., 2018; Wu et al., 2018; Grapov et al., 2018). For instance, computational chemists borrowed methods from neural language models (Bahdanau et al., 2014) to encode SMILES strings of molecules and predict chemical properties of molecules (Jastrzębski et al., 2016; Goh et al., 2017; Schwaller et al., 2018b). Once a gold standard in sequence modeling, recurrent neural networks were initially employed as SMILES encoders (Goh et al., 2017; Bjerrum, 2017; Segler et al., 2017). However, it has been recently shown that convolutional architectures are superior to RNNs for sequence modeling (Bai et al., 2018), and specifically for modeling the SMILES string encoding of compounds (Kimber et al., 2018). It is noteworthy that these findings are in agreement with our model comparison results that reveal convolutional architectures are superior for SMILES sequence modeling. Most recently, Chang et al. (2018) adopted deep learning methods to develop a pan-drug model for predicting IC50 drug sensitivity of drug-cell line pairs. Utilizing > 30,000 binary features (∼3,000 for the molecular drug fingerprint and the rest for a genomic fingerprint), they employ a model ensemble of 5 deep convolutional networks (4 are completely linear) with convolutions applied separately to each of the genomic and molecular features before the encodings are merged. While working towards a common goal, our approaches are vastly different. Our method presents several key advantages: First, our algorithm ingests raw information (SMILES string representation) which in turn enables data augmentation and boosts model performance (Bjerrum, 2017). Secondly, we apply convolutions only on SMILES representations for which convolutions are meaningful (i.e., convolutions combine information from various molecular substructures). In accordance with Costello et al. (2014), we use transcriptomic features (gene expression profiles) instead of genomic features. Moreover, we combine transcriptomic and molecular information using a contextual attention encoder that renders our model transparent and interpretable, a feature that is paramount in precision medicine and has only recently started to be tackled (Yang et al., 2018). An additional key advantage of our approach is our strict splitting strategy and evaluation criterion. While previous works relied on lenient splitting strategies that ensured no drug-cell line *pair* in the test data was seen during training, we adopt a more stringent splitting strategy and deprive the model training of all drugs and cell lines that are present in the test dataset. Our strict training and evaluation strategy results in a significantly more challenging problem but in turn ensures the model is learning generalizable molecular substructures with anticancer properties as opposed to memorizing drug sensitivity from cell-drug pairs that it has encountered during training. A model that has been trained with such a criterion will generalize better to completely unseen drugs and cell lines thus paving the way for both, in silico validation of de novo drug candidates in pharmaceutics and selection of a suitable therapy in personalized medicine. A lenient split, on the other hand, may facilitate drug repositioning, as it performs best when drug and cell line have been encountered during training.

1.3 Scope of the presented work

In this work we build upon our previous work on multimodal drug sensitivity prediction using attention-based encoders (Oskooei et al., 2018a), and propose a novel best-performing architecture, an attention-based multiscale convolutional encoder. In addition, we perform a thorough validation of the attention weights given by our proposed MCA model. We combine 1) cell line data, 2) molecular structure of compounds and 3) prior knowledge of protein interactions to predict drug sensitivity. Specifically, for 1) we explore the usage of gene expression profiles and for 2) we explore different neural architectures in combination with our devised contextual attention architecture to encode raw SMILES of anticancer drugs in the context of the cell that they are acting on (see Figure 1). We show

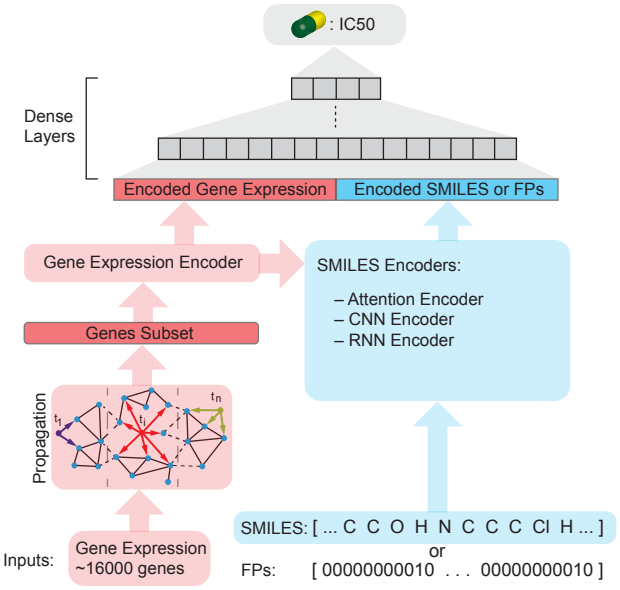

Figure 1. **The multimodal end-to-end architecture of the proposed encoders.** General framework for the explored architectures. Each model ingests a cell-compound pair and makes an IC50 drug sensitivity prediction. Cells are represented by the gene expression values of a subset of 2,128 genes, selected according to a network propagation procedure. Compounds are represented by their SMILES string (apart from the baseline model that uses 512-bit fingerprints). The gene-vector is fed into an attention-based gene encoder that assigns higher weights to the most informative genes. To encode the SMILES strings, several neural architectures are compared and used in combination with the gene expression encoder in order to predict drug sensitivity.

that attention-based SMILES encoders significantly surpass a baseline feedforward model utilizing Morgan (circular) fingerprints (Rogers & Hahn, 2010) (p < 1e-6 on RMSE). Using our multiscale convolutional attentive (MCA) encoder, we show that we achieve superior IC50 prediction performance on the GDSC database (Iorio et al., 2016) compared with the existing methods (Menden et al., 2013; Chang et al., 2018). Utilizing SMILES representations is highly desirable, as they are ubiquitously available and more interpretable than traditional fingerprints. Furthermore, our contextual attention mechanism emerges as the key component of our proposed SMILES encoder, as it helps validate our findings by explaining the model's inner working and reasoning process, many of which are in agreement with domain-knowledge on biochemistry of cancer cells.

