# TransCDR: a deep learning model for enhancing the generalizability of drug activity prediction through transfer learning and multimodal data fusion.

## Background and Motivation

Precise and reliable drug response prediction is crucial in precision medicine. While numerous models have been designed to leverage drug and cancer cell line representations for predicting cancer drug responses (CDR), their effectiveness can be enhanced by overcoming challenges such as limited data modalities, suboptimal fusion strategies, and poor generalizability to novel drugs or cell lines.

## Objectives

Our plan is to reproduce the TransCDR model to the best of our ability and experiment on datasets supplied by the authors of the paper, including GDSC, CCLE, TCGA, and MSigDB. We plan to implement and observe the interaction between the fusion of multi-modality features of drugs and cell lines to test the accuracy of drug response prediction. Specifically, drug/cell-line attention fusion modules will be tested. We will reproduce empirical evaluation metrics illustrated in the paper. We plan to study the how graph embeddings of drugs can improve the performance of the model with respect to IC50 values or sensitive states of drugs on cell lines. The role of Gene Set Enrichment Analysis will be studied. We will scrutinize the authors' methodology for quantifying the contribution of the most impactful drug notations and omics profiles on predictive performance. Due to the scale of the research paper and its implementation complexity, we will only cover the most important aspects of its capabilities.

## Methodology

### Data Acquisition

We will use the same datasets as in this article, GDSC (Genomics of Drug Sensitivity in Cancer) and CCLE (Cancer Cell Line Encyclopedia), TCGA (The Cancer Genome Atlas Program), and MsigDB (Molecular Signatures Database).

### Tools and Technologies

We plan to use primarily PyTorch to restore the code infrastructure. We will follow the Github documentation to perform data extraction and data integration with the model.

### Reproduction Plan

First, we plan to collect data and preprocess the data, extracting useful information. Then, we start model training and download necessary preprocessed models. Here we will try the same parameters as the ones selected by the paper. Our reproduction pipeline will largely follow the instructions outlined in the GitHub page: CV10 for TransCDR, Training the final TransCDR, Test the trained TransCDR on CCLE, Screening Drugs for TCGA Patients, Predicting CDRs of a drug. Once the training and testing processes are complete, we will start to analyze and verify the performance metrics of the model, including RMSE, R2, etc. Perform feature analysis to detect the impact of the features of various data sources on prediction results. We will try to change some parameters and try to obtain external test sets to see if there is any deviation from the experimental results of the paper. If there are deviations, we will try to investigate further and arrive at a plausible explanation.

## Expected Outcomes

We consider that if we restore the experiment successfully, the results given should be similar to the paper. If there are significant deviations, we will study implicated parameters such as computation resources, preprocessing methods and data differences to rationalize the differences. After that, we will experiment with other factors that may affect the data and test on external datasets.

## Tentative Timeline

- By Feb 23: Finish literature review and analysis of the paper. Prepare to present.
- By March 9: Finish studying the data encodings. Set up coding environment.
- By March 23: Finish following the GitHub instructions and start tuning parameters.
- By March 30: Finish the generation of charts, metrics, and code refactoring.
- By April 6: Arrive at the final report before submission.

## References

Xia, X., Zhu, C., Zhong, F. & Liu, L. (2024). TransCDR: a deep learning model for enhancing the generalizability of drug activity prediction through transfer learning and multimodal data fusion. BMC Biology, 22(1), 227. https://doi.org/10.1186/s12915-024-02023-8

PyTorch. (n.d.). *PyTorch: An open-source machine learning framework*. https://pytorch.org/

Wellcome Sanger Institute. (n.d.). *CancerRxGene: Genomics of drug sensitivity in cancer*. https://www.cancerrxgene.org/

Broad Institute. (n.d.). *Cancer Cell Line Encyclopedia (CCLE)*. https://sites.broadinstitute.org/ccle/

National Cancer Institute. (n.d.). *The Cancer Genome Atlas (TCGA)*. https://www.cancer.gov/ccg/research/genome-sequencing/tcga

Broad Institute & UC San Diego. (n.d.). *Molecular Signatures Database (MSigDB)*. https://www.gsea-msigdb.org/gsea/msigdb/