UNVEILING THE TRUE IMPACT OF DRUG AND CELL LINE REPRESENTATIONS IN DRUG SYNERGY PREDICTION

by
EMINE BEYZA ÇANDIR SOYDEMIR

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfilment of
the requirements for the degree of Master of Science

Sabancı University
December 2024

ABSTRACT

UNVEILING THE TRUE IMPACT OF DRUG AND CELL LINE REPRESENTATIONS IN DRUG SYNERGY PREDICTION

EMINE BEYZA ÇANDIR SOYDEMIR

COMPUTER SCIENCE AND ENGINEERING M.Sc. THESIS, DECEMBER 2024

Thesis Supervisor: Assoc. Prof. ÖZNUR TAŞTAN

Keywords: Drug Synergy, Deep Learning, Generalization, One-Hot-Enoding

Drug combination therapy holds promise as an effective strategy for treating complex diseases such as cancer. However, due to the vast combinatorial space of drug combinations, experimental screening of all of them is not feasible. Computational models have been developed to prioritize drug pairs that could work synergistically to accelerate experimental screening efforts. These models are trained on large datasets of previously reported drug combination measurements and use rich representations of drugs and cell lines that encode chemical, structural, and biological properties.

In this thesis, we first aimed to improve upon our previous synergy predictor, MatchMaker, by incorporating richer biological information such as pathways and mechanism of action or alternative drug representations. Despite all our efforts, none of the models could perform better. Motivated by these findings, we tested a more straightforward approach by replacing detailed feature representations with one-hot encodings of drugs and cell lines. Surprisingly, these models stripped of chemical and biological information can come very close to the results trained with rich biological and chemical information.

Here, in this thesis, we systematically experimented with published synergy prediction models by replacing drug representations and cell line features with a simple one-hot encoding of drugs and cell lines in various evaluation settings. Regardless of the drug input feature or the architecture, we observe that the simple one-hot encoding baseline performs similarly in all models. This unexpected result suggests

that the representations serve as simple identifiers and models that capture general co-variation patterns of synergy measurements rather than learning chemical or biological information. This could be why the models do not generalize well to new drugs and cell lines. While synergy prediction models are still beneficial in deciding on what pairs to test within a panel of drugs and cell lines, these results demonstrate that alternative approaches are needed for developing synergy prediction models that could work across new drugs, cell lines, and patients.

ÖZET

İLAÇ SINERJISI TAHMININDE İLAÇ VE HÜCRE HATTI TEMSILLERININ GERÇEK ETKISININ ORTAYA ÇIKARILMASI

EMİNE BEYZA ÇANDIR SOYDEMİR

BİLGİSAYAR BİLİMİ VE MÜHENDİSLİĞİ YÜKSEK LİSANS TEZİ, ARALIK 2024

Tez Danışmanı: Assoc. Prof. ÖZNUR TAŞTAN

Anahtar Kelimeler: İlaç Sinerjisi, Derin Öğrenme, Genelleme, Tekil Kodlama

İlaç kombinasyon terapisi, kanser gibi karmaşık hastalıkların tedavisinde etkili bir strateji olarak umut vadetmektedir. Ancak, ilaç kombinasyon uzayı çok geniş olduğundan, tüm kombinasyonların deneysel olarak değerlendirilmesi mümkün değildir. Bu nedenle, deneysel tarama çabalarını hızlandırmak için sinerjistik çalışabilecek ilaç çiftlerini önceliklendiren hesaplamalı modeller geliştirilmiştir. Bu modeller, daha önce raporlanmış ilaç kombinasyonu ölçümlerini içeren büyük veri kümeleri üzerinde eğitilmektidr ve kimyasal, yapısal ve biyolojik özellikleri kodlayan zengin ilaç ve hücre hattı temsilleri kullanmaktadır.

Bu tezde, öncelikle sinerji tahmin modeli MatchMaker'ı, yolaklar ve etki mekanizmaları gibi daha zengin biyolojik bilgileri veya alternatif ilaç temsillerini dahil ederek geliştirmeyi amaçladık. Ancak, tüm çabalarımıza rağmen hiçbir model daha iyi performans gösteremedi. Bu bulgulardan hareketle, daha basit bir yaklaşım denedik ve detaylı özellik temsillerini ilaçların ve hücre hatlarının tek-seçim kodlamalarıyla değiştirdik. Kimyasal ve biyolojik bilgiler içermeyen bu modeller, beklenmedik bir şekilde, zengin biyolojik ve kimyasal bilgilerle eğitilen modellerin sonuçlarına oldukça yakın bir performans sergileyebildi.

Bu tezde, yayınlanmış sinerji tahmin modellerini sistematik olarak inceledik ve çeşitli değerlendirme kurulumunda ilaç ve hücre hattı temsillerini basit tek-seçim kodlamalarıyla değiştirdik. Kullanılan ilaç giriş özelliği veya model mimarisi ne olursa olsun, basit tek-seçim kodlama yaklaşımının tüm modellerde benzer şekilde perfor-

mans gösterdiğini gözlemledik. Bu beklenmedik sonuç, temsillerin kimyasal veya biyolojik bilgiyi öğrenmek yerine yalnızca basit tanımlayıcılar olarak hizmet ettiğini ve modellerin kısa yoldan sinerji ölçümlerinin genel ortak varyasyon örüntülerini yakaladığını göstermektedir. Bu durum, modellerin yeni ilaçlar ve hücre hatlarına iyi genelleme yapamamasının bir nedeni olabilir. Sinerji tahmin modelleri, ilaç ve hücre hattı panelinde hangi çiftlerin test edileceğine karar verirken hala yararlı olsalar da, bu sonuçlar, yeni ilaçlar, hücre hatları ve hastalar için çalışabilecek sinerji tahmin modellerinin geliştirilmesi için alternatif yaklaşımlara ihtiyaç olduğunu ortaya koymaktadır.

ACKNOWLEDGMENTS

First and foremost, I want to thank my supervisor, Assoc. Prof. Öznur Taştan, for always being there to guide, support, and encourage me throughout my master's journey. Her thoughtful insights, patience, and caring approach have been truly invaluable. Beyond being an exceptional educator, she has set a remarkable example as a woman, inspiring me both academically and personally.

I'm also thankful to Halil İbrahim Kuru, author of MatchMaker, for openly sharing his expertise and experiences. His contributions have greatly enhanced my understanding and strengthened the foundation of this work.

My sincere thanks go to my friends, Selcan, Talha, Buse and Ekin, who cheered me on during the toughest moments. I am also thankful to Berkay and Emine Ayşe for patiently answering my endless questions and never hesitating to help.

I'm deeply grateful to my dear family, my mother, my father, my brothers , and my sister-in-law, for their constant belief. I'm also thankful for my niece, Elif. Her presence reminds me of the reasons I pursued this path in the first place: to hopefully make the world a better place for others, including her.

Finally, to my soulmate, İhsan Soydemir, I can only say thank you for consistently inspiring me to be better, for guiding me with your wisdom, and for supporting me without any conditions. Your love and understanding have meant more to me than words can say.

Dedicated
to my family

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACC: Accuracy

BACC: Balanced Accuracy

CI: Confidence Interval

CLE: Cell Line Gene Expression Profiles

CNNs: Convolutional Neural Networks

CNV: Copy Number Variation

DGE: Differential Gene Expression Profiles

DNN: Deep Neural Network

DSNs: Drug Specific Subnetworks

ECFP: Extended Connectivity Fingerprints

ELU: Exponential Linear Unit

FC: Fully Connected

FCNN: Fully Connected Neural Network

GAT: Graph Attention Network

GDSC: Genomics of Drug Sensitivity in Cancer

GCN: Graph Convolutional Network

GNN: Graph Neural Networks

HTS: High-Throughput Screening

KAPPA: Cohen's Kappa

LCO: Leave-Cell-Line-Out

LDO: Leave-Drug-Out

LODO: Leave-One-Drug-Out

LPO: Leave-Pair-Out

LTO: Leave-Triple-Out

MLP: Multilayer Perceptron

MoA: Mechanism of Action

MSE: Mean Squared Error

MSigDB: Molecular Signatures Database

MUT: Mutation Features

OHE: One-Hot-Encoded

PCA: Principal Component Analysis

PCC: Pearson's Correlation Coefficient

PPI: Protein-Protein Interaction

PR AUC: Precision-Recall Area Under Curve

ReLU: Rectified Linear Unit

RMSE: Root Mean Squared Error

ROC AUC: Receiver Operating Characteristic Area Under Curve

SE: Standard Error

SCC: Spearman's Correlation Coefficient

SELFIES: SELF-referencing Embedded Strings

SMILES: Simplified Molecular-Input Line-Entry System

SPN: Synergy Prediction Subnetwork

UMLS: Unified Medical Language System

1. INTRODUCTION

Drug combination therapies are an alternative to monotherapies for treating several diseases, including cancer(Mokhtari et al., 2017a; Al-Lazikani et al., 2012), bacterial infections(Tamma et al., 2012), and several others(Möttönen et al., 1999; Gradman et al., 2010). Synergistic drug combinations provide higher efficacy and reduced side effects and resistance(Mokhtari et al., 2017b). While high-throughput screening of drug combinations across cell lines is possible, these technologies still can only cover a small portion of all combinations (Iorio, 2016; Jaaks et al., 2022a). The vast number of possible drug combinations and cell lines makes exhaustive clinical testing infeasible. To tackle this challenge, several computational models have been developed to assist experimental efforts(Preuer et al., 2017; Zhang et al., 2022; Wang et al., 2021; Hu et al., 2022a; Kuru et al., 2022; Rafiei et al., 2023; Li et al., 2024b). By predicting synergistic scores for drug combinations, these models allow prioritization of which drug combinations to test.

In a typical cell-line-based synergy prediction model, each input example consists of a triplet: a drug pair and the cell line on which the two drugs's synergistic effect is measured. Models numerically represent these triplets using various descriptors for drug structures and cell lines. While some models cast the problem as a classification task, synergistic vs antagonistic, others would formulate it as a regression task where a synergy score is estimated directly.

Recent advancements in drug synergy prediction have introduced various computational approaches, often leveraging deep learning techniques. DeepSynergy (Preuer et al., 2017), one of the pioneering deep learning-based models for drug synergy prediction, outperformed traditional machine learning approaches such as Gradient Boosting Machines, Random Forests, and Support Vector Machines. Since then, numerous deep learning models have been introduced(Abbasi and Rousu, 2024). The summary of a subset of these models can be seen in Table 1.1. Commonly used architectures in this field include Deep Neural Networks (DNNs), Graph Neural Networks (GNNs), and Transformers. The models typically use fully connected networks to process one-dimensional representation of drugs and cell lines (Preuer

et al., 2017; El Khili et al., 2023; Kuru et al., 2022). Alternative approaches include GNNs that represent drug structures as graphs, where atoms are represented as nodes and chemical bonds as edges. Models like DeepDDS and JointSyn leverage this approach(Wang et al., 2021; Li et al., 2024b). Transformers, such as those employed in DTSyn (Hu et al., 2022b) and DFFNDDS(Xu et al., 2023), are increasingly used to capture complex relationships between drugs and biological information. Some models exclusively focus on synergy prediction, while others adopt a multitask approach, predicting single-drug responses or drug-drug interaction classifications alongside synergy scores. For example, the MARSY model predicts synergy scores and single-drug responses for both drugs in a combination(El Khili et al., 2023).

These models utilize a wide range of drug features, including chemical descriptors (e.g., fingerprints, SMILES), pharmacological properties, and gene expression profiles after drug treatment. In addition, information on drug-target relationships and drug interaction networks is often incorporated. Cell line descriptions primarily rely on gene expression profiles(Abbasi and Rousu, 2024; Wang et al., 2021).

Table 1.1 Summary of Computational Models for Drug Synergy Prediction

Method	Model Type	Drug Features	Cell Line Features	Reference
DeepSynergy	DNN	Chemical Structures	CLE	(Preuer et al., 2017)
AuDNNSynergy	AE, DNN	Fingerprints	Genomic Mutations, CNV	(Zhang et al., 2021)
DeepDDS	GNN, DNN	Chemical Structures	CLE	(Wang et al., 2021)
DCE-DForest	BERT, Deep Forest	SMILES	CLE	(Zhang et al., 2022)
DTSyn	Transformer	SMILES, PPI	CLE	(Hu et al., 2022a)
MatchMaker	DNN	Chemical Structures	CLE	(Kuru et al., 2022)
PRODeepSyn	GCN	Fingerprints	CLE, Genomic Mutations, PPI	(Wang et al., 2022)
SynPathy	DNN	Chemical Structures	Pathway	(Tang and Gottlieb, 2022)
TranSynergy	DNN, Transformer	SMILES	CLE	(Liu and Xie, 2021)
CCSynergy	DNN	Fingerprints	Multi-omics	(Hosseini and Zhou, 2023)
DeepTraSynergy	Multitask, Transformer	SMILES	PPI	(Rafiei et al., 2023)
DEML	DNN, Ensemble	Fingerprints	CLE	(Wang et al., 2023b)
DFFNDDS	BERT, DNN	SMILES, Fingerprints	CLE	(Xu et al., 2023)
Forsyn	Deep Forest	SMILES	CLE	(Wu et al., 2023)
GAECDS	Graph AE, CNN	SMILES	CLE	(Li et al., 2023)
HyperSynergy	GNN, DNN	SMILES	CLE	(Zhang et al., 2023)
MARSY	DNN, Multitask	CLE	CLE	(El Khili et al., 2023)
JointSyn	GNN, DNN	Chemical Structures, SMILES	CLE	(Li et al., 2024b)

Our main goal in this study was to improve upon previous work by incorporating richer biological information into the models to attain better predictive performance. Throughout the study, we observe that the existing models in the literature suffer generalization across unseen drugs and cell lines. While many approaches demonstrate sufficient performance with randomly split datasets, their performance significantly reduces the use of cell lines and drugs. This leads us to investigate the real generalization power of the models systematically.

1.1 Main Contribution

In this thesis, we provide evidence that many of the models in the literature for synergy prediction may not be effectively learning from the input biological and chemical features but are learning from co-variance patterns in the data. We demonstrate that by replacing the feature representations in the models using one-hot encoded features versus complex biological descriptors the models rely on. This could be why models fail to generalize to new drugs and cell lines.

1.2 Thesis Organization

The thesis is organized as follows:

- Chapter 2 presents a comprehensive background on drug synergy research. It includes experimental screening methods and synergy scoring models. Chapter also explains evaluation setups and focuses on data splitting strategies and evaluation metrics. Finally, it provides drug and cell line representations, as well as the data resources used in the field.
- Chapter 3 reviews related works on drug synergy prediction models. It discusses the architectures and methodologies of models. The Section 3.2 examines how these models generalize to unseen drugs and cell lines.
- Chapter 4 presents attempts made to improve the performance of the MatchMaker model. These include regularization techniques, enhanced drug representations, and the integration of biological information. Incorporating additional biological information covers pathway data and mechanisms of action.
- Chapter 5 evaluates the use of one-hot encoded (OHE) features in drug synergy prediction. It details the datasets, data splitting strategies, and feature replacement procedures. Experiments are performed on MatchMaker, DeepSynergy, MARSY, JointSyn, and DeepDDS models.
- Chapter 6 concludes the findings and suggests future research directions.

2. BACKGROUND

2.1 Experimental Screening of Drug Combinations

Traditional methods for assessing drug combinations use manual techniques like tissue culture cytotoxicity assays. These require a significant amount of effort and approximately one to two weeks to complete. Disease-relevant cell lines are cultured, and IC50 values are estimated from studies or literature. Drug combinations are applied, and effects are measured using assays like cell viability or proliferation. Although they are reliable, traditional approaches have limitations in terms of time, cost, and scalability(Chou, 2008). These have led to the adoption of high-throughput screening (HTS) techniques (Iorio, 2016).

Experimental screening is fundamental in drug synergy research, enabling the assessment of drug combinations for their effects on cell lines. In a typical setup, disease-relevant cell lines, such as cancer cell lines, are cultured in appropriate media and seeded into high-density microplates. Drugs are prepared in dose gradients and applied in combinations to create a dose-response matrix, allowing systematic evaluation of multiple concentration pairs. Following incubation under controlled conditions, various assays can be used, such as cell viability, apoptosis, or proliferation. Data from these experiments are normalized and analyzed to generate dose-response curves, which are further input to synergy models to calculate synergy scores(O'Neil et al., 2016).

High-throughput combination screening employs automation and advanced data analytics to test thousands of drug combinations rapidly(He et al., 2018). Datasets such as DrugComb(Zheng et al., 2021) and NCI Almanac (Holbeck et al., 2017) have emerged as key resources in the field. DrugComb integrates data from multiple studies and uses various mathematical models to calculate synergy scores. NCI

Almanac focuses on FDA-approved drugs, providing IC50 values and ComboScores. These resources facilitate the evaluation of drug combinations across diverse cell lines, significantly expanding the scope of synergy research.