2 Methods

2.1 Data

Throughout this work, we employed drug sensitivity data from the the publicly available Genomics of Drug Sensitivity in Cancer (GDSC) database (Iorio et al., 2016). The database includes the screening results of more than a thousand genetically profiled human pan-cancer cell lines with a wide range of anticancer compounds (both chemotherapeutic drugs and targeted therapeutics). The drug sensitivity values were represented by half maximal inhibitory concentration (IC50, i.e., the micromolar concentration of a drug necessary to inhibit 50% of the cells) on the log-scale. We focused on targeted drugs with publicly available molecular structure (208 compounds of 265 in total) and retrieved the molecular structure of the compounds from `PubChem` (Kim et al., 2018) and the `LINCS` database. From the collected canonical SMILES, Morgan fingerprints were acquired using `RDKit` (512-bit with radius 2). Exploiting the property that most molecules have multiple valid SMILES strings, the data augmentation strategy proposed by Bjerrum (2017) was adopted to represent each anticancer compound with 32 different SMILES strings. We chose to represent each cell by its transcriptomic profile as it has been demonstrated that transcriptomic data are more predictive of drug sensitivity when compared to other omic data (Costello et al., 2014). As such, all available RMA-normalized gene expression data were retrieved from the GDSC database resulting in transcriptomic profiles of 985 cell lines in total.

2.2 Network propagation

Each of the 985 cell lines was initially represented by the expression levels of 17,737 genes which we then reduced to a subset of 2128 genes through network propagation over the STRING protein-protein interaction (PPI) network (Szklarczyk et al., 2014), a comprehensive PPI database including interactions from multiple data sources. Following the procedure described in Oskooei et al. (2018b), STRING was used to incorporate intracellular interactions in our model by adopting a network propagation scheme for each drug, where the weights associated with each of the reported targets were diffused over the STRING network (including interactions from all the evidence types) leading to an importance distribution over the genes (i.e., the vertices of the network). Our adopted weighting and network propagation scheme consisted of the following steps: we first assigned a high weight (W = 1) to the reported drug target genes while assigning a very small positive weight (ε = 1e−5) to all other genes. Thereafter, the initialized weights were propagated over STRING. This process was meant to integrate prior knowledge about molecular interactions into our weighting scheme, and simulate the propagation of perturbations within the cell following the drug administration. Let us denote the initial weights as W_0 and the string network as $S = (P, E, A)$, where P are the protein vertices of the network, E are the edges between the proteins and A is the weighted adjacency matrix. The smoothed weights are determined from an iterative solution of the propagation function (Oskooei et al., 2018b):

$$W_{t+1} = \alpha W_t A' + (1 - \alpha) W_0 \quad (1)$$

where D is the degree matrix and A' is the normalized adjacency matrix, obtained from the degree matrix D:

$$A' = D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \tag{2}$$

The diffusion tuning parameter, α ($0 \leq \alpha \leq 1$), defines how far the prior knowledge weights can diffuse through the network. In this work, we used $\alpha = 0.7$, as recommended in the literature for the STRING network (Hofree et al., 2013). Adopting a convergence rule of $e = (W_{t+1} - W_t) < 1e{-}6$, we solved Equation 1 iteratively for each drug and used the resultant weights distribution to determine the top 20 highly ranked genes for each drug. By selecting the top 20 genes for every drug, it was possible to compile an interaction-aware subset of genes (2,128 genes in total). This subset containing the most informative genes was then used to profile each cell line in the dataset before it was fed into our models. We limited the selection to the top 20 genes for every drug to guarantee a trade-off between topology-awareness and the number of features describing the biomolecular profile. We then paired all screened cell lines and drugs to generate a pan-drug dataset of cell-drug pairs and the associated IC50 drug response. Due to missing values in the GDSC database, pairing of the 985 cell lines with the 208 drugs resulted in 175,603 pairs which could be augmented to more than 5.5 million data points following SMILES augmentation (Bjerrum, 2017).

2.3 Model architectures

The majority of previous efforts in drug sensitivity prediction focused on traditional molecular descriptors (fingerprints). Morgan fingerprints have been shown to be a highly informative representation for many chemical prediction tasks (Unterthiner et al., 2014; Chang et al., 2018). We explored several neural network SMILES encoder architectures to investigate whether the molecular information of compounds, in the context of drug sensitivity prediction, can be learned directly from the raw SMILES rather than using engineered fingerprints. As such, all explored encoder architectures were compared against a baseline model that utilized 512-bit Morgan fingerprints. The general architecture of our models is shown in Figure 1.

Deep baseline (DNN). The baseline model is a 6-layered DNN with [512, 256, 128, 64, 32, 16] units and a sigmoid activation. The hyperparameters for the baseline model were optimized via a cross-validation scheme (see subsection 2.4). 512-bit Morgan fingerprints and gene expression profiles (filtered using the network propagation described in subsection 2.1) were concatenated into a joint representation from the first layer onwards.

SMILES models (commonalities). To investigate which model architecture best learns the molecular information of compounds, we explored various SMILES encoders. All SMILES-based models ingest the expression profiles and the SMILES text encodings for the structure of the compounds. The SMILES sequences were tokenized using a regular expression (Schwaller et al., 2018a). The resulting atomic sequences were zero-padded and represented as $E = \{e_1, ..., e_T\}$, with learned embedding vectors $e_i \in \mathbb{R}^H$ for each dictionary token (see Figure 2A). Each cell line,

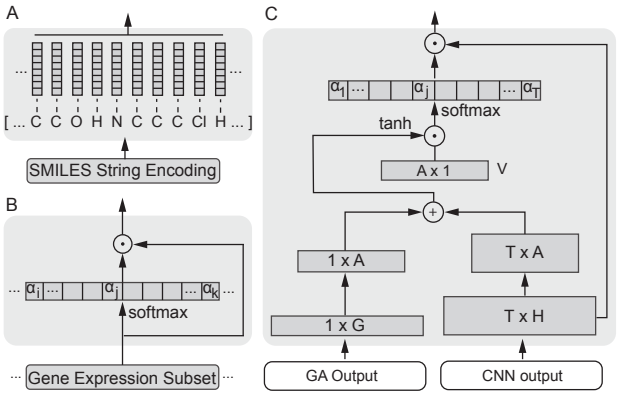

Figure 2. **Key layers employed throughout the SMILES encoder.** A) An embedding layer transforms raw SMILES strings into a sequence of vectors in an embedding space. B) An attention-based gene expression encoder generates attention weights that are in turn applied to the input gene subset via a dot product. C) A contextual attention layer ingests the SMILES encoding (either raw or the output of another encoder, e.g., CNN, RNN and so on) of a compound and genes from a cell to compute an attention distribution (α_i) over all tokens of the SMILES encoding, in the context of the genetic profile of the cell. The attention-filtered molecule represents the most informative molecular substructures for IC50 prediction, given the gene expression of a cell.

