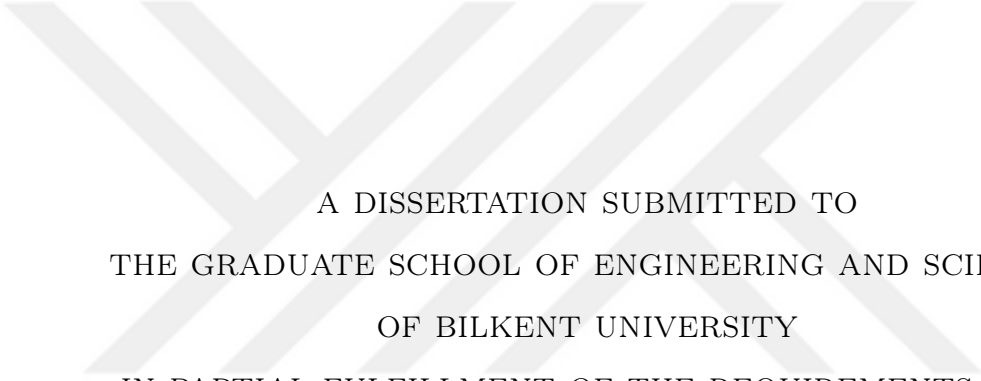


DEEP LEARNING-BASED METHODS FOR DRUG SYNERGY PREDICTION



A DISSERTATION SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
COMPUTER ENGINEERING

By
Halil İbrahim Kuru
June 2025

DEEP LEARNING-BASED METHODS FOR DRUG SYNERGY
PREDICTION

By Halil İbrahim Kuru

June 2025

We certify that we have read this dissertation and that in our opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

A. Ercüment Çiçek(Advisor)

Öznur Taştan(Co-Advisor)

Selim Aksoy

Can Alkan

Arzucan Özgür

Tunca Doğan

Approved for the Graduate School of Engineering and Science:

Orhan Arıkan
Director of the Graduate School

ABSTRACT

DEEP LEARNING-BASED METHODS FOR DRUG SYNERGY PREDICTION

Halil İbrahim Kuru

Ph.D. in Computer Engineering

Advisor: A. Ercüment Çiçek

Co-Advisor: Öznur Taştan

June 2025

Combination therapy involves administering multiple drugs concurrently and offers enhanced efficacy, reduced toxicity, and mitigation of drug resistance. However, as the number of possible drug combinations scales exponentially with the size of the drugs, the identification of drugs that would work synergistically remains an experimental challenge. This thesis presents a structured progression of three deep learning frameworks that advance *in-silico* drug-drug synergy prediction by improving predictive accuracy, enhancing clinical relevance, and deepening interpretability.

The first model, MatchMaker, learns drug representations conditioned on gene expression profiles to capture cellular context jointly with the chemical structure. Benchmarking on the DrugComb dataset demonstrates that MatchMaker reduces the mean squared error of predicting drug pair synergy scores on cell lines by over one-third relative to the state-of-the-art models. Matchmaker establishes a robust and scalable baseline for high-throughput combinatorial screening on cell lines.

The models trained on cell lines like MatchMaker can predict synergistic drug pairs, but do not model patient heterogeneity. Building upon its foundation, the second model—the Personalized Deep Synergy Predictor (PDSP)—addresses the critical need for patient-specific drug synergy prediction for the first time in the literature. Patient-level synergy data is scarce and prohibits training deep learning models. PDSP employs a multi-task learning strategy, coupling drug synergy prediction with a monotherapy response prediction task for which patient-level data is more abundant. PDSP applies transfer learning to fine-tune model parameters using such *ex-vivo* monotherapy data from individual patients. In leukemia samples, PDSP achieves a 27% increase in accuracy over state-of-the-art baselines such as MatchMaker with a low false-positive rate. PDSP demonstrated the feasibility of personalized combination prediction with limited patient-specific measurement data for the first time.

Both Matchmaker and PDSP use aggregated synergy scores that are calculated over the dose-response matrix, where rows and columns corresponds to different dosages of the given drug combination, and each entry in the matrix contains a measurement of the biological response (i.e., cell viability or inhibition) when we combine the two drugs at specific doses. The third model, DeepSynBa, goes beyond the single exaggerated synergy scores to the entire dose-response landscape. By embedding a two-dimensional generalized Hill equation within a deep learning architecture, DeepSynBa decouples the distinct dimensions of synergistic potency and efficacy. It enables flexible, post hoc computation of a range of drug interaction metrics and delivers mechanistic insight into the nature and extent of the interaction. On the NCI-ALMANAC dataset, DeepSynBa reduces dose-surface prediction error by 40–60% relative to the state-of-the-art tensor factorization methods and accurately localizes dose regions of maximal synergy.

Collectively, these three models present a coherent trajectory—from scalable general drug combination prediction models to personalized predictors, to dose-resolved and mechanistically grounded inference. Beyond oncology, the underlying methodological innovations—spanning multi-task learning, transfer learning, and mechanistic parameterization—are readily generalizable to other therapeutic domains such as infectious, metabolic, and neurological disorders, where combination regimens are standard care. The approaches developed in this work represent a significant step toward accelerating the high-throughput screening efforts in identifying synergistic drug pairs. These models also pave the way to the clinical deployment of AI-driven decision-support systems capable of recommending what to combine, for whom, and at which doses. In doing so, they promise to accelerate the rational design of combination therapies that are more effective, safer, and more precisely tailored to individual patients.

Keywords: *In-silico* Drug Synergy Prediction, Deep Learning, Combination Therapy, Personalized Medicine.

ÖZET

İLAÇ SİNERJİSİ TAHMİNİ İÇİN DERİN ÖĞRENME TABANLI YÖNTEMLER

Halil İbrahim Kuru

Bilgisayar Mühendisliği, Doktora

Tez Danışmanı: A. Ercüment Çiçek

İkinci Tez Danışmanı: Öznur Taştan

Haziran 2025

Kombinasyon tedavisi, birden fazla ilacın eş zamanlı uygulanmasını içerir ve artırılmış etkinlik, azalmış toksisite ile ilaç direncinin engellenmesi gibi avantajlar sunar. Ancak, olası ilaç kombinasyonlarının sayısı ilaç setinin büyüklüğüyle üstel olarak arttığından, hangi ilaç çiftlerinin sinerjik çalışacağını belirlemek deneysel açıdan ciddi bir güçlük oluşturmaktadır. Bu tez, öngörü doğruluğunu artıran, klinik geçerliliği güçlendiren ve yorumlanabilirliği derinleştiren üç derin öğrenme çerçevesinin yapılandırılmış bir ilerleme dizisini sunarak *in-silico* ilaç-ilac sinerjisi tahminini ileriye taşımaktadır.

İlk model olan MatchMaker, kimyasal yapı bilgisiyle hücresel bağlamı birlikte yakalayabilmek için gen ifade profillerine koşullanmış ilaç temsilleri öğrenir. Drug-Comb veri kümesi üzerinde yapılan kıyaslamalar, MatchMaker'ın hücre hatlarındaki ilaç çifti sinerji skorlarını tahmin ederken ortalama karesel hatayı çağdaş en iyi modellere kıyasla üçte birden fazla azalttığını göstermektedir. Böylece MatchMaker, hücre hatlarında yüksek verimli kombinatoriyal taramalar için sağlam ve ölçeklenebilir bir başlangıç noktası sunar.

MatchMaker gibi hücre hatlarında eğitilmiş modeller sinerjik ilaç çiftlerini öngörebilse de hasta heterojenliğini modelleyemez. Bu eksikliği gidermek üzere geliştirilen ikinci model — Personalized Deep Synergy Predictor (PDSP) — literatürde ilk kez hasta-spesifik ilaç sinerjisi tahmini ihtiyacını ele alır. Hasta düzeyinde sinerji verilerinin kıt olması, derin öğrenme modellerinin doğrudan eğitilmesini engeller. PDSP, ilaç sinerjisi tahminini, hasta düzeyinde daha bol bulunan tek-ilac yanıtı tahmini göreviyle birleştiren çok görevli (multi-task) bir öğrenme stratejisi uygular. Model parametreleri, bireysel hastalardan elde edilen bu tür *ex-vivo* monoterapi verileriyle transfer öğrenme yoluyla ince ayar yapılır. Lösemi örneklerinde PDSP, MatchMaker gibi çağdaş en iyi modellere kıyasla doğruluğu %27 oranında artırırken düşük yalancı-pozitif oranını korur. PDSP,

sınırlı hasta-spesifik ölçüm verileriyle kişiselleştirilmiş kombinasyon tahmininin uygulanabilirliğini ilk kez göstermiştir.

Hem MatchMaker hem de PDSP, verilen ilaç kombinasyonunun farklı dozlarına karşılık gelen satır ve sütunlardan oluşan doz-yanıt matrisinin tamamı üzerinden hesaplanan toplu sinerji skorlarını kullanır; matrisin her bir girdisi belirli dozlarda iki ilacın birlikte uygulanması sonrası biyolojik yanıtı (örneğin hücre canlılığı veya inhibisyon) içerir. Üçüncü model olan DeepSynBa, tek bir özet sinerji skorunun ötesine geçerek tüm doz-yanıt manzarasını (landscape) tahmin eder. İki boyutlu genelleştirilmiş Hill denklemini derin öğrenme mimarisine gömerek DeepSynBa, sinerjik potentlik ve etkinliğin (potency–efficacy) farklı boyutlarını birbirinden ayırır. Bu sayede çeşitli ilaç etkileşim metrikleri esnek biçimde sonradan hesaplanabilir ve etkileşimin niteliği ile kapsamı hakkında mekanistik içgörüler sunulur. NCI-ALMANAC veri kümesi üzerinde DeepSynBa, doz-yüzey tahmin hatasını çağdaş tensör faktörizasyon yöntemlerine göre %40–60 oranında azaltır ve en yüksek sinerjiye sahip doz bölgelerini doğru biçimde saptar.

