

Toward Explainable Anticancer Compound Sensitivity Prediction via Multimodal Attention-Based Convolutional Encoders

Matteo Manica,^{†,‡,§} Ali Oskooei,^{†,‡} Jannis Born,^{†,‡,⊥,‡,§} Vigneshwari Subramanian,[§] Julio Sáez-Rodríguez,^{||} and María Rodríguez Martínez^{*,†}

[†]IBM Research, 8803 Zürich, Switzerland

[‡]ETH Zürich, 8092 Zürich, Switzerland

[⊥]University of Zürich, 8006 Zürich, Switzerland

[§]RWTH Aachen University, 52056 Aachen, Germany

^{||}Heidelberg University, 69047 Heidelberg, Germany

Supporting Information

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 encoder significantly outperforms a baseline model trained on Morgan fingerprints and a selection of encoders based on SMILES, as well as the previously reported state-of-the-art for multimodal drug sensitivity prediction ($R^2 = 0.86$ 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 to 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.

KEYWORDS: drug sensitivity prediction, computational systems biology, deep learning, machine learning, drug discovery, multiscale, multimodal, attention, CNN, RNN, explainability, interpretability, molecular networks, molecular fingerprints, GDSC, SMILES, gene expression, drug discovery, drug sensitivity, anticancer compounds, IC50, EC50, lead discovery, personalized medicine, precision medicine

1. INTRODUCTION

1.1. Motivation. Discovering novel compounds with a desired efficacy and improving existing therapies are key bottlenecks in the pharmaceutical industry and fuel the largest R&D business spending of any industry, accounting for 19% of the total R&D spending worldwide.^{1,2} 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 (5212 of 15 267 drugs).³ Despite enormous scientific and technological advances in recent years, serendipity still plays a major role in anticancer drug discovery⁴ without a

Received: May 13, 2019

Revised: September 16, 2019

Accepted: October 16, 2019

Published: October 16, 2019

systematic way to accumulate and leverage years of R&D to achieve higher success rates. 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.⁵ 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.⁶ In this work, we present a multimodal approach 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.^{7–11} While the majority of them have focused on the analysis of unimodal datasets (genomics or transcriptomics, e.g., De Niz et al.,⁶ Tan,¹² Tan et al.,¹³ and Turki and Wei¹⁴), a handful of previous 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,¹⁵ kernelized Bayesian matrix factorization,¹⁶ Pearson correlation-based similarity networks,¹⁷ a Kronecker product kernel in conjunction with support vector machines (SVMs),¹⁸ autoencoders in combination with elastic net and SVMs,¹⁹ matrix factorization,²⁰ trace norm regularization,²¹ link predictions,²² and collaborative filtering.^{23,24} 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 interaction (PPI) networks.²⁵ For example, in a network-based per-drug approach integrating these data sources, Zhang et al.²⁶ 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,²⁷ but it has recently been argued that the usage of engineered features constrains the learning ability of machine learning algorithms.¹ 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.^{28–30} For instance, computational chemists borrowed methods from neural language models³¹ to encode SMILES³² strings of molecules and predict chemical properties of molecules.^{1,33,34} [SMILES (simplified molecular-input line-entry system) is an equivalent expression of molecules through text sequences; e.g., benzene is C1 = CC = CC = C1.] Once a gold standard in sequence modeling, recurrent neural networks (RNNs) were initially employed as SMILES encoders.^{1,35,36} However, it has been recently shown that convolutional architectures are superior to RNNs for sequence modeling,³⁷ and specifically for modeling

the SMILES string encoding of compounds.³⁸ It is noteworthy that these findings are in agreement with our model comparison results that reveal convolutional architectures as superior for SMILES sequence modeling.

Most recently, Chang et al.³⁹ adopted deep learning methods to develop a pan-drug model for predicting IC50 [half maximal inhibitory concentration, i.e., the micromolar concentration of a drug necessary to inhibit 50% of the cells] drug sensitivity of drug–cell line pairs. Utilizing >30 000 binary features (~3000 for the molecular drug fingerprint and the rest for a genomic fingerprint), they employed a model ensemble of five deep convolutional networks (four are completely linear) with convolutions applied separately to each of the genomic and molecular features before the encodings were merged. While we are working toward 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.³⁵ Second, by applying convolutions on SMILES, we can learn spatially meaningful filters, in contrast to applying them on molecular fingerprints. In accordance with Costello et al.,⁸ we use transcriptomic features (gene expression profiles) instead of genomic features since they have higher predictive power. 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.⁴⁰ 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 pharmaceuticals 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 the 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 encoders⁴¹ and propose a novel best-performing architecture, an attention-based multi-scale 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 that attention-based SMILES encoders significantly surpass a baseline feedforward model utilizing Morgan (circular) fingerprints.⁴²