2.1.1 Synergy Scoring Models

Several mathematical models are used to calculate synergy scores from the dose-response matrix to provide a quantitative summary of two drugs' interaction:

- Loewe Additivity: The Loewe additivity model describes the expected outcome of a drug combination based on the additive effects of each drug when used individually. The model assumes that the drugs do not interact with themselves. According to Loewe (Loewe, 1953), the condition for additivity is expressed as:

$$\frac{d_1}{D_1} + \frac{d_2}{D_2} = 1 \tag{2.1}$$

  where for two drug1 and drug2 the doses in combination are represented as d_1 and d_2, and D_1 and D_2 are the doses of drugs required to achieve the same effect when utilized alone. Deviations from this expected effect indicate synergy (positive deviation) or antagonism (negative deviation.)

- Bliss Independence: This model assumes that drugs act independently, with the combined effect calculated as the product of their individual effects. It predicts the expected combined effect based on the effects of each drug acting alone. (Bliss, 1939).

$$E_{\text{Bliss}} = E_1 + E_2 - E_1 E_2 \tag{2.2}$$

  where E_1 and E_2 are individual effects of drugs in a pair. If the observed effect of the drug combination exceeds the Bliss-predicted effect, they are synergistic ($E_{\text{observed}} > E_{\text{Bliss}}$), otherwise antagonistic.

- ZIP (Zero Interaction Potency): ZIP integrates the Loewe and Bliss models to evaluate deviations from a non-interaction reference. This model is particularly useful for more complex interactions (Yadav et al., 2015).

(2.3)

$$E_{D_1D_2} = \frac{\left(\frac{[D_1]}{EC_{50,D_1}}\right)^{\lambda_{D_1}}}{1+\left(\frac{[D_1]}{EC_{50,D_1}}\right)^{\lambda_{D_1}}} + \frac{\left(\frac{[D_2]}{EC_{50,D_2}}\right)^{\lambda_{D_2}}}{1+\left(\frac{[D_2]}{EC_{50,D_2}}\right)^{\lambda_{D_2}}} - \frac{\left(\frac{[D_1]}{EC_{50,D_1}}\right)^{\lambda_{D_1}}}{1+\left(\frac{[D_1]}{EC_{50,D_1}}\right)^{\lambda_{D_1}}} \frac{\left(\frac{[D_2]}{EC_{50,D_2}}\right)^{\lambda_{D_2}}}{1+\left(\frac{[D_2]}{EC_{50,D_2}}\right)^{\lambda_{D_2}}}$$

where $[D_1]$ and $[D_2]$ are the concentrations of drugs. EC_{50,D_1} and EC_{50,D_2} represent the concentrations at which the half-maximal response is observed. λ_{D_1} and λ_{D_2} are the slope parameters of the dose–response curves(Vlot et al., 2019).

- ComboScore: This score is used in the NCI Almanac to evaluate drug combination effects based on IC50 values and dose-response data. It provides a general metric for synergy and combination efficacy (Holbeck et al., 2017).

$$Y_{\text{expected}} = \begin{cases} \min\left(Y_{D_1}, Y_{D_2}\right) \text{ if } Y_{D_1} \text{ or } Y_{D_2} < 0, \\ Y_{D_1} \times Y_{D_2}/100 \text{ otherwise} \end{cases} \tag{2.4}$$

where Y_{D_1} and Y_{D_2} represent the observed growth fractions of drugs in a pair. The ComboScore is determined by summing the differences between the observed and expected growth fractions for all dose combinations.

These models provide a continuous score. Based on these scores and set thresholds, it is possible to classify drug combinations into synergistic, additive, or antagonistic categories, supporting both experimental and computational approaches in drug discovery.

Despite the power of high-throughput approaches, challenges such as data variability, high costs, and scalability limitations persist. To address these issues, leveraging this HTS as training data for machine learning models is a promising strategy to predict synergy for unexplored drug combinations. This integration enhances the efficiency and scope of drug synergy research, bridging the gap between experimental and computational methodologies (O'Neil et al., 2016; Yadav et al., 2015).

2.1.2 Drug Synergy Prediction Problem Statement

The drug synergy prediction task aims to identify effective drug combinations for specific cell lines, enabling the discovery of therapies that enhance efficacy while minimizing side effects. For a given drug pair (i, j) and a cell line k, the goal is to predict a synergy score $y_{i,j,k}$, which can be:

- A continuous value ($y_{i,j,k} \in \mathbb{R}$) for regression models, representing the magnitude of synergy.
- A binary label ($y_{i,j,k} \in \{0, 1\}$) for classification models, indicating whether the interaction is synergistic.

Drugs i and j are represented by feature vectors ($\mathbf{x}_i$, $\mathbf{x}_j$), which may include diverse data types such as molecular descriptors, graphs, or one-hot encoded representations of drug identities. Similarly, the cell line k is represented by a feature vector ($\mathbf{c}_k$), derived from gene expression profiles, pathway-level features, or one-hot encoded representations of cell line identifiers.

This prediction problem is challenging due to the heterogeneous nature of the input datasets, usually culled from different sources and the complexity of the underlying biology of the drug interactions in the cellular processes.

2.2 Evaluation Setups

Because the models take into account triplets, assessing the predictive performance of the models necessitates careful evaluation set ups (Whalen et al., 2021). Dependencies between examples, such as shared drugs or cell lines, can result in overly optimistic performance estimates. To address these limitations, tailored splitting strategies are required (Baumann and Baumann, 2014; Preuer et al., 2017).

2.3 Data Splitting Strategies

- Leave-Triple-Out (LTO): In this strategy, drug–drug–cell line triples are randomly excluded from the training set. The test set includes entirely unseen drug–drug–cell line triplets. However, individual drug, drug pair or cell lines within these combinations may still appear with another partner in the training data. The purpose of LTO is to assess the model's ability to predict entirely novel drug-pair-cell line combinations.

- Leave-Pair-Out (LPO): In this strategy, drug pairs present in the test set are not included in the training set. However, individual drugs within these pairs can still appear in the training data, paired with other drugs. There is no restriction on cell lines; the test set may include cell lines seen during training. The purpose of LPO is to evaluate the model's ability to predict interactions between drug pairs it has not encountered before, while still utilizing the knowledge of each drug's behavior from other pairings. This approach tests the model's capacity to generalize to new drug combinations based on familiar components.

- Leave-Cell-Line-Out (LCO): This strategy excludes all data related to specific cell lines from the training set. Consequently, none of the cell lines in the test set are shared with the training data. There are no restrictions on drugs or drug pairs. The purpose of LCO is to test the model's ability to generalize its predictions to new biological contexts represented by unseen cell lines. This is crucial for understanding how well the model can adapt to different cellular environments, which is important for applications like personalized medicine where patient-specific cell responses are considered.

- Leave-One-Drug-Out (LODO): In this method, at least one drug in each drug pair within the test set is completely absent in the training triplets. The other drug in the pair may still appear in the training set, interacting with different drugs. There is no restriction on cell lines; the test set may contain cell lines seen during training. The purpose of LODO is to evaluate the model's ability to predict interactions when it has incomplete information—specifically when one drug is entirely new to the model. This scenario reflects real-world situations where a new drug is introduced, and the model must rely on existing knowledge of known drugs to make predictions.

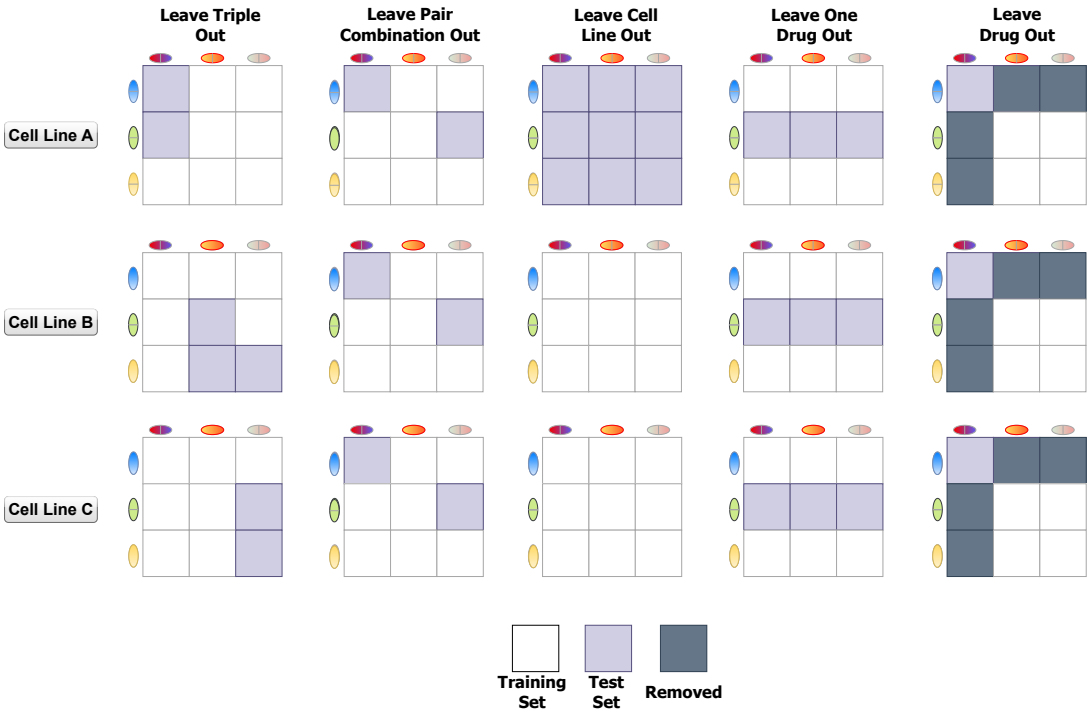
- Leave-Drug-Out (LDO): In this strategy, if a drug is seen in the training data,

this drug and all of its drug interactions are excluded in the test set. As a result, the model did not see any of the drugs in the test set during training. Implementing this split often reduces the dataset since removing drugs and their associated data decreases the amount of training information available. There is no restriction on cell lines; the test set can include cell lines present in the training data. The purpose of LDO is to assess the model's ability to handle entirely new drugs, evaluating how well it can predict interactions involving drugs it has no prior knowledge of.

- Group k-Fold Cross-Validation: Ensures that related entities (e.g., cell lines from the same tissue) are grouped into a single fold, avoiding information leakage that could artificially inflate performance metrics. For instance, grouping all pairs involving a specific drug within a fold ensures independence between training and testing datasets.

These structured splitting strategies are essential for assessing model performances in different real-world scenarios and their true generalization capacities, and are shown in Figure 2.1

Figure 2.1 Illustration of data split strategies. Each strategy shows how drug pairs, individual drugs, and cell lines are included or excluded in the training and test sets. This figure does not illustrate Group k-Fold Cross-Validation.



2.3.1 Evaluation Metrics

Proper metrics are required to evaluate the performance of models accurately. The choice of metric depends on whether the task is regression or classification:

- For Regression Tasks:
  - Mean Squared Error (MSE): Measures the average squared difference between predicted and true synergy scores. Lower values indicate better performance.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.5}$$

    where n is the number of observations, y_i is the actual value for the i-th observation, and $\hat{y}_i$ is the predicted value for the i-th observation.

  - Spearman's Rank Correlation Coefficient (SCC): Assesses the monotonic relationship between predicted and true values, capturing ranking performance.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{2.6}$$

    where n is the number of observations, and d_i is the difference between the ranks of the predicted and true values for the i-th observation. The SCC value ranges from -1 to $+1$, where:
    - $+1$: Perfect positive monotonic relationship,
    - 0: No monotonic relationship,
    - -1: Perfect negative monotonic relationship.
  - Pearson Correlation Coefficient (PCC): Measures the linear relationship between predictions and true values, often used alongside MSE for comprehensive evaluation.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{2.7}$$

    where n is the number of observations, x_i and y_i are the i-th predicted

and true values, respectively, and $\bar{x}$ and $\bar{y}$ are the mean of the predicted and true values, respectively.

The PCC value ranges from -1 to $+1$, where:

* $+1$: perfect positive linear correlation,
* 0: no linear correlation,
* -1: perfect negative linear correlation.

- For Classification Tasks:
  - Definitions of Key Terms:
    * TP (True Positives): Correctly predicted positive cases.
    * FP (False Positives): Incorrectly predicted positive cases.
    * FN (False Negatives): Positive cases incorrectly predicted as negative.
    * TN (True Negatives): Correctly predicted negative cases.
  - ROC AUC: Measures the trade-off between sensitivity (true positive rate) and specificity (false positive rate).

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.8}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{2.9}$$

The ROC AUC represents the area under the ROC curve, with values ranging from 0 to 1:

* 1: Perfect classification,
* 0.5: Random guessing,
* < 0.5: Poor classification.

  - Precision-Recall AUC (PR AUC): Evaluates the trade-off between precision and recall(true positive rate), particularly useful for imbalanced

datasets where the minority class is of primary interest.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.10}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.11}$$

The PR AUC is computed as the area under the Precision-Recall curve. Higher values indicate better performance.

These metrics provide complementary insights into model performance, ensuring reliable evaluation across different tasks.

2.4 Drug Representations

Drug representations are fundamental in predicting synergy effects. Traditional approaches often employ chemical fingerprints, such as Extended Connectivity Fingerprints (ECFPs), to encode molecular structures into fixed-length vectors. These fingerprints capture key chemical features, including functional groups and structural motifs(Rogers and Hahn, 2010; Pattanaik and Coley, 2020). Additionally, chemical descriptors generated by tools like ChemoPy(Cao et al., 2013a), PyBioMed(Dong et al., 2018) provide quantitative measures of properties such as hydrophobicity, molecular weight, and polarity.

In recent years, graph-based representations have gained attention. Here, drugs are modeled as molecular graphs, atoms as nodes, and bonds as edges(Yang et al., 2019). Models like DeepDDS(Wang et al., 2021) and AttenSyn(Wang et al., 2023a) use GNNs to learn detailed, topological insights from these graphs, improving the capture of molecular relationships. The graphs are generated from the SMILES(Weininger, 1988) of the drugs using the RDKit tool(Landrum, 2016).

Transformer-based large language models such as MolFormer(Ross et al., 2022), ChemFormer(Irwin et al., 2022), and SELFormer(Yüksel et al., 2023) process molecular strings to produce embedded structures that reflect both local and global chemical contexts. Although these representations have not yet been used for the drug

synergy prediction task, other benchmark tasks from MoleculeNet have shown success(Ross et al., 2022).

2.5 Cell Line Representations

Cell line representations aim to characterize the biological context in which drug interactions occur. Most commonly used feature is the cell line gene expression profiles(CLE). Gene expression levels in a cell line refer to the amount of RNA produced by specific genes, indicating their activity(Iorio, 2016; Teo et al., 2016; Charafe-Jauffret et al., 2005).

Genomic mutations (MUT) and copy number variations (CNV) are also used as genomic data in cell line representation. The integration of multi-omic data has broadened the scope of cell line representations. The omic datasets contain transcriptomics, proteomics, and epigenomics to provide a more comprehensive cellular profile. Research suggests that utilizing more than one complementary omic types can yield more robust predictive performance(Wang et al., 2022).

Approaches like TranSynergy use protein-protein interaction (PPI) networks, representing cellular components and their interactions as graphs(Liu and Xie, 2021). These graph-oriented frameworks complement multi-omic integration, offering richer representations of the biological contexts underlying drug synergy.

2.6 Drug Synergy Data Resources

In this section, we provide information on the dataset most used in synergy prediction tasks, and summarize in Table 2.1. O'Neil dataset(O'Neil et al., 2016), published in 2016, includes 263 drugs, 81 cell lines, and 369,776 combinations. It was among one of the earliest datasets systematically evaluating drug combinations in cancer cell lines.

Another dataset is the NCI Almanac dataset(Holbeck et al., 2017), released in 2017, contains 104 drugs tested across 60 cell lines, with a total of 304,549 combinations.

It has been widely used in tasks to predict synergy scores.

One of the most extensively used datasets, DrugComb, combines high-throughput screening studies into a standardized format. It was initially introduced in 2019 with 2,276 drugs, 93 cell lines, and 437,923 combinations(Zagidullin et al., 2019). It was expanded in 2021, with the latest version including 8,397 drugs, 2,320 cell lines, and 739,964 combinations(Liu et al., 2019).

The DrugCombDB(Liu et al., 2019), published in 2020, integrates data for 2,887 drugs and 124 cell lines, comprising 448,555 combinations. It aggregates experimental results from various sources. The SYNERGxDB dataset(Seo et al., 2020), released in 2021, consists of 1,977 drugs, 151 cell lines, and 477,839 combinations. It also combines multiple high-throughput drug combination studies. It provides molecular profiles and tools for predicting biomarkers and effective drug combinations.

Notably, there is significant overlap between datasets such as DrugComb, DrugCombDB, and SYNERGxDB, as these databases partly incorporate data from earlier studies, including NCI Almanac and O'Neil(Pan et al., 2023).

Lastly, the GDSC-combo dataset(Jaaks et al., 2022b) includes 2,025 drug combinations screened across 125 cell lines, resulting in 108,259 drug-cell line pairs. It integrates genomic and molecular features to enhance drug synergy analysis.