Bu üç model birlikte tutarlı bir yol haritası sunar: ölçeklenebilir genel kombinasyon tahmincilerinden kişiselleştirilmiş öngörülere, oradan da doz-çözünürlüklü ve mekanistik temelli çıkarımlara. Onkoloji alanının ötesinde, çok görevli öğrenme, transfer öğrenme ve mekanistik parametrelendirmeyi kapsayan temel yöntemsel yenilikler; kombinasyon tedavisinin standart olduğu enfeksiyöz, metabolik ve nörolojik hastalıklar gibi diğer terapötik alanlara kolaylıkla uyarlanabilir. Bu çalışmada geliştirilen yaklaşımlar, sinerjik ilaç çiftlerini belirlemede yüksek verimli tarama çalışmalarını hızlandırmaya yönelik önemli bir adım teşkil eder. Ayrıca, neyin, kimin için ve hangi dozda birleştirileceğini önerebilen yapay zekâ destekli karar destek sistemlerinin klinik uygulamasının önünü açar. Böylece, daha etkili, daha güvenli ve bireylere daha hassas uyarlanmış kombinasyon tedavilerinin rasyonel tasarımı hızlandırmayı vaat ederler.

Anahtar sözcükler: *In-silico* İlaç Sinerjisi Tahmini, Derin Öğrenme, Kombinasyon Tedavisi, Kişiselleştirilmiş Tıp.

Acknowledgement

First and foremost, I would like to express my deepest gratitude to my advisors, Ercüment Çiçek and Öznur Taştan, for their invaluable support, guidance, and mentorship throughout my doctoral journey. Their encouragement, scientific insight, and patience have played a central role in shaping this thesis.

I am also sincerely thankful to Can Alkan, Arzucan Özgür, Selim Aksoy, and Tunca Doğan for their thoughtful feedback, constructive suggestions, and for generously accepting to serve on my thesis committee. Their input has greatly enriched the quality and direction of my research.

I would like to extend my appreciation to Magnus Rattray, Marta Milo, Haoting Zhang, and Carl Henrik Ek for the collaboration on the DeepSynBa project. I am also grateful to Deniz Cem Önduygu for his wonderful work on the pipeline and result figures.

To my beloved wife Didem and my daughter Sare—thank you for bringing joy, strength, and purpose into my life. Your unwavering love and support have made me feel like the luckiest person in the world, and this thesis would not have been possible without your patience, encouragement, and belief in me.

Finally, I dedicate this work to the memory of my mother, who lost her life to cancer. Words cannot express my love, gratitude, and longing for her. I will always strive to honor her, and I hope that the scientific contributions in this thesis can, in some small way, contribute to advancing cancer research and treatment.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Contributions	3
2	Background	7
2.1	Large-scale experimental drug-combination screens	7
2.2	Synergy Scoring Models	8
2.3	Data representations for drugs and cell lines	9
2.4	Evaluation Scenarios	11
3	Related Work	12
3.1	Traditional Machine Learning Methods	13
3.2	Deep Learning for Synergy Score Prediction	14
3.3	Dose-Response Matrix Prediction	16
4	MatchMaker: A Novel Deep Learning Approach to Drug Synergy Prediction	19
4.1	Methodology	19
4.1.1	Problem Definition and Formulation	19
4.1.2	Drug Combination Dataset	20
4.1.3	Filtering and Preprocessing	21
4.1.4	Input Representations	24
4.1.5	The MatchMaker Architecture	26
4.1.6	Training Objective and Loss Function	29
4.1.7	Compared Methods and Baselines	30
4.2	Results	32

4.2.1	Experimental Setup	32
4.2.2	Performance Comparison	35
4.2.3	Robustness and Stability of Performance	44
4.2.4	MatchMaker Performance Across Cell Lines and Drug Pairs	46
4.2.5	Novel Predictions	51
4.2.6	Evaluation on Other Datasets	53
5	Personalized Drug Synergy Prediction (PDSP): Towards Patient-Specific Models	57
5.1	Methodology	57
5.1.1	Problem Definition and Formulation	57
5.1.2	Dataset Construction and Preprocessing	60
5.1.3	The Model Architecture	62
5.1.4	Model Training and Fine-Tuning Strategy	67
5.1.5	Compared Methods and Baselines	71
5.2	Results	72
5.2.1	Experimental Setup	72
5.2.2	Performance Benchmark	72
6	DeepSynBa: A deep learning method to predict dose-response matrices for drug combinations	84
6.1	Methodology	85
6.1.1	Problem Formulation	85
6.1.2	Dataset and Processing	86
6.1.3	DeepSynBa Architecture	87
6.1.4	Training details	92
6.2	Results	94
6.2.1	Case Studies	94
6.2.2	Compared Methods	96
6.2.3	Performance Evaluation Setup	97
6.2.4	Drug-Response Matrix Prediction Performance	98
6.2.5	Synergy Score Prediction Performance	105
7	Conclusion	109

7.1	MatchMaker: Advancing Accuracy and Generalization in Synergy Prediction	109
7.2	PDSP: Personalizing Drug Synergy Prediction via Transfer Learning	110
7.3	DeepSynBa: Predicting Dose-Response Surfaces and Disentangling Synergy Mechanisms	112
7.4	Integration with Personalized Medicine and Future Directions . .	114

A	Code	132
----------	-------------	------------



List of Figures

4.1	The distribution of Loewe scores in the processed DrugComb dataset. The final dataset comprises 286,421 drug combination experiments—each involving a pair of drugs and a cell line—spanning 3,040 unique drugs tested across 81 cell lines.	23
4.2	Overview of the MatchMaker architecture. The drug’s chemical features are concatenated with the cell line’s gene expression profile for each drug-cell line pair and passed through drug-specific subnetworks (DSNs). These DSNs independently learn drug representations conditioned on the cellular context. The resulting representations are then combined and input into the Synergy Prediction Network (SPN), which consists of fully connected layers that predict the Loewe synergy score. The model is trained end-to-end by minimizing a weighted mean squared error (MSE) between the predicted and true Loewe values.	27
4.3	Comparison of model performance using mean squared error (MSE), Pearson correlation, and Spearman correlation coefficients. Among the evaluated models, MatchMaker achieved the best results with an MSE of 79.49, a Pearson correlation of 0.79, and a Spearman correlation of 0.74.	38

4.4	Receiver Operating Characteristic (ROC) curves for all evaluated methods in the classification task. Combinations with Loewe scores greater than t are classified as positive (synergistic), while those with scores less than $-t$ are classified as negative (antagonistic). The figure also reports the corresponding Area Under the Curve (AUC) and partial AUC (pAUC) values computed up to a false positive rate (FPR) of 0.1.	42
4.5	Area Under the Receiver Operating Characteristic (ROC) Curve for MatchMaker, along with the confidence interval, in the classification task using a threshold 10.	46
4.6	Cell line-specific MSE performance of MatchMaker is shown. . . .	48
4.7	Cell line specific performances. The figures show the (a) MSE, (b) Pearson, and (c) Spearman correlation performances for each cell line.	49
5.1	Overview of the PDSP framework pipeline: 1) The architecture and training procedure of the PDSP model, which is initially trained on drug combination data using cell line features. 2) Fine-tuning of the pretrained PDSP model. The pretrained PDSP model is adapted by updating the drug encoders using patient-specific single-drug response data and gene expression profiles, either collectively across all patients (Step 2.a) or individually for each patient (Step 2.b). 3) Synergy prediction is then performed using the fine-tuned models, enabling personalized evaluation of drug combinations.	59
5.2	Classification results on patient drug combination synergy data. All models are pretrained on DrugComb using cell line gene expression profiles. PDSP ^{st1} and PDSP ^{st2} represent personalized variants fine-tuned using Strategy 1 and Strategy 2, respectively. In the figure, correct predictions are shown with blue rectangles, while incorrect predictions are highlighted in red. Synergistic outcomes are marked with a plus (+) symbol, and antagonistic outcomes with a minus (−) symbol.	80

6.1	Overview of the DeepSynBa architecture. Drug and cell line features are first concatenated and passed through a series of fully connected layers to generate a unified latent representation. This representation is then processed by multiple prediction heads that output eight parameters defining the dose-response surface. During training, DeepSynBa uses the mean function from the SynBa likelihood model [1] to reconstruct the full dose-response matrices based on these predicted parameters. Model optimization is performed by minimizing the mean squared error (MSE) between the predicted and experimentally measured dose-response matrices. .	88
6.2	Predicted dose-response outputs from DeepSynBa for two distinct drug-cell line combinations, alongside comparisons with corresponding ground-truth experimental measurements. Panels (a)–(d) illustrate the combination of Drug 1 (Bleomycin Sulfate) and Drug 2 (Cabazitaxel) on cell line 786-0, while panels (e)–(h) depict the combination of Drug 1 (Azacitidine) and Drug 2 (Triethylenemelamine) on cell line MDA-MB-468. (a, e): Predicted dose-response surfaces with measured responses overlaid in red. (b, f): Contour plots showing predicted drug effects across varying dosage levels. (c, g): Dose-response curves at fixed concentrations of Drug 2 (in μM), comparing model predictions with experimental values. (d, h): Dose-response curves at fixed concentrations of Drug 1 (in μM), comparing predicted and measured responses. .	91
6.3	The distributions of RMSE, Pearson, and Spearman correlation metrics for dose-response surface predictions, shown across different cell lines (top panel) and drugs (bottom panel).	101
6.4	The ground-truth and the predicted response distributions for different drug type combinations.	106

