23B2444

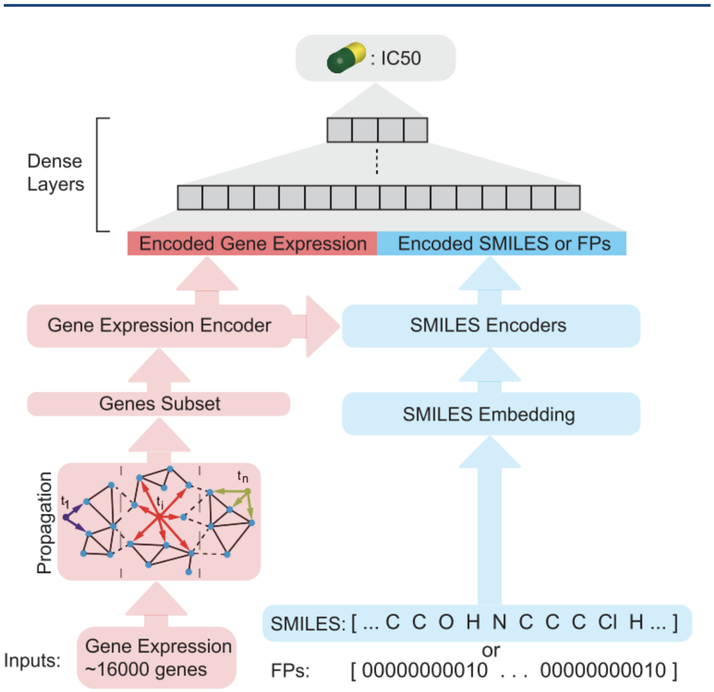
Harsh Kumar Singh

Report on PaccMann (reward function) -SURP 2025

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

Motivation~ The paper is motivated by the high cost, low efficiency, and randomness in anticancer drug discovery, combined with the need for precision medicine due to biological variability. The authors propose a multimodal attention-based deep learning model that systematically combines chemical and genomic data to predict drug sensitivity and aid in personalized therapy design and rational drug development.

Overview~



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. Left side inputs 16,000 gene expression values from a cancer cell line. **PPI Propagation**: Uses the PPI network to filter out unimportant genes based on their **connectivity**

and interaction patterns. Cells are represented by the gene expression values of a subset of 2128 genes, selected according to a network propagation procedure.

SMILES string is embedded as a **sequence of atoms**. 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.

Datasets used~

1)**GDSC (Genomics of Drug Sensitivity in Cancer)** 985 cancer cell lines from different cancer types ,we know the gene expressions. Alos they were tested with different drugs.
2)Drug sensitivity values were measured by IC50 values.
3) SMILES = a text format that represents the atoms and bonds in a molecule .Embedding them into morgan fingerprints(binary representation of molecuels) . For data augmentation ,each molecule can be represented by different smiles strings so helps in generalization.

Transcriptomics (gene expression of 17k genes) of 985 cell lines from different cancer types .

Network Propagation~

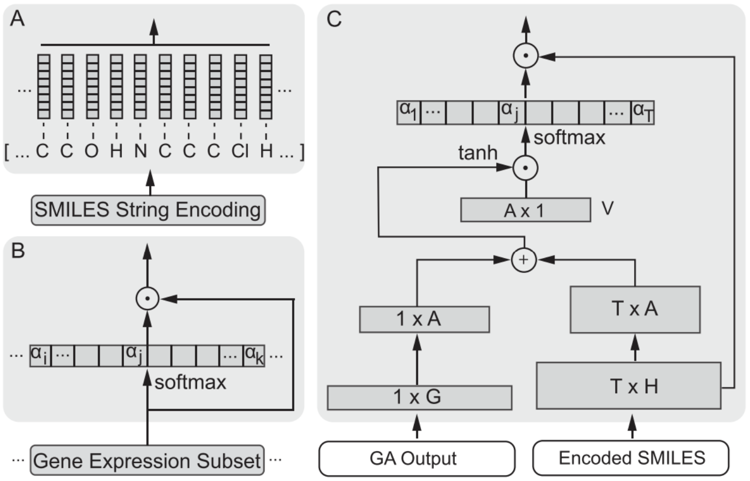
Since the transcriptome of of 17k genes (features) can cause overfitting due to large feature numbers. We need dimension reduction of these features into smaller and useful ones. This is done using PPI (Protein-protein interaction) network, network propagation over the STRING protein− protein interaction (PPI) network44 (a comprehensive PPI database including interactions from multiple data sources) leaving a subset of 2128 genes. We adopted network propagation where weights were distributed for each target for every drug. Initialising with w=1 for the reported drug test genes and low value of w to the others. Then we iterated through all weights $W(t+1)= \alpha W(t)A` +(1- \alpha)W(0)$

where D is the degree matrix and A' is the normalized adjacency matrix, obtained from the degree matrix D: $A`=D^{(-1/2)} A D(1/2)$ …where diffusion tuning parameter-> α ($0 \le \alpha \le 1$), d $\alpha = 0.7$, as recommended in the literature for the STRING network.

