CSC4046

Individual Research Project

Interpretable Deep Learning for Robust Drug Mechanism Prediction in High-Throughput Gene Expression Data

Registration No.: SC/2020/11730

Student Name: Navinda Hewawickrama

Supervisor(s): Dr.Sugandima Vidanagamachchi

October 2025

Bachelor of Computer Science (Special) Degree
Department of Computer Science, University of Ruhuna

1 Introduction

1.1 Introduction

Data-driven mechanisms of action (MoA) prediction is made possible by the large-scale, huge drug-cell response profiles provided by high-throughput perturbational transcriptomics, especially the LINCS L1000 collection. However, in practice, these datasets are difficult to use consistently because of a combination of inherent technical/inferential noise (most landmark genes are inferred, while only 978 are directly measured), batch and assay artefacts, and frequent gaps or inconsistencies in metadata. These factors collectively weaken model generalisation and reduce reproducibility.However, integrating transcriptome signals with external data like drug chemical structure, protein-protein interaction networks, and Gene Ontology hierarchies is both beneficial and challenging, requiring careful representation learning and cross-modal alignment[12, 11, 6, 14]

In order to address these problems, this project (i) thoroughly examines previous and present computational pipelines for L1000-based MoA prediction in order identify failure modes and best practices (ii) In order to improve interpretability and generalisation, a modular, reproducible framework that simplifies data-centric cleaning and metadata consistency, integrates robust/model-centric training procedures, and incorporates external biological knowledge is proposed. Explainable AI is a key focus, in addition to attaining high accuracy, we want to make model reasoning auditable by connecting predictions to pharmacological substructures, genes, and pathways, making the resulting MoA hypotheses easier to evaluate and scientifically significant[2, 10, 9]

1.2 Problem Definition

The main obstacle to using the LINCS L1000 dataset for drug development is that, although many computer models are capable of achieving high predicted accuracy, their interpretability and robustness are frequently lacking, which restricts their application in clinical settings. There is a need for a comprehensive solution because previous research has frequently addressed these issues separately.
These are our original research questions:

- **R1:** What techniques have been used to predict drug mechanisms of action from LINCS gene expression data, and how do they balance accuracy, scalability, and interpretability, in high-dimensional biomedical contexts?

- **R2:** How do data quality issues (e.g., noise, inconsistencies) affect the performance and generalizability of predictive models, and what computational methods have been used to address these challenges?

- **R3:** Are there frameworks that effectively integrate external knowledge (e.g., drug structures, biological networks) with transcriptomic data, and how do they impact model interpretability, generalizability, and explainability?

R1 compares techniques under accuracy,scalability,interpretability trade-offs. R2 quantifies the impact of data quality and evaluates mitigation strategies. R3 studies how integrating external knowledge improves interpretability and generalization.
These problems motivate our research questions:

1. The Robustness Problem (R2/R1): Standard machine learning models are highly sensitive to the technical noise, batch effects, and data inconsistencies inherent in the L1000 platform. This leads to models that do not generalize well to new data, making their predictions unreliable. Models, classical and deep, are sensitive to structured technical and annotation noise in L1000 (e.g., deconvolution errors, inferred genes, batch effects, metadata gaps), which degrades out-of-distribution (OOD) generalization and reliability.

2. The Context Problem (R3/R1): Models relying only on gene expression data lack biological context. Their predictions are often correlational rather than mechanistic, failing to explain why a drug has a certain effect based on underlying biological pathways.Expression-only models capture correlations but lack mechanistic context. Without integrating drug structure and biological networks (PPIs, Gene Ontology), predictions do not explain *why* an effect occurs.

3. The Trust Problem (R1/R3): The increasing complexity of deep learning models results in "black boxes".For clinicians and biologists, a prediction is not actionable if its underlying reasoning cannot be understood, audited, and trusted. State-of-the-art models are often black boxes. Actionable use requires understandable explanations that are faithful to the model, biologically grounded (pathways/targets), and reproducible.

