Deep Learning for Drug Response Prediction with Gene Expression Data

Sardar Jaffar Ali[1], Muhammad Omer[2], Duc Tai Le[3], Syed M. Raza[4], and Hyunseung Choo[3,*]

[1]Dept. of AI Systems Engineering, Sungkyunkwan University, Suwon, South Korea
[2]Dept. of Computer Science and Engineering, Sungkyunkwan University, Suwon, South Korea
[3]Dept. of Electrical and Computer Engineering, Sungkyunkwan University, Suwon, South Korea
[4]James Watt School of Engineering, University of Glasgow, Glasgow, United Kingdom
[*]Corresponding author {choo@skku.edu}

***Abstract*—Accurately predicting drug responses based on individual patient profiles is a critical challenge in personalized medicine, primarily due to the complex biological variability involved. This paper presents a deep learning framework for predicting changes in gene expression, providing insights into how drugs impact cells at the molecular level. Using data from the Kaggle competition, several models have been evaluated, including LSTM, GRU, Transformer, and Autoencoder architectures. Among these, the 3-stacked GRU with Attention demonstrated superior performance, achieving the highest sign accuracy of 79% and the lowest mean absolute error across diverse biological conditions. The robust performance of the model highlights the effectiveness of attention mechanisms in capturing critical patterns in gene expression data.**

***Index Terms*—Drug response prediction, Gene expression analysis, Personalized medicine**

I. Introduction

The prediction of drug responses is a critical area of research in personalized medicine, aiming to tailor treatments based on individual patient profiles. Accurate prediction models can significantly enhance the efficacy of treatments, reduce adverse effects, and ultimately improve patient outcomes [1]. With the increasing availability of genomic data, there is a growing opportunity to leverage this information to predict how patients will respond to various drugs [2]. Traditional methods often fall short in handling the complexity and variability inherent in biological systems. As a result, there is a pressing need for advanced computational approaches that can integrate and analyze high-dimensional data to provide reliable predictions. One promising avenue for achieving this is through the use of deep learning techniques, which can model the intricate relationships within the data [3].

Numerous studies have explored predicting drug responses using genomic and transcriptomic data. Early efforts relied on bulk RNA-seq, aggregating gene expression from entire tissue samples, which can mask cellular heterogeneity [4]. Recent advances have incorporated machine learning and deep learning to analyze these high-dimensional datasets, improving predictive performance [5], [6]. However, many models fail to account for the diversity of cell types and their specific responses. Single-cell RNA sequencing (scRNA-seq) has begun to address this gap, as it predicts drug responses in heterogeneous tumor samples by accounting for cellular diversity [7]. Despite these advancements, accurate predictions whether a drug will have a positive or negative effect on specific cell types remains challenging, highlighting the need for more refined computational models.

This paper presents a comparative analysis of different models for the task of drug response prediction using data from the Kaggle competition "Open Problems - Single Cell Perturbations" [8]. The dataset contains measurements from Peripheral Blood Mononuclear Cells (PBMCs) treated on 96-well plates. Each plate includes two columns dedicated to positive controls (dabrafenib and belinostat) and one column for negative control (DMSO). The positive controls are selected for their significant impact on transcription, while DMSO serves as the solvent control. The remaining wells are allocated to 72 different compounds, with the dataset covering two different compound plates per donor, totaling six plates. The task is to predict the drug response, either positive, negative, or no response of the drug molecule on given 18211 genes. This design ensures a comprehensive assessment of drug effects across multiple compounds and conditions, providing a robust foundation for evaluating the predictive accuracy of various computational models. The comparative analysis aims to identify the most effective approaches for accurate predictions of drug responses in different cell types, thereby contributing to advancements in personalized medicine.

II. Data and model description

A. Data description

The dataset used in this research is obtained from the Kaggle competition on single-cell perturbations. The dataset consists of single-cell gene expression profiles from human PBMCs treated with 144 compounds from the Library of Integrated Network-Based Cellular Signatures (LINCS) Connectivity Map dataset. These compounds were selected for their diverse transcriptional signatures, and the experiments were conducted on PBMCs from three healthy donors. Each 96-well plate used in the experiment included positive controls (dabrafenib and belinostat) and negative control (DMSO), with the remaining wells allocated to the test compounds.