We used the resultant weights distribution to determine the top 20 highly ranked genes for each drug. The union of these top ranked genes for 208 drugs gives the number of 2,128 which is resultant useful no. of genes for out model. 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 >5.5 million data points following SMILES augmentation.

Model Architecture~ There are many **alternative ways** (not combined) to encode the **drug's SMILES representation** in the overall PACMAN pipeline.
They are **compared independently** to see which encoder works best for representing molecular structure.

1) Baseline model A **6-layer Deep Neural Network (DNN)** with layer sizes: [512, 256, 128, 64, 32, 16] used with sigmoid activation function.Input is given as combination of gene expression + morgan fingerprint 512 bit and out is predicted IC50.
2) Commonalities of SMILES encoder: common components or shared architecture that are used across **all** the SMILES encoder models (like bRNN, SCNN, MCA, etc.). so Here firstly we tokenize the SMILES means model can interpret functional groups and not just separate characters like (N+ etc) These tokens are embedded into vectors(like embedding in NLP) ,further zero padded to make equal lengths of vector. The genetic subset i.e result of network propagation is fed to gene attention encoder which distributes attention weights.SMILES encoder followed by dense layer (dropout =0.5)for regularization.Linear activation (rather than sigmoid) to avoid restricting the values between 0 and 1.
3) Bidirectional Recurrent (bRNN) : Concatene the final states of forward and backward GRU-RNN and fed to dense layer for IC50 prediction.
4) The SCNN model: learns to represent molecules from SMILES using layered convolutions, compresses embeddings, and captures **global molecular features** — offering a faster alternative to bRNN while maintaining strong performance.



- Self Attention: "Self-attention" means that **each token (e.g., atom) in the input sequence looks at all other tokens (including itself)** to decide how much attention (importance) to give them — hence, it attends to **itself** and others. This mechanism was adopted for encoding SMILES string to explain and interpret the results in context of biological and chemical knowledge. The model will ensure final results will show which part of molecule is more important and useful. Attention weights which are attached to different tokens are computed for this task->

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

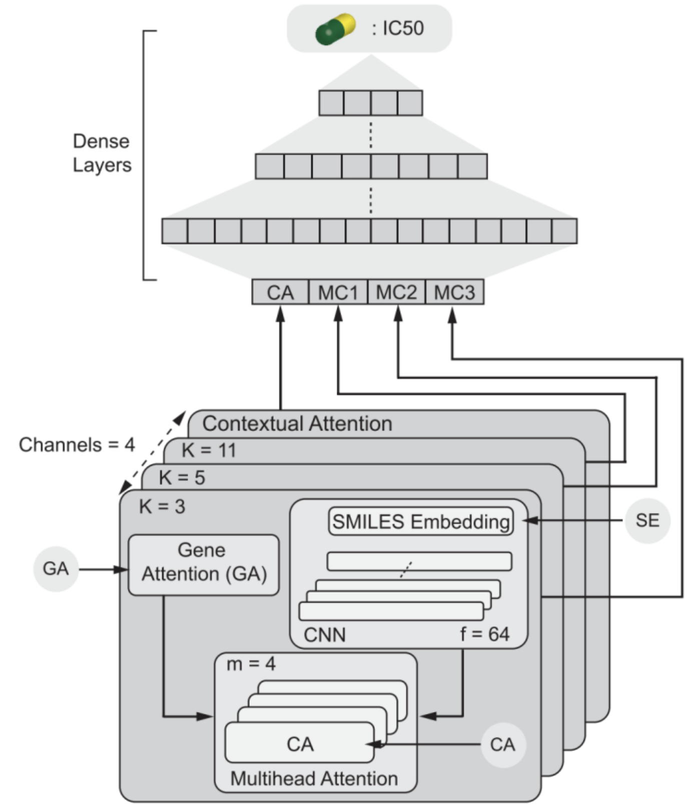
α is the final attention weight for i th token normalized by softmax. We is weight matrix projecting embeddings ,Si is the encoding of ith token of molecule, b is bias matrix.U is learnable attention vector of ith token

- Contextual Attention: Now the effect of molecule doesn't not just depend on structure alone but the biological context of drug with cell line as well. The attention weights attached to different tokens of molecule along with Gene expression subset as context, determines their role in affecting the particular cancer cell line and improves personalized treatment.