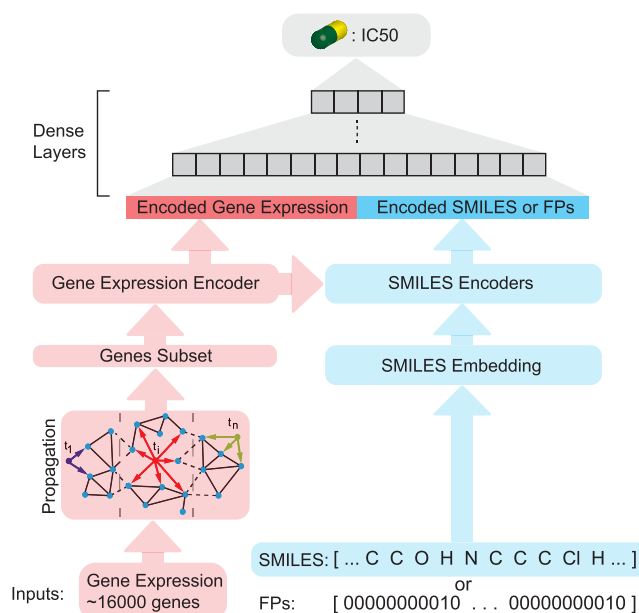


Figure 1. Multimodal end-to-end architecture of the proposed encoders. General framework for the explored architectures. Each model ingests a cell–compound pair and makes an IC₅₀ drug sensitivity prediction. Cells are represented by the gene expression values of a subset of 2128 genes, selected according 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 (for details see section 2) and used in combination with the gene expression encoder in order to predict drug sensitivity.

Using our multiscale convolutional attentive (MCA) encoder, we show that we achieve superior IC₅₀ prediction performance on the GDSC database⁴³ compared with the existing methods.^{15,39} 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 publicly available Genomics of Drug Sensitivity in Cancer (GDSC) database.⁴³ 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) on the log-scale. From the collected canonical SMILES, Morgan fingerprints were acquired using RDKit (512-bit with radius 2). Due to the nature of the SMILES language, most molecules can be validly represented through different SMILES strings (e.g., C(=O)=O and O=C=O both represent carbon dioxide). Exploiting this property, Bjerrum³⁵ proposed an effective data augmentation strategy which is adopted herein. 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.⁸ 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. Since each of the 985 cell lines in GDSC was initially represented by the expression levels of 17 737 genes an informed feature reduction was indispensable as we found it computationally intractable to process the high-dimensional raw data. To that end, we employed network propagation over the STRING protein–protein interaction (PPI) network⁴⁴ (a comprehensive PPI database including interactions from multiple data sources) leaving a subset of 2128 genes. Following the procedure described in Oskooei et al.,²⁵ 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 ($\epsilon = 1 \times 10^{-5}$) 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:²⁵

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

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

$$A' = D^{-1/2} A D^{-1/2} \quad (2)$$

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.⁴⁵ Adopting a convergence rule of $e = (W_{t+1} - W_t) < 1 \times 10^{-6}$, we solved eq 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 (2128 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. The selection was limited 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.³⁵

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.^{39,46} We explored several neural network SMILES encoder architectures to investigate whether the molecular information on 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 six-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) starting from the model proposed by Menden et al.,¹⁵ wherein 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 onward.

Commonalities of SMILES Encoders. To investigate which model architecture best learns the molecular information on compounds, we explored various SMILES encoders. Next to the expression profiles they ingest the SMILES text encodings for the structure of the compounds. The raw SMILES strings were tokenized to individual atoms using the regular expression from Schwaller et al.⁴⁷ For example the SMILES string of dinitrogen tetroxide ([N+](=O)[N+](=O)[O-][O-]) consists of 26 characters, but is decomposed into 14 entities ([N+] (=O) [N+] (=O) [O-] [O-]). This ensured that small functional units of the molecule (such as [NH] or [N+]) were represented as single entities to the model. The resulting atomic sequences were zero-padded and represented as $E = \{e_1, \dots, e_T\}$, with learned embedding vectors $e_i \in \mathbb{R}^H$ for each dictionary token (see Figure 2A). Each cell line, 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 ($p_{\text{drop}} = 0.5$) 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 hindering 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 two-layered bidirectional recurrent neural network (bRNN) with gated recurrent units (GRUs).⁴⁸ 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