979-8-3315-0694-0/25/$31.00 ©2025 IEEE

TABLE I
INPUT FEATURES SAMPLE

Cell_Type	**SM_Name**	**SM_LINCS_ID**	**SMILES**	**SMILES_Len**
NK cells	Clotrimazole	LSM-5341	Clc1ccccc1C(c1ccccc1)(c1ccccc1)n1ccnc1	38
T cells CD4+	Clotrimazole	LSM-5341	Clc1ccccc1C(c1ccccc1)(c1ccccc1)n1ccnc1	38
T cells CD8+	Clotrimazole	LSM-5341	Clc1ccccc1C(c1ccccc1)(c1ccccc1)n1ccnc1	38
T regulatory cells	Clotrimazole	LSM-5341	Clc1ccccc1C(c1ccccc1)(c1ccccc1)n1ccnc1	38
NK cells	Mometasone Furoate	LSM-3349	C[C@@H]1C[C@H]2[C@@H]3CCC4=CC(=O)C=C[C@]4(C)[C]	101

TABLE II
LABLES: FOLD CHANGE IN 18,211 GENE EXPRESSIONS RECORDED AFTER 24 HOURS OF DRUG TREATMENT

A1BG	A1BG-AS1	A2M	...	ZYX	ZZEF1
0.10472	-0.077524	-1.625596	...	0.221377	0.368755
0.915953	-0.88438	0.371834	...	1.096702	-0.869887
-0.387721	-0.305378	0.567777	...	0.078439	-0.259365
0.232893	0.129029	0.336897	...	0.216139	-0.085024
4.290652	-0.063864	-0.017443	...	-0.122193	0.676629

The PBMCs included various cell types such as T cells, B cells, NK cells, and myeloid cells, identified through single-cell RNA sequencing (scRNA-seq). The dataset parameters include cell type, compound name (sm_name), simplified molecular-input line-entry system (SMILES) representations, and time point (timepoint_hr). Gene expression data is provided as raw counts and normalized counts. Differential expression (DE) analysis was performed using the Limma model, with pseudobulked counts data and technical covariates (library ID, plate, donor) to estimate the impact of each compound on gene expression. DE values are available for all 18,211 genes included in the dataset. LFC is the estimated log-fold change in expression between the treatment and control condition where positive LFC means the gene goes up in the treatment condition relative to the control.

The input features for our model include numerical features such as Smile_len, and categorical features like cell type, sm_name, sm_lincs_ID, and SMILES representation of compounds, as illustrated in Table I. Categorical features are encoded using ordinal encoding, while numerical features are utilized in their original form. The model predicts 18,211 gene expression values, measured after 24 hours of drug treatment, serving as target labels for evaluation, as shown in Table II.

B. Models architecture

In this work, several deep learning models are compared for the task of predicting drug responses based on single-cell gene expression profiles from human PBMCs. The models under evaluation include Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), 3-stacked GRU, 3-stacked GRU with Attention (GRU-ATTN), Transformer networks, and Autoencoder architectures. Each model offers distinct advantages: LSTMs and GRUs are well-suited for sequential data processing, while stacked GRUs aim to capture deeper hierarchical representations. Transformers excel in capturing global dependencies and have shown promise in various natural language processing tasks, potentially applicable to gene expression sequences. Variation autoencoders, on the other hand, leverage latent variable modeling to capture intrinsic features and nonlinear relationships within the data.

The presented models predict gene expression levels for 18,211 genes by processing multiple input features, including cell type, drug name, drug ID, SMILES, and SMILES length. Each feature is ordinally encoded and transformed by individual embedding layers into dense vector representations of dimension 16. The embeddings, which capture meaningful relationships between the features, are concatenated into a unified representation of size (5,16) and fed into one of the models described in Table III. The output from the model is fed into a dense layer, followed by a dropout layer with a dropout rate of 0.5 to prevent overfitting. At last, the processed information is passed through an output layer of size 18,211 to generate the final prediction. The model parameters, such as learning rate, epochs, and batch size, are fine-tuned for each of the six models to optimize performance. A detailed description of all the models is provided in Table III.

III. PERFORMANCE EVALUATION

A. Performance Metrics

All the models are compared for the task of predicting the drug response on a given cell type. To assess the effectiveness of each model, the following evaluation metrics are employed.