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

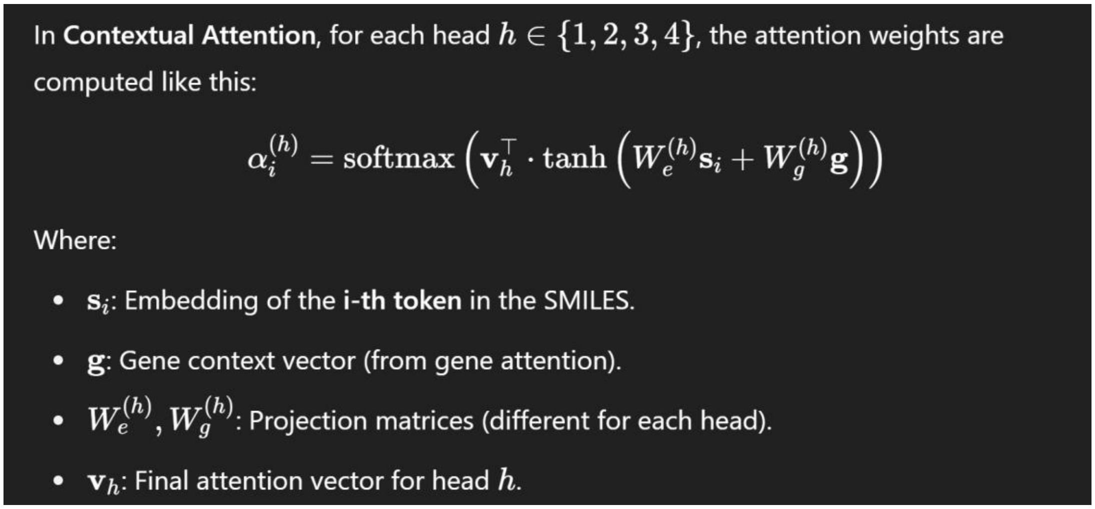
We is weight for encoding the molecular structure into attention space. Wg is the attention weight for molecule but with biological context(gene profile of target cells). Ui is final contextual attention wight of i th token of drug molecule.

Activation function is tanh. V(T) is transpose of learnable attention vector which scores the molecule on overall information.
We can apply softmax later for normalization.



- Multiscale Convolutional Attention (MCA): We have self attention and contextual attention for individual molecular structure but position of different function groups and tokens also carry importance while studying drugs sensitivity. Input will be embedded vectors of SMILES string of molecule . Filtered Gene expression will passed through gene attention.
  For multiscale feature extraction form SMILES we pass embeddings through 3 different sizes of kernals where size 3 captures short range dependencies, size 5 captures mid range and 11 sized kernals capture long range dependencies. Each applies 64 (f = 64) filters for rich feature mapping.
  Each convolutional output goes through **4 contextual attention heads** (m=4) attention weights for SMILES tokens are not just based on the molecule , they are **personalized for each cell line**. One extra (4th) channel **skips the CNN** part and directly applies attention on raw SMILES embeddings to preserve the original information. All 4 channels (3 CNN + 1 skip) after attention are concatenated together. These are marked as MC1, MC2, MC3

and CA at the bottom of the pyramid. This fused feature vector is passed through **stacked Dense Layers** (MLP) to predict the final output: **IC50** (how effective the drug is).



In **Contextual Attention**, for each head $h \in \{1, 2, 3, 4\}$, the attention weights are computed like this:

$$\alpha_i^{(h)} = \text{softmax}\left(\mathbf{v}_h^\top \cdot \tanh\left(W_e^{(h)}\mathbf{s}_i + W_g^{(h)}\mathbf{g}\right)\right)$$

Where:

- $\mathbf{s}_i$: Embedding of the **i-th token** in the SMILES.
- $\mathbf{g}$: Gene context vector (from gene attention).
- $W_e^{(h)}, W_g^{(h)}$: Projection matrices (different for each head).
- $\mathbf{v}_h$: Final attention vector for head h.

Model evaluation: To evaluate how well our model(MCA) predicts drug sensitivity . In strict split certain number of drugs and cell lines are kept for final testing and never seen during training of model. This helps in generalization and evaluation of real world unseen drugs.
208 drugs , 985 cell lines, which constitutes of 175603 pairs out of these 10% of drugs and cell lines kept separate from training dataset.25-fold cross validation scheme helps is employed where 4% part of dataset is kept for validation while rest is used in training. Because you only use pairs where both drug and cell line are in the same group, many combinations can't be used. In lineant split , 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 were normalized to the range **[0, 1].** Gene expression data was standardized **(i.e., converted to mean = 0 and standard deviation = 1**) based on the training data.