Table 2.1 Summary of Datasets for Drug Synergy Prediction

Dataset	**# of Drugs**	**# of Cell Lines**	**# of Combinations**	**Reference**
O'Neil	39	38	22,737	(Iorio, 2016)
NCI Almanac	104	60	304,549	(Holbeck et al., 2017)
DrugComb v.1	2,276	93	437,932	(Zagidullin et al., 2019)
DrugComb	8,397	2,320	739,964	(Zheng et al., 2021)
DrugCombDB	2,887	124	448,555	(Liu et al., 2019)
SynergXDB	1,977	151	22,507	(Seo et al., 2020)
GDSC-combo	65	125	108,259	(Jaaks et al., 2022b)

3. RELATED WORKS

3.1 Overview of Architectures

Deep learning architectures have been widely applied to predicting synergistic drug combinations. These architectures vary based on their design principles, training methodologies, and input types. Their input representations range from simple one-dimensional feature vectors to complex graph-structured data. They also differ in how they capture interactions between drugs and cell lines.

Deep Neural Networks (DNNs/FCNNs): Many early models rely on fully connected neural networks to process one-dimensional inputs such as molecular fingerprints, chemical descriptors, and CLE profiles. These networks combine multiple linear layers and nonlinear activations, allowing them to learn patterns from integrated drug and cell line features. Examples include models like DeepSynergy(Preuer et al., 2017) and AuDNNsynergy(Zhang et al., 2021), as well as later approaches like CC-Synergy(Hosseini and Zhou, 2023) and MARSY(El Khili et al., 2023). Some methods adopt subnetworks for each drug to ensure permutation-invariant predictions, enabling more robust handling of pairwise drug inputs(Kuru et al., 2022).

Convolutional Neural Networks (CNNs): Though not as common in drug synergy tasks, there are some models that adapt CNNs. CNNs leverage convolutional filters to extract local patterns. For example, the GAECDS model utilizes CNN in addition to GNN and DNN(Li et al., 2023).

Graph Neural Networks (GNNs): GNNs are a good option for modeling molecular structures and biological networks. GNNs can learn detailed topological and chemical information directly from the graph-structured inputs. Models like DeepDDS(Wang et al., 2021) extract structural drug information using drug graphs, while PRODeepSyn(Wang et al., 2022) builds more complex graphs to encode interactions

between drugs and cell lines.

Attention Mechanisms and Transformers: Attention based models, including Transformers, have gained popularity for their ability to handle with long-range dependencies and combine information from different sources. These architectures process features from various sources, such as drugs, targets, and cell lines, more flexibly. Models like SynergyX(Guo et al., 2024) employ Transformers to handle diverse omic data, while CancerGPT(Li et al., 2024a) uses large language models trained on scientific literature of drug combinations.

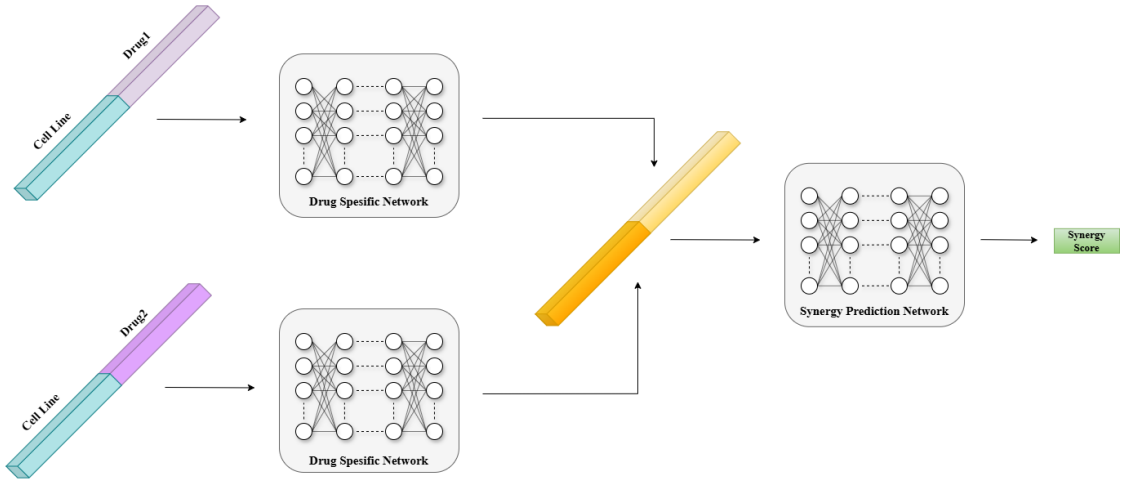
3.1.1 Drug Synergy Prediction Models Tested

3.1.1.1 MatchMaker

MatchMaker is one of the state-of-the-art models designed for drug synergy prediction. It is a deep neural network based model designed to predict the synergy score of drug pairs in cell lines. The model takes three inputs, chemical descriptors of each drug pair, and CLE profile of a cell line. The architecture of MatchMaker(Figure 3.1) consists of two Drug Specific Subnetworks (DSN) and one Synergy Prediction Subnetwork (SPN). Each DSN processes the chemical descriptors of a drug together with the CLE profile of the cell line to create a latent representation of the drug on the cell line profile. These DSNs share the same architecture and consist of three fully connected (FC) layers with the rectified linear unit (ReLU) activation in the first two layers, linear activation in the last one, and dropout applied to the first two layers with probability 0.2 and 0.5, respectively.

The outputs of the two DSNs are combined and fed into a three-layer FC network, the SPN. SPN also uses ReLU activation in the first two layers, linear activation in the output layer, and dropout with probability 0.5 applied after the first ReLU layer. SPN predicts the Loewe score for a given drug pair and cell line.

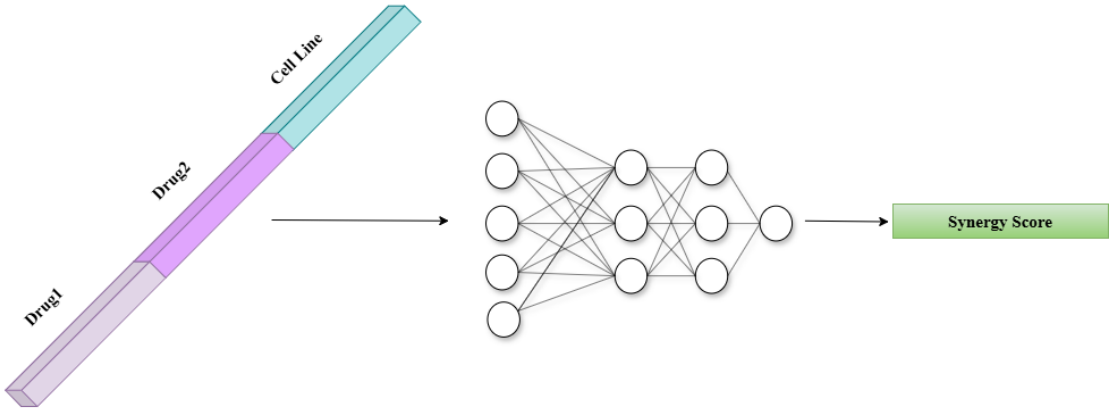
Figure 3.1 Architecture of the MatchMaker model. The model comprises two Drug Specific Subnetworks and one Synergy Prediction Subnetwork. Drugs are represented by chemical descriptors, and cell lines are represented by CLE profiles. Each DSN learns the representation of the drugs on the cell line. The SPN combines these representations and predicts the synergy score.



3.1.1.2 DeepSynergy

DeepSynergy is a deep learning model developed to predict drug synergy scores by utilizing the input features of two drugs and a cell line. The model takes concatenated input vectors containing the chemical descriptors of two drugs and the CLE profile of the cell line, as shown in Figure 3.2. DeepSynergy uses a structure with three hidden layers. A dropout rate of 0.2 is applied to the input layer, while a higher rate of 0.5 is used for the hidden layers. The hidden layers utilize the ReLU activation function for processing, and the output layer employs a linear activation function to generate the predicted synergy score. This structure focuses on extracting significant interactions between drugs and cell lines. This architecture has been widely recognized in the field as a baseline model for drug synergy prediction and serves as a benchmark for evaluating the prediction performance in related research.

Figure 3.2 Architecture of the DeepSynergy model. The model takes concatenated input vectors consisting of the two drugs and the cell line. Chemical descriptors are used for drug representations and CLE profiles for the cell line. The output layer employs a linear activation function to predict the synergy score.



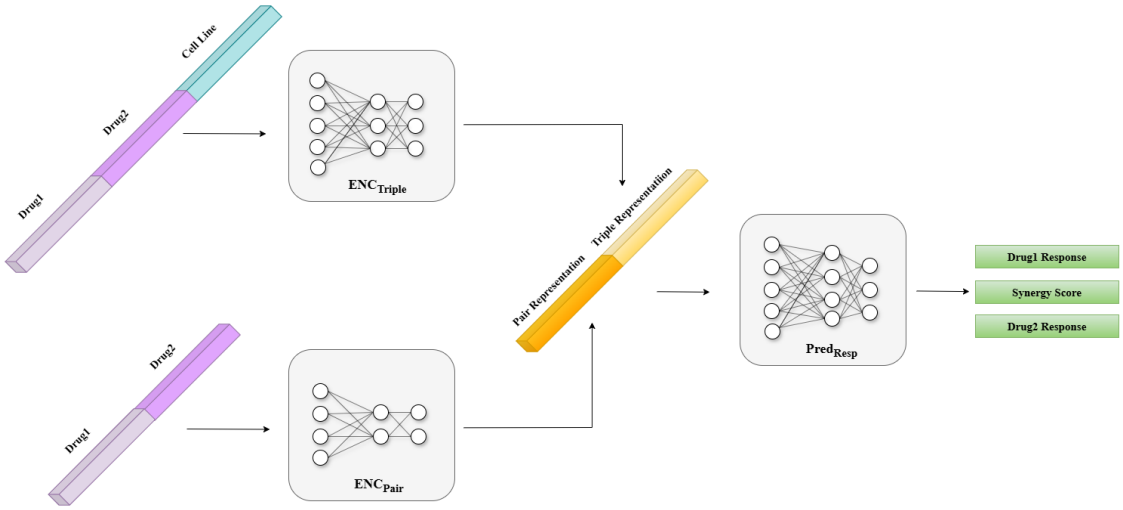
3.1.1.3 MARSY

MARSY is a deep learning model designed to predict drug synergy scores in cell lines by leveraging two parallel encoders and a multitask predictor. The architecture of MARSY can be seen in Figure 3.3. MARSY represents drugs using differential gene expression (DGE) signatures measured in MCF7 and PC3 cell lines after drug treatment, and it represents the untreated CLE profile of cell lines.

The first encoder, ENCPair, processes the concatenated feature vectors of the drug pair using two fully connected layers with a dropout rate of 0.2. The second encoder, ENCTriple, processes the concatenated features of the drug pair and the CLE profiles, also with two layers and the same dropout rate. Both encoders use linear activation for the first layer and ReLU activation for the second.

The embeddings from these encoders are concatenated and passed to the multitask predictor, PREDResp, which has two hidden layers and an output layer of size three. The predictor applies ReLU activation in the hidden layers and linear activation in the output layer, with dropout regularization at 0.5. This multitask setup simultaneously predicts the drug pair's synergy score and individual drug responses.

Figure 3.3 Architecture of the MARSY model. ENCPair encodes the concatenated feature vectors of the drug pair, while ENCTriple encodes the combined features of the drug pair and cell line. The embeddings are concatenated and passed to PREDResp. The multitask predictor simultaneously estimates the synergy score and individual drug responses Drugs are represented with DGE, while CLE profiles are used for cell lines.



3.1.1.4 JointSyn

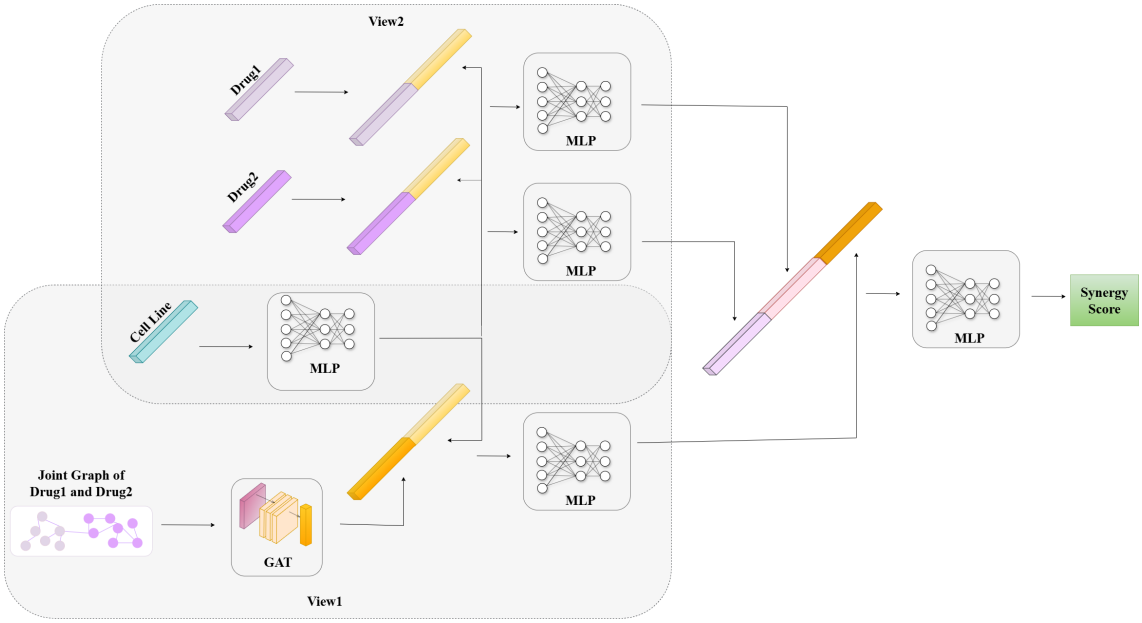
JointSyn is a deep learning model designed to predict drug synergy scores by combining multiple representations of drug pairs and cell lines. The model takes three inputs, a joint graph of the drug combination, Morgan fingerprints of the drugs, and the cell line's expression profile. It uses two views and a prediction network for synergy prediction, as detailed in Figure 3.4 .

View 1 constructs a joint graph to capture interactions between drug molecules. A three-layer Graph Attention Network (GAT) processes this graph, generating a combined drug embedding. exponential linear unit (ELU) activations are applied to each GAT layer, and dropout regularization with a rate of 0.2 is used. Simultaneously, a two-layer multilayer perceptron (MLP) processes the CLE profile. Similar to the GAT layers, ReLU activations and 0.2 dropout are applied here as well.

View 2 learns embeddings for individual drugs on the cell line by combining Morgan fingerprints with the cell line embedding. These are processed through an MLP, which also uses ReLU activations and dropout with a rate of 0.2, to form representations of the drugs and the cell line.

Prediction Network integrates the embeddings from both views. A three-layer MLP combines these representations to predict the synergy score. ReLU activations are used in the hidden layers, with a dropout rate of 0.2. The output layer employs either a softmax or linear activation for classification or regression tasks.

Figure 3.4 Architecture of the JointSyn model. The model includes multiple inputs: a joint graph of the drug combination, drug fingerprints, and the cell line's CLE profile. View 1 processes the joint graph with a GAT and the cell line with an MLP. Then concatenates that embedding to feed them through another MLP. View 2 integrates fingerprints and cell line embeddings processed by an MLP. The Prediction Network combines embeddings from both views through an MLP to predict synergy scores.



3.1.1.5 DeepDDS

DeepDDS is a deep learning model designed to predict synergistic drug combinations. The model processes molecular graphs of drug pairs and CLE profiles of cancer cell lines to predict synergy label.

This model employs a GNN for drug feature extraction. Specifically, it utilizes the GAT as the primary method, although a Graph Convolutional Network (GCN) was also tested. GAT layers, with multihead attention mechanisms, process the molecular graphs, capturing higher-level node features. The GAT configuration includes two hidden layers, and uses the ELU activation function. Dropout regularization

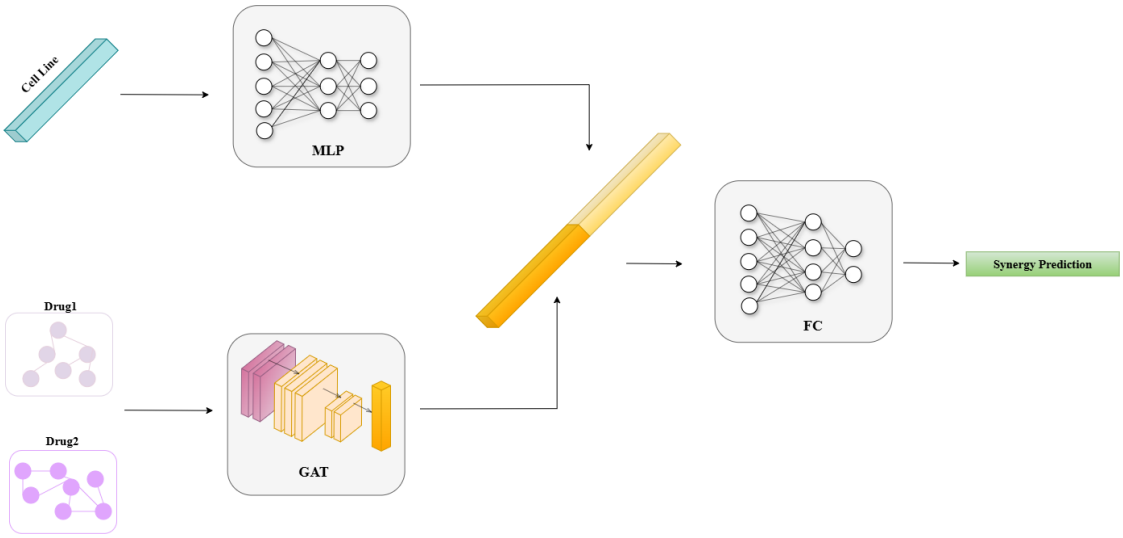
with a rate of 0.2 is applied to the GAT layers.