List of Tables

4.1	The details of the DrugComb Dataset	22
4.2	The data statistics of the studies after preprocessing the DrugComb Dataset. DrugComb curates the NCATS studies from NIH NCATS Tripod, and we refer all other studies with their references.	24
4.3	Drug chemical structure features collected from ChemoPy	26
4.4	Hyperparameters for DeepSynergy. The input data is normalized either by <i>tanh_norm</i> or <i>tanh</i> as given at <i>Preuer et al.</i> . <i>tanh</i> first standardizes and then applies the hyperbolic tangent. <i>tanh_norm</i> first standardizes, applies hyperbolic tangent, and standardizes again.	33
4.5	Hyperparameters for TreeCombo. The input data is normalized with <i>tanh_norm</i> . The number of estimator parameters is set to 1000 as given at <i>Janizek et al.</i> . Maximum tree depth and learning rate ranges used the same as <i>Janizek et al.</i>	33
4.6	Hyperparameters for Random Forest. The input data is normalized with <i>tanh_norm</i>	34
4.7	Hyperparameters for ElasticNet. The input data is normalized with <i>tanh_norm</i>	34
4.8	Comparison of model performance in the classification task using three evaluation metrics: Area Under the ROC Curve (AUC), partial AUC up to a false positive rate of 0.1 (pAUC), and Area Under the Precision-Recall Curve (AUPR). Combinations with Loewe scores above the threshold t are labeled as positive (synergistic), while those with scores below $-t$ are labeled as negative (antagonistic).	45

4.9	The top predicted drug pairs based on MSE.	50
4.10	List of high-confidence false positives identified as novel synergistic drug pair predictions. These combinations exhibit an actual Loewe score below -10 and a MatchMaker-predicted Loewe score above +10 across all 10 random model initializations. The following five columns report additional sensitivity and synergy metrics: CSS, S, ZIP, BLISS, and HSA, with the latter four sourced from the DrugComb database. The final column summarizes the level of support from these alternative metrics, where the number of stars reflects how many of them also indicate a synergistic interaction.	51
4.11	O’Neil and NCI ALMANAC datasets details	53
4.12	Performance comparison of the methods in the O’Neil and NCI ALMANAC datasets based on MSE, Pearson, and Spearman correlation coefficients.	53
5.1	Summary of the DrugComb dataset and patient drug combination datasets following preprocessing. Each data instance represents a combination of two drugs and a single cell line (or patient sample).	61
5.2	Comparison of model performance in synergy prediction using mean squared error (MSE), Pearson correlation, and Spearman correlation coefficients.	74
5.3	Top 10 cell line performance of different models	76
5.4	Comparison of classification performance across models using various evaluation metrics. Drug combination experiments with a Bliss score greater than 0 are labeled as positive (synergistic), while those below 0 are considered negative (antagonistic).	77
6.1	The search space used for hyperparameter and training parameter optimization of the DeepSynBa architecture. This includes ranges and candidate values for network components, as well as optimization settings explored during model tuning.	93

6.2	The comparison of model performance across three evaluation scenarios: (a) new drug combination, (b) new cell line, and (c) new drug. Detailed descriptions of these scenarios are provided in Section 6.2.3.	99
6.3	The pairwise t-test p -value results for drugs and cell lines	102
6.4	Performance results across different tissue types under the new drug combination scenario.	103
6.5	Performance results across different drug combination categories under the new drug combination scenario.	104
6.6	Performance comparison using Loewe and Bliss synergy scores. Ground-truth Loewe/Bliss values are computed from experimentally measured dose-response matrices. For DeepSynBa, ComboFM, and ComboLTR, predicted synergy scores are derived from their respective predicted dose-response matrices. In contrast, MatchMaker and DeepSynergy are trained directly on Loewe/Bliss scores as regression targets.	107

Chapter 1

Introduction

1.1 Motivation

Complex diseases such as cancer, diabetes, and HIV pose significant challenges to treatment due to their complex molecular nature. These diseases often involve multiple cellular mechanisms that interact in complicated ways, making it difficult to achieve effective therapy with single-target drugs. It is particularly true in cancer, where drug resistance mechanisms frequently evolve, rendering single-drug therapies ineffective over time [2, 3]. Resistance arises due to various factors, including genetic mutations, tumor heterogeneity, and adaptive cellular responses, leading to a pressing need for new treatment strategies. In response, combination therapies, which involve concurrently using multiple drugs, have become an essential tool in the fight against complex diseases. These therapies are particularly beneficial because they offer several advantages over single-drug treatments, including enhanced efficacy, reduced toxicity, and overcoming drug resistance [4, 5].

In recent years, drug combinations have become the standard treatment strategy for many diseases beyond cancer, such as HIV [6], diabetes [7], and, more recently, COVID-19 [8]. By using multiple drugs that act on different molecular

targets, these therapies can create a synergistic effect, where the combined impact of the drugs is greater than the sum of their individual effects [9]. Synergy between drugs not only enhances therapeutic outcomes but also reduces the risk of adverse side effects by lowering the dosage required for each drug [3]. Moreover, combinatorial strategies allow overcoming drug resistance, a critical challenge in cancer treatment. Despite these benefits, identifying effective drug combinations is a complex and resource-intensive process, requiring extensive experimental testing.

The drug discovery process is already a complex and costly endeavor, with many candidate drugs failing during clinical trials due to issues of efficacy or safety [10, 11]. The challenge becomes even greater when attempting to discover effective drug combinations. The number of possible drug pairings grows exponentially as more drugs are considered, making exhaustive experimental testing impractical even with high-throughput screening technologies [12, 13]. In cancer, where the genetic and phenotypic diversity among patients further complicates treatment, the task of identifying patient-specific drug combinations is even more difficult [14]. It has prompted the development of computational approaches to predict drug synergies and prioritize combinations for experimental validation.

Computational methods have emerged as essential tools for drug discovery, offering a way to sift through the vast numbers of potential drug combinations without exhaustive laboratory testing. Early approaches to predicting drug synergy relied on classical machine learning models, mathematical frameworks, and network-based techniques [15, 16]. These methods typically use various data types, including chemical properties of drugs, gene expression profiles from cancer cell lines, drug-target interactions, and protein-protein interaction networks [17, 18]. While these models have shown promise, their performance is often limited by the quality and completeness of the available data. Biological networks, for example, are frequently incomplete, and gene expression profiles from cell lines may not fully capture the complexities of patient-specific biology [19].

In response to these limitations, recent advances in deep learning have opened up new possibilities for drug synergy prediction. Deep learning models, which learn complex representations from large datasets, have performed better than

earlier methods. The availability of large-scale drug synergy datasets, such as the AstraZeneca-Sanger [20] and NCI-ALMANAC [21] datasets, has fueled the development of more sophisticated models capable of making accurate predictions based on chemical structures, gene expression data, and other biological features.

One of the key challenges in drug synergy prediction is learning how the individual characteristics of two drugs interact within a particular biological context, such as a cancer cell line. Traditional models, like DeepSynergy [22] and TreeCombo [23], concatenate the features of the two drugs into a single input, which is then fed into a prediction model. While this approach can capture some of the interactions between the drugs, it fails to fully exploit the potential of learning drug-specific representations.

1.2 Contributions

To address the limitations of traditional models, which are not able to learn drug-specific representations, we developed *MatchMaker* [24], a deep neural network-based model that learns separate representations for each drug in a pair and then combines these representations to predict drug synergy scores. The model consists of two parallel subnetworks, each of which processes one drug in the pair. We design these subnetworks to learn drug-specific features based on chemical structure and untreated gene expression profiles of the relevant cancer cell lines. The resulting drug-specific embeddings are then combined in a third subnetwork, which predicts the synergy score for the drug pair. This approach allows MatchMaker to capture the unique properties of each drug, leading to more accurate predictions of drug synergy.

In our experiments, MatchMaker achieved the state-of-the-art performance on the largest publicly available drug combination dataset [25], reducing the Mean Squared Error (MSE) by 33% and improving Pearson and Spearman correlation coefficients by 13% and 14%, respectively, compared to previous models. MatchMaker can identify novel drug synergies not annotated in the ground truth dataset

but supported by literature, highlighting its potential to uncover new therapeutic combinations. We publish MatchMaker in the IEEE/ACM Transactions on Computational Biology and Bioinformatics, and the work is also presented at the joint Intelligent Systems for Molecular Biology (ISMB) / European Conference on Computational Biology (ECCB) 2021, as part of the Machine Learning in Computational and Systems Biology (MLCSB) track.