represented by the genetic subset selected through network propagation, is fed to the gene attention encoder (see Figure 2B). A single dense softmax layer with the same dimensionality as the input produces an attention weight distribution over the genes and filters them in a dot product, ensuring most informative genes are given a higher weight for further processing. The resulting gene attention weights render the model interpretable, as they identify genes that drive the sensitivity prediction for each cell line. This architecture was also investigated for the deep baseline model but discarded due to inferior performance. All SMILES encoders were followed by a set of dense layers (as shown in Figure 1) with dropout (p_{drop} = 0.5) for regularization and sigmoid activation function. The regression was completed by a single neuron with linear activation (rather than sigmoid) to avoid restricting the values between 0 and 1 and hinder the learning process of the network as a result.

Bidirectional recurrent (bRNN). RNNs have traditionally been the first-line approach for sequence encoding. To investigate their effectiveness in encoding SMILES, we adopted a 2-layered bidirectional recurrent neural network (bRNN) with gated recurrent units (GRU) (Cho et al., 2014).

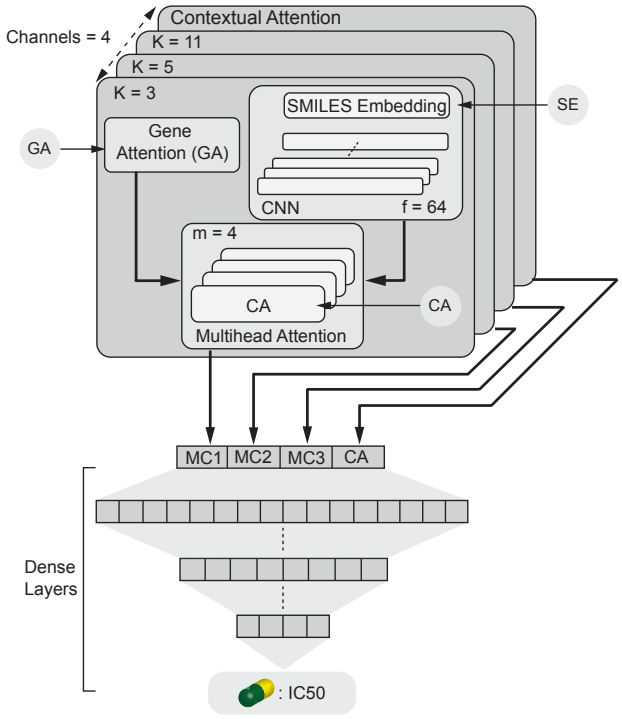

Figure 3. **Model architecture of the multiscale convolutional attentive (MCA) encoder.** The MCA model employed 3 parallel channels of convolutions over the SMILES sequence with kernel sizes K and one residual channel operating directly on the token level. Each channel applied a separate gene attention layer, before (convoluted) SMILES and filtered genes were fed to a multihead of 4 contextual attention layers. The output of these 16 layers were concatenated and resulted in an IC50 prediction through a stack of dense layers. For CA, GA and SE, see Figure 2.

The final states of the forward and backward GRU-RNN were concatenated and fed to the dense layers for IC50 prediction.

Stacked convolutional encoder (SCNN). Next, we employed an encoder with four layers of stacked convolutions and sigmoid activation function. 2D convolution kernels in the first layer collapsed the embedding vectors' hidden dimensionality while subsequent 1D convolutions extracted increasingly long-range dependencies between different parts of the molecule. As a result, similarly to the bRNN, any output neuron of the SCNN SMILES encoder had integrated information from the entire molecule.

Self-attention (SA). We investigated several encoders that leveraged neural attention mechanisms, originally introduced by Bahdanau et al. (2014). Interpretability is paramount in healthcare and drug discovery (Koprowski & Foster, 2018). As such, neural attention mechanisms are central in our models as they enable us to explain and interpret the observed results in the context of underlying biological and chemical processes. Our first attention configuration is a self-attention (SA) mechanism adapted from document classification (Yang et al., 2016) for encoding SMILES strings. The SMILES attention weights α_i were computed per atomic token as:

$$\alpha_i = \frac{\exp(u_i)}{\sum_j^T u_j} \quad \text{where} \quad u_i = V^T \tanh(W_e s_i + b) \qquad (3)$$

The matrix $W_e \in \mathbb{R}^{A \times H}$ and the bias vector $b \in \mathbb{R}^{A \times 1}$ are learned in a dense layer. s_i is an encoding of the i-th token of the molecule, in the most basic case simply the SMILES embedding e_i. In all attention mechanisms, the encoded smiles are obtained by filtering the inputs with the attention weights.

Contextual-attention (CA). Alternatively, we devised a contextual attention (CA) mechanism that utilizes the gene expression subset G as a context (Figure 2C). The attention weights α_i are determined according to the following equation :

$$u_i = V^T \tanh(W_e s_i + W_g G) \quad \text{where} \quad W_g \in \mathbb{R}^{A \times |G|} \qquad (4)$$

First, the matrices W_g and W_e project both genes G and the encoded SMILES tokens s_i into a common attention space, A. Adding the gene context vector to the projected token ultimately yields an α_i that denotes the relevance of a compound substructure for drug sensitivity prediction, given a gene subset G.