1) Sign accuracy: It measures the ability of the models to correctly predict the direction (positive, negative, or no response) of gene expression changes. Each model predicts 18,211 gene expressions, and the accuracy is computed as Eq. (1). Correct predictions occur when the model forecasts a positive change for genes where the actual change is positive, and vice versa.

$$Sign\ accuracy = \frac{Number\ of\ correctly\ predicted\ signs}{Total\ gene\ expressions\ (18,211)} \tag{1}$$

2) Mean Absolute Error (MAE): MAE measures the average absolute difference between the predicted and actual gene expression values. Evaluating MAE for both correct and incorrect predictions helps to understand the accuracy of the models in different scenarios. MAE for correct predictions is calculated where the model correctly predicts the direction of gene expression change. Likewise, MAE for incorrect predictions is calculated where the model fails to predict the correct direction. To sum up, the total average MAE is computed across all predictions, providing an overall measure of prediction accuracy.

Furthermore, to analyze model performance across different cell types, MAE for each cell type is computed separately.

TABLE III
COMPARISON OF CONFIGURATION AND HYPERPARAMETERS OF DIFFERENT MODELS

Feature / Hyperparameter	GRU	LSTM	3-Stacked GRU	3-Stacked GRU-ATTN	Transformer	Autoencoder
Vocabulary size	1000					
Embedding dimension	16					
Embedding layers	Embedding (1000, 16)					
Concatenation output shape	(None, 5, 15)					
RNN / Transformer layers	GRU (128)	LSTM (16)	GRU (128), GRU (64)	GRU (64, with Attention)	MultiHeadAttention (2 heads)	Encoder (Dense 256, 64) Decoder (Dense 64, 256)
Latent space dimension	-	-	-	-	-	64
Dense layer (Size, Activation)	(256, tanh)	(256, tanh)	(256, tanh)	(256, tanh)	(256, tanh)	Dense (128, relu)
Dropout layer	Dropout rate (0.5)					
Output layer	Dense (18,211)					
Learning rate	0.001				0.003	0.003
Optimizer	Adam					
Epochs	300				400	500
Batch size	30	30	30	20	50	70

TABLE IV
SIGN ACCURACY OF DIFFERENT MODELS

Model	Sign Accuracy (%)
LSTM	64.5
GRU	70.6
3-Stacked GRU	73.7
3-Stacked GRU-ATTN	**79**
Transformer	69.3
Autoencoder	65.4

This analysis highlights which models are better at predicting gene expression changes in specific cellular environments. At last, the MAE for each drug is also computed to evaluate how well each model predicts gene expression changes induced by different drugs. This metric helps to identify which models perform better for specific drugs.

B. Results

The results for sign accuracy shown in Table IV indicate that the 3-Stacked GRU-ATTN model achieved the highest sign accuracy at 79%, significantly outperforming the other models. This highlights the effectiveness of attention mechanisms in enhancing model performance by allowing the model to focus on the most relevant parts of the input. The GRU (70.6%) and 3-Stacked GRU (73.7%) also performed well, suggesting that GRU-based architectures are suitable for this non-sequential data task. Conversely, the LSTM (64.5%) and Autoencoder (65.4%) had the lowest accuracies, indicating limited effectiveness for this task. The Transformer model achieved a moderate accuracy of 69%, suggesting further tuning.

Fig. 1 presents MAE across six models employed for gene expression prediction. Among these, the model with 3-stacked GRU-ATTN achieved the lowest MAE for correct predictions, indicating its superior accuracy in identifying the direction of gene expression changes. The LSTM model exhibited the lowest MAE for incorrect predictions, implying a more refined ability to discern and rectify prediction errors. These variances could be attributed to the architectural complexities of the models, such as the incorporation of attention mechanisms in the 3-stacked GRU model and the inherent memory retention of LSTM models. Such nuanced comparisons underscore the critical role of model design in optimizing predictive accuracy for genomic studies and computational biology.