For cell line feature extraction, the model employs a MLP, which processes CLE profiles. The MLP consists of two hidden layers, applying ReLU activations and a dropout rate of 0.2.

The extracted embeddings from the GAT and MLP are concatenated and fed into a fully connected network for final classification, as represented in Figure 3.5. This network includes three hidden layers, employing ReLU activations and a softmax function in the output layer for classification.

While both GAT and GCN architectures were tested by DeepDDS, they reported that the GAT-based implementation performed better than GCN. Therefore, GAT was preferred while using DeepDDS.

Figure 3.5 Architecture of the DeepDDS model. The model uses a GAT for extracting drug features from molecular graphs and an MLP for processing CLE profiles. The embeddings from GAT and MLP are concatenated and passed through a fully connected network to predict synergy labels.



3.2 Generalization of Models to Unseen Drugs and Cell Lines

The table 3.1 highlights the current generalization problem in drug synergy prediction. Even with advanced architectures and data integration strategies, current models struggle to perform well on splits that involve novel drugs or cell lines(Abbasi

Table 3.1 Performance of Models on Different Stratified Splits

Method	**LTO**	**LPO**	**LCO**	**LODO**	**LDO**	**Metric**
DeepSynergy	–	255.49	405.40	435.92	–	MSE
DeepDDS	0.93	0.89	–	0.73	–	ROC AUC
DTSyn	–	0.78	0.82	0.73	–	ROC AUC
DFFNDDS	0.92	0.81	0.82	0.65	–	ROC AUC
Forsyn	–	0.44	0.45	0.36	–	PR AUC
MARSY	5.36	5.62	–	–	–	RMSE
JointSyn	0.89	0.86	0.67	–	0.19	PCC
SynergyX	79.55	90.43	–	–	–	MSE

and Rousu, 2024; Li et al., 2024b). This challenge shows the need for better feature representations, data augmentation techniques, and incorporation of external biological knowledge (e.g., protein-protein interaction networks or pathway data, mechanism of action).

Overall, some models demonstrate strong results under easier splits, such as LTO and LPO. However, the reduction in performance on LODO and LCO splits highlights the need for new approaches to overcome generalization barriers.

4. ATTEMPTS TO IMPROVE MATCHMAKER PERFORMANCE

The MatchMaker model, our previous work, demonstrated strong performance in drug synergy prediction, outperforming even some recent, more complex models(Abbasi and Rousu, 2024). Our main goal was further to enhance the overall predictive capabilities of the MatchMaker model. To achieve this, we explored various data splitting strategies and evaluated its performance on unseen drugs and cell lines.

4.1 Methods

4.1.1 Input Features & Dataset

For the cell line representations, we used the untreated cell line gene expression profiles following the same procedure as in the original MatchMaker paper. We obtained the RMA-normalized gene expression profiles from Genomics of Drug Sensitivity in Cancer (GDSC) in 2024(Iorio, 2016). We used the expression profiles of landmark genes, consisting of 972 genes. The list of these landmark genes was obtained from Subramanian (2017).

The MatchMaker model utilizes chemical descriptors to represent drugs, capturing structural and physicochemical properties. In the original MatchMaker paper, chemical descriptors were generated using ChemoPy(Cao et al., 2013b). However, as we used an updated version of the DrugComb dataset in this study, we needed to generate feature descriptors for new drugs. To address this, we replaced ChemoPy with PyBioMed(Dong et al., 2018), a more recent library that performs property cal-

culations while offering improved functionality and compatibility. Using PyBioMed, we generated a total of 367 features, details of which are provided in Table A.1.

In the original implementation of the MatchMaker model, the authors utilized the December 2019 version (v1.4) of the DrugComb dataset. After applying their filtering criteria, they obtained a dataset of 286,421 triplets (combinations of two drugs and a cell line) involving 3,040 drugs tested across 81 cell lines. For our experiments, we expanded this approach by employing the latest version of the DrugComb dataset, which includes 739,964 drug-cell line combinations involving 8,397 drugs tested on 2,320 cell lines. This updated dataset provides a broader range of drug interactions.

We filtered this dataset by selecting only those drugs with accessible structural information in the PubChem database and cell lines with gene expression data from GDSC(Iorio, 2016). After filtering, we obtained 426,386 combinations covering 3,057 drugs and 167 cell lines. However, when applying the LDO split we faced limitations in utilizing the entire dataset. The nature of the LDO split inherently reduces the amount of data available for training because it excludes all interactions involving the test drugs from the training set. Consequently, this restriction reduced the dataset to 151,050 samples for the LDO experiments.

4.1.2 Dataset Split Strategies

To systematically evaluate the model's performance, we partitioned the dataset into training, validation, and testing sets with a distribution of approximately 60% for training, 20% for validation, and 20% for testing. The following sections outline the data split stragies used to evaluate the model's performances in different scenarios:

- Leave-Pair-Out (LPO): Drug pairs in the test set are entirely excluded from the training set. However, individual drugs from these pairs may still appear in other combinations within the training data. This tests the model's ability to predict unseen drug interactions.
- Leave-Cell-Line-Out (LCO): All data associated with specific cell lines are removed from the training set. Consequently, none of the cell lines in the test set are present in the training data. This assesses the model's capability to generalize to entirely new biological contexts.
- Leave-One-Drug-Out (LODO): At least one drug in each drug pair within the test set is entirely absent from the training data. The other drug in the

pair may still appear in the training set in different contexts. This scenario evaluates how well the model can predict interactions involving new drugs by leveraging knowledge of known drugs.

- Leave-Drug-Out (LDO): Specific drugs and all their associated combinations are completely excluded from the training set. None of the drugs in the test set appear in the training data, making this the most challenging split for assessing the model's ability to generalize to entirely new drugs.

These strategies simulate real-world scenarios where the model encounters unseen drugs, cell lines, or drug combinations. For further details, refer to Section 2.3 and Figure 2.1.

4.1.3 Applying Regularization

Deep learning methods have shown exceptional potential to capture complex biological interactions(Esteva et al., 2019). However, this complexity often increases the risk of overfitting, where the model learns patterns specific to the training data. To mitigate this issue, regularization techniques are employed to constrain the model's capacity. This improves the model to focus on the most relevant patterns and the ability to generalize to unseen data. Among these techniques, dropout and L2 regularization are widely utilized for reducing overfitting(Kukačka et al., 2017).

4.1.3.1 Dropout Regularization

Dropout is a regularization method that reduces overfitting in neural networks by randomly dropping neurons during training. This process is controlled by the dropout rate, which determines the proportion of neurons deactivated in each iteration. This prevents overfitting of specific neurons and improves performance(Srivastava et al., 2014). The original architecture of MatchMaker already incorporates dropout regularization. The inDrop parameter applies the dropout rate to the first layer of each subnetwork, while the drop parameter applies it to subsequent hidden layers. We increased these dropout rates and conducted experiments for each data split method to assess their impact on model performance.

4.1.3.2 L2 Regularization

L2 regularization, also known as weight decay or ridge regression, reduces overfitting by adding a penalty term to the loss function based on the sum of the squared weights(Goodfellow et al., 2016; Ng, 2004). This approach discourages the model from assigning excessively large weights to any single feature, promoting better performance on unseen data. The regularized loss function is expressed as:

$$L_{\text{total}} = L_{\text{original}} + \lambda \sum_i w_i^2 \tag{4.1}$$

where L_{original} is the original loss, λ is the regularization strength, and w_i are the model weights. We applied L2 regularization to the MatchMaker model to further reduce overfitting and improve generalization across different data splits.

4.1.4 Improving Drug Representations

For all experiments with the MatchMaker model, we used CLE for cell line representations. These profiles provide a standardized and consistent representation of cell lines, ensuring compatibility across all analyses. For drug representations, we utilized chemical structure descriptors, which capture the structural, topological, and physicochemical properties of drugs. This approach aligns with the original implementation of the MatchMaker model and ensures that both cell line and drug features are represented in a manner that supports interpretable predictions.

To enhance the quality of the chemical descriptors and improve the model's ability to handle unseen drugs, we applied Principal Component Analysis (PCA) to reduce the number of descriptors to 50 components. This reduction retained more than 95% of the variance.

We also used transformer-based molecular language models to further enhance the model's performance on unseen drugs. These models learn high-quality molecular embeddings by analyzing large datasets and chemical notations. They have been very effective in predicting molecular properties for drug discovery and material science (Zhang et al., 2024). Both PCA and transformer-based embeddings were specifically aimed at making the model better at handling unseen drugs and cell

lines.

4.1.4.1 Using MoLFormer Embeddings

For our experiments, we incorporated embeddings generated by MoLFormer, a transformer-based molecular language model trained on 1.1 billion SMILES sequences. MoLFormer uses rotary positional embeddings and a linear attention mechanism to efficiently capture both chemical and structural information. MolFormer outperforms traditional models in various molecular property prediction tasks(Ross et al., 2022). We specifically used the MoLFormer-XL-both-10pct model provided by IBM on Hugging Face[1], which is trained on 10% of the original dataset. By encoding drug SMILES strings using this pre-trained model, we obtained enhanced drug feature representations for our experiments.

4.1.4.2 Using SELFormer Embeddings

We also use the SELFormer embeddings to enhance drug representations. SELFormer is a transformer-based chemical language model that uses SELFIES (SELF-referencing Embedded Strings) (Krenn et al., 2020), a molecular notation system designed to encode molecules as character strings. Unlike SMILES, SELFIES ensures that every generated string corresponds to a valid molecule, offering reliabile molecular representations. SELFormer trained on a dataset of two million drug-like compounds, and has demonstrated superior performance compared to SMILES-based methods(). For this study, we obtained the SELFormer model from its official GitHub repository[2] and created drug embeddings by first converting SMILES to SELFIES and then encoding them using SELFormer.

4.1.5 Incorporating Richer Biological Information

[1](https://huggingface.co/ibm/MoLFormer-XL-both-10pct)

[2](https://github.com/HUBioDataLab/SELFormer)

Utilizing that chemical features alone might be insufficient to capture the complexity of drug synergy. We integrated additional biological data into the model to enhance its predictive capabilities. By combining chemical and biological information, the model could gain a deeper understanding of the interactions that drive synergy.

4.1.5.1 Incorporating Pathway Information

Incorporating biological pathway information can improve a model's ability to capture the complex interactions underlying cellular processes(Kaynar et al., 2023). We utilized pathway data from the Molecular Signatures Database (MSigDB), specifically the 50 Hallmark gene sets(Liberzon et al., 2015). This pathway information was integrated into the model through a custom neural network layer that transforms gene expression profiles into pathway related features. The transformation is defined as:

$$\mathrm{ReLU}(\mathbf{X} \cdot (\mathbf{P} \odot \mathbf{W}) + \mathbf{b}) \tag{4.2}$$

where $\mathbf{X}$ represents the input gene expression matrix, $\mathbf{P}$ is the pathway association matrix, $\mathbf{W}$ and $\mathbf{b}$ are trainable weights and biases, and $\odot$ denotes element-wise multiplication. The pathway-specific features were combined with the drug features, to enable the model to capture interactions between cellular pathways and drug properties. This integration aimed not only to enhance predictive performance but also to provide insights into the biological mechanisms underlying cellular responses.

4.1.5.2 Mechanism of Action Features

We incorporated Mechanism of Action (MoA) information in addition to chemical descriptors to further enrich drug representations. MoA data were sourced from the DrugBank database(Wishart et al., 2017), which provides comprehensive textual descriptions of how a drug interacts with its target to produce a therapeutic effect. For example, the MoA description for dopamine states:

> "Dopamine is a precursor to norepinephrine in noradrenergic nerves and is also a neurotransmitter in certain areas of the central nervous system.

> Dopamine produces positive chronotropic and inotropic effects on the myocardium, resulting in increased heart rate and cardiac contractility. This is accomplished directly by exerting an agonist action on beta-adrenoceptors and indirectly by causing the release of norepinephrine from storage sites in sympathetic nerve endings. In the brain, dopamine acts as an agonist to the five dopamine receptor subtypes (D1, D2, D3, D4, D5)."

These detailed textual descriptions were transformed into numerical embeddings using SapBERT which a biologically pre-trained language model built upon the PubMedBERT(Liu et al., 2021). SapBERT was trained on the Unified Medical Language System (UMLS) dataset(Bodenreider, 2004). We combined the MoA embeddings with chemical structure features by concatenating them, creating a unified representation that incorporates both biological and chemical information. This combined representation aims to allow the model to better capture the complexity of drug interactions.

4.2 Results and Discussion

4.2.1 Impact of Regularization Methods

To improve MatchMaker's performance, we experimented with increasing the dropout regularization rates and applying L2 regularization. The Table 4.1, table shows the evaluation of the performance of different regularization strategies on the MatchMaker model using the DrugComb dataset across various splits. The first row for each data split method, represents the results obtained with the original MatchMaker hyperparameters. The subsequent rows show the outcomes after increasing the dropout rates and applying L2 regularization approaches.

Among the methods, increasing dropout rate generally showed the most consistent improvements, reducing the MSE across all splits while maintaining or slightly improving Spearman and Pearson correlation coefficients. L2 Regularization, on the other hand, demonstrated inconsistent results. Both strategy showed no significant

benefit compared to the original configuration.

Notably, the LODO and LDO splits, which involve testing on unseen drugs, exhibited significantly higher MSE and lower SCC and PCC values than the LPO splits. This indicates that the current regularization strategies fail to adequately address the generalization challenge based on these splits. While increased dropout shows some promise in reducing MSE, the improvements are not sufficient to overcome the difficulties of predicting drug synergy for unseen drugs or cell lines.

Table 4.1 Performance Evaluation of Regularization Hyperparameters on the MatchMaker Model Using the DrugComb Dataset

Split	Regularization Method	MSE	SCC	PCC	Regularization Hyperparameters
LPO	Original	99.4	0.71	0.75	inDrop = 0.2, drop = 0.5
	Increased Dropout	97.29	0.71	0.75	inDrop = 0.3, drop = 0.7
	L2 Regularization	100.83	0.69	0.74	inDrop = 0.2, drop = 0.5, L2 factor = 0.01
LCO	Original	174.07	0.47	0.53	inDrop = 0.2, drop = 0.5
	Increased Dropout	165.50	0.49	0.55	inDrop = 0.3, drop = 0.7
	L2 Regularization	170.30	0.49	0.55	inDrop = 0.2, drop = 0.5, L2 factor = 0.01
LODO	Original	216.26	0.39	0.40	inDrop = 0.2, drop = 0.5
	Increased Dropout	214.61	0.38	0.40	inDrop = 0.3, drop = 0.7
	L2 Regularization	211.71	0.40	0.42	inDrop = 0.2, drop = 0.5, L2 factor = 0.01
LDO	Original	248.98	0.17	0.15	inDrop = 0.2, drop = 0.5
	Increased Dropout	246.54	0.15	0.15	inDrop = 0.3, drop = 0.7
	L2 Regularization	261.31	0.15	0.14	inDrop = 0.2, drop = 0.5, L2 factor = 0.01

4.2.2 Impact of Alternative Drug Representations

The results in the Table 4.2 highlight the varying performance of different drug feature types in the MatchMaker model across distinct split methods. Chemical descriptors, originally used in the MatchMaker model, provide a strong baseline. This is particularly evident in simpler splits like LPO and LCO, where they achieve competitive correlation metrics and reasonable MSE values. Applying PCA to chemical descriptors shows little improvement. This indicates that dimensionality reduction does not significantly enhance predictive power. Transformer-based models, such as MoLFormer and SELFormer, demonstrate inconsistent results. While MoLFormer achieves the lowest MSE in the LPO split, it struggles to generalize in more challenging splits as showing low correlation values. On the other hand, SELFormer performs similarly to chemical descriptors. However it excels in the LDO split, achieving the highest Spearman and Pearson correlations.

Despite utilizing a wide range of drug representations, including chemical descriptors, PCA-applied chemical structures, and transformer-based embeddings, the per-

formance across different split methods remains remarkably similar. Metrics such as MSE, Spearman, and Pearson correlations show only minor variations, regardless of the feature type used. Addressing this issue may require more advanced strategies, such as incorporating additional data sources or redesigning the predictive framework.

Table 4.2 Performance Comparison of the MatchMaker Model Using Different Features on the DrugComb Dataset Across Different Split Methods

Split Method	Drug Feature Type	MSE	SCC	PCC
LPO	MatchMaker-Chemical Descriptors	99.4	0.71	0.75
	PCA-applied Chemical Structures	99.85	0.71	0.74
	MoLFormer	98.21	0.70	0.75
	SELFormer	99.48	0.71	0.75
LCO	MatchMaker-Chemical Descriptors	174.07	0.47	0.53
	PCA-applied Chemical Structures	170.84	0.48	0.54
	MoLFormer	171.74	0.47	0.54
	SELFormer	171.87	0.47	0.54
LODO	MatchMaker-Chemical Descriptors	216.26	0.39	0.40
	PCA-applied Chemical Structures	213.58	0.38	0.40
	MoLFormer	225.72	0.31	0.35
	SELFormer	218.91	0.35	0.39
LDO	MatchMaker-Chemical Descriptors	248.98	0.17	0.15
	PCA-applied Chemical Structures	249.03	0.15	0.14
	MoLFormer	251.56	0.01	0.09
	SELFormer	236.31	0.19	0.23

Notes: Chemical descriptors were originally used in the MatchMaker model as its primary feature representation. These descriptors serve as the baseline for comparison with alternative feature representations.