Thus, the main issue this study attempts to solve is the absence of a single computational framework that simultaneously guarantees resilience against data noise, incorporates outside biological knowledge for mechanistic background, and makes its predictions easily interpretable.

1.3 Objective of the Research

The primary objective of this research is to design, develop, and evaluate a novel, end-to-end computational framework that significantly improves the robustness, biological relevance, and interpretability of drug Mechanism of Action (MoA) prediction from LINCS L1000 data.To achieve this overarching goal, the research is guided by the following specific objectives (derived from the corresponding research questions R1, R2, and R3).

Objective 1: To Define Optimal Performance and Interpretability Trade-offs (R1) First, to conduct a systematic literature review of existing methods,including classic machine learning (ML), Multi-Layer Perceptrons (MLPs), Graph Neural Networks (GNNs), and attention/Transformer architectures, to establish a performance baseline, critically analyze their strengths and weaknesses, and identify best practices for working with LINCS L1000 data. This objective focuses on balancing accuracy, scalability, and interpretability within high-dimensional biomedical contexts, providing a robust performance baseline for the final framework.

Objective 2: To Enhance Robustness and Mitigate Data Quality Issues (R2) To design a data preprocessing pipeline that reduces and mitigates noise and corrects inconsistencies, such as technical/inferential noise, batch effects, and metadata inconsistencies, in the L1000 dataset. This involves integrating external knowledge from biological databases (e.g., protein-protein interaction networks) using graph-based methods to enrich the gene expression features and fill metadata gaps.

Objective 3: To Ground Predictions in Mechanistic Biological Context (R3) To overcome the "Context Problem", we aim to design and implement knowledge-infused deep learning architectures that integrate external biological knowledge. This involves

leveraging graph-based methods to fuse multi-modal data, such as drug chemical structure, Protein-Protein Interaction (PPI) networks, and Gene Ontology (GO) hierarchies, thereby enhancing mechanistic interpretability and generalizability.
Objective 4: To Deliver Transparent and Auditable Explanations (XAI) To integrate Explainable AI (XAI) techniques into the framework. The goal is not only to achieve high accuracy but also to make the model's predictions transparent, allowing researchers to understand the underlying biological drivers (e.g., key genes, pathways) behind its decisions.

1.4 Scope of the Research

In particular, the LINCS L1000 dataset analysis is the main focus of this study. There are more transcriptome datasets, but they are considered outside the purview of this project's main analysis and framework creation. The main objective is to anticipate the Mechanism of Action (MoA) of a medicine. External knowledge will only be incorporated into the suggested computational framework from publically accessible biological resources, such as the Gene Ontology (GO) hierarchy and the STRING database for protein-protein interactions.
The primary deliverables for this project are:

- A comprehensive Systematic Literature Review (SLR) manuscript that establishes a baseline of current methods.
- A novel computational framework that integrates data preprocessing, knowledge infusion, and a predictive deep learning model.
- A trained and evaluated deep learning model for MoA prediction, complete with interpretability analyses.
- A final research thesis detailing the project's methodology, results, and conclusions.

Some key assumptions are made when operating the project.
(i) that the biological annotations from external databases are sufficiently accurate to provide meaningful context
(ii) that a learnable signal correlating to drug MoA exists within the noisy L1000 data and
(iii) that the 978 landmark genes serve as an effective proxy for the full transcriptome
The research is bound by several constraints, including the computational resources (GPU/CPU time) available for training complex models, and the inherent limitations in the "ground truth" MoA labels, which can be incomplete or ambiguous.

2 Literature Review

This review looks at methods for predicting drug mechanisms of action (MoA) from LINCS/CMap (L1000) transcriptomic data, organized along three themes that showcase our research questions:
(R1) Techniques and the accuracy,scalability,interpretability trade-off
(R2) Data-quality issues and mitigation; and
(R3) Integration of external biological knowledge for contextual, auditable predictions.