While MatchMaker performs well in predicting synergies for cancer cell lines, one of the major challenges in applying these predictions to clinical settings is the variation in patient responses to drug combinations. While useful as model systems, cancer cell lines do not fully capture human patients’ genetic and phenotypic diversity. This variation complicates the task of predicting how a particular drug combination will perform in an individual patient [19]. To bridge this gap, we developed *Personalized Deep Synergy Predictor (PDSP)* [26], an extension of MatchMaker designed to predict patient-specific drug synergies.

PDSP uses a transfer learning approach to adapt the knowledge learned from large-scale cell line data to patient-specific predictions. In the first stage, PDSP is trained on a large cancer cell line drug synergy dataset. In the second stage, the model is fine-tuned using patient-derived gene expression data and single-drug response measurements, which are more widely available than drug combination data for individual patients. This two-stage process allows PDSP to leverage the extensive data from cell lines while adapting to the specific characteristics of individual patients.

To improve the personalization of synergy predictions, we apply two fine-tuning strategies: (i) using whole patient data to adjust the model for broader patient populations, and (ii) patient-specific fine-tuning, which tailors the model to the individual patient’s data. In our experiments, PDSP demonstrated the ability to accurately predict drug synergies for individual patients, offering a new approach to personalized cancer treatment. When tested on leukemia patient data, PDSP improves synergy prediction accuracy by 27% compared to non-personalized models. Its ability to detect antagonistic combinations, critical for avoiding toxic regimens, emerged by 40%, with precision increasing from 47% (MatchMaker)

to 83% in patient-specific fine-tuning. No prior models have integrated large-scale synergy data with patient-specific measurements in this way, making PDSP a significant contribution to the field. PDSP is published in Bioinformatics by Oxford University Press in 2024. In 2023, we presented PDSP at the 15th RECOMB Satellite Workshop on Computational Cancer Biology (RECOMB-CCB), held in Istanbul, Turkey, where it was among the accepted full papers for oral presentation.

Although PDSP demonstrates the potential of transfer learning to tailor synergy predictions to individual patients, it retains a critical limitation of its predecessor, MatchMaker, which is the dependence on single-score synergy outputs. Although computationally efficient, these aggregated scores do not capture the nuanced interplay between efficacy (maximum achievable therapeutic effect) and potency (the dosage required to achieve a specific effect) at varying dosage levels. This gap highlights the need for models capable of resolving dose-specific drug interactions. To address this challenge, we introduce the DeepSynBa model to predict the complete dose-response matrix.

In contrast to traditional single-score approaches, DeepSynBa represents a different view on synergy prediction by predicting full dose-response matrices. This framework models the complete response surface of drug combinations, explicitly decoupling efficacy from potency. Such granularity offers two key advantages: (i) enhanced reliability: By analyzing the entire dosage spectrum, DeepSynBa reduces prediction uncertainty inherent to single-score models; and (ii) Actionable insights: The resolved dose-response profiles guide experimental design, enabling researchers to identify optimal dosage ranges that maximize efficacy while minimizing off-target toxicity. This innovation advances computational synergy prediction and bridges the gap between in-silico models and clinically actionable therapeutic strategies.

This thesis presents three novel models that advance *the-state-of-the-art* in drug synergy prediction. First, *MatchMaker* significantly improves synergy prediction by learning drug-specific representations, achieving superior performance on large-scale datasets compared to previous models. Second, *PDSP* extends the

capabilities of MatchMaker to the realm of personalized medicine, using transfer learning to fine-tune the model for patient-specific predictions. Finally, *Deep-SynBa* marks a significant advancement in drug synergy prediction by moving beyond scalar synergy scores to model complete dose-response surfaces, thereby allowing dose-aware, mechanistically interpretable, and clinically actionable insights into drug interactions. Together, these models represent an important step toward personalized combination therapies, which address the scalability of drug combination discovery and the need for individualized treatment strategies in cancer.

Chapter 2

Background

2.1 Large-scale experimental drug-combination screens

High-throughput screening of drug pairs has produced large datasets of combination responses in cancer cell lines. For example, the NIH’s NCI-ALMANAC effort systematically tested over 5,000 pairs of FDA-approved cancer drugs (104 distinct agents) across the NCI-60 panel of human tumor cell lines [21]. Likewise, O’Neil et al. (2016) [27] screened 583 paired combinations among 38 drugs on 39 diverse cancer cell lines (including breast, lung, colon, melanoma, ovarian, prostate, and leukemia) using a 4×4 dose–response matrix for each pair. In 2019, AstraZeneca and collaborators released a large combination screen of 910 drug pairs across 85 cancer cell lines, each tested in a 5×5 matrix of doses [28]. Each experiment measures cell viability or growth inhibition after drug exposure, yielding a two-dimensional dose–response surface for each drug pair and cell line. These datasets comprise thousands of combination scenarios (unique drug–drug–cell line tuples) and millions of individual dose–response measurements.

Typically, each screen uses panels of molecularly characterized cell lines (e.g., NCI-60 or CCLE) and measures viability (or inhibition) relative to untreated

controls. Dose–response matrices allow modeling of interaction across a range of doses. These high-quality screens are essential benchmarks and training sets for computational models of drug synergy. By comparing observed combination effects to null models (additivity or independence), researchers identify synergistic (greater-than-additive) and antagonistic (less-than-additive) interactions, and prioritize combinations for further study or clinical evaluation.

2.2 Synergy Scoring Models

Drug synergy means that a drug combination produces a more substantial effect than expected from the single drugs alone. Clinically, synergy is essential because it may allow lower doses (reducing toxicity) or overcome resistance [4, 5]. More formally, two drugs are synergistic if the observed effect of the combination exceeds the predicted effect under a "no-interaction" model. Here we focus on three reference models that we use in this thesis:

- **Bliss Independence [29]:** Assumes each drug acts independently on cell viability. If v_A and v_B are the fractional viabilities under drug A and B alone (at given doses), the expected viability under Bliss is $v_{A\&B}^{\text{exp}} = v_A \times v_B$. Equivalently, if E_A and E_B are fractional inhibitions, then $E_{A+B}^{\text{exp}} = E_A + E_B - E_A E_B$. A Bliss synergy score can be defined as the difference $E_{\text{observed}} - E_{A+B}^{\text{exp}}$ (or similarly in terms of viability). Positive excess (observed effect $>$ expected) indicates synergy. (The NCI-ALMANAC ComboScore uses a variant of this approach [21])
- **Loewe Additivity [30]:** Assumes the two drugs are similar in mechanism and asks: If drug 1 produces effect E at dose D_{E1} and drug 2 at dose D_{E2} , then any additive combination should lie on the isobole defined by $\frac{d_1}{D_{E1}} + \frac{d_2}{D_{E2}} = 1$ for effect E [31]. This equation defines the expected effect, and the deviations from the expected effect indicate synergy when positive, and antagonism when negative.

- **NCI ComboScore [21]:** The NCI-ALMANAC study defined a quantitative ComboScore by extending the Bliss concept over the whole dose-response surface. For each dose pair (p, q) , let Y_{obs} be the observed growth fraction (viability) and $Y_{exp} = Y(A_p) \times Y(B_q)$ the Bliss-predicted growth. Then, the ComboScore becomes the sum of the difference between the expected and observed responses at all tested dose combinations. It is formally defined as: $\sum_{p,q} (Y_{exp}(p, q) - Y_{obs}(p, q))$. A positive ComboScore (more expected growth than observed) means greater-than-additive killing (synergy), while a negative score means antagonism. In ALMANAC, each drug pair-cell line combination has a single ComboScore summarizing the surface.

All of these models rely on accurate single-drug dose-response curves. Dose-response data are usually fit by a Hill (sigmoid) model:

$$E(d) = E_{min} + \frac{E_{max} - E_{min}}{1 + (\frac{IC_{50}}{d})^h} \quad (2.1)$$

where $E(d)$ is effect (inhibition) at dose d , E_{min} the minimum effect, E_{max} the maximum effect, h the Hill slope, and IC_{50} the half-maximal inhibitory concentration. The IC_{50} is the concentration at which the drug produces 50% inhibition of the measured response. It is a standard potency metric in pharmacology (often expressed as $-\log_{10}(IC_{50})$).

2.3 Data representations for drugs and cell lines

Machine-learning models for drug synergy require numerical features for drugs and cell lines. Common representations include:

- **Drug Features:** Drugs can be described by their chemical structures encoded as SMILES strings [32] or molecular graphs. These are converted into descriptors or fingerprints (binary vectors) that capture substructures.

For example, Morgan fingerprints [33] are often used, where each bit indicates the presence of a particular chemical substructure. Other descriptors include physical-chemical properties (molecular weight, lipophilicity, H-bond donors, etc.) or learned molecular embeddings. Public chemical databases supply this data: PubChem (NIH) [34] and ChEMBL (EBI) [35] contain millions of compounds with structures and bioactivity data, including all FDA-approved drugs. These repositories provide standard identifiers, SMILES, and target annotations that can be used to featurize drugs for modeling. Some open-source tools (i.e., ChemoPy [36]) also generate such molecular features by using SMILES representations of the molecules.

- **Cell line features:** Molecular and genomic profiles represent each cancer cell line. Common features include baseline gene expression (e.g., from RNA-seq or microarrays), mutation status of cancer-associated genes, and copy-number variation (CNV) profiles. For instance, the Cancer Cell Line Encyclopedia (CCLE) measured gene expression and DNA copy number in 947 lines [37], and the Genomics of Drug Sensitivity in Cancer (GDSC) project [38] assayed drug responses in ~ 700 lines while integrating mutations and expression. One might use key oncogenes/tumor suppressors' expression levels and binary mutation indicators (e.g., KRAS, TP53) as features. Some approaches also incorporate protein-protein interaction or pathway information. Databases like CCLE/DepMap and GDSC provide precompiled genomic profiles for hundreds of cell lines, enabling models to leverage tissue of origin and molecular context.