4.2.3 Impact of Incorporating more Biological Information

4.2.4 Incorporating Pathway Information

We aimed to improve the MatchMaker model by integrating biological pathway features from MSigDB. However, as shown in Table 4.3, this addition did not yield the expected performance improvement. While there were slight increases in MSE and minimal changes in SCC and PCC, the pathway data ultimately failed to provide a meaningful advantage for predicting synergy scores. These results indicate that

adding pathway information as biological context does not effectively address the lack of improvement.

Table 4.3 Performance Comparison of the MatchMaker Model Using Pathway Information on the DrugComb Dataset Across Different Split Methods

Split Method	Cell Line Feature Type	MSE	SCC	PCC
LPO	MatchMaker - CLE	99.4	0.71	0.75
	Pathway-Enriched Features	100.2	0.70	0.74
LCO	MatchMaker - CLE	174.07	0.47	0.53
	Pathway-Enriched Features	178.06	0.46	0.52
LODO	MatchMaker - CLE	216.26	0.39	0.40
	Pathway-Enriched Features	222.63	0.37	0.37
LDO	MatchMaker - CLE	248.98	0.17	0.15
	Pathway-Enriched Features	253.47	0.19	0.15

Notes: Cell line gene expression levels were originally used in the MatchMaker model as its primary feature representation. These features serve as the baseline for comparison with pathway-enriched cell line representations.

4.2.5 Incorporating Mechanism of Action

The results in the Table 4.4 demonstrate the inconsistent impact of incorporating MoA features into the MatchMaker model alongside chemical descriptors. In simpler splits like LPO, MoA features slightly decrease performance, increasing the MSE and lowering the correlation metrics compared to using chemical descriptors alone. In the LCO split, MoA features showed small improvements, which indicates they may help in capturing drug combination effects. For the LODO split, MoA features made a bigger impact, lowering MSE and improving correlations significantly. This shows that MoA features can capture biological mechanisms that help with generalization. However, in the hardest split, LDO, where entirely new drugs are introduced, both feature types performed poorly, and the MoA features didn't provide much improvement. Overall, while MoA features help in some situations, they do bot improve performance for all scenarios.

Table 4.4 Performance Comparison of the MatchMaker Model Using Mechanism of Action Features on the DrugComb Dataset Across Different Split Methods

Split Method	Drug Feature Type	MSE	SCC	PCC
LPO	MatchMaker - Chemical Descriptors	101.18	0.71	0.75
	Mechanism of Action Enriched Features	116.67	0.65	0.72
LCO	MatchMaker - Chemical Descriptors	166.67	0.45	0.54
	Mechanism of Action Enriched Features	165.32	0.47	0.56
LODO	MatchMaker - Chemical Descriptors	257.55	0.26	0.29
	Mechanism of Action Enriched Features	229.79	0.35	0.44
LDO	MatchMaker - Chemical Descriptors	262.76	0.09	0.14
	Mechanism of Action Enriched Features	261.81	0.07	0.08

Notes: Chemical descriptors were originally used in the MatchMaker model as its primary feature representation. These descriptors serve as the baseline for comparison with MoA-enriched drug representations.

5. ASSESSING THE BENEFIT OF USING CHEMICAL AND BIOLOGICAL INFORMATION

Our primary objective was to improve the performance of the MatchMaker model. To achieve this, we explored various approaches, including regularization techniques, improved drug representations, and the integration of additional biological information. Despite these efforts, the model's performance did not significantly improve and remained consistent across the tested strategies.

The lack of improvement raised questions about how the MatchMaker model captures the drug pair-cell line synergy relationship from the features. It led us to hypothesize that the model might not fully utilize the provided chemical and genetic features. To test this hypothesis, we replaced the original drug and cell line features with one-hot encoded (OHE) representations. These representations contain only identity information. Upon observing that the one-hot encoded models' performances are on par with the models that include biological and chemical information, we subsequently expanded our analysis and systematically experimented with other well-known drug synergy prediction models to investigate whether other models behave the same.

5.1 Methods

5.1.1 Models Compared With

We compared different models that are well-recognized and/or recent. To observe the performance of models when using OHE representations, we experimented with five

state-of-the-art drug synergy prediction models. In our selection, we also paid attention that diverse drug representations are used in each model: MatchMaker(Kuru et al., 2022) , DeepSynergy (Preuer et al., 2017), MARSY(Li et al., 2024b), and DeepDDS(Wang et al., 2021). Each model uses different approaches and features for drug synergy prediction, while all utilize gene expression profiles for cell line features. Below, we provide a summary of these approaches. The details about the architectures are provided in Chapter 3 and their datasets are detailed in the subsequent sections:

- DeepSynergy, one of the pioneer models in this field, uses a fully connected neural network that takes as input the concatenated feature vectors of two drugs and the cell line. Drug features include three types of chemical features: extended connectivity fingerprints (ECFP6) with 1309 features, physicochemical properties with 802 features, and toxicophore features with 2276 features. For cell lines, employs gene expression profiles of untreated cells(Iorio, 2016), filtered to include 3984 genes. The profiles were processed using FARMS(Hochreiter et al., 2006) for normalization and summarization, retaining only informative genes(Preuer et al., 2017).

- MatchMaker, our earlier model employs two parallel drug-specific subnetworks alongside a Synergy Prediction Network. Each DSN processes the chemical features of one drug and the gene expression features of the cell line. The outputs of the DSNs are concatenated and fed into the SPN to predict synergy scores. In this work, we represent drugs using 367 chemical descriptors calculated with the PyBioMed library. These descriptors encode the chemical structure of each drug. For cell lines, baseline gene expression profiles of untreated cells were used, consisting of 972 landmark genes. The data were obtained from the ArrayExpress database (accession number E-MTAB-3610)(Iorio, 2016) and normalized using the RMA method(Kuru et al., 2022).

- MARSY is a very recent model. It generates representations for the drug pair and the drug–cell line interaction through separate encoders. These representations are combined in a multitask predictor to output synergy scores and single-drug responses. MARSY represents drugs using their differential gene expression(DGE) signatures obtained from the LINCS database(Subramanian, 2017). The signatures are measured in two cell lines, MCF7 and PC3, 24 hours after drug treatment. Each drug is represented by a concatenation of 978 landmark genes from both cell lines, resulting in 1956 features per drug. For cell lines, MARSY uses baseline gene expression profiles of untreated cells from the CCLE from CellMiner database(Reinhold et al., 2012). After filtering out

lowly expressed and low-variance genes, the final representation includes 4639 genes(El Khili et al., 2023).

- DeepDDS is an GNN-based model designed to learn drug representations from molecular graphs created from SMILES strings. In these graphs, nodes represent atoms, and edges represent bonds. Each node is described by a binary vector containing five atomic properties, including atom type and aromaticity. For cell lines, DeepDDS uses baseline gene expression profiles of untreated cells, filtered to include 954 genes. These genes were selected by intersecting CCLE expression data with LINCS landmark genes and removing noncoding RNA transcripts(Wang et al., 2021).

- JointSyn, which was published in 2024, is one of the latest models achieving competitive results against other state-of-te-art methods. The model integrates a joint graph of drug combinations, drug features, and cell line representations into a dual-view architecture. These views extract embeddings for the drug combination and cell line, which are passed to a prediction network to estimate synergy scores (Figure 3.4). JointSyn uses molecular graphs and Morgan fingerprints to represent drugs. Molecular graphs are derived from SMILES strings using RDKit(Landrum, 2016), where atoms are represented as nodes and bonds as edges. Each node is characterized by a 78-dimensional atomic feature vector computed with DeepChem(Ramsundar et al., 2019). Morgan fingerprints, with 1309 features, capture additional structural characteristics. For cell lines, JointSyn employs baseline gene expression profiles with 2087 genes, filtered from the CCLE database based on drug sensitivity relevance(Li et al., 2024b).

The Table 5.1 highlights the diverse methodologies used across models to encode drug and cell line features.

5.1.2 Datasets The Models Employ

Synergy datasets consist of drug pairs, cell lines, and their associated synergy scores, providing the foundation for training and evaluating predictive models. In this study, we primarily used the datasets reported in the original publications for each model. However, for the MatchMaker model, we opted for an updated version of the DrugComb dataset. This updated dataset offers a more comprehensive coverage of drug

Table 5.1 Summary of Drug and Cell Line Features Used by Each Model

Model	Drug Features	Cell Line Features
MatchMaker	Chemical descriptors of drugs (367 features).	Untreated gene expression profiles with 972 landmark genes.
DeepSynergy	ECFP6 molecular fingerprints (1309 features), physicochemical properties (802 features), and toxicophore features (2276 features).	Untreated gene expression profiles with 3984 genes.
MARSY	Differential gene expression profiles measured in two cell lines (MCF7 and PC3) after drug treatment, with 978 landmark genes per cell line (total 1956 features per drug).	Untreated gene expression profiles with 4639 genes.
JointSyn	Molecular graphs: Molecular graphs derived by SMILES; nodes represent atoms, and edges represent bonds, with 78-dimensional atomic feature vectors. Morgan fingerprints: ECFP6 molecular fingerprints (1309 features).	Untreated gene expression profiles with 2087 genes.
DeepDDS	Molecular graphs derived by SMILES; nodes represent atoms, and edges represent bonds, with node features as binary vectors containing 5 atom-related properties.	Untreated gene expression profiles with 954 genes.

combinations and cell lines. These improvements make it particularly suitable for the leave-drug-out split, where a subset of the dataset has to be discarded. Additionally, while the NCI Almanac dataset was not provided in the original publication, we obtained and processed it for our experiments. Table 5.2 summarizes the characteristics of the datasets used, including the synergy score metrics, the number of combinations, drugs, and cell lines.

Table 5.2 Summary of Synergy Score Metrics and Datasets for Each Model

Models	Synergy Score	# of Combinations	# of Drugs	# of Cell Lines
MatchMaker - DrugComb	Loewe	426,386	3,057	167
MatchMaker - NCI Almanac	ComboScore	264,528	99	54
MARSY - DrugComb	Zip	86,348	670	75
DeepSynergy - O'Neil	Loewe	23,052	38	39
JointSyn - O'Neil	Loewe	12,033	38	34
DeepDDS - O'Neil	Binarized classification label	12,415	36	31

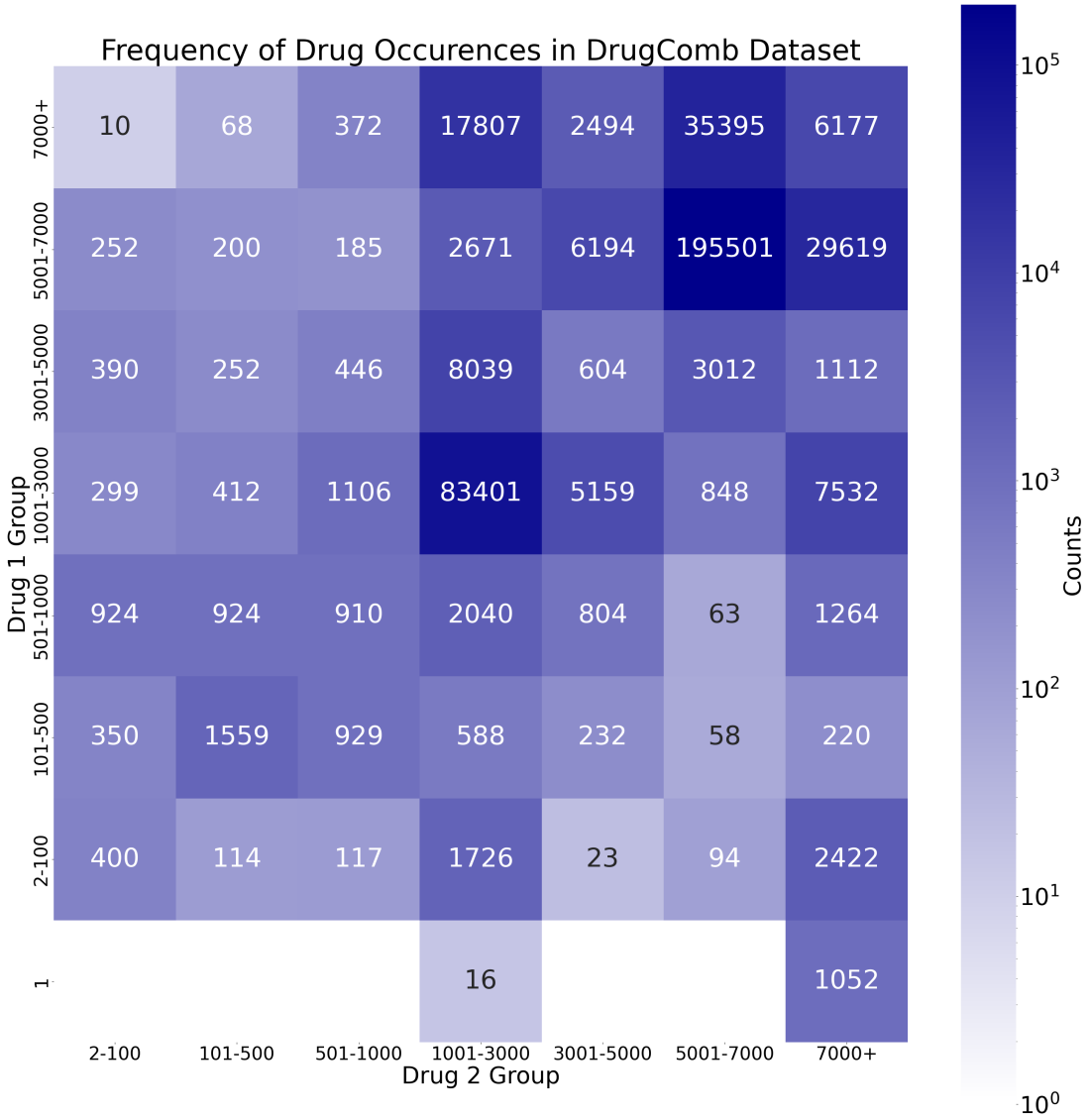
Notes: While some models utilize datasets from same sources, the sizes of these datasets vary according to the distinct filtering criteria employed in each study.

We used the DrugComb and NCI Almanac datasets for the MatchMaker model. The DrugComb dataset comprises 739,964 combinations of 8,397 drugs and 2,320 cell lines. We filtered this data set as consisting of only drugs if their structural information is available in the PubChem database and cell lines if their gene expres-

sion data are accessible from GDSC(Iorio, 2016). Following to filtering process, we obtained 426,386 combinations covering 3,057 drugs and 167 cell lines.

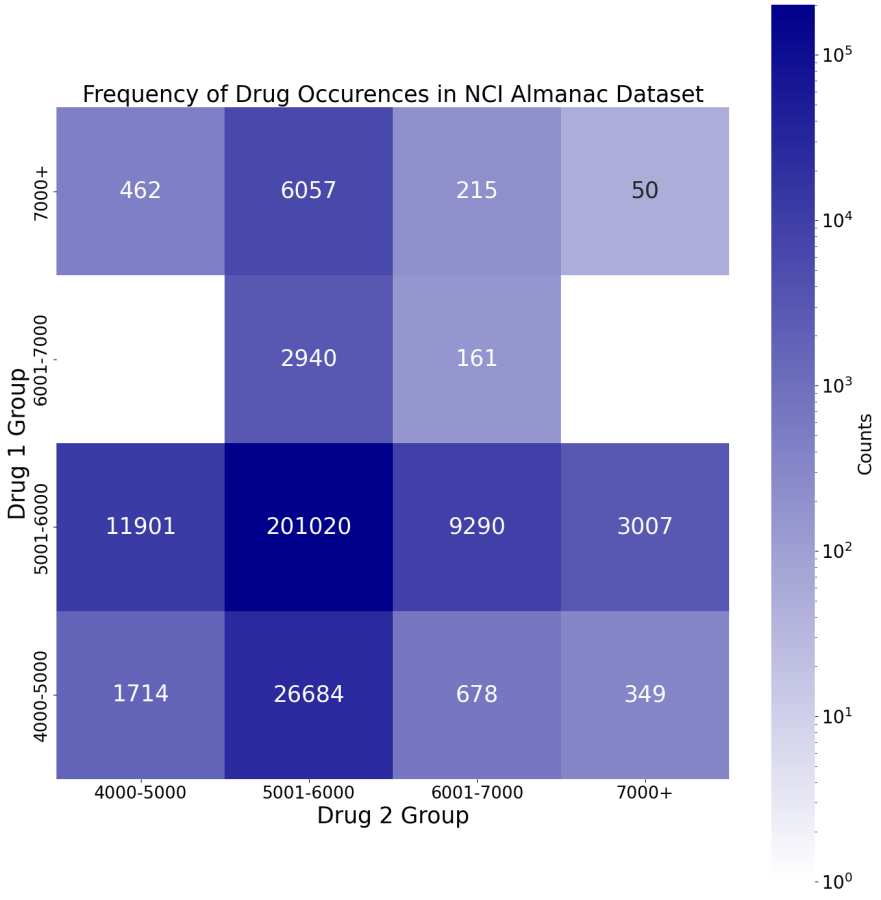
DrugComb dataset demonstrates a significant imbalance in the frequency with which drugs appear across combinations. Approximately two-thirds of the drugs are present in only one or two combinations within the dataset. In contrast, some drugs appear in more than ten thousand combinations. This imbalance also extends to drug pairs. As illustrated in Figure 5.1, one drug in a pair may appear in just a single combination, while the other in more than 7,000 combinations. Such imbalances pose significant challenges for model training, evaluation, and generalization, particularly when predicting interactions involving rare drugs.