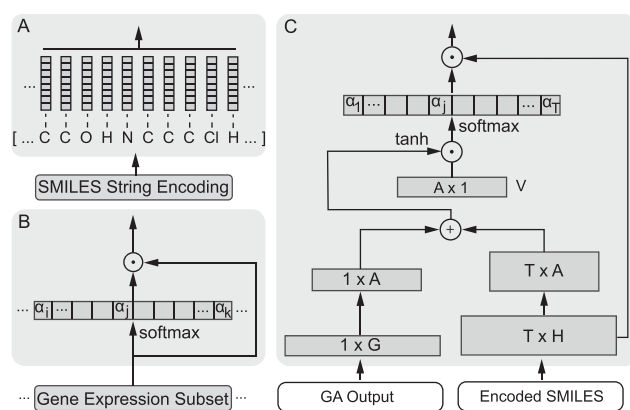


Figure 2. Key layers employed throughout the SMILES encoder. (A) SMILES Embedding (SE): An embedding layer transforms raw SMILES strings into a sequence of vectors in an embedding space. (B) Gene attention (GA): An attention-based gene expression encoder generates attention weights that are in turn applied to the input gene subset via a dot product. (C) Contextual attention (CA): 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.

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.³¹ Interpretability is paramount in healthcare and drug discovery.⁴⁹ 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 classification⁵⁰ for encoding SMILES strings. The SMILES attention weights α_i were computed per atomic token as

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

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:

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

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 models operate 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

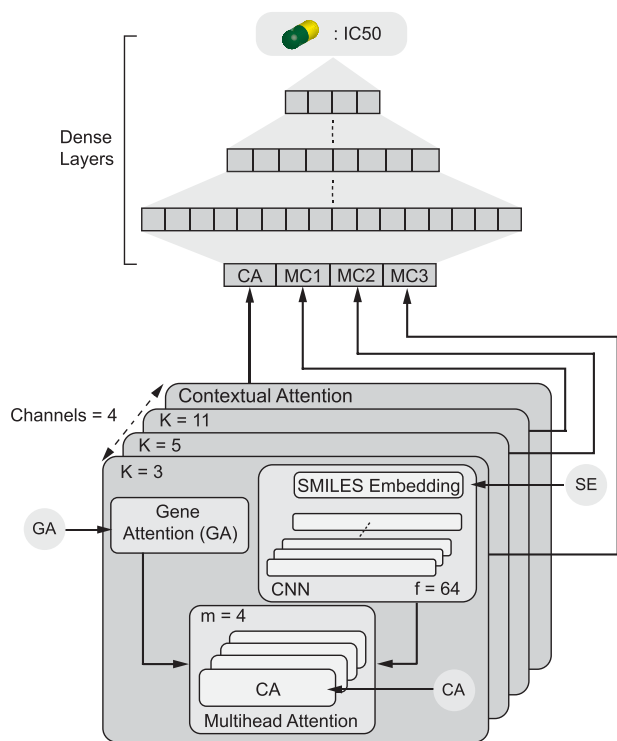


Figure 3. Model architecture of the multiscale convolutional attentive (MCA) encoder. The MCA model employed three parallel channels of convolutions over the SMILES sequence with kernel sizes K and one residual channel operating directly on the token level. Each channel applied a separate gene attention layer, before (convolved) SMILES and filtered genes were fed to a multihead of four contextual-attention layers. The outputs 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.

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 that attaches to the receptor's binding site. 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.⁵¹ 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.,⁵² 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 multihead attention approach, counteracts the tendency of the softmax to filter out the vast majority of the sequence steps.⁵³ 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 was concatenated before being 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 (1) because 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 the 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. This 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 ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$) and a decreasing learning rate.⁵⁴ 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 models, while $A = f = 64$ for MCA. In the final dense layers of all models, we employed dropout ($p_{\text{drop}} = 0.5$), batch normalization, and a sigmoid activation. All models were trained with a batch size of 2048 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

Table 1. Performance of the Explored Architectures on Test Data Following 25-Fold Cross-Validation^a

encoder type	drug structure	standardized RMSE median \pm IQR
deep baseline (DNN)	fingerprints	0.122 \pm 0.010
bidirectional recurrent (bRNN)	SMILES	0.119 \pm 0.011
stacked convolutional (SCNN)	SMILES	0.130 \pm 0.006
self-attention (SA)	SMILES	0.112* \pm 0.009
contextual attention (CA)	SMILES	0.110* \pm 0.007
multiscale convolutional attentive (MCA)	SMILES	0.109* \pm 0.009
MCA (prediction averaging)	SMILES	0.104** \pm 0.005

^aThe median RMSE and the 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).