2.1 Early Approaches: Signature Similarity & Classical ML (R1)

According to the Connectivity Map concept, medications with comparable gene-expression profiles have similar molecular patterns. Large retrieval operations were made possible by the availability of scalable profiles (MODZ-aggregated signatures) by the next-generation CMap/L1000[12]. To improve signal quality, Characteristic Direction (CD) signatures replaced naive fold-change or t-statistics, yielding more accurate reversal/connection performance[2]. Traditional machine learning (ML) baselines (RF, SVM, regularised logistic/XGBoost) became competitive as datasets increased. Optimised pipelines like WRFEN, XGBoost demonstrate that resilient learners and good feature engineering may compete with deeper models on certain specific tasks[5].
Strengths: simple, fast, interpretable (feature importance); scalable screening. **Weaknesses:** Has limitations in nonlinearity, struggle with high-dimensional patterns and cross-context generalization.

2.2 Deep Learning: MLPs, GNNs, Attention/Transformers (R1)

Benchmarks show multilayer perceptrons (MLPs) typically perform better than classical baselines on L1000 tasks, but transparency is the cost of that performance[6]. Knowledge-aware Graph CNNs (GNNs) constrain learning to biological topology (e.g., PPI), promoting pathway-aligned features—promising on large data, but sometimes outperformed by MLPs on smaller collections[6]. Recent work shows and uses attention/Transformers for cross-context modeling. DeepCE aligns drug substructures with gene effects, improving accuracy and offering substructure, gene attributions[7], while MultiDCP uses knowledge-aware Transformers to generalize across cell types and even outperform noisy experimental baselines in some settings[14].
Strengths: higher accuracy, flexible representation learning, phas the otential of cross-context generalization. **Weaknesses:** opacity (“black boxes”), data-hungry, sensitive to dataset noise if not handled explicitly.

2.3 Data Quality in L1000 & Mitigation (R2)

L1000 measures 978 landmark genes and infers the remainder computationally, introducing inferential noise alongside batch/assay artefacts and metadata inconsistencies[12, 6]. Peak-calling/deconvolution instabilities in the original pipeline have motivated GMM/AGMM and Bayesian deconvolution, which reduce errors[8]. Replicate-correlation filtering and better signature generation (e.g., CD) increase signal-to-noise; model-guided augmentation can salvage high-quality replicates otherwise flagged as noisy[13, 2, 7]. On the model side, robust losses, co-teaching/curriculum learning, and semi/self-supervised pretraining improve resilience to label and feature noise.
Strengths: measurable SNR gains; improved reproducibility and downstream accuracy.
Weaknesses: heterogeneous, structured noise remains challenging; label/metadata gaps are non-trivial to repair at scale.

2.4 Knowledge Integration for Context & Interpretability (R3)

Expression alone provides correlations, not mechanisms. Integrating drug structure and biological networks (PPIs, Gene Ontology) adds context[4]. Multimodal fusion strategies (early/intermediate/late) show that intermediate or attention-based fusion often outperforms naive concatenation[10]. Cross-attention aligns modalities, while knowledge-infused architectures (e.g., GNNs on PPI, GO-hierarchy layers/DrugCell-style designs) embed biological priors so that internal activations have pathway-level meaning[3, 6, 9]. Surveys frame this direction as essential to balance accuracy with actionable transparency[1]. **Strengths:** improved accuracy and out-of-distribution robustness; pathway/target-grounded explanations ("glass-box"). **Weaknesses:** integration pipelines are complex; depend on quality/coverage of external knowledge bases.

2.5 Comparative Summary

Table 1: Methods landscape for L1000-based MoA prediction: typical setup, strengths, and weaknesses.