Figure 5.1 Heatmap representing the frequency of drug occurrences and their pairwise groupings in the DrugComb dataset. Drug frequencies are categorized into distinct groups based on their occurrence counts in triplets.



As the imbalanced nature of the DrugComb dataset might be leading to certain issues, we extended our experiments to include the NCI Almanac dataset, which offers a more balanced representation of drug combinations. The original NCI Almanac dataset comprised 304,549 combinations of 104 drugs and 60 cell lines. After applying the same filtering criteria as for DrugComb, we retained 264,528 combinations involving 99 drugs and 54 cell lines. As illustrated in Figure 5.2, over 75% of the drug pairs in the filtered dataset feature both drugs appearing between 5,000 and 6,000 times. Additionally, every drug in the dataset appears in at least 4,000 combinations, resulting in a more evenly distributed dataset. This balance makes the NCI Almanac dataset a valuable resource for evaluating drug synergy models under more uniform conditions.

Figure 5.2 Heatmap representing the frequency of drug occurrences and their pairwise groupings in the NCI ALMANAC dataset. Drug frequencies are categorized into distinct groups based on their occurrence counts.



For DeepSynergy, we used the O'Neil dataset, available on the DeepSynergy website [1], which includes 23,052 combinations of 38 drugs and 39 cell lines.

[1] https://www.bioinf.jku.at/software/DeepSynergy/

The MARSY model was trained on a filtered version of the DrugComb dataset provided by its authors. This dataset contained 86,348 combinations involving 670 drugs and 75 cell lines.

The JointSyn model was evaluated on a subset of the O'Neil dataset, as detailed in its publication, containing 12,033 combinations of 38 drugs and 34 cell lines. Similarly, the DeepDDS model used another subset of the O'Neil dataset, which included 12,415 combinations involving 36 drugs and 31 cell lines.

These datasets also report various synergy scores for a given drug interactions. The DrugComb dataset, used for MatchMaker and MARSY, incorporates different synergy scores. MatchMaker uses the Loewe Additivity score, which evaluates drug synergy by analyzing dose-response matrices. Scores above zero indicate synergy, while scores below zero suggest antagonism. MARSY, on the other hand, uses the ZIP score from the DrugComb dataset. ZIP measures synergy by assessing deviations from expected dose-response curves, assuming minimal changes for non-interacting drugs.

The O'Neil dataset, used for DeepSynergy and JointSyn, also employs the Loewe Additivity score. However, in the case of DeepDDS, the dataset applies a threshold. Combinations with a synergy score greater than 10 are classified as synergistic, while scores below zero are categorized as antagonistic.

Finally, the NCI Almanac dataset, used for MatchMaker, employs the ComboScore instead of Loewe. ComboScore evaluates drug synergy by summing the differences between observed and expected growth fractions across all dose combinations.

5.1.3 Data Splitting Strategies

For evaluating the models, structured data splitting strategies were employed, as detailed in Section 2.3 and illustrated in Figure 2.1. These include:

- Leave-Triple-Out (LTO): Entirely unseen drug–drug–cell line triples in the test set.
- Leave-Pair-Out (LPO): Unseen drug pairs in the test set, but individual drugs may appear in the training set.
- Leave-Cell-Line-Out (LCO): Unseen cell lines in the test set, testing generalization to new biological contexts.

- Leave-One-Drug-Out (LODO): At least one drug in the test set is entirely absent from training data.
- Leave-Drug-Out (LDO): Completely excludes certain drugs and all their interactions from training.

5.1.4 Testing Feature Learning with One-Hot Encoded Representations

We trained the models using OHE representations, which preserve only identity information without incorporating detailed chemical, biological, or genetic attributes. These OHE representations remove the influence of feature-based insights, allowing the models to operate based only on associations between drugs and cell lines. This approach evaluates whether the models' performance depends on detailed feature representations or simpler identity-based patterns.

To ensure a fair comparison and assess the impact of OHE representations, we followed the original architectures, hyperparameters, datasets, and splitting methods as reported in the publications of the selected models. These architectures are detailed in Section 3.1.1. Initially, the models were trained and tested with their original drug and cell line features. Subsequently, the same models were retrained and tested using OHE representations in place of the original features. This setup allowed a direct comparison of the models' performance across different types of input representations.

5.1.4.1 Procedure For Replacing Original Features with One-Hot Encoded Representations

In the OHE method, drugs and cell lines are represented with binary feature vectors. For drugs, each column corresponds to a specific drug, where the active drug's column is set to 1 and all others to 0. Similarly, for cell lines, each column corresponds to a specific cell line, which is set to 1 if it is active and 0 otherwise.

The primary objective of these experiments was to assess how the models perform when original feature representations are replaced with OHE representations. To ensure fair evaluation, the models' original architectures, hyperparameters, datasets, and split methods were preserved. This approach allowed us to isolate the impact

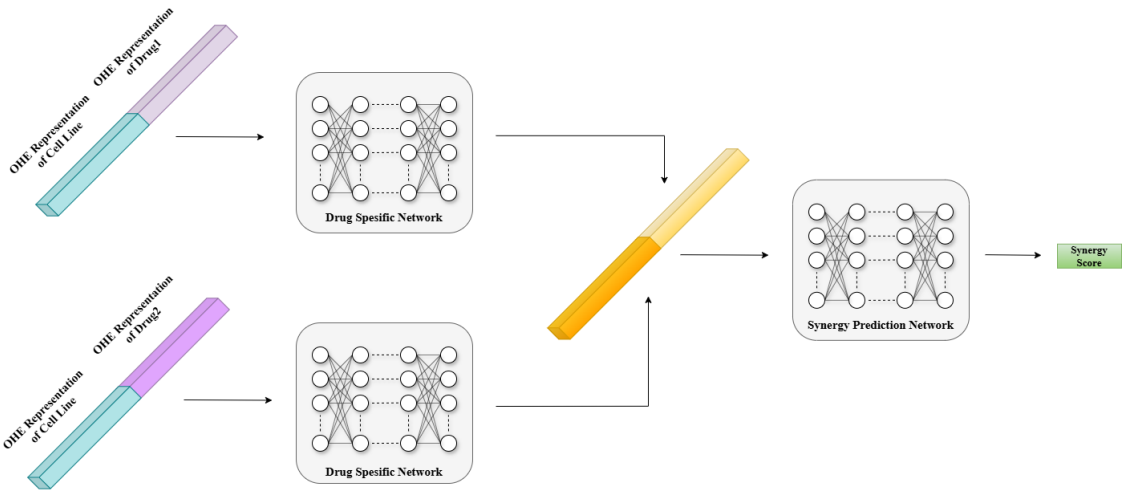
of OHE representations on model performance while maintaining consistency across experimental setups. In the following sections, we detail for each model what specific actions are performed.

5.1.4.2 Application to MatchMaker

The original MatchMaker architecture, illustrated in Figure 3.1, was trained using chemical structure and CLE features. In the OHE experiments, the chemical structure features were replaced with OHE representations of drugs, and CLE features were replaced with OHE representations of cell lines. These experiments were conducted using the original MatchMaker GitHub repository [2], ensuring consistency with the published model. The architecture for OHE experiments is shown in Figure 5.3.

MatchMaker was evaluated using the DrugComb and NCI Almanac datasets. While the original MatchMaker study reported results using the LPO split, we extended the evaluation to include all four split methods: LPO, LCO, LODO, and LDO. Each split followed a distribution of approximately 60% for training, 20% for validation, and 20% for testing.

Figure 5.3 Architecture of the MatchMaker model for OHE Experiment. The model comprises two Drug Specific Subnetworks and one Synergy Prediction Subnetwork. Drugs and cell lines are represented using one-hot-encoding. Each DSN learns the representation of the drugs on the cell line. The SPN combines these representations and predicts the



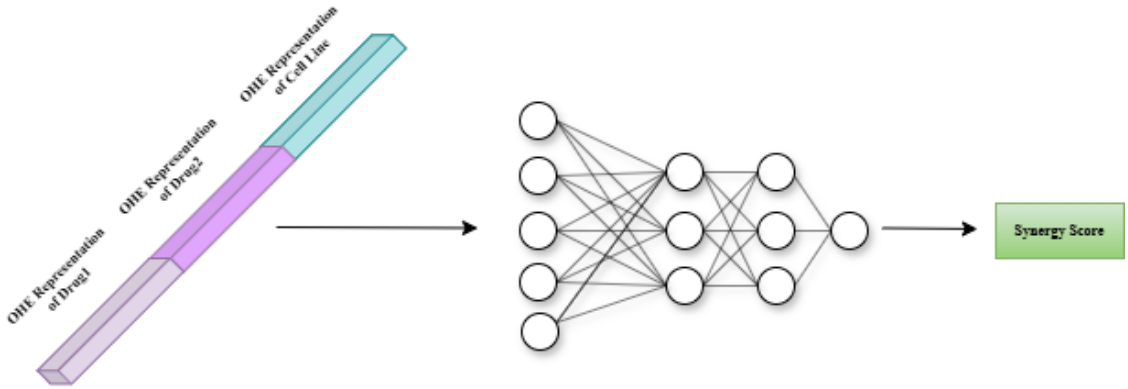
[2] https://github.com/tastanlab/matchmaker/tree/master

5.1.4.3 Application to DeepSynergy

The DeepSynergy model, as shown in Figure 3.2, was initially trained on its original dataset and input features, as outlined in its publication. For the OHE experiments, the original input features (Fingerprints for drugs and CLE for cell lines) were replaced with OHE representations for both drugs and cell lines. The model architecture, hyperparameters, and evaluation setup were consistent with the original study as the codebase[3] provided. Figure 5.4 illustrates the DeepSynergy architecture used in the OHE experiments.

The O'Neil dataset was used for training. 5-fold cross-validation for training and testing. During training, 60% of the data was used for hyperparameter selection, 20% for validation, and the remaining 20% was reserved for unbiased testing. LPO split strategy was used for this model.

Figure 5.4 Architecture of the DeepSynergy model for OHE Experiment. The model takes concatenated input vectors consisting of the two drugs and the cell line. One-hot-encoded representation are used for drugs and cell lines. The output layer employs a linear activation function to predict the synergy score.



5.1.4.4 Application to MARSY

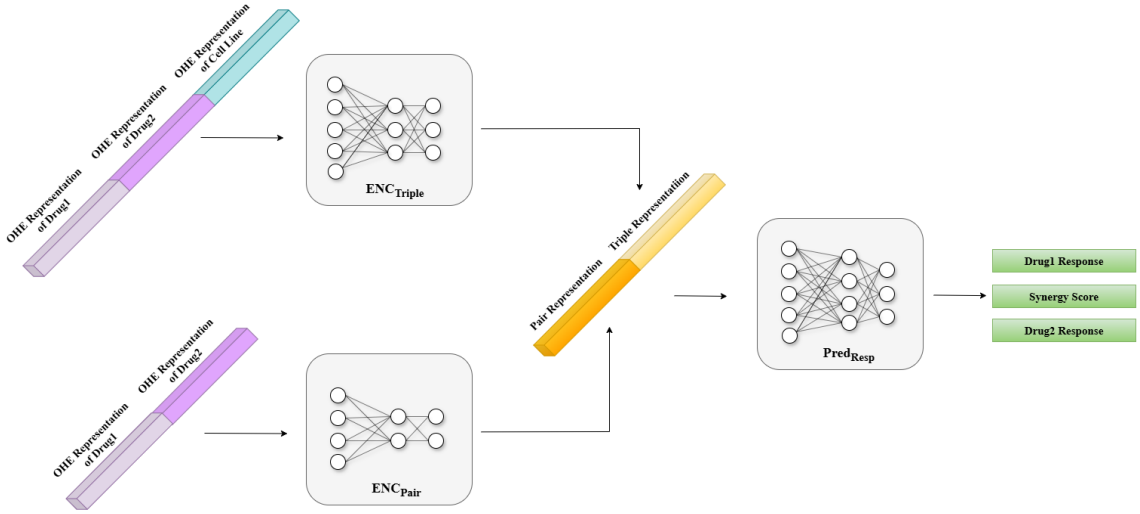
MARSY could not be retrained with the original drug and cell line features (DGE for drugs and CLE for cell lines) due to the large vector size of 8551, which caused an out-of-memory error during training. Instead, the results reported in the paper

[3]https://github.com/KristinaPreuer/DeepSynergy/tree/master

were used. For OHE experiments, input features were replaced with OHE representations, as shown in Figure 5.5. The model code was obtained from the authors' repository[4]. The only change made for this experiment is to replace the drug and cell line representations with OHE.

The DrugComb dataset was used for MARSY. Experiments employed the provided five-fold cross-validation splits with the LPO strategy for consistency.

Figure 5.5 Architecture of the MARSY model for OHE Experiment. ENCPair encodes the concatenated feature vectors of the drug pair, while ENCTriple encodes the combined features of the drug pair and cell line. The embeddings are concatenated and passed to PREDResp. The multitask predictor simultaneously estimates the synergy score and individual drug responses Drugs are represented with one-hot-encoding.

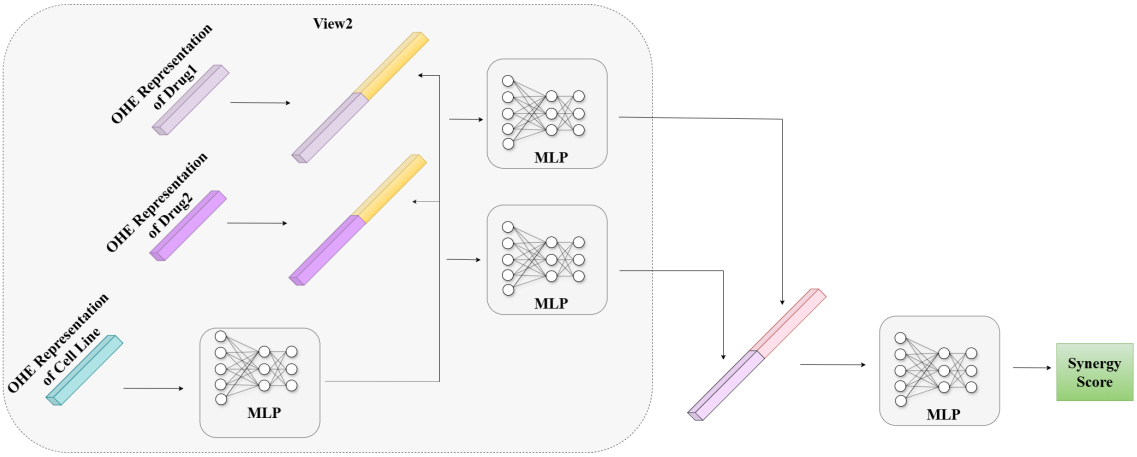


5.1.4.5 Application to JointSyn

The original JointSyn implementation3.4, with its graphs, could not be retrained due to incompatibilities among the graph libraries utilized in the framework. Instead, the published results used for original drug and cell line features. In OHE experiments, graph-based features were excluded, and Morgan fingerprints and cell line features were replaced with OHE representations. The modified architecture for OHE experiments is shown in Figure 5.6.

[4]https://github.com/Emad-COMBINE-lab/MARSY/tree/main

Figure 5.6 Architecture of the JointSyn model for OHE Experiment. View 2 integrates drugs and cell line embeddings processed by an MLP. Both drug and cell lines are represented with one-hot-encoding. The Prediction Network combines embeddings from both views through an MLP to predict synergy scores.



To ensure fairness, the implementation adhered to the original paper's setup, except for the graph exclusion in OHE experiments. The JointSyn code was obtained from the official GitHub repository[5]. Both the original and OHE experiments used the O'Neil dataset. We followed the authors' protocol for five-fold cross-validation with an LTO split, repeating the procedure 10 times for evaluation.