Multiscale convolutional attention (MCA). In their simplest form, the attention mechanisms of the SA and CA model operates directly on the embeddings, disregarding positional information and long-range dependencies. Instead they exploit the frequency counts on individual tokens (atoms, bonds). Interestingly, the attention models nevertheless outperform the bRNN and SCNN which integrated information from the entire molecule. In order to combine the benefits of the attention-based models, i.e., interpretability with the ability of sequence encoders to extract both local and long-range dependencies, we devised the multiscale convolutional attentive (MCA) encoder shown in Figure 3. Using MCA, the SMILES string of a compound is analyzed using three separate channels, each convolving the SMILES embeddings with a set of f kernels of sizes $[H, 3]$, $[H, 5]$ and $[H, 11]$ and ReLU activation. The efficacy of a drug may be tied primarily to the occurrence of a specific molecular substructure. MCA is designed to capture substructures of various size using its variable kernel size. For instance, a particular kernel could detect a steroid structure, typical across anticancer molecules (Gupta et al., 2013). Following the multiscale convolutions, the resulting feature maps of each channel were fed into a contextual attention layer that received the filtered genes as context. Similarly to Vaswani et al. (2017), we employed $m = 4$ contextual attention layers for each channel, in order to allow the model to jointly attend several parts of the molecule. The multi-head attention approach, counteracts the tendency of the softmax

to filter out the vast majority of the sequence steps (Li & Maki, 2018). In a fourth channel, the convolutions were skipped (residual connection) and the raw SMILES embeddings were directly fed to the parallel CA layers. The output of these $4m$ layers were concatenated before given to the stack of dense feedforward layers.

2.4 Model evaluation

Strict split. To benchmark the different proposed architectures, a strict data split approach was adopted to ensure neither the cell lines nor the compound structures within the validation or test datasets have been seen by our models prior to validation or testing. This is in contrast to previously published pan-drug models which have explored only a lenient splitting strategy, where both compound and cell-line of any sample in the test dataset were encountered during training. In our data split strategy, 10% subsets of the total number of 208 compounds and 985 cell lines from the GDSC database were set aside to be used as an unseen test dataset to evaluate the trained models. The remaining 90% of compounds and cell lines were then used in a 25-fold cross-validation scheme for model training and validation. In each fold, 4% of the drugs and 4% of cell lines were separated and used to generate the validation dataset and the remaining drugs and cell lines were paired and fed to the model for training. In practice, this strategy deprived the model from a significant proportion of samples which were not sorted into any of training, validation or testing data. We decided to choose 25-fold cross-validation because: 1) this number is large enough to employ tests of statistical significance across different models and 2) To increase the size of the training set and in turn improve the performance of the trained models by decreasing the number of pairs that were excluded from training set (i.e., the validation set).

Lenient split. To compare our model with prior works that chose a less strict data split strategy, we adopted a similar strategy that rather than depriving the model from both the cells and drugs in the test set, ensured no cell-drug pair in the test set has been seen before. The new split consisted of a standard 5-fold cross-validation scheme, wherein 10% of the pairs (175,603 pairs from 985 cell lines and 208 drugs) were set aside for testing. IC50 values of the training data were normalized to $[0, 1]$ and the same transformation was applied to validation and test data. Gene expression values in the training set were standardized and the same transformation was applied to the gene expression in the validation and test sets.

2.5 Training procedure

All described architectures were implemented in `TensorFlow 1.10` with a MSE loss function that was optimized with Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 1\text{e}{-8}$) and a decreasing learning rate (Kingma & Ba, 2014). An embedding dimensionality of H = 16 was adopted for all SMILES encoders. The attention dimensionality was set to A = 256 for the SA and CA model, while $A = f$ = 64 for MCA. In the final dense layers of all models we employed dropout (p_{drop} = 0.5), batch normalization and a sigmoid activation. All models were trained with a batch size of 2,048 for a maximum of 500k steps on a cluster equipped with `POWER8` processors and an `NVIDIA Tesla P100`.

3 Results

3.1 Model performance comparison on strict split

Table 1 compares the test performance of all models trained using a 25-fold cross-validation scheme. As shown in Table 1, the MCA model yielded the best performance in predicting drug sensitivity (IC50) of unseen drugs-cell line pairs within the test dataset. Since IC50 was normalized to [0,1], the observed RMSE implies an average deviation of 10.4% of the predicted IC50 values from the true values. Interestingly, the bRNN SMILES encoder matched, but did not surpass the performance of the baseline model (DNN). The SCNN encoder which combined and encoded information from across the entire SMILES sequence, performed significantly worse than the baseline, as assessed by a one-sided Mann-Whitney-U test (U = 126, p < 2e-4). We therefore hypothesize that local features of the SMILES sequence (such as counts of atoms and bonds) contain information most predictive of a drug's efficacy. Attention-based models that operated directly on the SMILES embeddings (SA, CA), performed significantly better than all previous models (e.g., CA vs. DNN: U = 42, p < 9e-8, SA vs. DNN: U = 82, p < 5e-6). Surprisingly, neither complementing the SMILES embedding with positional encodings (similarly to Vaswani et al. (2017)) nor complementing the bRNN encoder with attention was found to improve the model performance. Ultimately, the MCA model was devised to combine token-level information (benefical for the attention-

Table 1. **Performance of the explored architectures on test data following 25-fold cross-validation**. The median Root Mean Square Error (RMSE) and the Interquartile Range (IQR) between predicted and true IC50 values on test data of all 25 folds are reported. Interestingly, attention-based models outperform all other models including models trained on fingerprints with a statistically significant margin (* indicating a significance of $p < 0.01$ compared to the DNN encoder, ** to the MCA).

Encoder type	Drug structure	Standardized RMSE Median ± IQR
Deep baseline (DNN)	Fingerprints	0.122 ± 0.010
Bidirectional recurrent (bRNN)	SMILES	0.119 ± 0.011
Stacked convolutional (SCNN)	SMILES	0.130 ± 0.006
Self-attention (SA)	SMILES	0.112* ± 0.009
Contextual attention (CA)	SMILES	0.110* ± 0.007
Multiscale convolutional attentive (MCA)	SMILES	0.109* ± 0.009
MCA (prediction averaging)	SMILES	**0.104**** ± **0.005**

only models) with spatially more holistic chemical features within the same model. By architecture, some convolution kernels in the MCA could for example develop a sensitivity for a pyrimidine ring, potentially indicative of a tyrosine kinase inhibitor (such as Gefitinib, Afatanib or Erlotinib); an enzyme which inhibits phosphorylation of epidermal growth factor receptors (EGFR) to suppress tumor cell proliferation (Jiao et al., 2018). The MCA model also outperformed the baseline model significantly (U = 136, p < 3e-4) like an improved version employing prediction averaging across the 20 best checkpoints did (U = 10, p < 2e-9 and U = 152, p < 9e-4 comparing to the plain MCA). In general, we observed a strong variability across the folds, leading us to report median as a more robust measure of performance than the mean across the folds. The variability across the folds stemmed from the strict splitting strategy (see subsection 2.4) that resulted in training, validation and test datasets that were significantly different from one another. In conclusion, our results suggest that in order to effectively capture the mode of action of a compound, we require information from a combination of token-level (i.e., atom or bond level) and longer range dependencies across the SMILES sequence.