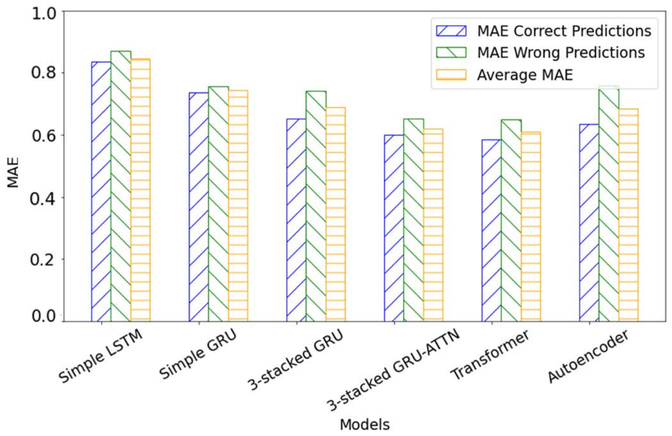

Fig. 1. MAE comparison for correct predictions, wrong predictions, and total MAE across different models.

Comparison of MAE among the top-performing models in Fig. 2 reveals varying predictive accuracies across different drugs. The 3-Stacked GRU and its attention-enhanced variant consistently demonstrate strong performance across numerous drugs, indicating robust predictive capabilities. Meanwhile, the Transformer model shows comparable performance but exhibits noticeable variations, suggesting its effectiveness depends on specific drug contexts. This analysis highlights the distinct abilities of the models in predicting drug responses, pivotal for advancing precision medicine. Similarly, evaluation of MAE across six cell types, depicted in Fig. 3 shows the 3-Stacked GRU model maintains competitive MAE values across all types, underscoring its robust predictive accuracy across diverse biological contexts.

IV. DISCUSSION AND CONCLUSION

The 3-stacked GRU with Attention model demonstrated superior performance in predicting drug-induced gene expression changes, achieving the highest sign accuracy (79%) and the lowest total MAE (0.620) among all models tested.

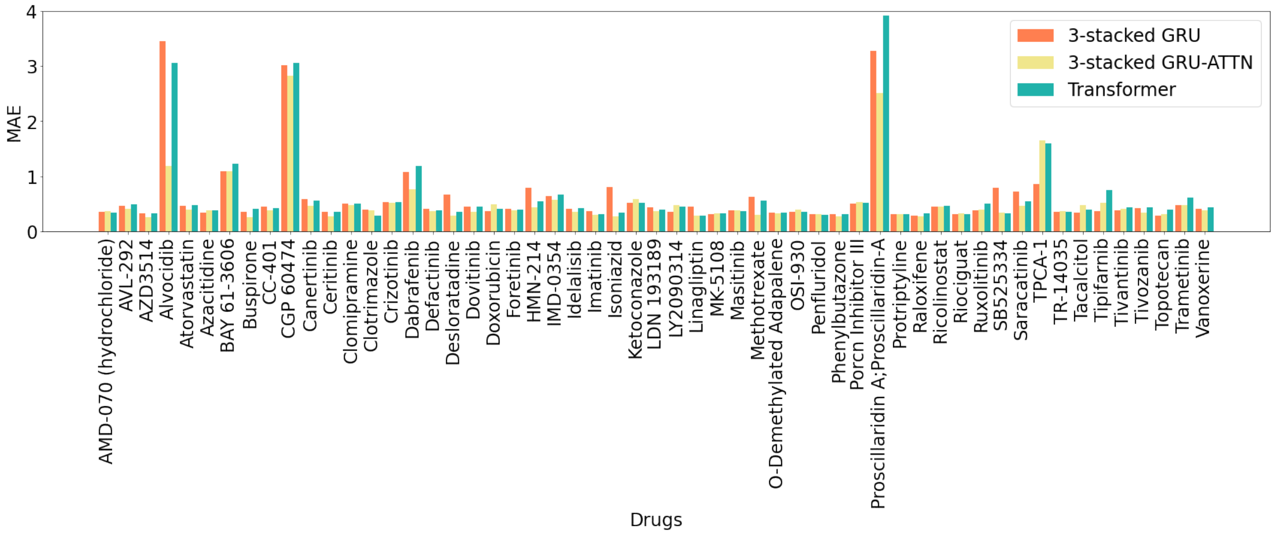

Fig. 2. Comparison of MAE with respect to correctly predicted sign in top three performing models.