using a 25-fold cross-validation scheme. As shown in Table 1:performance, 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 ($U = 126$, $p < 2 \times 10^{-4}$). 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: $U = 42$, $p < 9 \times 10^{-8}$, SA vs DNN: $U = 82$, $p < 5 \times 10^{-6}$). Surprisingly, neither complementing the SMILES embedding with positional encodings (similarly to Vaswani et al.⁵²) nor complementing the bRNN encoder with attention was found to improve the model performance. Ultimately, the MCA model, a development of the CA model (itself a progression from SA), was devised to combine token-level information (beneficial for the attention-only SA and CA 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, Afatinib, or Erlotinib), an enzyme which inhibits phosphorylation of epidermal growth factor receptors (EGFR) to suppress tumor cell proliferation.⁵⁵

The resulting MCA model also outperformed the baseline model significantly ($U = 136$, $p < 3 \times 10^{-4}$). In accordance with previous theoretical and empirical works demonstrating the superiority of model ensembles over single models in predictive accuracy,⁵⁶ we utilized a prediction averaging technique to further boost performance. Pooling the predictions of 20 MCA models (from different time points

during training), we obtained a RMSE of 0.104 on test data that not only significantly outperformed the baseline, but also the single MCA model ($U = 10$, $p < 2 \times 10^{-9}$ to the baseline and $U = 152$, $p < 9 \times 10^{-4}$ comparing to the plain MCA). While model averaging is known to be an efficient way to improve accuracy, the ensemble size of 20 was set to compromise computational cost and performance. 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 that had been adopted in previous works.³⁹ This allowed for a more meaningful comparison between the performance of our models with previous state of the art. As Figure 4 shows, the

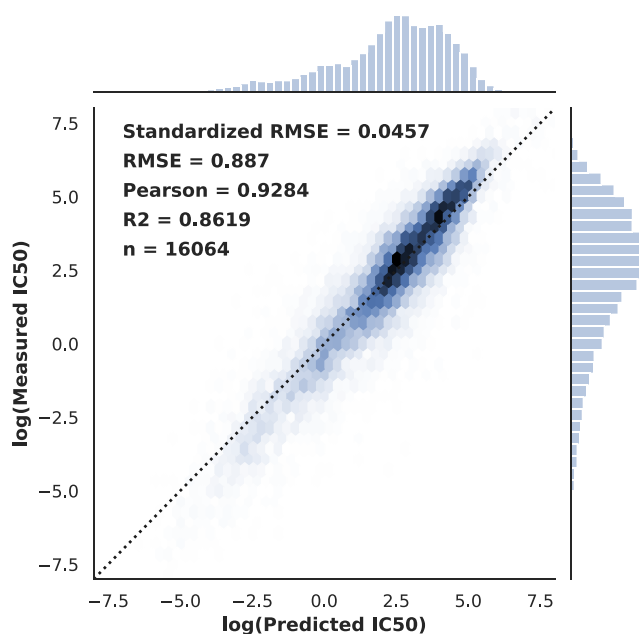


Figure 4. Test performance of MCA on lenient splitting. Scatter plot of correlation between true and predicted drug sensitivity by a late-fusion model ensemble of all five folds. The model was fitted in log space.

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.

Comparing to previous pan-drug models, we surpass the results of Menden et al.¹⁵ and the recently presented CDRscan³⁹ model [they achieved a RMSE of 1.07 and R^2 of 0.84 despite using more than 1 order of magnitude more features than our model's cell biology] while leveraging model interpretability as follows.

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,⁵⁷ an established index for evaluating drug similarity based on fingerprints.⁵⁸ This approach resulted in a Pearson correlation of $\rho = 0.64$ ($n = 22500$, $p < 1 \times 10^{-50}$). 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.^{59,60} 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.⁶¹ Programmed cellular death is a key molecular process elicited by anticancer compounds which validates the attention mechanism as being in accordance with prior knowledge from cell biology.

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 studied drugs are tyrosine kinase inhibitors that are predominantly applied in hematopoietic 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 = 81 nM) being superior to Masitinib (223 nM). 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 are presented the most attended genes of the studied leukemia cell line and their STRING protein neighborhoods. Interestingly, the *DDR1* protein is a member of receptor tyrosine kinases (RTKs), the same group of cell membrane receptors that both Imatinib and Masitinib