3.2 Model validation on lenient data split

In addition to the performance evaluations in Table 1, we evaluated the MCA model using a less strict data split strategy comparable to what had been adopted in previous works (Chang et al., 2018). This allowed for a more meaningful comparison between the performance of our models with existing prior art. As Figure 4 shows, the MCA model ensemble achieved a RMSE of 0.887 on the log-IC50-scale, corresponding to a deviation of 4.6%. The explained variance of 86.19% suggests that our model learned to a significant extent to predict the IC50 value of an unknown pair, when both the cell line and drugs in the test set were not excluded from the training set. In addition, we observed that MCA's performance on the lenient split was on par or superior to that of existing pan-drug models (Menden et al., 2013; Chang et al., 2018).

3.3 Attention analysis

Drug structure attention. To quantify and analyze the drug attention on a large scale, we retrieved attention profiles for a panel of drug-cell line pairs where each drug has been evaluated for all the cell lines in the set. The selected panel consisted of 150 drugs and 200 cell lines. For each drug, we defined a matrix of pairwise Euclidean distances between the attention profiles of the treated cell lines. The resulting distance matrix quantifies the variation in attention profiles of a drug as a function of the treated cell lines. We then computed, for each pair of drugs, the Frobenius distance between the attention distance matrices defined above. Finally, we evaluated the correlation between the Frobenius distances of each pair of drugs and their Tanimoto coefficient (Tanimoto, 1958), an established index for evaluating

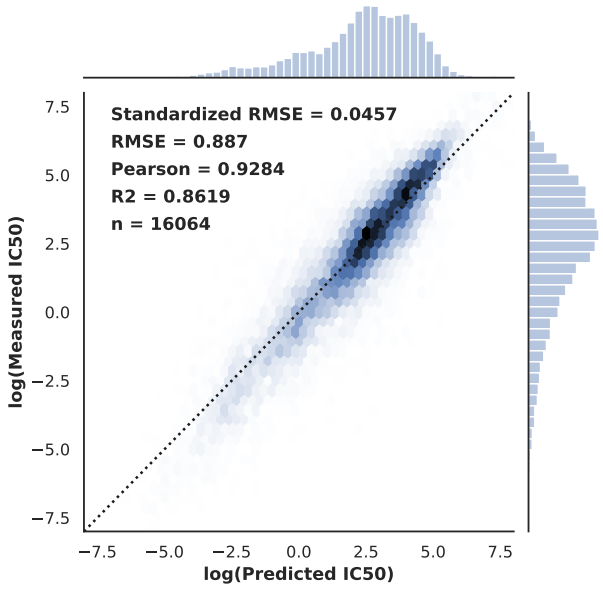

Figure 4. **Test performance of MCA on lenient splitting.** Scatter plot of correlation between true and predicted drug sensitivity by a late-fusion model ensemble of all five folds. RMSE refers to the plotted range, since the model was fitted in log space.

drug similarity based on fingerprints (Bajusz et al., 2015). This approach resulted in a Pearson correlation of ρ =0.64 (n = 22500, p < 1e-50). The fact that the attention similarity of any two drugs is highly correlated with their structural similarity indicates that the model indeed learns valuable insights on structural properties of compounds.

Gene attention. In order to thoroughly validate the gene attention weights, we computed the attention profiles of all cell lines in the test data, averaged the attention weights and filtered them by discarding genes with negligible attention values ($a_i < \frac{1}{K}$, where K is the number of genes in the panel). Based on the resulting subset of 371 highly attended genes, we performed a pathway enrichment analysis using Enrichr (Chen et al., 2013; Kuleshov et al., 2016). The goal was to identify relevant processes highlighted by the genes the model learned to focus on. The analysis revealed a significant activation (adjusted p < 0.004) of the apoptosis signaling pathway in PANTHER (Mi et al., 2016). In essence, drug sensitivity prediction is connected to apoptosis and our analysis suggests that the model learns to focus on genes associated with key molecular processes elicited by anticancer compounds, e.g., programmed cellular death.

A case study: two TK inhibitors. As a further validation, we analyzed in detail the neural attention mechanism of the best MCA model (lenient split) for two very similar anticancer compounds (Imatinib and Masitinib) which only differ in one functional group: a Thiazole ring for Masitinib instead of a Piperazine ring for Imatinib. Both

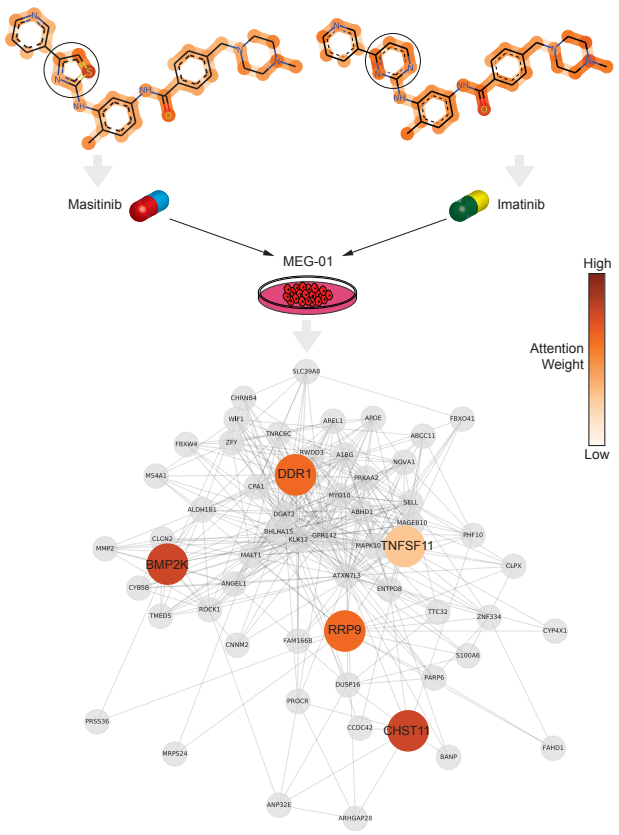

Figure 5. **Neural attention on molecules and genes.** The molecular attention maps on the top demonstrate how the model's attention is shifted when the Thiazole group is replaced by a Piperazine group. The change in attention across the two molecules is particularly concentrated around the affected rings, signifying that these functional groups play an important role in the mechanism of action for these tyrosine-kinase-inhibitors when they act on a chronic myelogenous leukemia (CML) cell line. The gene attention plot at the bottom depicts the most attended genes of the CML cell line, all of which can be linked to leukemia (details see text).