Family	**Typical setup**	**Strengths**	**Weaknesses**
Signature similarity (CMap/CD)	MODZ/CD signatures; retrieval/reversal scoring [12, 2]	Simple, scalable; quick screening	Correlational; limited mechanism; sensitivity to noise
Classical ML (RF/SVM/XGBoost)	Engineered features from L1000; optimized pipelines [5]	Fast; decent accuracy; interpretable features	Limited nonlinearity; weaker cross-context performance
MLPs (deep baselines)	Expression-only deep encoders [6]	Strong accuracy; flexible	Black-box; noise-sensitive without robust training
GNNs / GO-structured nets	PPI/GO priors; message passing; intrinsic interpretability [6, 9]	Mechanism grounding; pathway-level attributions	Data/knowledge hungry; integration complexity
Attention/ Transformers / multi-modal fusion	Cross-attention; expression+structure; knowledge-aware encoders [7, 14, 10, 3]	SOTA accuracy; better OOD; substructure–gene alignment	Computationally heavy; explanation validation required

2.6 Gaps and Opportunities

(i) **Quantitative validation of explanations** remains limited—many works rely on anecdotal literature matches rather than faithfulness tests and perturbation-based validation. (ii) **Generalization to unseen contexts** (cell types, doses, time points) still lags, especially under structured, real-world noise. (iii) **Seamless knowledge ingestion** (ID harmonization; stable fusion of structure/PPIs/GO) requires standardized, reproducible pipelines. These gaps motivate our objectives: standardized data/metadata handling (R2), knowledge-infused architectures (R3), and systematic baselines with robustness and interpretability evaluations (R1–R3).

3 Methodology

This research adopts the **Design Science Research Methodology (DSRM)**, a well-established framework for creating and evaluating novel artifacts in information systems and computer science. The primary goal of DSRM is to develop solutions to real-world problems through the design, implementation, and evaluation of a new technological artifact. In the context of this project, the **artifact** is the proposed computational framework for robust and interpretable Mechanism of Action (MoA) prediction.

The DSRM process is iterative and follows a set of logical steps, which are mapped to this research as follows:

- **1. Problem Identification and Motivation:** This initial step involves defining the research problem and justifying the value of a solution. This was accomplished through the comprehensive literature review (Chapter 2), which identified the key challenges in the current state-of-the-art: a lack of robustness to data noise, limited biological context, and poor model interpretability, which together hinder the clinical translation of existing models.

- **2. Define the Objectives for a Solution:** Based on the identified problems, this step defines the goals for the new artifact. As outlined in Section 1.3, the objectives for the proposed framework are to achieve high predictive accuracy while simultaneously ensuring robustness, integrating external biological knowledge, and providing transparent, interpretable predictions.

- **3. Design and Development:** This is the core construction phase where the artifact is created. This research will involve the design and implementation of a multi-stage software pipeline in Python. Key development activities will include:

  - Implementing a data-centric preprocessing module to normalize data and mitigate batch effects.
  - Constructing a knowledge graph from external databases (e.g., STRING, Gene Ontology).
  - Building a novel deep learning architecture that integrates the knowledge graph and transcriptomic data.
  - Integrating Explainable AI (XAI) algorithms into the output layer of the model.

- **4. Demonstration:** In this step, the artifact is used to solve the problem. The developed framework will be applied to the LINCS L1000 dataset to predict the MoA for a curated set of chemical compounds, demonstrating its end-to-end functionality.

- **5. Evaluation:** This step involves measuring how well the artifact achieves its objectives. The framework will be rigorously evaluated using a two-pronged approach:

  - **Quantitative Evaluation:** The model's predictive performance (e.g., F1-score, accuracy) will be measured and compared against the baseline models identified in the literature review.

  - **Qualitative Evaluation:** The XAI outputs will be used to conduct case studies on specific drug predictions. The goal is to demonstrate that the model's explanations are biologically plausible and provide novel insights that align with known pharmacology.