This performance was consistent across various cell types, particularly NK cells and T regulatory cells, underscoring its robustness in diverse biological contexts. In comparison, the Transformer model showed moderate performance with potential for optimization, while the LSTM and Autoencoder models were less effective, reflecting their limitations in capturing complex interactions in this dataset. These findings highlight the efficacy of attention mechanisms and hierarchical representations in enhancing predictive accuracy for drug response prediction.

This work establishes a robust framework for leveraging deep learning models to predict drug responses at a cellular level, with implications for personalized medicine. By systematically evaluating multiple architectures, the work identifies the strengths and limitations of each approach, paving the way for improved precision in treatment strategies. The 3-stacked GRU-ATTN stands out as a promising model for addressing the complexities of gene expression prediction, offering a pathway to refine drug development and treatment personalization. Future research could explore optimizing underperforming models and integrating additional biological features to further enhance prediction accuracy.

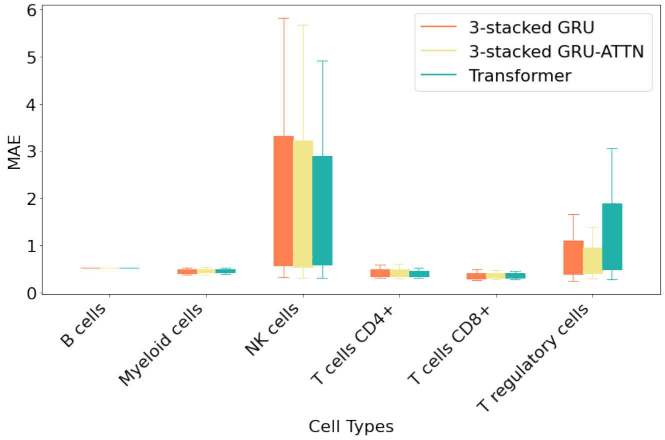

Fig. 3. Min, Max, and Average MAE for correct predictions, wrong predictions, and total MAE for different cell types.

Acknowledgments

This work was partly supported by the Korea government (MSIT), IITP, Korea, under the ICT Creative Consilience program (RS-2020-II201821, 40%), AI Innovation Hub (RS-2021-II212068, 30%), and AI Graduate School Program (Sungkyunkwan University, RS-2019-II190421, 30%).

References

[1] Y. Chen and L. Zhang, "How much can deep learning improve prediction of the responses to drugs in cancer cell lines?" *Briefings in bioinformatics*, vol. 23, no. 1, p. bbab378, 2022.

[2] K. B. Johnson, W.-Q. Wei, D. Weeraratne, M. E. Frisse, K. Misulis, K. Rhee, J. Zhao, and J. L. Snowdon, "Precision medicine, ai, and the future of personalized health care," *Clinical and translational science*, vol. 14, no. 1, pp. 86–93, 2021.

[3] B. M. Kuenzi, J. Park, S. H. Fong, K. S. Sanchez, J. Lee, J. F. Kreisberg, J. Ma, and T. Ideker, "Predicting drug response and synergy using a deep learning model of human cancer cells," *Cancer cell*, vol. 38, no. 5, pp. 672–684, 2020.

[4] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin *et al.*, "The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, no. 7391, pp. 603–607, 2012.

[5] M. P. Menden, F. Iorio, M. Garnett, U. McDermott, C. H. Benes, P. J. Ballester, and J. Saez-Rodriguez, "Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties," *PLoS one*, vol. 8, no. 4, p. e61318, 2013.

[6] Y. Chang, H. Park, H.-J. Yang, S. Lee, K.-Y. Lee, T. S. Kim, J. Jung, and J.-M. Shin, "Cancer drug response profile scan (cdrscan): a deep learning model that predicts drug effectiveness from cancer genomic signature," *Scientific reports*, vol. 8, no. 1, p. 8857, 2018.

[7] A. F. Aissa, A. B. Islam, M. M. Ariss, C. C. Go, A. E. Rader, R. D. Conrardy, A. M. Gajda, C. Rubio-Perez, K. Valyi-Nagy, M. Pasquinelli *et al.*, "Single-cell transcriptional changes associated with drug tolerance and response to combination therapies in cancer," *Nature communications*, vol. 12, no. 1, p. 1628, 2021.

[8] Kaggle, "Open problems - single cell perturbations," 2024, accessed: 2024-07-08. [Online]. Available: https://www.kaggle.com/competitions/open-problems-single-cell-perturbations/overview