studied drugs are tyrosine-kinase-inhibitors that are predominantly applied in haematopoietic and lymphoid tissue. Generally, their IC50 values are highly correlated, particularly for their target cell lines ($\rho = 0.72$). Figure 5 depicts the attention over both molecules when paired with cell line MEG-01 (COSMIC ID 1295740, a type of chronic myelogenous leukemia). Leukemia is targeted quite successfully by both drugs, with Imatinib (IC50 = 81nM) being superior to Masitinib (223nM). Comparing the attention weights on both molecules depicted in Figure 5 reveals that the attention weights on the affected functional groups (encircled) are drastically different in the two compounds whereas the remaining regions of the both molecules are primarily unaffected. The localized discrepancy in attention centered at the affected rings suggests that these substructures are of primary importance to the model in predicting the sensitivity of the MEG-01 cell line to Imatinib and Masitinib.

At the bottom of Figure 5 the most attended genes of the studied leukemia cell line and their STRING protein neighborhoods are presented. Interestingly, the DDR1 protein is a member of receptor tyrosine kinases (RTKs), the same group of cell membrane receptors that both Imatinib and Masitinib inhibit (Kim et al., 2011). DDR1 gene is highly expressed in various cancer types, such as in chronic lymphocytic leukaemia (Barisione et al., 2017). In addition, BMP2K gene has been recently shown to be implicated in chronic lymphocytic leukemia (CLL) (Pandzic et al., 2016), while CHST11 has long been known to be deregulated in CLL (Schmidt et al., 2004). TNFSF11 encodes RANKL which is part of a prominent cancer signalling pathway (Renema et al., 2016) and TNFSF11 has been reported to be the most overexpressed gene in a sample of $n = 129$ acute lymphoblastic leukemia (ALL) patients (Heltemes-Harris et al., 2011). RRP9 has been shown to be crucial in treating ALL (Rainer et al., 2012)). In conclusion, the prior knowledge from the cancer literature validate our findings and indicate that the genes that were given the highest attention weights by our model are indeed crucial players in the progression and treatment of leukemia.

4 Discussion

We presented an attention-based multimodal neural approach for explainable drug sensitivity prediction using a combination of 1) SMILES string encoding of drug compounds 2) transcriptomics of cancer cells and 3) intracellular interactions incorporated into a PPI network. In an extensive comparative study of SMILES sequence encoders, We demonstrated that using the raw SMILES string of drug compounds, we were able to surpass the predictive performance reached by a baseline model utilizing Morgan fingerprints. In addition, we demonstrated that the attention-based SMILE encoder architectures, especially the newly proposed MCA, performed the best while producing results that were verifiably explainable. The validity of the drug attention has been corroborated by demonstrating its strong correlation with a well established structure similarity measure. To further improve the explainabiltiy of our models, we devised a gene attention mechanism that acts on genetic profiles and focuses on genes that are most informative for IC50 prediction. We validated the correctness of the gene attention weights performing a pathway enrichment analysis over all the cell lines contained in GDSC and finding a significant enrichment of apoptotic processes. In a case study on a leukemia cell line we have showcased how our model is able to focus on relevant compounds' structural elements and consider genes relevant for the disease of interest.

A key feature of our models was the strict training and evaluation strategy that set our work apart from previous art. In our strict model evaluation approach, cells and compounds

were split in training, validation and test dataset before building the pairs, ensuring neither cells nor compounds in the validation or test datasets were ever seen by the trained model, thus depriving the model from a significant portion of available samples. Despite this unforgiving evaluation criterion, our best model (MCA) achieved an average standard deviation of 0.11 in predicting normalized IC50 values for unseen drug-cell pairs. Furthermore, in a separate comparative study on the same dataset, this time with a lenient data split and model evaluation criterion, we demonstrated that our MCA model outperformed previously reported state-of-the-art results by achieving a RMSE of 0.89 and a R2 of 86%. We envision our attention-based approach to be of great utility in personalized medicine and de novo anticancer drug discovery where explainable prediction of drug sensitivity is paramount. Furthermore, having established a solid multimodal predictive model we have paved the way for future directions such as: 1) Drug repositioning applications as our model enables drug sensitivity prediction for any given drug-cell line pair. 2) Leveraging our model in combination with recent advances in small molecule generation using generative models (Kadurin et al., 2017; Blaschke et al., 2017) and reinforcement learning (Popova et al., 2018) to design novel disease-specific, or even patient-specific compounds. This opens up a scenario where personalized treatments and therapies can become a concrete option for patient care in cancer precision medicine.

5 Availability of software and materials

The data in TFRecord format used in the benchmark studies conducted in this work can be downloaded at the following url: `https://ibm.biz/paccmann-data`. Alternatively the reader can access the raw cell line data from GDSC (Iorio et al., 2016) and the compound structural information from `PubChem` (Kim et al., 2018) and the `LINCS` database.

The implementation of the models used in the benchmark is available in the form of a toolbox on Github a the following link: `https://github.com/drugilsberg/paccmann`.

Furthermore, the best MCA model has been deployed as a service on IBM Cloud. Users can access the app and provide a compound in SMILES format to obtain a prediction of its efficacy in terms of IC50 on 970 cell lines from GDSC. The results on drug sensitivity together with the top-attended genes can be examined in a tabular format and downloaded for further analysis. The service is open access and users can register directly on the web application a the following url: `https://ibm.biz/paccmann-aas`.

Acknowledgments

The project leading to this publication has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826121

References

Ali, M. and Aittokallio, T. Machine learning and feature selection for drug response prediction in precision oncology applications. *Biophysical reviews*, pp. 1–9, 2018.