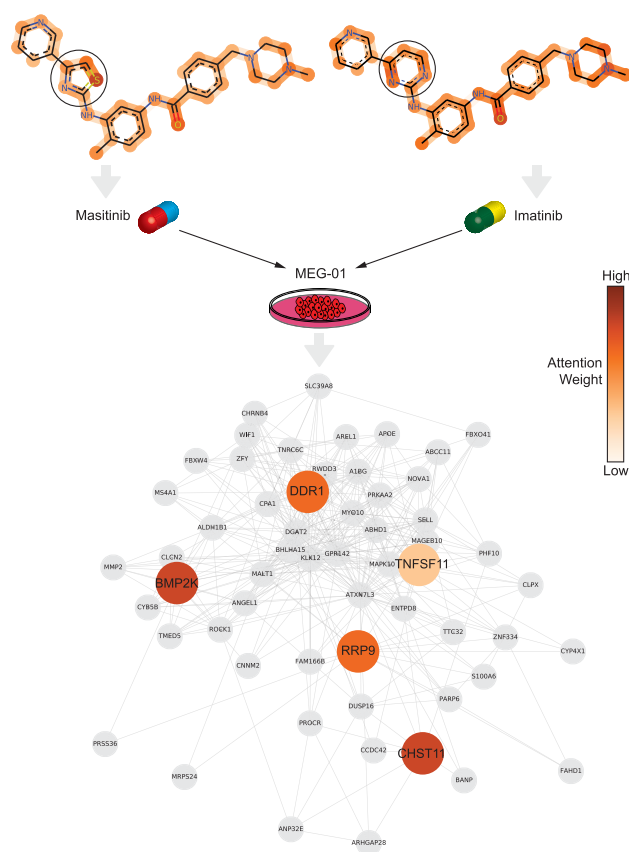


Figure 5. Neural attention on molecules and genes. The molecular attention maps on the top demonstrate how the model's attention is shifted when the thiazole group is replaced by a piperazine group. The change in attention across the two molecules is particularly concentrated around the affected rings, 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).

inhibit.⁶² *DDR1* gene is highly expressed in various cancer types, such as in chronic lymphocytic leukemia.⁶³ In addition, *BMP2K* gene has been recently shown to be implicated in chronic lymphocytic leukemia (CLL),⁶⁴ while *CHST11* has long been known to be deregulated in CLL.⁶⁵ *TNFSF11* encodes RANKL, which is part of a prominent cancer signaling pathway,⁶⁶ and *TNFSF11* has been reported to be the most overexpressed gene in a sample of $n = 129$ acute lymphoblastic leukemia (ALL) patients.⁶⁷ *RRP9* has been shown to be crucial in treating ALL.⁶⁸ In conclusion, the prior knowledge from the cancer literature validates our findings and indicates 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 showed 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 explainability 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 by 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. Following a propagation technique over the STRING PPI network, our model explored the 2128 most informative instead of all 17 737 genes. Utilizing the full set of genes instead would render model training computationally intractable; but alternative feature reduction techniques that do not neglect the majority of genes, such as deriving single-sample signature scores for all relevant pathways,⁶⁹ were not yet explored herein. The apparent benefit of our gene-based approach is the gene attention mechanism which would be dropped in an approach purely based on pathway activity scores. However, extending our model with an additional input channel for pathway scores and an associated pathway attention mechanism could greatly complement the representation of the tumor cell.

A key feature of our models was the strict training and evaluation strategy that set our work apart from previous approaches. In our strict model evaluation approach, cells and compounds were split in training, validation and test datasets 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 R^2 of 86%. A valid concern is regarding the choice of IC50 as cell response metric. We acknowledge that choice being simplistic, but we emphasize that our method is data driven and the large public databases (GDSC, CCLE etc.) do not share any information about cell growth rates. However, an interesting future endeavor is to incorporate the full dose–response curve instead of only exploring the IC50 point estimates. This would not only greatly increase the amount of available data points but also to assay and employ the model more rigorously and precisely as the user could also vary drug concentration.

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, and (2) leveraging our model in combination with recent advances in small-molecule generation using generative models^{70,71} and reinforcement learning⁷² 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.

■ ASSOCIATED CONTENT

📄 Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.molpharmaceut.9b00520.

Details of data splits, CNV inclusion, and comparisons with other regression models; the best trained model in compressed format; the processed data following both the strict split and the lenient split strategies; and a list of genes selected via network propagation (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: tte@zurich.ibm.com.

ORCID

Matteo Manica: 0000-0002-8872-0269

Jannis Born: 0000-0001-8307-5670

Notes

The authors declare no competing financial interest.

Availability of Software and Materials. The data in TFRecord format used in the benchmark studies conducted in this work can be downloaded at <https://ibm.biz/paccmann-data>. Alternatively the reader can access the raw cell line data from GDSC⁴³ and the compound structural information from PubChem⁷³ and the LINCS database. The implementation of the models used in the benchmark is available in the form of a toolbox on GitHub at <https://github.com/drugilberg/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 at <https://ibm.biz/paccmann-aas>.

#M.M., A.O., and J.B. share first authorship.

■ ACKNOWLEDGMENTS

The authors would like to thank Dr. Maria Gabriani for her continuous support and useful discussions. The projects leading to this publication have received funding from the European Union's Horizon 2020 research and innovation program under grant agreements no. 668858 and no. 826121.