- **6. Communication:** The final step is to communicate the research to a wider audience. The findings, the framework's design, and its utility will be formally documented and communicated through the final research thesis and potentially a journal or conference publication.

4 Progress

4.1 Current Progress

Significant progress has been made in both the theoretical and practical phases of this research project. The work completed to date has established a strong foundation for the development of the novel computational framework.

- **Systematic Literature Review (SLR) Completed:** A comprehensive SLR has been successfully completed, and a full manuscript detailing the findings has been drafted. This review systematically analyzed 26 primary research articles, establishing a performance baseline for existing models, identifying key data-related challenges, and mapping the current landscape of interpretability techniques. The insights from this review directly inform the design and objectives of the proposed framework.

- **Framework Development Initiated:** The foundational work for the software implementation has commenced. This preparatory phase is critical for ensuring a modular, reproducible, and well-documented project. Key achievements include:
  - **Technology Stack Selection:** The core technologies have been selected, including Python as the primary language, PyTorch for deep learning, PyG (PyTorch Geometric) for graph neural networks, and RDKit for handling chemical structures.
  - **Project Scaffolding:** A modular codebase structure has been designed and created. This separates concerns for data preprocessing, model architecture, training, and evaluation, facilitating organized development.
  - **Version Control and Experiment Tracking:** A Git repository has been established for robust version control. Furthermore, a strategy for documenting and maintaining different baseline models from the literature has been implemented to ensure reproducible and fair comparison of results.

4.2 Future Time Plan

The remainder of the research will focus on the development, evaluation, and documentation of the proposed framework. The following is a projected timeline to guide the project toward completion.

- **Phase 1: Data Pipeline and Feature Engineering (3 Months: Nov 2025 - Jan 2026)**

  - Implement the full data ingestion and cleaning pipeline for the LINCS L1000 dataset.
  - Develop the module for constructing and integrating the biological knowledge graph.
  - Generate and validate the final, enriched feature sets for the modeling phase.

- **Phase 2: Model Development and Training (4 Months: Feb 2026 - May 2026)**
  - Implement and train the baseline models identified in the SLR for direct comparison.
  - Design and implement the novel knowledge-infused deep learning architecture.
  - Conduct extensive training experiments, including hyperparameter tuning and optimization.

- **Phase 3: Evaluation and Analysis (2 Months: June 2026 - July 2026)**
  - Perform a rigorous quantitative evaluation of the final model against the baselines.
  - Conduct a qualitative evaluation using the integrated XAI techniques to produce and analyze case studies for specific drug predictions.
  - Synthesize all results and draw final conclusions.

- **Phase 4: Thesis Finalization (3 Months: Aug 2026 - Oct 2026)**
  - Write the remaining chapters of the final research thesis (Methodology, Results, Discussion).
  - Revise, proofread, and format the complete manuscript.
  - Prepare for final submission and presentation of the research.

5 Summary

This research project addresses critical limitations in the computational analysis of the LINCS L1000 dataset for drug discovery. While existing models can predict a drug's Mechanism of Action (MoA), they often lack robustness against data noise and are too opaque to be trusted in a clinical context. The primary contribution of this work is the design and development of a unified, end-to-end computational framework that overcomes these issues. The proposed framework is built on three pillars: (1) a robust data preprocessing pipeline to mitigate technical noise, (2) the integration of external biological knowledge graphs to provide mechanistic context, and (3) the implementation of Explainable AI (XAI) to ensure that all predictions are transparent and interpretable. The ultimate goal is to produce a more reliable and scientifically valuable tool for accelerating drug discovery.

The project is currently at a key transition point. The foundational phase, a comprehensive Systematic Literature Review, has been completed, and a full manuscript has been drafted. This review has successfully established performance baselines and identified the specific gaps in the literature that the proposed framework will address. Based on these findings, the initial technical work for the framework has commenced, including the selection of the technology stack, the creation of a modular codebase, and the implementation of version control. The project is now moving from the literature analysis phase into the core development and implementation of the novel framework, as detailed in the future time plan.