Ammad-Ud-Din, M., Georgii, E., Gonen, M., Laitinen, T., Kallioniemi, O., Wennerberg, K., Poso, A., and Kaski, S. Integrative and personalized qsar analysis in cancer by kernelized bayesian matrix factorization. *Journal of chemical information and modeling*, 54(8):2347–2359, 2014.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Bai, S., Kolter, J. Z., and Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Bajusz, D., Rácz, A., and Héberger, K. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7(1):20, 2015.

Barisione, G., Fabbi, M., Cutrona, G., De Cecco, L., Zupo, S., Leitinger, B., Gentile, M., Manzoni, M., Neri, A., Morabito, F., et al. Heterogeneous expression of the collagen receptor ddr1 in chronic lymphocytic leukaemia and correlation with progression. *Blood cancer journal*, 7(1):e513, 2017.

Bjerrum, E. J. Smiles enumeration as data augmentation for neural network modeling of molecules. *arXiv preprint arXiv:1703.07076*, 2017.

Blaschke, T., Olivecrona, M., Engkvist, O., Bajorath, J., and Chen, H. Application of generative autoencoder in de novo molecular design. *arXiv preprint arXiv:1711.07839*, nov 2017. URL `http://arxiv.org/abs/1711.07839`.

Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S., and Pujadas, G. Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63, 2015.

Chang, Y., Park, H., Yang, H.-J., Lee, S., Lee, K.-Y., Kim, T. S., Jung, J., and Shin, J.-M. Cancer drug response profile scan (cdrscan): A deep learning model that predicts drug effectiveness from cancer genomic signature. *Scientific reports*, 8(1): 8857, 2018.

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., and Ma'ayan, A. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC bioinformatics*, 14(1):128, 2013.

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., and Blaschke, T. The rise of deep learning in drug discovery. *Drug discovery today*, 2018.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Costello, J. C., Heiser, L. M., Georgii, E., Gönen, M., Menden, M. P., Wang, N. J., Bansal, M., Hintsanen, P., Khan, S. A., Mpindi, J.-P., et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature biotechnology*, 32(12):1202, 2014.

De Niz, C., Rahman, R., Zhao, X., and Pal, R. Algorithms for drug sensitivity prediction. *Algorithms*, 9(4):77, 2016.

Ding, M. Q., Chen, L., Cooper, G. F., Young, J. D., and Lu, X. Precision oncology beyond targeted therapy: Combining omics data with machine learning matches the majority of cancer cells to effective therapeutics. *Molecular Cancer Research*, 16(2): 269–278, 2018.

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. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570, 2012.

Geeleher, P., Cox, N. J., and Huang, R. S. Cancer biomarker discovery is improved by accounting for variability in general levels of drug sensitivity in pre-clinical models. *Genome biology*, 17 (1):190, 2016.

Goh, G. B., Hodas, N. O., Siegel, C., and Vishnu, A. Smiles2vec: An interpretable general-purpose deep neural network for predicting chemical properties. *arXiv preprint arXiv:1712.02034*, 2017.

Grapov, D., Fahrmann, J., Wanichthanarak, K., and Khoomrung, S. Rise of deep learning for genomic, proteomic, and metabolomic data integration in precision medicine. *Omics: a journal of integrative biology*, 22(10):630–636, 2018.

Gupta, A., Kumar, B. S., and Negi, A. S. Current status on development of steroids as anticancer agents. *The Journal of steroid biochemistry and molecular biology*, 137:242–270, 2013.

Hargrave-Thomas, E., Yu, B., and Reynisson, J. Serendipity in anticancer drug discovery. *World journal of clinical oncology*, 3(1):1, 2012.

Heltemes-Harris, L. M., Willette, M. J., Ramsey, L. B., Qiu, Y. H., Neeley, E. S., Zhang, N., Thomas, D. A., Koeuth, T., Baechler, E. C., Kornblau, S. M., et al. Ebf1 or pax5 haploinsufficiency synergizes with stat5 activation to initiate acute lymphoblastic leukemia. *Journal of Experimental Medicine*, 208(6):1135–1149, 2011.

Hofree, M., Shen, J. P., Carter, H., Gross, A., and Ideker, T. Network-based stratification of tumor mutations. *Nature methods*, 10(11):1108, 2013.

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. A landscape of pharmacogenomic interactions in cancer. *Cell*, 166(3):740–754, 2016.

Jastrzębski, S., Leśniak, D., and Czarnecki, W. M. Learning to smile (s). *arXiv preprint arXiv:1602.06289*, 2016.

Jiao, Q., Bi, L., Ren, Y., Song, S., Wang, Q., and Wang, Y.-s. Advances in studies of tyrosine kinase inhibitors and their acquired resistance. *Molecular cancer*, 17(1):36, 2018.

Kadurin, A., Aliper, A., Kazennov, A., Mamoshina, P., Vanhaelen, Q., Khrabrov, K., and Zhavoronkov, A. The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget*, 8(7): 10883–10890, feb 2017. ISSN 1949-2553. doi: 10.18632/oncotarget.14073.

Kalamara, A., Tobalina, L., and Rodriguez, J.-S. How to find the right drug for each patient? advances and challenges in pharmacogenomics. *Current Opinion in Systems Biology*, 2018.

Kim, H.-G., Hwang, S.-Y., Aaronson, S. A., Mandinova, A., and Lee, S. W. Ddr1 receptor tyrosine kinase promotes prosurvival pathway through notch1 activation. *Journal of Biological Chemistry*, 286(20):17672–17681, 2011.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., et al. Pubchem 2019 update: improved access to chemical data. *Nucleic acids research*, 2018.

Kimber, T. B., Engelke, S., Tetko, I. V., Bruno, E., and Godin, G. Synergy effect between convolutional neural networks and the multiplicity of smiles for improvement of molecular prediction. *arXiv preprint arXiv:1812.04439*, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Koprowski, R. and Foster, K. R. Machine learning and medicine: book review and commentary, 2018.

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, 44(W1):W90–W97, 2016.

Li, V. and Maki, A. Feature contraction: New convnet regularization in image classification. In *BMVC*, 2018.