■ REFERENCES

- (1) Goh, G. B.; Hodas, N. O.; Siegel, C.; Vishnu, A. SMILES2Vec: An Interpretable General-Purpose Deep Neural Network for Predicting Chemical Properties. *arXiv:1712.02034 [stat.ML]*, arXiv preprint, 2017. <https://arxiv.org/abs/1712.02034>.
- (2) Petrova, E. *Innovation and marketing in the pharmaceutical industry*; Springer, 2014; pp 19–81.

- (3) Lloyd, I.; Shimmings, A.; Scrip, P. S. Pharma R&D Annual Review 2018. <https://pharmaintelligence.informa.com/resources/product-content/pharma-rd-annual-review-2018> (accessed June 25, 2018).
- (4) Hargrave-Thomas, E.; Yu, B.; Reynisson, J. Serendipity in anticancer drug discovery. *World J. Clin. Oncol.* **2012**, *3* (1), 1.
- (5) Geeleher, P.; Cox, N. J.; Huang, R. S. Cancer biomarker discovery is improved by accounting for variability in general levels of drug sensitivity in pre-clinical models. *Genome Biol.* **2016**, *17*, 190.
- (6) De Niz, C.; Rahman, R.; Zhao, X.; Pal, R. Algorithms for drug sensitivity prediction. *Algorithms* **2016**, *9*, 77.
- (7) Ali, M.; Aittokallio, T. Machine learning and feature selection for drug response prediction in precision oncology applications. *Biophys. Rev.* **2019**, *11*, 31.
- (8) Costello, J. C.; et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.* **2014**, *32*, 1202.
- (9) Garnett, M. J.; et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **2012**, *483*, 570.
- (10) Kalamara, A.; Tobalina, L.; Saez-Rodriguez, J. How to find the right drug for each patient? Advances and challenges in pharmacogenomics. *Curr. Opin. Syst. Biol.* **2018**, *10*, 53.
- (11) Yang, W.; et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* **2012**, *41*, D955–D961.
- (12) Tan, M. Prediction of anti-cancer drug response by kernelized multi-task learning. *Artificial intelligence in medicine* **2016**, *73*, 70–77.
- (13) Tan, M.; Özgül, O. F.; Bardak, B.; Ekşioğlu, I.; Sabuncuoğlu, S. Drug response prediction by ensemble learning and drug-induced gene expression signatures. *arXiv:1802.03800*, arXiv preprint, 2018. <https://arxiv.org/abs/1802.03800>.
- (14) Turki, T.; Wei, Z. A link prediction approach to cancer drug sensitivity prediction. *BMC Syst. Biol.* **2017**, *11*, 94.
- (15) Menden, M. P.; et al. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS One* **2013**, *8*, No. e61318.
- (16) Ammad-Ud-Din, M.; et al. Integrative and personalized QSAR analysis in cancer by kernelized Bayesian matrix factorization. *J. Chem. Inf. Model.* **2014**, *54*, 2347–2359.
- (17) Zhang, N.; et al. Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model. *PLoS Comput. Biol.* **2015**, *11*, No. e1004498.
- (18) Wang, Y.; Fang, J.; Chen, S. Inferences of drug responses in cancer cells from cancer genomic features and compound chemical and therapeutic properties. *Sci. Rep.* **2016**, *6*, 32679.
- (19) Ding, M. Q.; Chen, L.; Cooper, G. F.; Young, J. D.; Lu, X. Precision oncology beyond targeted therapy: Combining omics data with machine learning matches the majority of cancer cells to effective therapeutics. *Mol. Cancer Res.* **2018**, *16*, 269–278.
- (20) Wang, L.; Li, X.; Zhang, L.; Gao, Q. Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. *BMC Cancer* **2017**, *17*, 513.
- (21) Yuan, H.; Paskov, I.; Paskov, H.; González, A. J.; Leslie, C. S. Multitask learning improves prediction of cancer drug sensitivity. *Sci. Rep.* **2016**, *6*, 31619.
- (22) Stanfield, Z.; Coşkun, M.; Koyutürk, M. Drug response prediction as a link prediction problem. *Sci. Rep.* **2017**, *7*, 40321.
- (23) Liu, H.; Zhao, Y.; Zhang, L.; Chen, X. Anti-cancer drug response prediction using neighbor-based collaborative filtering with global effect removal. *Mol. Ther.–Nucleic Acids* **2018**, *13*, 303–311.
- (24) Zhang, L.; Chen, X.; Guan, N.-N.; Liu, H.; Li, J.-Q. A hybrid interpolation weighted collaborative filtering method for anti-cancer drug response prediction. *Front. Pharmacol.* **2018**, *9*, 01017.