References

[1] Yuen Ler Chow, Shantanu Singh, Anne E Carpenter, and Gregory P Way. Predicting drug polypharmacology from cell morphology readouts using variational autoencoder latent space arithmetic. *PLoS computational biology*, 18(2):e1009888, 2022.

[2] Qiaonan Duan, St Patrick Reid, Neil R Clark, Zichen Wang, Nicolas F Fernandez, Andrew D Rouillard, Ben Readhead, Sarah R Tritsch, Rachel Hodos, Marc Hafner, et al. L1000cds2: Lincs l1000 characteristic direction signatures search engine. *NPJ systems biology and applications*, 2(1):1–12, 2016.

[3] Mazen Elabd and Sardar Jaf. A simple attention-based mechanism for bimodal emotion classification. *arXiv preprint arXiv:2407.00134*, 2024.

[4] Chenglin Liu, Jing Su, Fei Yang, Kun Wei, Jinwen Ma, and Xiaobo Zhou. Compound signature detection on lincs l1000 big data. *Molecular BioSystems*, 11(3):714–722, 2015.

[5] Jiaxing Lu, Ming Chen, and Yufang Qin. Drug-induced cell viability prediction from lincs-l1000 through wrfen-xgboost algorithm. *BMC bioinformatics*, 22(1):13, 2021.

[6] Matthew BA McDermott, Jennifer Wang, Wen-Ning Zhao, Steven D Sheridan, Peter Szolovits, Isaac Kohane, Stephen J Haggarty, and Roy H Perlis. Deep learning benchmarks on l1000 gene expression data. *IEEE/ACM transactions on computational biology and bioinformatics*, 17(6):1846–1857, 2019.

[7] Thai-Hoang Pham, Yue Qiu, Jucheng Zeng, Lei Xie, and Ping Zhang. A deep learning framework for high-throughput mechanism-driven phenotype compound screening and its application to covid-19 drug repurposing. *Nature machine intelligence*, 3(3):247–257, 2021.

[8] Yue Qiu, Tianhuan Lu, Hansaim Lim, and Lei Xie. A bayesian approach to accurate and robust signature detection on lincs l1000 data. *Bioinformatics*, 36(9):2787–2795, 2020.

[9] Bikash Ranjan Samal, Jens Uwe Loers, Vanessa Vermeirssen, and Katleen De Preter. Opportunities and challenges in interpretable deep learning for drug sensitivity prediction of cancer cells. *Frontiers in Bioinformatics*, 2:1036963, 2022.

[10] Sören Richard Stahlschmidt, Benjamin Ulfenborg, and Jane Synnergren. Multimodal deep learning for biomedical data fusion: a review. *Briefings in bioinformatics*, 23(2):bbab569, 2022.

[11] Vasileios Stathias, Amar Koleti, Dušica Vidović, Daniel J Cooper, Kathleen M Jagodnik, Raymond Terryn, Michele Forlin, Caty Chung, Denis Torre, Nagi Ayad,

et al. Sustainable data and metadata management at the bd2k-lincs data coordination and integration center. *Scientific data*, 5(1):1–14, 2018.

[12] Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452, 2017.

[13] Bence Szalai, Vigneshwari Subramanian, Christian H Holland, Róbert Alföldi, László G Puskás, and Julio Saez-Rodriguez. Signatures of cell death and proliferation in perturbation transcriptomics data—from confounding factor to effective prediction. *Nucleic Acids Research*, 47(19):10010–10026, 2019.

[14] You Wu, Qiao Liu, Yue Qiu, and Lei Xie. Deep learning prediction of chemical-induced dose-dependent and context-specific multiplex phenotype responses and its application to personalized alzheimer's disease drug repurposing. *PLoS computational biology*, 18(8):e1010367, 2022.