Liu, H., Zhao, Y., Zhang, L., and Chen, X. Anti-cancer drug response prediction using neighbor-based collaborative filtering with global effect removal. *Molecular Therapy-Nucleic Acids*, 13:303–311, 2018.

Lloyd, I., Shimmings, A., and Scrip, P. S. Pharma r&d annual review 2017. *Available at: pharmaintelligence. informa. com/resources/product-content/pharma-rd-annual-review-2018. Accessed [June 25, 2018]*, 2017.

Menden, M. P., Iorio, F., Garnett, M., McDermott, U., Benes, C. H., Ballester, P. J., and Saez-Rodriguez, J. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS one*, 8(4):e61318, 2013.

Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., and Thomas, P. D. Panther version 11: expanded annotation data from gene ontology and reactome pathways, and data analysis tool enhancements. *Nucleic acids research*, 45(D1): D183–D189, 2016.

Oskooei, A., Born, J., Manica, M., Subramanian, V., Sáez-Rodríguez, J., and Martínez, M. R. Paccmann: Prediction of anticancer compound sensitivity with multi-modal attention-based neural networks. *arXiv preprint arXiv:1811.06802*, 2018a.

Oskooei, A., Manica, M., Mathis, R., and Martínez, M. R. Network-based biased tree ensembles (NetBiTE) for drug sensitivity prediction and drug sensitivity biomarker identification in cancer. *arXiv preprint arXiv:1808.06603*, 2018b.

Pandzic, T., Larsson, J., He, L., Kundu, S., Ban, K., Akhtar-Ali, M., Hellström, A. R., Schuh, A., Clifford, R., Blakemore, S. J., et al. Transposon mutagenesis reveals fludarabine-resistance mechanisms in chronic lymphocytic leukemia. *Clinical Cancer Research*, pp. clincanres–2903, 2016.

Petrova, E. Innovation in the pharmaceutical industry: The process of drug discovery and development. In *Innovation and marketing in the pharmaceutical industry*, pp. 19–81. Springer, 2014.

Popova, M., Isayev, O., and Tropsha, A. Deep reinforcement learning for de novo drug design. *Science advances*, 4(7):eaap7885, 2018.

Rainer, J., Lelong, J., Bindreither, D., Mantinger, C., Ploner, C., Geley, S., and Kofler, R. Research resource: transcriptional response to glucocorticoids in childhood acute lymphoblastic leukemia. *Molecular endocrinology*, 26(1):178–193, 2012.

Renema, N., Navet, B., Heymann, M.-F., Lezot, F., and Heymann, D. Rank–rankl signalling in cancer. *Bioscience reports*, 36(4): e00366, 2016.

Rogers, D. and Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, May 2010. ISSN 1549-9596. doi: 10.1021/ci100050t. URL https://doi.org/10.1021/ci100050t.

Schmidt, H. H., Dyomin, V. G., Palanisamy, N., Itoyama, T., Nanjangud, G., Pirc-Danoewinata, H., Haas, O. A., and Chaganti, R. Deregulation of the carbohydrate (chondroitin 4) sulfotransferase 11 (chst11) gene in a b-cell chronic lymphocytic leukemia with at (12; 14)(q23; q32). *Oncogene*, 23(41):6991, 2004.

Schwaller, P., Gaudin, T., Lanyi, D., Bekas, C., and Laino, T. Found in translation: predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chemical science*, 9(28):6091–6098, 2018a.

Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Bekas, C., and Lee, A. A. Molecular transformer for chemical reaction prediction and uncertainty estimation. *arXiv preprint arXiv:1811.02633*, 2018b.

Segler, M. H., Kogej, T., Tyrchan, C., and Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS central science*, 4(1):120–131, 2017.

Stanfield, Z., Coşkun, M., and Koyutürk, M. Drug response prediction as a link prediction problem. *Scientific reports*, 7:40321, 2017.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., et al. String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, 43(D1):D447–D452, 2014.

Tan, M. Prediction of anti-cancer drug response by kernelized multi-task learning. *Artificial intelligence in medicine*, 73:70–77, 2016.

Tan, M., Özgül, O. F., Bardak, B., Ekşioğlu, I., and Sabuncuoğlu, S. Drug response prediction by ensemble learning and drug-induced gene expression signatures. *arXiv preprint arXiv:1802.03800*, 2018.

Tanimoto, T. T. Elementary mathematical theory of classification and prediction. *IBM Technical Report*, 1958.

Turki, T. and Wei, Z. A link prediction approach to cancer drug sensitivity prediction. *BMC systems biology*, 11(5):94, 2017.

Unterthiner, T., Mayr, A., Klambauer, G., Steijaert, M., Wegner, J. K., Ceulemans, H., and Hochreiter, S. Deep learning as an opportunity in virtual screening. In *Proceedings of the deep learning workshop at NIPS*, volume 27, pp. 1–9, 2014.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Wang, L., Li, X., Zhang, L., and Gao, Q. Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. *BMC cancer*, 17(1):513, 2017.

Wang, Y., Fang, J., and Chen, S. Inferences of drug responses in cancer cells from cancer genomic features and compound chemical and therapeutic properties. *Scientific reports*, 6:32679, 2016.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Yang, M., Simm, J., Lam, C. C., Zakeri, P., van Westen, G. J., Moreau, Y., and Saez-Rodriguez, J. Linking drug target and pathway activation for effective therapy using multi-task learning. *Scientific reports*, 8, 2018.

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J. A., Thompson, I. R., et al. Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research*, 41(D1):D955–D961, 2012.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1480–1489, 2016.

Yuan, H., Paskov, I., Paskov, H., González, A. J., and Leslie, C. S. Multitask learning improves prediction of cancer drug sensitivity. *Scientific reports*, 6:31619, 2016.

Zhang, F., Wang, M., Xi, J., Yang, J., and Li, A. A novel heterogeneous network-based method for drug response prediction in cancer cell lines. *Scientific reports*, 8(1):3355, 2018a.

Zhang, L., Chen, X., Guan, N.-N., Liu, H., and Li, J.-Q. A hybrid interpolation weighted collaborative filtering method for anticancer drug response prediction. *Frontiers in Pharmacology*, 9, 2018b.

Zhang, N., Wang, H., Fang, Y., Wang, J., Zheng, X., and Liu, X. S. Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model. *PLoS computational biology*, 11(9):e1004498, 2015.