Combining these features with the synergy scores above allows computational methods to learn patterns (e.g., drug A and B are synergistic specifically in lines overexpressing gene X). The CCLE and GDSC resources are frequently used for such modeling because they link drug response (IC_{50} values) to genomic data [38, 37]. Likewise, standardized drug representations from PubChem or ChEMBL ensure that chemical similarities and target annotations are captured consistently.

2.4 Evaluation Scenarios

Assume that each data instance in the dataset is structured as a triplet comprising two drugs and one cell line, formally represented as (i, j, k) , where i and j denote the drugs, and k represents the cell line. The robust evaluation of computational models for drug synergy prediction requires careful design of validation strategies. Unlike standard machine learning tasks, where random splits are sufficient, drug combination prediction presents specific challenges related to generalization across drugs, cell lines, and dosage conditions. Improper splitting can lead to inflated performance estimates due to information leakage. Therefore, appropriate cross-validation designs must reflect the intended application scenario of the model, such as predicting the synergy of unseen drug pairs, in novel cell lines, or both. We use the following scenarios in our study.

- **New drug combination:** The individual drug pairs denoted as $\langle i, j \rangle$ are uniquely assigned to one of the data partitions—training, validation, or testing. It ensures that any given combination is only encountered in a single split, preventing data leakage and enabling an unbiased assessment of the model’s capacity to generalize to previously unseen drug pairings.
- **New cell line:** It involves grouping drug combinations according to their associated cell lines. Each cell line k is included in a single partition, such that all drug combinations involving that cell line are contained exclusively within the training, validation, or testing set. This stratification tests the model’s ability to extrapolate its predictions to new cellular contexts not represented during training.
- **New drug:** The test set comprises drug combinations with at least one compound absent from all combinations in the training and validation sets. Similarly, combinations in the validation set must involve at least one drug not present in the training set. This approach simulates real-world situations in which models are applied to predict the effects of newly developed or previously uncharacterized compounds.

Chapter 3

Related Work

Early computational approaches for synergy prediction leveraged traditional machine learning and statistical models, often hand-crafting features from chemical structures, genomic profiles, or protein interaction networks [18, 39, 15, 40, 41, 42, 17, 43, 44, 28, 45, 46]. In recent years, deep learning methods have achieved the state-of-the-art performance by automatically learning complex representations of drugs and biological context [22, 24, 47, 48, 49, 50]. Moreover, the field has seen a broadening of scope: From predicting a single synergy score for a drug pair in a given cell line, towards predicting the full dose-response matrix of a drug combination [51, 52, 53], and from models trained on generic cancer cell lines towards personalized models tuned for individual patients. Below, we provide a comprehensive literature review, organizing methods by their methodological paradigms. We discuss traditional machine-learning and matrix factorization approaches, modern deep learning strategies, entire dose-response surface prediction, multi-task and transfer learning techniques for personalized models, and emerging methods incorporating graph neural networks (GNNs), attention mechanisms, and multi-omics data. We highlight the strengths and limitations of each class of methods and how they have evolved to address the challenges of drug synergy prediction.

3.1 Traditional Machine Learning Methods

Traditional machine learning (ML) techniques are among the first to predict synergistic drug combinations. These methods typically require predefined feature vectors for drugs (e.g. chemical descriptors) and for the biological context (e.g. cell line gene expression profiles), and then learn a mapping to synergy outcomes using algorithms like random forests (RF) [54, 55], or gradient-boosted trees [45, 56]. Early studies explored the various engineered features beyond chemical structure, including drug–target interaction networks, signaling networks, disease ontologies, and omics data. For example, Li et al. (2018) [57] design a model that propagates influence through a cancer gene network to predict synergy, integrating drug–target networks and gene expression; this method successfully identified synergistic pairs by capturing network effects. Other researchers formulate synergy prediction as a standard classification or regression task. One study [58] trained an RF model on known patient-derived combination responses, using as features the predicted dose–response curves of each drug. This RF-based approach prioritizes combinations for acute myeloid leukemia patients by modeling each drug’s efficacy in the patient and then extrapolating to pairs. Such approaches, however, do not leverage large public combo datasets – they are constrained to the small patient-specific data available and cannot easily generalize beyond the given context.

A major benchmark in this category is TreeCombo, an extreme gradient boosting model (XGBoost) introduced by Janizek et al. (2018) [23]. TreeCombo was trained on a large drug-screening dataset with cell line gene expression and produced better accuracy than an earlier deep neural network on the same data. Similarly, Sidorov et al. (2019) applied XGBoost to the NCI-ALMANAC dataset (a large matrix of drug pairs tested across cancer cell lines). They achieved robust performance in identifying synergistic vs. non-synergistic pairs [45]. These boosting and tree-based models have the advantage of handling diverse feature types and capturing non-linear interactions to some extent, while still being faster to train and more interpretable (via feature importance) than very deep models. Indeed, a recent systematic evaluation article [59] finds that a well-tuned random

forest or boosted tree model can reach reasonably high predictive power when combined with biologically informed features. Yet, they still cannot get the level of deep neural networks regarding overall accuracy.

Despite their successes, traditional ML approaches face important limitations. First, they are heavily based on hand-made features. Although many studies engineered extensive feature sets (chemical fingerprints, gene expression signatures, network metrics, etc.), not all of these data are available or reliable in every context. In practice, models that depended on comprehensive molecular interaction networks suffered when those networks were incomplete or noisy [60]. In contrast, simpler choices of characteristics (chemical structure combined with cell baseline gene expression) proved to be surprisingly effective baselines [23, 22, 24]. Second, shallow ML models do not natively account for dose-dependence or higher-order interactions. They treat synergy as a single label per drug pair in a given experimental setup. This means information about how the effect changes with different dose combinations is lost. Third, traditional models require more data preprocessing and may not scale well as the feature space grows. Nevertheless, they played a crucial role in establishing the feasibility of *in silico* synergy prediction. Methods like TreeCombo demonstrated that even without deep architectures, one could achieve meaningful predictive performance by leveraging gene expression and chemical features. These early methods also provided a baseline for later deep models to surpass. Indeed, the general trend has been that deep learning (with sufficient data) can outperform conventional ML by a significant margin, although careful feature engineering and ensembling can narrow the gap [59]

3.2 Deep Learning for Synergy Score Prediction

Deep learning has emerged as the dominant approach for drug synergy prediction in the last few years, due to its ability to integrate multiple data modalities and capture non-linear interactions at scale. Deep learning models for synergy scores generally take as input features for two drugs and the cellular context, and output a predicted synergy metric (such as Bliss or Loewe synergy) for that combination.

A landmark in this area was DeepSynergy by Preuer et al. (2018), which used a multi-layer perceptron to combine Morgan fingerprints of two compounds with the gene expression profile of a cancer cell line [22]. Remarkably, DeepSynergy achieved the state-of-the-art performance at the time despite using only chemical structure and baseline gene expression – no network or pathway features. This demonstrated that a sufficiently expressive neural network can learn the relevant biological interactions from data, alleviating the need for extensive manual feature engineering. Following DeepSynergy, we introduce the first model in this thesis, MatchMaker [24], as an improved deep framework that learns separate drug representations conditioned on the cell line, rather than simply concatenating drug and cell features. In MatchMaker’s design, each drug is passed through its subnetwork with the cell’s gene expression profile; the outputs (drug-specific embeddings) are combined to predict the synergy score. The success of MatchMaker underscored the value of contextualized drug representations – essentially a form of interaction modeling within the network – as opposed to a fixed representation per drug. Other deep learning models have followed, each introducing novel architectural modules or additional data sources. For example, AuDNNsynergy incorporates multi-omics features (mutations, copy number variations, etc.) alongside gene expression and chemical structure, using an autoencoder-based deep network to fuse these inputs [47]. Liu and Xie (2021) introduce TranSynergy, an interpretable DNN that integrates known signaling pathways [48]. It predicts synergy and simultaneously tries to deconvolve which biological pathways contribute to the interaction. This approach yields mechanistic insights (e.g., identifying pathway-level contributions) while maintaining strong predictive performance.

GNN architectures dominate recent synergy prediction models. For example, DeepDDS [49] uses graph attention networks (GAT) to encode each drug’s molecular graph and fuses these with cell-line gene expression profiles via an MLP. MGAE-DC [50] learns drug embeddings by building three graphs per cell line (synergistic, additive, and antagonistic edges) and employing a multi-channel graph autoencoder with attention to fuse cell-line-specific and shared features. CGMS [61] represents each drug pair and cell line as a heterogeneous complete

graph (drug and cell-line nodes) and applies a heterogeneous graph attention network. CGMS is trained with multi-task learning (synergy and sensitivity) and yields order-independent synergy scores. Other GNN-based models include PathSynergy [62], which integrates drug features, cell-line data, drug-target interactions, and signaling pathway maps into a GNN framework, and GraphTransSynergy [63], which uses a graph-transformer to extract drug-pair features and a BiLSTM for cell-line gene expression. These methods can naturally encode chemical structure and molecular interactions. However, GNNs often require large labeled datasets and can struggle with generalizing to new drugs or cell types if not trained carefully.