- (25) Oskooei, A.; Manica, M.; Mathis, R.; Martínez, M. R. Network-based Biased Tree Ensembles (NetBiTE) for Drug Sensitivity Prediction and Drug Sensitivity Biomarker Identification in Cancer. *arXiv:1808.06603 [q-bio.QM]*, arXiv preprint, 2018. <https://arxiv.org/abs/1808.06603>
- (26) Zhang, F.; Wang, M.; Xi, J.; Yang, J.; Li, A. A novel heterogeneous network-based method for drug response prediction in cancer cell lines. *Sci. Rep.* **2018**, *8*, 3355.
- (27) Cereto-Massagué, A.; et al. Molecular fingerprint similarity search in virtual screening. *Methods* **2015**, *71*, 58–63.
- (28) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The rise of deep learning in drug discovery. *Drug Discovery Today* **2018**, *23*, 1241.
- (29) Grapov, D.; Fahrman, J.; Wanichthanarak, K.; Khoomrung, S. Rise of deep learning for genomic, proteomic, and metabolomic data integration in precision medicine. *Omics: a journal of integrative biology* **2018**, *22*, 630–636.
- (30) Wu, Z.; et al. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **2018**, *9*, 513–530.
- (31) Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473 [cs.CL]*, arXiv preprint, 2014. <https://arxiv.org/abs/1409.0473>.
- (32) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **1988**, *28*, 31–36.
- (33) Jastrzębski, S.; Leśniak, D.; Czarnecki, W. M. Learning to SMILE (S). *arXiv:1602.06289 [cs.CL]*, arXiv preprint, 2016. <https://arxiv.org/abs/1602.06289>
- (34) Schwaller, P.; et al. Molecular transformer for chemical reaction prediction and uncertainty estimation. *arXiv:1811.02633 [physics.chem-ph]*, arXiv preprint, 2018. <https://arxiv.org/abs/1811.02633>.
- (35) Bjerrum, E. J. SMILES enumeration as data augmentation for neural network modeling of molecules. *arXiv:1703.07076 [cs.LG]*, arXiv preprint, 2017. <https://arxiv.org/abs/1703.07076>.
- (36) Segler, M. H.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* **2018**, *4*, 120–131.
- (37) Bai, S.; Kolter, J. Z.; Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv:1803.01271 [cs.LG]*, arXiv preprint, 2018. <https://arxiv.org/abs/1803.01271>.
- (38) Kimber, T. B.; Engelke, S.; Tetko, I. V.; Bruno, E.; Godin, G. Synergy Effect between Convolutional Neural Networks and the Multiplicity of SMILES for Improvement of Molecular Prediction. *arXiv:1812.04439 [cs.LG]* arXiv preprint, 2018. <https://arxiv.org/abs/1812.04439>.
- (39) Chang, Y.; Park, H.; Yang, H.-J.; Lee, S.; Lee, K.-Y.; Kim, T. S.; Jung, J.; Shin, J.-M.; et al. Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature. *Sci. Rep.* **2018**, *8*, 8857.
- (40) Yang, M.; Simm, J.; Lam, C. C.; Zakeri, P.; van Westen, G. J. P.; Moreau, Y.; Saez-Rodriguez, J. Linking drug target and pathway activation for effective therapy using multi-task learning. *Sci. Rep.* **2018**, *8*, 8322.
- (41) Oskooei, A. et al. PaccMann: Prediction of anticancer compound sensitivity with multi-modal attention-based neural networks. *arXiv:1811.06802 [cs.LG]*, arXiv preprint, 2018. <https://arxiv.org/abs/1811.06802>.
- (42) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
- (43) Iorio, F.; et al. A landscape of pharmacogenomic interactions in cancer. *Cell* **2016**, *166*, 740–754.
- (44) Szklarczyk, D.; et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **2015**, *43*, D447–D452.
- (45) Hofree, M.; Shen, J. P.; Carter, H.; Gross, A.; Ideker, T. Network-based stratification of tumor mutations. *Nat. Methods* **2013**, *10*, 1108.
- (46) Unterthiner, T.; et al. Deep learning as an opportunity in virtual screening. *Proceedings of the Deep Learning Workshop at NIPS*, 2014:19.
- (47) Schwaller, P.; Gaudin, T.; Lanyi, D.; Bekas, C.; Laino, T. Found in Translation: predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* **2018**, *9*, 6091–6098.