5.1.4.6 Application to DeepDDS

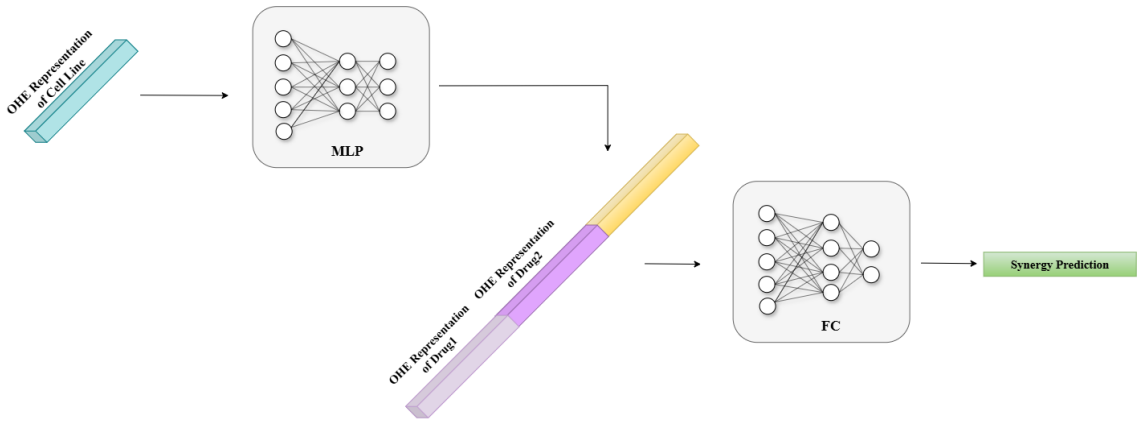
DeepDDS was evaluated using both its original features (molecular graphs and CLE) and OHE representations. The implementation and code were sourced from the authors' repository [6]. DeepDDS includes implementations for both GAT and GCN architectures. Since the paper reported better performance, the GAT architecture was used (Figure 3.5). For OHE experiments, graph-based features were excluded, and drug representations were replaced with OHE features. The modified OHE architecture is shown in Figure 5.7.

The experiments were conducted on the O'Neil dataset. The LTO split was applied as per the original work, with random splits saved and reused for consistency. A five-fold cross-validation protocol was employed for all experiments.

[5]https://github.com/LiHongCSBLab/JointSyn/tree/main

[6]https://github.com/Sinwang404/DeepDDs/tree/master

Figure 5.7 Architecture of the DeepDDS model for OHE Experiment. Cell line embeddings from the MLP and drug representations are concatenated and passed through a fully connected network to predict synergy labels. One-hot-encoded representations are used for both drugs and cell lines.



5.1.4.7 Computational Setup

We performed all experiments on systems equipped with Tesla V100-PCIE-32GB and Tesla V100S-PCIE-32GB GPUs. The Tesla V100-PCIE-32GB featured 31.74 GiB of memory, 80 cores, and a memory bandwidth of 836.37 GiB/s, while the Tesla V100S-PCIE-32GB offered the same memory and cores but with an enhanced bandwidth of 1.03 TiB/s. These systems also included CPUs with frequencies ranging from 2.29 GHz to 3.59 GHz, supporting AVX2, AVX512F, and FMA instructions. TensorFlow with CUDA 10.1 and cuDNN 7 was used for training, ensuring consistent performance across sessions.

5.2 Results & Discussion

The use of drug and cell line features is based on the assumption that deep learning models learn the synergistic relationships between drugs and cell lines by leveraging these features. Therefore, when a model encounters a drug or cell line it has never seen during training, it is expected to make accurate predictions by analyzing these features. However, according to various data reported in the literature and our

experiments with the MatchMaker model, we observed issues during the generalization phase. Attempts to overcome this with different methods were unsuccessful. This led us to question whether the models are effectively capturing the chemical and biological information about synergistic relationships from the features and the quality of the representations used. To investigate this, we experimented with five different synery model where we replaced their features with OHE features.

5.2.1 Performance with Drug & Cell Line Features vs. OHE Representations for MatchMaker

In the experiments conducted with the DrugComb dataset, we compared the MatchMaker model's performance using chemical descriptors and gene expression levels against OHE representations for drugs and cell lines. As shown in Table 5.3, the results across all split methods were very similar for both feature types.

Table 5.3 Performance Comparison of MatchMaker Model Using Drug & Cell Line Features vs OHE Representations on DrugComb Dataset Across Different Split Methods

Split Method	**Feature Type**	**MSE**	**SCC**	**PCC**
LPO	Drug & Cell Line Features	99.4	0.71	0.75
	One-Hot-Encoded	101.36	0.69	0.74
LCO	Drug & Cell Line Features	174.07	0.47	0.53
	One-Hot-Encoded	168.46	0.47	0.54
LODO	Drug & Cell Line Features	216.26	0.39	0.40
	One-Hot-Encoded	223.61	0.30	0.34
LDO	Drug & Cell Line Features	248.98	0.17	0.15
	One-Hot-Encoded	237.26	0.11	0.12

Notes: Chemical descriptors were originally used to represent drug features, while cell line gene expression levels represented cell line features in the MatchMaker model. These representations serve as the baseline for comparison with one-hot-encoded drug and cell line representations.

The results in Table 5.4 for the NCI Almanac dataset show that models trained with original drug and cell line features and those trained with OHE representations perform similarly across all split methods.

In both datasets, the performance of OHE representations closely matches or, in some cases, slightly surpasses that of original features. This consistent pattern suggests that the models may not be learning from the intended features. Instead, it becomes evident that they rely on identity-based correlations, which are also well captured by OHE representations.

Table 5.4 Performance Comparison of MatchMaker Model Using Drug & Cell Line Features vs OHE Representations on NCI Almanac Dataset Across Different Split Methods

Split Method	Feature Type	MSE	SCC	PCC
LPO	Drug & Cell Line Features	2759.38	0.53	0.56
	One-Hot-Encoded	2728.38	0.54	0.56
LCO	Drug & Cell Line Features	2788.02	0.48	0.61
	One-Hot-Encoded	2805.84	0.49	0.61
LODO	Drug & Cell Line Features	3766.30	0.25	0.26
	One-Hot-Encoded	3726.41	0.27	0.29
LDO	Drug & Cell Line Features	4970.01	0.16	0.20
	One-Hot-Encoded	5212.88	0.24	0.32

Notes: Chemical descriptors were originally used to represent drug features, while cell line gene expression levels represented cell line features in the MatchMaker model. These representations serve as the baseline for comparison with one-hot-encoded drug and cell line representations.

5.2.2 Application on Other Models

The results in Figure 5.8 reveal a consistent trend across multiple models and datasets. The performance differences between using original drug and cell line features and OHE representations are minimal. This finding raises critical questions about the role of detailed features in these models and whether their predictions genuinely depend on the biological or chemical information provided.

For MatchMaker on the DrugComb dataset under the LPO split, the MSE difference between the original features and OHE representations was just 1.97%, indicating that the model performance remains largely unaffected by learning from the feature representation. This aligns with earlier observations where MatchMaker also displayed similar results across various drug representations, further suggesting that its predictive capabilities might rely more on inherent patterns in the data than on complex feature learning.

DeepSynergy exhibited a slightly larger deviation (-4.93%) on the O'Neil dataset with OHE representations, yet the performance remains competitive. Similarly, MARSY, tested on the DrugComb dataset, showed a marginal improvement (-4.40%) with OHE, suggesting that the model's dependency on intricate biological features may be limited.

For models like JointSyn and DeepDDS, which incorporate graph-based and embedding techniques, the performance gap was negligible as well. JointSyn reported only a 1.07% deviation in MSE under the LTO split, while DeepDDS showed a mere 0.61% deviation in AUC. This indicates that even for advanced architectures, OHE

representations are sufficient for comparable performance.

Figure 5.8 Comparison of Drug & Cell Line Features vs OHE Representations Across Different Models and Datasets. Drug and cell line features refer to the original representations of drugs and cell lines used in their respective models. MSE and SE_{mse} results for the evaluated models. For MARSY and JointSyn, SE_{mse} was derived from fold-based calculations (see Appendix, Standard Error of MSE, for details). For DeepDDS, ROC-AUC is reported.
*Results for MARSY and JointSyn using original features were sourced from their respective publications.

Model	Dataset	Split Method	Metric	Drug & Cell Line Features	One-Hot Encoding	(•-•)/• Percent Deviation (%)
Matchmaker	DrugComb	LPO	MSE	**99.40** ±1.17	101.36 ±1.14	1.97 ±0.50
DeepSynergy	O'Neil	LPO	MSE	209.44 ±10.52	**199.12** ±9.84	-4.93 ±2.06
MARSY	DrugComb	LPO	MSE	31.58 ±0.76*	**30.19** ±1.10	-4.40
JointSyn	O'Neil	LTO	MSE	**76.56** ±0.48*	77.38 ±2.21	1.07
DeepDDS	O'Neil	LTO	AUC	0.93	**0.94**	0.61

The performance and standard errors remained largely consistent regardless of whether original features or OHE representations were used. In addition to these results, the MARSY model also predicts single-drug responses and strong results were obtained with OHE despite the lack of any transcriptomic or other genomic information(Table A.4). This finding suggests that this issue may extend beyond the drug synergy prediction domain.

5.2.3 Possible Explanations

5.2.3.1 Dataset Biases

Synergy score datasets are structured as drug pair and cell line triples, where individual drugs or cell lines often appear in thousands of such combinations. This repeated presence helps the model take shortcuts and just focus on the identity of

the drugs and cell lines. For example, in the Leave Pair Out split, specific drug pairs in the test set are excluded from the training set. However, the individual drugs and cell lines within those pairs may still frequently appear in the training data. This overlap allows the model to repeatedly encounter the same drugs and cell lines, which leads the model towards learning correlations based on their occurrences rather than understanding true synergistic interactions.

Similarly, in the Leave Drug Out split, drugs in the test set are entirely unseen during training. Nevertheless, within the training set, the remaining drugs appear repeatedly across different drug pairs. This repetition in the dataset structure increases the likelihood of the model relying on occurrence patterns and take a shortcut rather than learning from the intended feature information.

5.2.3.2 Biological Process Complexity

Predicting drug synergy in cell lines involves modeling highly complex biological processes that occur sequentially. Initially, drugs enter the cells through mechanisms like membrane permeability or transporter proteins. Once internalized, they bind to specific biological targets such as proteins, enzymes, or receptors. This binding alters cellular signaling pathways, either activating or inhibiting them, which in turn impacts cellular functions such as gene expression, protein synthesis, and metabolic activities. These changes disrupt essential survival mechanisms of the cells, leading to outcomes like inhibited proliferation, apoptosis induction, or increased oxidative stress. Additionally, genetic and epigenetic factors, such as DNA methylation and histone modifications, may further enhance drug responsiveness. The combined effects of these processes ultimately result in phenotypic outcomes, including reduced cell growth, tumor shrinkage, or improved sensitivity to treatment. The current input features used in models may be insufficient to capture this biological complexity, potentially limiting the models' ability to predict drug synergy accurately on new drugs and cellular contexts. Incorporating more detailed information about the mechanisms of action of drugs holds promise for addressing this limitation. Such data could provide a more comprehensive understanding of the underlying interactions and improve the predictive capabilities of the models.

6. CONCLUSION

This thesis investigates whether the synergy prediction models benefit from complex chemical and biological features and how this sets limitations in improving the model performances on unseen drugs and cell lınes. We evaluated the performance of models utilizing detailed feature representations against those using only one-hot encoding for drugs and cell line identities. Surprisingly, we discovered that models show comparable performance even when replacing detailed features with basic one-hot encoded representations. This suggests that rather than capturing the biological or chemical complexities, the models could just be remembering correlation drugs and cell lines they have already encountered.

Moreover, the minimal impact of replacing detailed features with OHE representations highlights a broader concern regarding the generalization capabilities of these models. If their predictions are driven more by dataset-specific patterns rather than an understanding of feature-based relationships, their applicability to novel scenarios could be significantly constrained.

There are a few reasons why this could be happening. First, the data often contains the same drugs and cell lines appear repeatedly. This can lead the models to focus on these repeated patterns rather than learning the deeper biological relationships needed to predict synergy. This situation, also known as shortcut learning, describes the process by which an artificial intelligence model learns to solve a task by relying on spurious correlations that are present in the data rather than attributes that are directly connected to the task itself(Ong Ly et al., 2024; Geirhos et al., 2020). Second, the actual biology behind drug interactions is very complex and may not be fully captured by the features we currently use. As a result, the models may rely on essential signals instead of learning the real underlying mechanisms.

To improve this situation, future work could involve adding more meaningful biological or chemical information that helps models understand how drugs and cell lines interact. We can also change how we split the data, ensuring the model does not just memorize the same drugs and cell lines from training to testing sets. Finally, we can

explore new modeling approaches that encourage the model to learn the biological patterns rather than shortcuts.

In short, while today's models perform well on benchmark tests, they may not truly understand drug synergy. By focusing on richer features, better splitting strategies, and improved model designs, we can advance toward models that genuinely capture the underlying biology. This progress can ultimately enable the discovery of more effective drug combinations.

BIBLIOGRAPHY

Abbasi, F. and Rousu, J. (2024). New methods for drug synergy prediction: a mini-review.

Al-Lazikani, B., Banerji, U., and Workman, P. (2012). Combinatorial drug therapy for cancer in the post-genomic era. *Nature Biotechnology*, 30(7):679–692.

Baumann, D. and Baumann, K. (2014). Reliable estimation of prediction errors for qsar models under model uncertainty using double cross-validation. *Journal of cheminformatics*, 6:1–19.

Bliss, C. I. (1939). The toxicity of poisons applied jointly 1. *Annals of applied biology*, 26(3):585–615.

Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(90001):267D – 270.

Cao, D.-S., Xu, Q.-S., Hu, Q.-N., and Liang, Y.-Z. (2013a). Chemopy: freely available python package for computational biology and chemoinformatics. *Bioinformatics*, 29(8):1092–1094.

Cao, D.-S., Xu, Q.-S., Hu, Q.-N., and Liang, Y.-Z. (2013b). Chemopy: freely available python package for computational biology and chemoinformatics. *Bioinformatics*, 29(8):1092–1094.

Charafe-Jauffret, E., Ginestier, C., Monville, F., Finetti, P., Adélaïde, J., Cervera, N., Fekairi, S., Xerri, L., Jacquemier, J., Birnbaum, D., and Bertucci, F. (2005). Gene expression profiling of breast cell lines identifies potential new basal markers. *Oncogene*, 25(15):2273–2284.

Chou, T.-C. (2008). Preclinicalversusclinical drug combination studies. *Leukemia amp; Lymphoma*, 49(11):2059–2080.

Dong, J., Yao, Z.-J., Zhang, L., Luo, F., Lin, Q., Lu, A.-P., Chen, A. F., and Cao, D.-S. (2018). Pybiomed: a python library for various molecular representations of chemicals, proteins and dnas and their interactions. *Journal of Cheminformatics*, 10(1).

El Khili, M. R., Memon, S. A., and Emad, A. (2023). Marsy: a multitask deep-learning framework for prediction of drug combination synergy scores. *Bioinformatics*, 39(4).

Esteva, A., Robicquet, A., and Ramsundar, B. e. a. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24–29.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Gradman, A. H., Basile, J. N., Carter, B. L., and Bakris, G. L. (2010). Combination therapy in hypertension. *Journal of the American Society of Hypertension*, 4(2):90–98.

Guo, Y., Hu, H., Chen, W., Yin, H., Wu, J., Hsieh, C.-Y., He, Q., and Cao, J. (2024). Synergyx: a multi-modality mutual attention network for interpretable drug synergy prediction. *Briefings in Bioinformatics*, 25(2).

He, L., Kulesskiy, E., Saarela, J., Turunen, L., Wennerberg, K., Aittokallio, T., and Tang, J. (2018). *Methods for High-throughput Drug Combination Screening and Synergy Scoring*, pages 351–398. Springer New York, New York, NY.

Hochreiter, S., Clevert, D.-A., and Obermayer, K. (2006). A new summarization method for affymetrix probe level data. *Bioinformatics*, 22(8):943–949.

Holbeck, S. L., Camalier, R., Crowell, J. A., Govindharajulu, J. P., Hollingshead, M., Anderson, L. W., Polley, E., Rubinstein, L., Srivastava, A., Wilsker, D., Collins, J. M., and Doroshow, J. H. (2017). The national cancer institute almanac: A comprehensive screening resource for the detection of anticancer drug pairs with enhanced therapeutic activity. *Cancer Research*, 77(13):3564–3576.

Hosseini, S.-R. and Zhou, X. (2023). Ccsynergy: an integrative deep-learning framework enabling context-aware prediction of anti-cancer drug synergy. *Briefings in bioinformatics*, 24(1):bbac588.

Hu, J., Gao, J., Fang, X., Liu, Z., Wang, F., Huang, W., Wu, H., and Zhao, G. (2022a). Dtsyn: a dual-transformer-based neural network to predict synergistic drug combinations. *Briefings in Bioinformatics*, 23(5):bbac302.

Hu, J., Gao, J., Fang, X., Liu, Z., Wang, F., Huang, W., Wu, H., and Zhao, G. (2022b). Dtsyn: a dual-transformer-based neural network to predict synergistic drug combinations. *Briefings in Bioinformatics*, 23(5).

Iorio, F. (2016). A landscape of pharmacogenomic interactions in cancer. *Cell*, 166(3):740–754.

Irwin, R., Dimitriadis, S., He, J., and Bjerrum, E. J. (2022). Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):015022.