Several recent works leverage transformers and attention for synergy prediction. DeepTraSynergy [64] applies transformer encoders to learn drug features from interaction networks (drug-target, protein-protein, cell-target) in a multi-task framework (predicting synergy, toxicity, and drug-protein interactions). DFFNDDS [65] uses a pre-trained chemical language model (a BERT-based model [66] trained with SMILES representation of drugs) to encode drug features, and fuses these with cell-line features in a dual feature fusion network. JointSyn [67] uses dual views. One view constructs a joint graph of the two drugs plus the cell line (processed by a GAT), and the other view processes each drug’s fingerprint through a small MLP; both are concatenated with gene expression embeddings for prediction. These attention-based models capture global context (e.g., long-range interactions or sequence features) and often excel when multimodal data are available. A notable example is MGTNSyn [68], which reportedly uses a graph-transformer with relational attention to incorporate 3D molecular geometry. Transformers can handle sequential and structured input flexibly, but they are computationally intensive and may require careful tuning or pretraining.

3.3 Dose-Response Matrix Prediction

While many models output a single synergy score or label for a drug pair, there is a growing recognition that predicting the entire dose-response matrix of a

drug combination can be more informative. A dose–response matrix consists of the cellular responses for a range of concentrations of drug A (rows) and drug B (columns); from this matrix, one can compute synergy metrics like Bliss or Loewe scores at various dose levels. Traditional approaches that predict only a summary synergy score may miss concentration-dependent effects and can be misleading if the tested doses were suboptimal.

Classical approaches fit parametric functions (e.g., sigmoid/Hill curves) to experimental data, but modern ML methods aim to predict entire response curves or surfaces from features of drugs and cells. Multi-output Gaussian Processes (MOGPs) have been applied to model drug response across doses as a correlated function [69]. Rønneberg et al. (2023) [53] proposed PIICM, a permutation-invariant Gaussian process model that directly predicts the full dose–response surface for untested drug pairs. PIICM treats each dose–response curve as a function and uses Gaussian process regression to estimate the surface, while being invariant to the labeling of which drug is A vs B. Notably, PIICM does not incorporate any drug or cell features – it purely learns from the shapes of dose–response functions of known combinations. This yielded a powerful interpolation tool for drug pairs similar (in dose–response behavior) to those seen before. However, as the authors point out, a limitation is that PIICM cannot generalize to entirely new drugs or new cell types because it lacks any mechanism to use chemical or genomic information. In other words, it can only predict for drug–cell combinations within the convex hull of those it was trained on, restricting its practical utility for novel therapies.

Another class of approaches frames drug synergy prediction as a matrix-completion or factorization problem [51, 52]. Here, the idea is to model the (drug A, drug B, cell line) outcome space as a partially observed tensor (or set of matrices) and to infer missing entries through latent factors or statistical assumptions. These methods do not explicitly use high-dimensional feature vectors for drugs or cells; instead, they derive latent representations from the pattern of observed combination responses. A representative example is comboFM, proposed by Julkunen et al. (2020) [51], which uses a factorization machine model

to capture higher-order interactions in drug combination data. comboFM represents each data point in a three-way tensor – drug1, drug2, and cell context – and factorizes this tensor to learn latent embeddings for drugs and cells. Combining the learned latent factors by leveraging multi-way factorization, comboFM can generalize to unseen pairs (e.g., a new drug combination in a known cell line).

Building on this idea, Wang et al. (2021) [52] introduced comboLTR, which replaces the factorization machine with a latent tensor reconstruction approach. comboLTR still assumes a polynomial form of interactions but learns to reconstruct the combination outcome tensor through latent factors more flexibly than classical matrix factorization. Both comboFM and ComboLTR embrace the data sparsity of combination screens by learning compressed representations, effectively filling in missing synergy scores by discovering underlying patterns. These models are advantageous when one has a large but sparse dataset of combination measurements and wants to interpolate or extrapolate without necessarily using explicit chemical or genomic features. They also naturally handle cold start problems like a new drug or new cell line, if that entity’s latent factor can be inferred from some data (e.g., if the new drug has been tested in a few combinations, one can place it in the latent space).

Chapter 4

MatchMaker: A Novel Deep Learning Approach to Drug Synergy Prediction

MatchMaker is a deep learning framework designed to predict drug synergy by capturing how drug effects vary across cellular contexts. Unlike earlier models that concatenate drug and cell line features, MatchMaker employs two drug-specific subnetworks that are conditioned on the gene expression profile of the target cell line. This architecture allows the model to learn context-aware drug representations that reflect the dynamic nature of pharmacological interactions.

4.1 Methodology

4.1.1 Problem Definition and Formulation

Drug synergy prediction is modeled in this study as a supervised regression task. Our objective is to develop a predictive model that can estimate the degree of synergistic interaction between two drugs when applied to a particular cancer

cell line. Each data instance in the task is structured as a triplet comprising two drugs and one cell line, formally represented as (i, j, k) , where i and j denote the drugs, and k represents the cell line.

The goal is to predict a continuous-valued synergy score $y_{i,j,k} \in \mathbb{R}$ for each such triplet. This score reflects the level of interaction between the drugs i and j when co-administered to cell line k . The MatchMaker model specifically was trained to predict the Loewe Additivity Score [30], a well-established pharmacological metric used to quantify synergy. The Loewe score captures how much the observed combined effect of two drugs deviates from the expected additive effect. Positive values indicate synergy (greater-than-additive interaction), zero implies additivity, and negative values suggest antagonism (less-than-additive effect).

Our dataset comprises n unique triplets, each consisting of two drugs, denoted as i and j , and a corresponding cell line k . Each drug is represented by a feature vector, d_i and d_j , capturing its molecular and pharmacological characteristics. Similarly, each cell line is described by a feature vector c_k , encapsulating its genomic or transcriptomic profile. The primary objective is to predict the outcome variable $y_{i,j,k}$, which reflects the interaction or response associated with the drug pair (i, j) in the context of cell line k . This prediction task leverages both drug-specific and cell line-specific features, aiming to model complex biological interactions that influence drug synergy or efficacy. Such a framework is instrumental in advancing computational drug discovery and precision oncology, where accurate prediction of drug responses is crucial for therapeutic development.

4.1.2 Drug Combination Dataset

To train and evaluate the MatchMaker model, we utilized a large-scale pharmacogenomic resource known as DrugComb [25] database, a curated repository that aggregates results from multiple high-throughput drug combination screening studies such as NCI Almanac dataset [21], Forcina dataset [70], and O’Neil dataset [27].

We access DrugComb v1.4 (downloaded in December 2019, from <https://drugcomb.fimm.fi/>), which contains synergy measurements for 466,033 drug-pair experiments involving 4,150 unique compounds tested across 112 cell lines. This resource provides a rich foundation for model training, but requires careful preprocessing to ensure data consistency and compatibility with our feature extraction pipelines.

4.1.3 Filtering and Preprocessing

Although the raw DrugComb dataset is extensive, it contains many entries that could not be directly utilized due to missing or incomplete feature data. For example, some compounds lack detailed chemical structure information necessary for computing descriptors, and specific cell lines do not have gene expression data readily available.

To ensure each training example had all required inputs, we applied the following filtering steps:

4.1.3.1 Feature Availability Filter

- We exclude all drug pairs for which chemical structure data (required for descriptor computation) is not available in the PubChem database [34].
- Cell lines without gene expression profiles are also removed. Gene expression data is retrieved from the publicly available dataset curated by [71], which includes RMA-normalized baseline expression levels.
- After applying this filter, the dataset is reduced to approximately 335,692 valid triplets, comprising 3,040 drugs and 81 cell lines.

Table 4.1: The details of the DrugComb Dataset

	# of Combinations	# of Drugs	# of Cell lines
Raw dataset	466,033	4,150	112
After feature filtering	335,692	3,040	81
After duplicate removal	286,421	3,040	81

4.1.3.2 Handling Replicates and Inconsistencies

Drug combinations are often tested multiple times under similar conditions. While most combinations appear only once (single replicate), about 19% have between two and eight replicate entries. These replicates are not always consistent regarding synergy labels—some even conflicted in directionality, reporting both synergy and antagonism for the same combination. We apply the following filters to handle replicates.

- We remove 1,861 triplets that had perfectly split replicates (e.g., 50% synergy, 50% antagonism), since such cases do not provide a reliable supervisory signal.
- For the remaining replicates, we compute the average Loewe score across repeats and use that value as the target label, effectively summarizing the combination’s behavior.

4.1.3.3 Final Dataset Composition

After applying all filters and processing steps, the final curated dataset consists of 286,421 high-confidence drug–combination samples, spanning 3,040 distinct compounds and 81 human cancer cell lines. On average, each drug pair has synergy measurements across 25 cell lines, though this coverage varied. Table 4.1 summarizes the number of samples, drugs, and cell lines remaining after data processing and filtering. It is important to note that drug pairs are not uniformly tested across all cell lines; on average, each pair is evaluated on 25 cell lines. Figure 4.1 presents the distribution of Loewe scores.