- (48) Cho, K.; et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv:1406.1078 [cs.CL]*, arXiv preprint, 2014. <https://arxiv.org/abs/1406.1078>.
- (49) Koprowski, R.; Foster, K. R. Machine learning and medicine: book review and commentary. *BioMed. Eng.* **2018**, *17*, 17.
- (50) Yang, Z.; et al. Hierarchical attention networks for document classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* **2016**, 1480–1489.
- (51) Gupta, A.; Kumar, B. S.; Negi, A. S. Current status on development of steroids as anticancer agents. *J. Steroid Biochem. Mol. Biol.* **2013**, *137*, 242–270.
- (52) Vaswani, A.; et al. Attention is all you need. *Advances in Neural Information Processing Systems 30*, NIPS 2017; pp 5998–6008.
- (53) Li, V.; Maki, A. Feature Contraction: New ConvNet Regularization in Image Classification. *BMVC* 2018.
- (54) Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. *arXiv:1412.6980 [cs.LG]*, arXiv preprint, 2014. <https://arxiv.org/abs/1412.6980>.
- (55) Jiao, Q.; Bi, L.; Ren, Y.; Song, S.; Wang, Q.; Wang, Y.-s.; et al. Advances in studies of tyrosine kinase inhibitors and their acquired resistance. *Mol. Cancer* **2018**, *17*, 36.
- (56) Finlay, S. Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research* **2011**, *210*, 368–378.
- (57) Tanimoto, T. T. Elementary mathematical theory of classification and prediction. *IBM Technical Report*, 1958.
- (58) Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminf.* **2015**, *7*, 20.
- (59) Chen, E. Y.; et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinf.* **2013**, *14*, 128.
- (60) Kuleshov, M. V.; et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **2016**, *44*, W90–W97.
- (61) Mi, H.; et al. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.* **2017**, *45*, D183–D189.
- (62) Kim, H.-G.; Hwang, S.-Y.; Aaronson, S. A.; Mandinova, A.; Lee, S. W. DDR1 receptor tyrosine kinase promotes prosurvival pathway through Notch1 activation. *J. Biol. Chem.* **2011**, *286*, 17672–17681.
- (63) Barisione, G.; et al. Heterogeneous expression of the collagen receptor DDR1 in chronic lymphocytic leukaemia and correlation with progression. *Blood cancer journal* **2017**, *7*, e513.
- (64) Pandzic, T.; Larsson, J.; He, L.; Kundu, S.; Ban, K.; Akhtar-Ali, M.; Hellstrom, A. R.; Schuh, A.; Clifford, R.; Blakemore, S. J.; Strefford, J. C.; Baumann, T.; Lopez-Guillermo, A.; Campo, E.; Ljungstrom, V.; Mansouri, L.; Rosenquist, R.; Sjoblom, T.; Hellstrom, M. Transposon mutagenesis reveals fludarabine-resistance mechanisms in chronic lymphocytic leukemia. *Clin. Cancer Res.* **2016**, *22*, 6217.
- (65) Schmidt, H. H.; et al. Deregulation of the carbohydrate (chondroitin 4) sulfotransferase 11 (CHST11) gene in a B-cell chronic lymphocytic leukemia with at (12; 14)(q23; q32). *Oncogene* **2004**, *23*, 6991.
- (66) Renema, N.; Navet, B.; Heymann, M.-F.; Lezot, F.; Heymann, D. RANK–RANKL signalling in cancer. *Biosci. Rep.* **2016**, *36*, No. e00366.
- (67) Heltemes-Harris, L. M.; et al. Ebf1 or Pax5 haploinsufficiency synergizes with STAT5 activation to initiate acute lymphoblastic leukemia. *J. Exp. Med.* **2011**, *208*, 1135–1149.
- (68) Rainer, J.; et al. Research resource: transcriptional response to glucocorticoids in childhood acute lymphoblastic leukemia. *Mol. Endocrinol.* **2012**, *26*, 178–193.
- (69) Zhang, J. D.; Hatje, K.; Sturm, G.; Broger, C.; Ebeling, M.; Burtin, M.; Terzi, F.; Pomposiello, S. I.; Badi, L.; et al. Detect tissue heterogeneity in gene expression data with BioQC. *BMC Genomics* **2017**, *18*, 277.
- (70) Blaschke, T.; Olivecrona, M.; Engkvist, O.; Bajorath, J.; Chen, H. Application of generative autoencoder in de novo molecular design. *arXiv:1711.07839 [cs.LG]*, arXiv preprint, 2017. <https://arxiv.org/abs/1711.07839>.
- (71) Kadurin, A.; Aliper, A.; Kazennov, A.; Mamoshina, P.; Vanhaelen, Q.; Khrabrov, K.; Zhavoronkov, A.; et al. The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget* **2017**, *8*, 10883–10890.
- (72) Popova, M.; Isayev, O.; Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci. Adv.* **2018**, *4*, eaap7885.
- (73) Kim, D.; Hur, J.; Han, J. H.; Ha, S. C.; Shin, D.; Lee, S.; Park, S.; Sugiyama, H.; Kim, K. K. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **2018**, *46*, 10504.