Jaaks, P., Coker, E. A., Vis, D. J., Edwards, O., Carpenter, E. F., Leto, S. M., Dwane, L., Sassi, F., Lightfoot, H., Barthorpe, S., van der Meer, D., Yang, W., Beck, A., Mironenko, T., Hall, C., Hall, J., Mali, I., Richardson, L., Tolley, C., Morris, J., Thomas, F., Lleshi, E., Aben, N., Benes, C. H., Bertotti, A., Trusolino, L., Wessels, L., and Garnett, M. J. (2022a). Effective drug combinations in breast, colon and pancreatic cancer cells. *Nature*, 603(7899):166–173.

Jaaks, P., Coker, E. A., Vis, D. J., Edwards, O., Carpenter, E. F., Leto, S. M., Dwane, L., Sassi, F., Lightfoot, H., Barthorpe, S., van der Meer, D., Yang, W.,

Beck, A., Mironenko, T., Hall, C., Hall, J., Mali, I., Richardson, L., Tolley, C., Morris, J., Thomas, F., Lleshi, E., Aben, N., Benes, C. H., Bertotti, A., Trusolino, L., Wessels, L., and Garnett, M. J. (2022b). Effective drug combinations in breast, colon and pancreatic cancer cells. *Nature*, 603(7899):166–173.

Kaynar, G., Cakmakci, D., Bund, C., Todeschi, J., Namer, I. J., and Cicek, A. E. (2023). Pideel: metabolic pathway-informed deep learning model for survival analysis and pathological classification of gliomas. *Bioinformatics*, 39(11).

Krenn, M., Häse, F., Nigam, A., Friederich, P., and Aspuru-Guzik, A. (2020). Self-referencing embedded strings (selfies): A 100 *Machine Learning: Science and Technology*, 1(4):045024.

Kukačka, J., Golkov, V., and Cremers, D. (2017). Regularization for deep learning: A taxonomy. *arXiv preprint arXiv:1710.10686*.

Kuru, H. I., Tastan, O., and Cicek, A. E. (2022). Matchmaker: A deep learning framework for drug synergy prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(4):2334–2344.

Landrum, G. (2016). Rdkit: Open-source cheminformatics software.

Li, H., Zou, L., Kowah, J. A., He, D., Wang, L., Yuan, M., and Liu, X. (2023). Predicting drug synergy and discovering new drug combinations based on a graph autoencoder and convolutional neural network. *Interdisciplinary Sciences: Computational Life Sciences*, 15(2):316–330.

Li, T., Shetty, S., Kamath, A., Jaiswal, A., Jiang, X., Ding, Y., and Kim, Y. (2024a). Cancergpt for few shot drug pair synergy prediction using large pretrained language models. *npj Digital Medicine*, 7(1).

Li, X., Shen, B., Feng, F., Li, K., Tang, Z., Ma, L., and Li, H. (2024b). Dual-view jointly learning improves personalized drug synergy prediction. *Bioinformatics*, 40(10).

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J., and Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. *Cell Systems*, 1(6):417–425.

Liu, F., Shareghi, E., Meng, Z., Basaldella, M., and Collier, N. (2021). Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Liu, H., Zhang, W., Zou, B., Wang, J., Deng, Y., and Deng, L. (2019). Drugcombdb: a comprehensive database of drug combinations toward the discovery of combinatorial therapy. *Nucleic Acids Research*.

Liu, Q. and Xie, L. (2021). Transynergy: Mechanism-driven interpretable deep neural network for the synergistic prediction and pathway deconvolution of drug combinations. *PLoS computational biology*, 17(2):e1008653.

Loewe, S. (1953). The problem of synergism and antagonism of combined drugs arzneimittelforschung 3: 285–290. *Find this article online*.

Mokhtari, R. B., Homayouni, T. S., Baluch, N., Morgatskaya, E., Kumar, S., Das, B., and Yeger, H. (2017a). Combination therapy in combating cancer. *Oncotarget*, 8(23):38022–38043.

Mokhtari, R. B., Homayouni, T. S., Baluch, N., Morgatskaya, E., Kumar, S., Das, B., and Yeger, H. (2017b). Combination therapy in combating cancer. *Oncotarget*, 8(23):38022.

Möttönen, T., Hannonen, P., Leirisalo-Repo, M., Nissilä, M., Kautiainen, H., Korpela, M., Laasonen, L., Julkunen, H., Luukkainen, R., Vuori, K., et al. (1999). Comparison of combination therapy with single-drug therapy in early rheumatoid arthritis: a randomised trial. *The Lancet*, 353(9164):1568–1573.

Ng, A. Y. (2004). Feature selection, l1 vs. l2 regularization, and rotational invariance. In *Proceedings of the Twenty-first International Conference on Machine Learning (ICML'04)*, pages 78–85, New York. ACM.

Ong Ly, C., Unnikrishnan, B., Tadic, T., Patel, T., Duhamel, J., Kandel, S., Moayedi, Y., Brudno, M., Hope, A., Ross, H., and McIntosh, C. (2024). Shortcut learning in medical ai hinders generalization: method for estimating ai model generalization without external data. *npj Digital Medicine*, 7(1).

O'Neil, J., Benita, Y., Feldman, I., Chenard, M., Roberts, B., Liu, Y., Li, J., Kral, A., Lejnine, S., Loboda, A., Arthur, W., Cristescu, R., Haines, B. B., Winter, C., Zhang, T., Bloecher, A., and Shumway, S. D. (2016). An unbiased oncology compound screen to identify novel combination strategies. *Molecular Cancer Therapeutics*, 15(6):1155–1162.

Pan, Y., Ren, H., Lan, L., Li, Y., and Huang, T. (2023). Review of predicting synergistic drug combinations. *Life*, 13(9):1878.

Pattanaik, L. and Coley, C. W. (2020). Molecular representation: going long on fingerprints. *Chem*, 6(6):1204–1207.

Preuer, K., Lewis, R. P. I., Hochreiter, S., Bender, A., Bulusu, K. C., and Klambauer, G. (2017). Deepsynergy: predicting anti-cancer drug synergy with deep learning. *Bioinformatics*, 34(9):1538–1546.

Rafiei, F., Zeraati, H., Abbasi, K., Ghasemi, J. B., Parsaeian, M., and Masoudi-Nejad, A. (2023). Deeptrasynergy: drug combinations using multimodal deep learning with transformers. *Bioinformatics*, 39(8):btad438.

Ramsundar, B., Eastman, P., Walters, P., and Pande, V. (2019). *Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more*. " O'Reilly Media, Inc.".

Reinhold, W. C., Sunshine, M., Liu, H., Varma, S., Kohn, K. W., Morris, J., Doroshow, J., and Pommier, Y. (2012). Cellminer: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the nci-60 cell line set. *Cancer research*, 72(14):3499–3511.

Rogers, D. and Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754.

Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., and Das, P. (2022). Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264.

Seo, H., Tkachuk, D., Ho, C., Mammoliti, A., Rezaie, A., Madani Tonekaboni, S., and Haibe-Kains, B. (2020). Synergxdb: an integrative pharmacogenomic portal to identify synergistic drug combinations for precision oncology. *Nucleic Acids Research*, 48(W1):W494–W501.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Subramanian, A. (2017). A next generation connectivity map: L1000 platform and the first 1, 000, 000 profiles. *Cell*, 171(6):1437–1452.e17.

Tamma, P. D., Cosgrove, S. E., and Maragakis, L. L. (2012). Combination therapy for treatment of infections with gram-negative bacteria. *Clinical Microbiology Reviews*, 25(3):450–470.

Tang, Y.-C. and Gottlieb, A. (2022). Synpathy: Predicting drug synergy through drug-associated pathways using deep learning. *Molecular Cancer Research*, 20(5):762–769.

Teo, Z. L., Savas, P., and Loi, S. (2016). *Gene Expression Analysis: Current Methods*, page 107–136. Springer New York.

Vlot, A. H., Aniceto, N., Menden, M. P., Ulrich-Merzenich, G., and Bender, A. (2019). Applying synergy metrics to combination screening data: agreements, disagreements and pitfalls. *Drug Discovery Today*, 24(12):2286–2298.

Wang, J., Liu, X., Shen, S., Deng, L., and Liu, H. (2021). Deepdds: deep graph neural network with attention mechanism to predict synergistic drug combinations. *Briefings in Bioinformatics*, 23(1).

Wang, T., Wang, R., and Wei, L. (2023a). Attensyn: An attention-based deep graph neural network for anticancer synergistic drug combination prediction. *Journal of Chemical Information and Modeling*, 64(7):2854–2862.

Wang, X., Zhu, H., Jiang, Y., Li, Y., Tang, C., Chen, X., Li, Y., Liu, Q., and Liu, Q. (2022). Prodeepsyn: predicting anticancer synergistic drug combinations by embedding cell lines with protein–protein interaction network. *Briefings in bioinformatics*, 23(2):bbab587.

Wang, Z., Dong, J., Wu, L., Dai, C., Wang, J., Wen, Y., Zhang, Y., Yang, X., He, S., and Bo, X. (2023b). Deml: drug synergy and interaction prediction using ensemble-based multi-task learning. *Molecules*, 28(2):844.

Weininger, D. (1988). Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36.

Whalen, S., Schreiber, J., Noble, W. S., and Pollard, K. S. (2021). Navigating the pitfalls of applying machine learning in genomics. *Nature Reviews Genetics*, 23(3):169–181.

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., Pon, A., Knox, C., and Wilson, M. (2017). Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Research*.

Wu, L., Gao, J., Zhang, Y., Sui, B., Wen, Y., Wu, Q., Liu, K., He, S., and Bo, X. (2023). A hybrid deep forest-based method for predicting synergistic drug combinations. *Cell Reports Methods*, 3(2).

Xu, M., Zhao, X., Wang, J., Feng, W., Wen, N., Wang, C., Wang, J., Liu, Y., and Zhao, L. (2023). Dffndds: prediction of synergistic drug combinations with dual feature fusion networks. *Journal of Cheminformatics*, 15(1).

Yadav, B., Wennerberg, K., Aittokallio, T., and Tang, J. (2015). Searching for drug synergy in complex dose–response landscapes using an interaction potency model. *Computational and structural biotechnology journal*, 13:504–513.

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., and Barzilay, R. (2019). Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388.

Yüksel, A., Ulusoy, E., Ünlü, A., and Doğan, T. (2023). Selformer: molecular representation learning via selfies language models. *Machine Learning: Science and Technology*, 4(2):025035.

Zagidullin, B., Aldahdooh, J., Zheng, S., Wang, W., Wang, Y., Saad, J., Malyutina, A., Jafari, M., Tanoli, Z., Pessia, A., and Tang, J. (2019). Drugcomb: an integrative cancer drug combination data portal. *Nucleic Acids Research*, 47(W1):W43–W51.

Zhang, Q., Ding, K., Lyv, T., Wang, X., Yin, Q., Zhang, Y., Yu, J., Wang, Y., Li, X., Xiang, Z., et al. (2024). Scientific large language models: A survey on biological & chemical domains. *arXiv preprint arXiv:2401.14656*.

Zhang, Q.-Q., Zhang, S.-W., Feng, Y.-H., and Shi, J.-Y. (2023). Few-shot drug synergy prediction with a prior-guided hypernetwork architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9709–9725.

Zhang, T., Zhang, L., Payne, P. R., and Li, F. (2021). Synergistic drug combination prediction by integrating multiomics data in deep learning models. *Translational bioinformatics for therapeutic development*, pages 223–238.

Zhang, W., Xue, Z., Li, Z., and Yin, H. (2022). Dce-dforest: A deep forest model for the prediction of anticancer drug combination effects. *Computational and Mathematical Methods in Medicine*, 2022(1):8693746.

Zheng, S., Aldahdooh, J., Shadbahr, T., Wang, Y., Aldahdooh, D., Bao, J., Wang, W., and Tang, J. (2021). Drugcomb update: a more comprehensive drug sensitivity data repository and analysis portal. *Nucleic Acids Research*, 49(W1):W174–W184.

APPENDIX A

Chemical Features of MatchMaker

The chemical features of drugs are derived using the PyBioMed library, which offers a wide range of tools for calculating molecular descriptors. These descriptors include various chemical and physical properties of the molecules. A total of 12 distinct descriptor categories, comprising 367 individual features, are computed using the available functionalities in the library. Table A.1 provides a detailed breakdown of these descriptor categories and their corresponding feature counts.

Table A.1 Drug chemical structure features collected from PyBioMed.

Descriptor Category	Count
Kappa Descriptors	7
Charge Descriptors	25
Connectivity Descriptors	44
Constitution Descriptors	28
Geary Descriptors	32
MOE Descriptors	59
Moran Descriptors	32
Moreau-Broto Descriptors	32
Topology Descriptors	19
Molecular Properties	4
Basak Descriptors	21
Burden Descriptors	64

Performance Comparisons of Original Features vs. OHE Across Models

Table A.2 Performance Comparison of DeepSynergy Model Using Drug & Cell Line Features vs OHE Representations on O'Neil Dataset dataset with LPO split.

Feature Type	**MSE**	**PCC**	**SCC**
Drug & Cell Line Features	209.44	0.72	0.72
One-Hot Encoded	199.12	0.74	0.72

Notes: Chemical descriptors (ECFP6 molecular fingerprints, physicochemical properties, and toxicophore features) were originally used to represent drug features, while cell line gene expression levels represented cell line features in the DeepSynergy model. These representations serve as the baseline for comparison with one-hot encoded drug and cell line representations.

Table A.3 Performance Comparison of MARSY Model Using Drug & Cell Line Features vs OHE Representations on DrugComb Dataset dataset with LPO split. *Results using original features were sourced from its publications.

Feature Type	**MSE**	**PCC**	**SCC**
Drug & Cell Line Features	31.58*	0.88*	0.75*
One-Hot Encoded	30.19	0.88	0.75

Table A.4 MARSY single response prediction results from OHE experiment. RS1 and RS2 represent relative inhibition responses for drugs in a pair.

Metric	**MSE**	**PCC**	**SCC**
RS1	35.05	0.94	0.92
RS2	35.03	0.94	0.92

Table A.5 Performance Comparison of JointSyn Model Using Drug & Cell Line Features vs OHE Representations on O'Neil Dataset dataset with LTO split. *Results using original features were sourced from its publications.

Feature Type	**MSE**	$\mathbf{R^2}$	**PCC**
Drug & Cell Line Features	76.56*	0.78*	0.89*
One-Hot Encoded	77.38	0.78	0.88

Table A.6 Performance Comparison of DeepDDS Model Using Drug & Cell Line Features vs OHE Representations on O'Neil Dataset dataset with LTO split.

Feature Type	**AUC**	**PR_AUC**	**ACC**	**BACC**	**PREC**	**TPR**	**KAPPA**
Drug & Cell Line Features	0.93	0.93	0.85	0.85	0.86	0.83	0.70
One-Hot Encoded	0.94	0.93	0.86	0.86	0.86	0.84	0.72

Notes: Molecular graphs and atom-related properties were originally used to represent drug features, while cell line gene expression levels represented cell line features in the DeepDDS model. These representations serve as the baseline for comparison with one-hot encoded drug and cell line representations.

Standard Error of MSE

The standard error of the MSE (SE_{mse}) is calculated based on the standard deviation (σ) of the squared errors and the sample size N, using the formula:

$$SquarredError = (\text{truth} - \text{pred})^2$$

$$SE_{\text{mse}} = \frac{\sigma_{\text{squared\_error}}}{\sqrt{N}}$$

Standard Error for MARSY

For the MARSY model, the authors reported the RMSE across 5 folds as 5.62 ± 0.15. To compute the standard error of MSE (SE_{mse}), we must first convert RMSE to MSE:

1.1 Convert RMSE to MSE:

$$MSE = RMSE^2$$

1.2 Propagate the error during the RMSE-to-MSE conversion:

$$\sigma_{\text{mse}} = 2 \times RMSE \times \sigma_{\text{rmse}}$$

Substituting values:

$$\sigma_{\text{mse}} = 2 \times 5.62 \times 0.15 = 1.686$$

1.3 Calculate SE_{mse}:

$$SE_{\text{mse}} = \frac{\sigma_{\text{mse}}}{\sqrt{5}} = \frac{1.686}{\sqrt{5}} = 0.76$$

Standard Error for JointSyn

For the JointSyn model, the authors reported the MSE across 10 replicates as 76.562, along with the 95% confidence interval (CI): $75.612, 77.511$.

$$\text{CI} = \text{mean} \pm z \times SE$$

where $z = 1.96$ for a 95% confidence interval, the standard error SE can be derived as:

$$75.612 = 76.562 - (1.96 \times SE)$$

$$77.511 = 76.562 + (1.96 \times SE)$$

Solving for SE:

$$(77.511 - 75.612) = 2 \times (1.96 \times SE)$$

$$SE = \frac{1.899}{2 \times 1.96} = 0.48$$

Percent Deviation and Its Standard Error

The percent deviation is calculated as:

$$PercentDeviation = \frac{y_1 - y_2}{y_2}$$

where y_1 is the OHE value and y_2 is the reference.

To compute the standard error of percent deviation:

2.1 Compute the squared error difference:

$$\Delta = (\text{pred}_1 - \text{truth})^2 - (\text{pred}_2 - \text{truth})^2$$

2.2 Standard error of Δ:

$$SE_\Delta = \frac{\sigma_\Delta}{\sqrt{N}}$$

2.3 Scale by the reference MSE (MSE_2) to express as a percentage:

$$SE_{\text{percent deviation}} = \frac{SE_\Delta}{\text{MSE}_2} \times 100$$