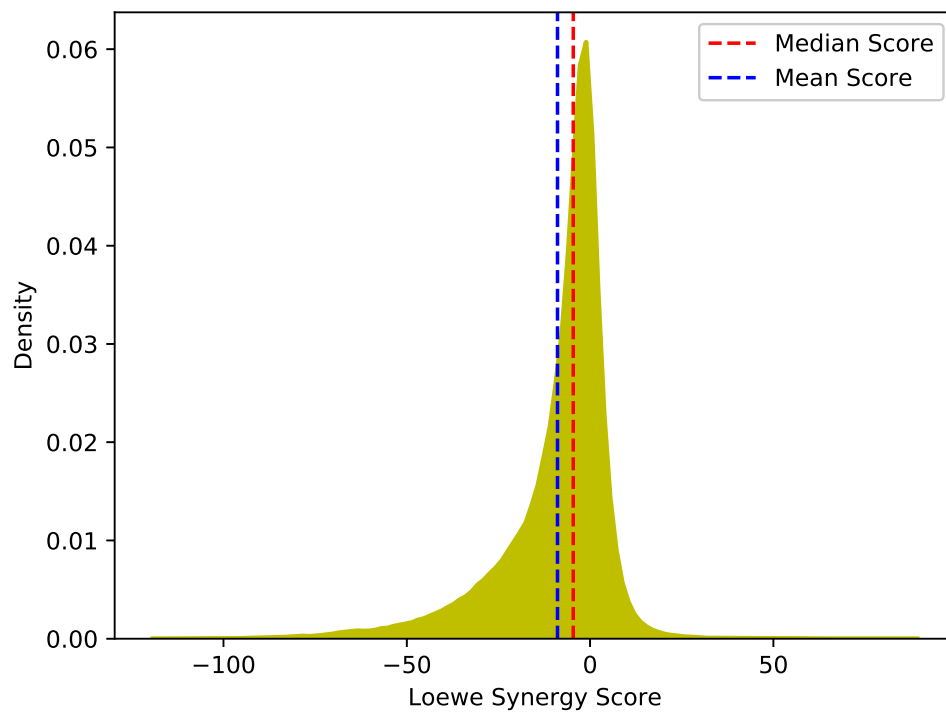


Figure 4.1: The distribution of Loewe scores in the processed DrugComb dataset. The final dataset comprises 286,421 drug combination experiments—each involving a pair of drugs and a cell line—spanning 3,040 unique drugs tested across 81 cell lines.

Table 4.2: The data statistics of the studies after preprocessing the DrugComb Dataset. DrugComb curates the NCATS studies from NIH NCATS Tripod, and we refer all other studies with their references.

Study	Number of drug combinations	Number of drugs	Number of cell lines	Size of full dose-response matrix
NCI ALMANAC [21]	266798	101	54	4x4, 4x6
FORCINA [70]	1804	1804	1	2x2
NCATS_2D_3D	10	5	1	10x10
NCATS_ES(FAKI/AURKI)	1834	1835	1	6x6
NCATS_ES(NAMPT+PARP)	44	30	1	6x6, 10x10
NCATS_HL	2385	1829	3	6x6, 10x10
ONEIL [27]	13456	36	28	5x5
YOHE [72]	90	25	1	10x10

Since DrugComb integrates drug combinations from multiple studies, we further examine the number of combinations within each study. Table 4.2 reports the number of combinations, unique cell lines, drugs, and dose-response matrix sizes per study, following the processing steps outlined in Table 4.1. As shown in Table 4.2, the NCI Almanac study contributes the majority of combinations and represents the largest dataset within DrugComb. Notably, dose-response matrix sizes vary across studies and even within a single study (e.g., the NCI Almanac includes 4×4 and 4×6 matrices).

4.1.4 Input Representations

To effectively predict drug synergy, MatchMaker relies on rich, informative feature representations for each component of a drug-drug-cell line triplet. These include gene expression profiles for cell lines and drug chemical descriptor features.

4.1.4.1 Cell Line Representations (Gene Expression Profiles)

We represent cell lines with gene expression profiles that capture their molecular state. The gene expression data is obtained from the publicly available dataset by [71] (accession number E-MTAB-3610 at ArrayExpress [73]), accessible through the CancerRXgene portal (downloaded on Dec 2019). These profiles were generated using Affymetrix microarrays, and we used the Robust Multi-array Average (RMA) normalized version to ensure comparability across cell lines.

To reduce dimensionality and focus on biologically relevant signals, we used the 972 landmark genes identified by the LINCS (Library of Integrated Network-Based Cellular Signatures) project [74]. These genes were chosen in LINCS due to their ability to explain a large portion of gene expression variation and are commonly used in systems biology and pharmacogenomics studies. Using a fixed subset ensures that the model is not overwhelmed by noise from uninformative genes and facilitates generalization across datasets. Thus, for every cell line k , we obtain a 972-dimensional feature vector denoted $c_k \in \mathbb{R}^{972}$, which serves as the molecular context in which drug effects are interpreted.

4.1.4.2 Drug Representations (Chemical Descriptors)

For each drug i (or j), we derive a detailed vector representation based on its chemical structure. This representation is computed using *ChemoPy* [36], a Python-based cheminformatics toolkit. *ChemoPy* extracts 541 molecular descriptors, encompassing a wide variety of features that include:

- **Constitutional descriptors:** e.g., number of atoms, molecular weight
- **Topological descriptors:** e.g., connectivity indices, kappa shape indices
- **Electronic properties:** e.g., polarizability, partial charges
- **Geometrical descriptors:** e.g., 3D structure-dependent features

The detailed chemical descriptors and the number of features for each descriptor are provided in Table 4.3.

These descriptors offer a comprehensive and interpretable view of a compound’s physicochemical and structural characteristics. Each drug is represented as $d_i \in \mathbb{R}^{541}$. When considering a drug pair $\langle i, j \rangle$, both d_i and d_j are used in conjunction with the same cell line vector c_k .

Table 4.3: Drug chemical structure features collected from ChemoPy

Chemical Feature Type	Feature Count
Constitutional descriptors	30
Connectivity descriptors	44
Kappa descriptors	7
Basak descriptors	21
E-state descriptors	245
Burden descriptors	32
Moreau-Broto autocorrelation descriptors	32
Moran autocorrelation descriptors	32
Geary autocorrelation descriptors	32
Molecular property descriptors	6
MOE-type descriptors	60

4.1.5 The MatchMaker Architecture

MatchMaker is a deep neural network-based model designed to predict drug synergy. Its architecture is depicted in Figure 4.2. The model takes as input (i) chemical descriptors for a given drug pair and (ii) the gene expression profile of an untreated cell line, and outputs a predicted Loewe synergy score [30]. The architecture comprises three neural subnetworks: The first two subnetworks learn representations of each drug, conditioned on the gene expression of the cell line. These representations are then concatenated and passed to a third subnetwork, which produces the final prediction. A detailed description of each component is provided below.

4.1.5.1 Drug Specific Subnetworks (DSN)

For each drug-pair-cell line triplet, MatchMaker employs two Drug-Specific Subnetworks (DSNs), denoted as DSN_i and DSN_j , corresponding to drugs i and j , respectively. Each DSN receives as input the chemical descriptor of a drug— d_i for DSN_i and d_j for DSN_j —along with the untreated gene expression profile of the k^{th} cell line, denoted as c_k . Both subnetworks share an identical architecture

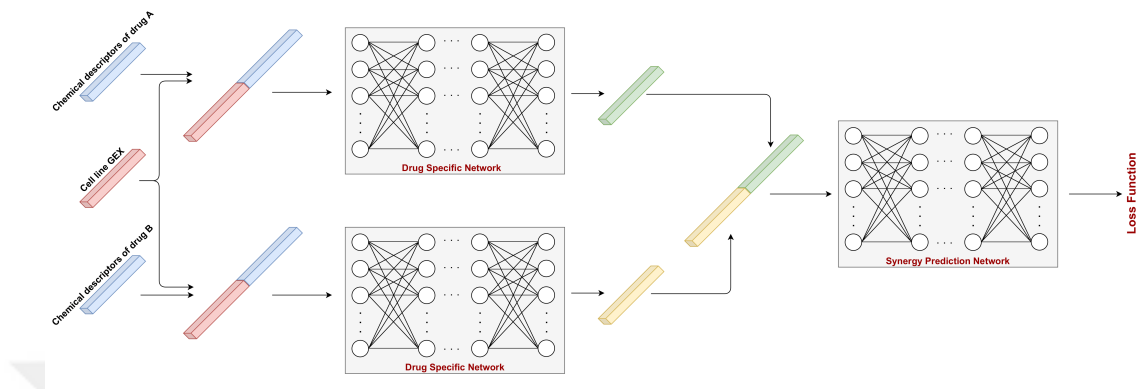


Figure 4.2: Overview of the MatchMaker architecture. The drug’s chemical features are concatenated with the cell line’s gene expression profile for each drug-cell line pair and passed through drug-specific subnetworks (DSNs). These DSNs independently learn drug representations conditioned on the cellular context. The resulting representations are then combined and input into the Synergy Prediction Network (SPN), which consists of fully connected layers that predict the Loewe synergy score. The model is trained end-to-end by minimizing a weighted mean squared error (MSE) between the predicted and true Loewe values.

comprising three fully connected (FC) layers, each followed by activation functions and dropout regularization [75]. Rectified Linear Unit (ReLU) activation [76] is applied to the first two FC layers, while the final layer uses a linear activation. Dropout is applied with probabilities of 0.2 and 0.5 at the first and second layers, respectively, and is omitted in the final layer. Each DSN outputs a vector embedding that represents the drug in the context of the given cell line, denoted as $x'_{i,k}$ and $x'_{j,k}$ for drugs i and j , respectively. The formulation of the DSN is summarized as follows:

$$\begin{aligned}
 O_{1,m,k} &= \text{Dropout}^{(p_1)} (\text{ReLU} (\text{FC}^{(n_1)} (\text{CAT}(d_m, c_k)))) \\
 O_{2,m,k} &= \text{Dropout}^{(p_2)} (\text{ReLU} (\text{FC}^{(n_2)}(O_{1,m,k}))) \\
 x'_{m,k} &= \text{Linear} (\text{FC}^{(n_3)}(O_{2,m,k}))
 \end{aligned}
 \tag{4.1}$$

Here, $x'_{m,k}$ denotes the latent representation of drug m in the context of cell line k , as produced by the corresponding subnetwork DSN_m . The notation $\text{FC}^{(*)}$ indicates a fully connected layer with $*$ neurons, while $\text{Dropout}^{(p)}$ denotes a dropout

layer with dropout probability p . *Linear* and *ReLU* refer to the linear and rectified linear unit activation functions, respectively. The operator CAT represents feature vector concatenation. The purpose of the DSN is to learn a *contextualized embedding* of each drug—that is, a representation that captures how a drug’s effect may vary depending on the gene expression profile of the associated cell line.

