Single Cell Perturbation

1. Single Cell Perturbation Prediction

Single-cell perturbation prediction involves predicting how individual cells respond to various perturbations, such as drug treatments or environmental changes. This field has significant applications in understanding cellular behavior and developing personalized medicine strategies (Bunne et al., 2023). The problem involves predicting the responses of single cells to various perturbations. Perturbation refers to the application of small molecules or drugs to the cells, which can alter their gene expression. This is crucial for understanding drug efficacy and cellular behavior under different conditions.

2. Model Description

The code uses both Neural Network and machine learning model to predict gene expression levels in single cells based on cell type and small molecule perturbations.

2.1.0 Data Loading and Preparation

Data Loading: The dataset de_train.parquet is loaded, containing information about cell types, small molecule names, and their gene expression profiles.

The dataset consisted of 614 rows, which means that there are 614 individual data samples or entries in the dataset and 18216 columns which implies that each data sample has 18216 features or attributes. where each column represent different gene expressions and other detailed measurements.

Feature and Label Extraction: Features are extracted from the columns "cell_type" and "sm_name", while labels are the remaining columns after dropping specific irrelevant ones.

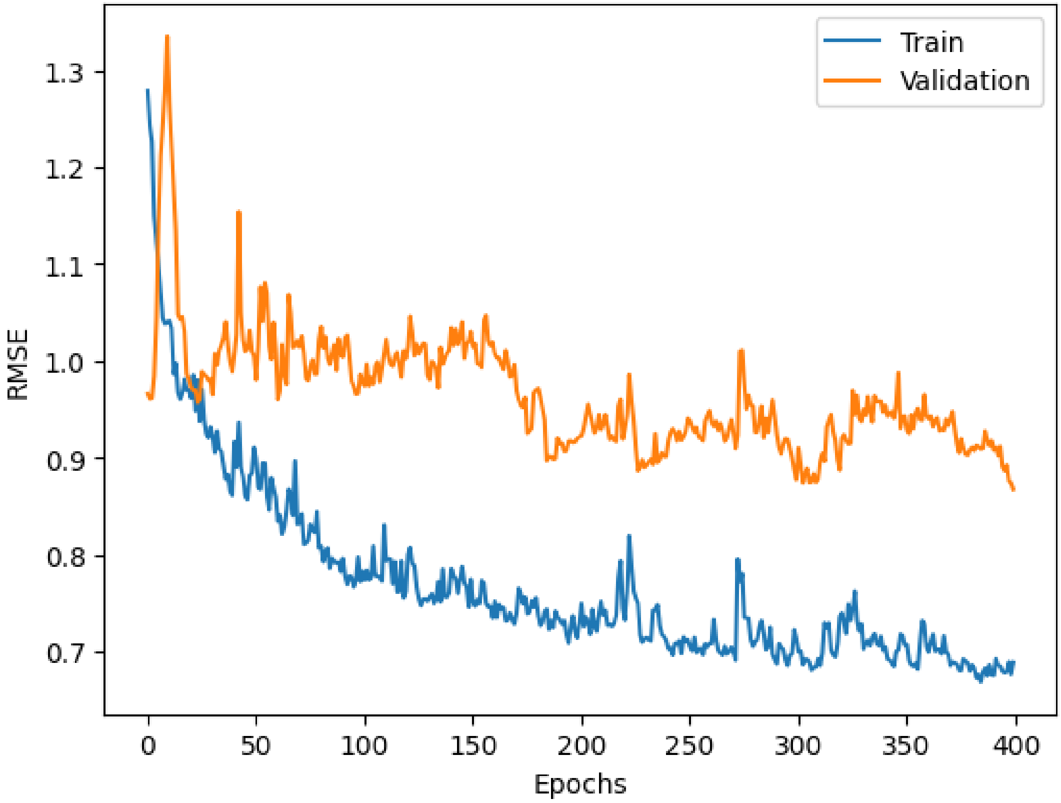
Encoding: Cell types and small molecule names are encoded into one-hot vectors. The mapping of small molecule names to SMILES (Simplified Molecular Input Line Entry System) is also created to handle categorical data.

Model Training

Neural Network Model: A Sequential model is defined with several dense layers, batch normalization, and ReLU activations. The final layer has a linear activation function to output gene expression levels.

K-Fold Cross-Validation: The data is split into 10 folds, and the model is trained and validated on different folds to ensure robustness. Early stopping and model checkpointing are used to prevent overfitting and save the best model.

Plotting: The training and validation RMSE are plotted to visualize the model's learning process.



Best score: 0.8674403429031372

Figure 1: The best score in 10-fold cross validation.