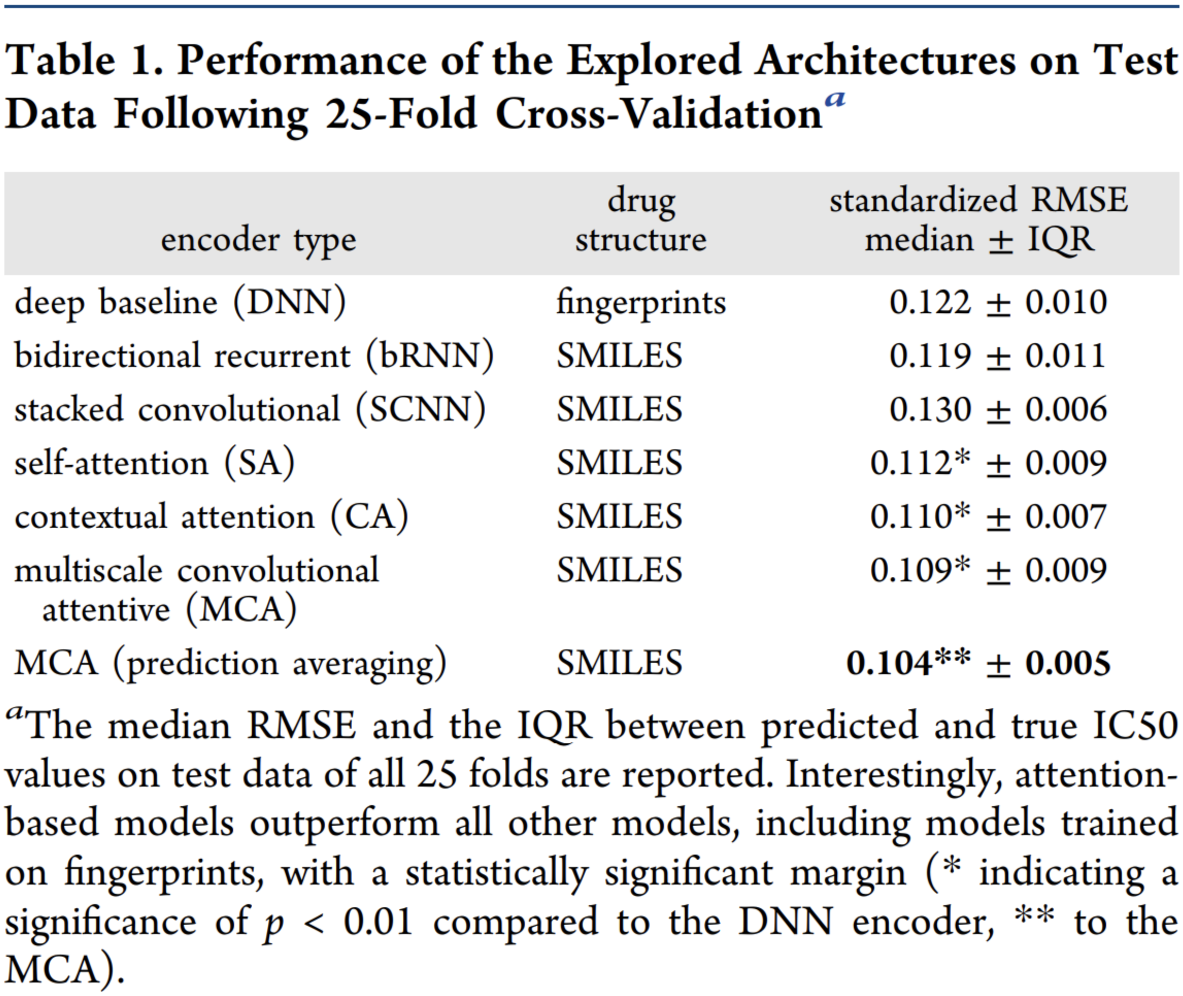


Table 1. Performance of the Explored Architectures on Test Data Following 25-Fold Cross-Validation[a]

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**

[a] The 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).

Results: DNN is baseline model . bRNN although with recurrent neural netwok for sequential information couldn't perform better. SCNN performance is even worse means the long range dependencies are not good measure for drug sensitivity. Attention mechanism reduces error which means short-range dependencies plays major role. Multiscale convolutional attention outperforms all the model the averaging of 20 such models gives more reduced error.this metrices makes it State-of-the art model for drug sensitivity prediction.The **MCA model with prediction averaging** is the best-performing architecture for drug sensitivity prediction, with the **lowest RMSE** and **statistical significance** over traditional methods. It combines the power of:

- CNNs (local + large-scale substructure detection),
- gene expression context,
- attention mechanisms (interpretability + relevance),
  to give **state-of-the-art results** on the GDSC dataset.