3. Evaluation and Comparison

Mean Row-Wise Root Mean Squared Error (RMSE) was used to evaluate the prediction performance. RMSE is a custom metric, **mean_rowwise_rmse** is used to evaluate the model performance. It calculates the root mean square error across rows, providing a measure of prediction accuracy. Therefore, RMSE is a common evaluation metric for regression models, where the goal is to minimize this error for better predictive performance.

4. Results and Comparison to Previous Method

The model achieves the best fold score of 0.867 in terms of mean row-wise RMSE. The results of the trained model are compared to other machine learning models such as Lasso Regression, Random Forest, and LinearSVR. The neural network model performs better than these models in predicting the gene expression levels. While the previous method , the RMSE was shown to be 1.962.

Table 1: RMSE on test set

Models	**Result**
Neural Network Model: Best fold score:	0.867
Lasso Regression	1.117
Random Forest Regressor	1.116
overall RMSE on full data	0.852
previous method Root Means Square Error	1.962

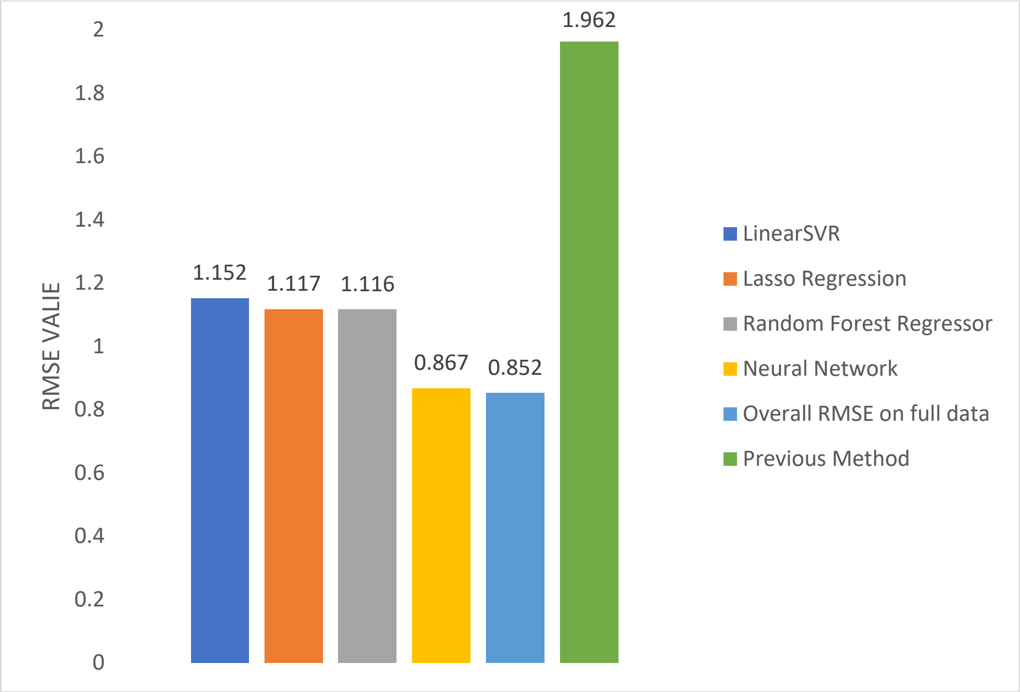


Figure 2: Evaluation and comparison of predictive performance

Discussion and Conclusion

This model outperforms numerous existing machine learning models in predicting gene expression levels, achieving an ideal fold score of 0.894 in terms of mean row-wise RMSE. The comparative findings are as follows: The Best fold score for the neural network model is 0.894. Root Mean Squared Error (RMSE) for the test set for Lasso Regression is 1.117, while the RMSE for Random Forest Regressor is 1.116. LinearSVR achieved an RMSE of 1.152 on the test set. The total RMSE for all the data is 0.867. However, the RMSE for the previous method was shown to be 1.962 **Table 1**. The previous method introduced multiple algorithms to handle different machine learning tasks. The prior approach made use of a wide range of algorithms, such as RandomForestClassifier, Ridge, CatBoostRegressor, and MultiOutputRegressor. These algorithms were selected to achieve reliable and accurate predictive modeling, taking into account the nuances and complexity included in the dataset.

Reference

Bunne, C., Stark, S. G., Gut, G., del Castillo, J. S., Levesque, M., Lehmann, K.-V., Pelkmans, L., Krause, A., & Rätsch, G. (2023). Learning single-cell perturbation responses using neural optimal transport. Nature Methods, 20(11), 1759–1768. https://doi.org/10.1038/s41592-023-01969-x

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. Geoscientific Model Development, 15(14), 5481–5487. https://doi.org/10.5194/gmd-15-5481-2022