4.1.5.2 Synergy Prediction Subnetwork (SPN)

Following the generation of contextualized embeddings for both drugs in a given cell line, MatchMaker employs a Synergy Prediction Network (SPN) to predict the drug pair’s synergy score. The SPN inputs the concatenated embeddings produced by the subnetworks DSN_i and DSN_j . This joint representation captures the interaction between the two drugs, each conditioned on the same cell line context. The SPN consists of a three-layer fully connected (FC) network. ReLU activation functions are applied to the first two layers, while a linear activation is used at the output layer. A dropout layer with a dropout probability of 0.5 is applied after the first ReLU activation to reduce overfitting. The SPN outputs the predicted Loewe synergy score for drug pair $\langle i, j \rangle$ in cell line k , denoted as $\hat{y}_{i,j,k}$. The computation performed by the SPN is formulated as follows:

$$\begin{aligned} O'_{3,i,j} &= \text{Dropout}^{(p_4)} (\text{ReLU} (\text{FC}^{(n_4)} (\text{CAT}(x'_{i,k}, x'_{j,k})))) \\ \hat{y}_{i,j,k} &= \text{Linear}(\text{FC}^{(1)} (\text{ReLU} (\text{FC}^{(n_5)} (O'_{3,i,j})))) \end{aligned} \quad (4.2)$$

This two-stage architecture (DSNs followed by SPN) differs from existing models like DeepSynergy and TreeCombo, which typically concatenate all features at the input level and process them with a single network. MatchMaker explicitly decomposes the task into context-aware drug embedding followed by drug–drug interaction modeling, leading to improved generalization and interpretability.

4.1.6 Training Objective and Loss Function

MatchMaker is trained to minimize the error between its predicted synergy score $\hat{y}_{i,j,k}$ and the ground truth Loewe synergy score $y_{i,j,k}$ for a given drug–drug–cell line combination. Like many deep regression models, it uses a mean-squared error (MSE) loss as its base objective function. However, a key innovation in the training procedure is using a weighted MSE loss (Equation 4.3) that emphasizes highly synergistic drug combinations.

This decision is motivated by the observation that high-synergy pairs are relatively rare but clinically significant. In real-world drug development, these rare combinations represent the most promising candidates for follow-up experimentation. Thus, the model is encouraged to prioritize learning from examples that could yield the greatest therapeutic impact. To achieve this, the weight of each trio in the loss function is determined proportionally to the distance of its actual Loewe score from the minimum Loewe score across the training dataset (Equation 4.4). The weighted mean squared error (MSE) is employed solely during training, while the validation and test folds are evaluated using the standard MSE.

Let the dataset D consist of N samples indexed by triplets (i, j, k) . For each sample, the weighted MSE loss is defined as:

$$MSE_w = \frac{1}{N} \sum_{i,j,k \in D} w_{i,j,k} * (y_{i,j,k} - \hat{y}_{i,j,k})^2 \quad (4.3)$$

Here, $w_{i,j,k}$ is a sample-specific (drug i , drug j , cell line k) weight designed to increase the contribution of high-synergy cases to the overall loss. The weighting function is defined as:

$$w_{i,j,k} = \ln(y_{i,j,k} - \min(\mathbf{y}) + e). \quad (4.4)$$

Where:

- $\mathbf{y} \in \mathbb{R}^N$ denotes true Loewe score vector in the training set
- $y_{i,j,k}$ is the true Loewe score for the corresponding trio
- e is Euler’s number (approximately 2.718) to ensure a positive argument inside the logarithm
- The logarithmic scale ensures that extremely high synergy cases receive proportionally more emphasis without allowing them to dominate training entirely

This formulation guarantees that all weights are positive and that weights scale smoothly with the magnitude of synergy, with higher scores producing larger weights. Significantly, the weighting affects only the training process. Evaluation metrics on validation and test sets are computed using the standard (unweighted) MSE, ensuring comparability across models.

4.1.7 Compared Methods and Baselines

To rigorously evaluate MatchMaker’s performance, we benchmarked it against the state-of-the-art deep learning methods and traditional machine learning models. We made comparisons under identical conditions: the same data splits, the same input features, and consistent evaluation metrics. All methods utilize a concatenated vector comprising the drug chemical structure and cell line features: $CAT(d_i, d_j, c_k)$. This approach contrasts with the MatchMaker architecture, which first generates drug embeddings for a specific cell line before concatenating them to predict synergy scores.

4.1.7.1 DeepSynergy

One of the first deep learning models for drug synergy prediction, DeepSynergy [22] uses a fully connected architecture that takes the concatenation of both drugs’ features and the cell line profile as input. While effective, it lacks the drug-specific

embedding mechanism that MatchMaker introduces. It treats all input features symmetrically without distinguishing drug roles, potentially limiting expressiveness.

4.1.7.2 TreeCombo

TreeCombo [23] is a gradient boosting decision tree model implemented using XGBoost [77]. It handles non-linear interactions between inputs well and performs strongly on tabular data. However, it cannot exploit hierarchical or contextual information like neural networks, which can explicitly model drug-specific interactions within the biological context.

4.1.7.3 Random Forest

As a classical ensemble learning approach, Random Forests [78] are effective in many low-to-moderate dimensional data scenarios and provide robustness to overfitting. However, like TreeCombo, they do not offer the depth of representation or context awareness required for modeling subtle drug-cell line interactions.

4.1.7.4 Elastic Net

The Elastic Net [79] combines L_1 and L_2 regularization penalties in a linear regression model. It provides interpretability and sparsity, but cannot model non-linear or higher-order interactions, making it a valuable but limited baseline for a task of this complexity.

4.1.7.5 Cell Line Mean Predictor (Naïve Baseline)

To establish a performance floor, a trivial baseline is included: For any given triplet, it predicts the label as the mean of the synergy scores for all training

examples that involve the cell line k : $\hat{y}_{i,j,k} = \frac{\sum_{\langle i,j \rangle \in D^k} y_{i,j,k}}{|D^k|}$ where D^k contains all experiments performed on cell line k . This ignores drug information entirely and helps assess how much information is gained by modeling drug effects.

4.2 Results

4.2.1 Experimental Setup

To assess performance, the dataset is divided into training, validation, and test sets with proportions of 60%, 20%, and 20%, respectively. To ensure a fair comparison, the samples are stratified into folds using the new drug combination approach, as outlined in Section 2.4. This approach ensures that each drug pair appears in only one of the training, validation, or test sets. Hyperparameters are tuned on the validation set, and final performance metrics are reported on the test set. All algorithms are evaluated across the training, validation, and test splits.

Additionally, we re-evaluate MatchMaker using 5-fold cross-validation (CV) to assess the robustness of the results across different data splits. In each CV iteration, three folds are used for training, one for validation, and one for testing, rotating the folds for each run. This procedure results in five trained models, with a single prediction for each drug pair and cell line combination, as each example is only present in the test fold once. The CV procedure is repeated 10 times with different random fold splits, yielding 10 predictions per drug pair and cell line combination.

The models compared in this study are trained on the same dataset used to develop and evaluate MatchMaker. Various hyperparameter combinations are tested for training DeepSynergy, including different data normalization techniques $\in \{\text{tanh}, \text{tanh_norm}\}$, varying numbers of layers $\in \{2, 3\}$, and different numbers of neurons $\in \{512, 1024, 2048, 4096\}$ (see Table 4.4). For the TreeCombo model,

the number of estimators is set to 1000, as in [23]. The *maximum tree depth* parameter is tuned to values $\in \{4, 6, 8, 10, 12\}$, and the *learning rate* is set to $\in \{0.05, 0.10, 0.15\}$, based on recommendations by Janizek *et al.* (see Table 4.5). For the Random Forest model, the best performing *maximum tree depth* $\in \{4, 6, 8, 10, 12\}$ and *number of trees* $\in \{10, 100, 500, 1000\}$ parameters are selected (Table 4.6). Finally, for the ElasticNet model, the *L1 ratio* is considered for values $\in \{0.20, 0.40, 0.60, 0.80\}$, and the α parameter is evaluated for values $\in \{0.001, 0.01, 0.1, 1.0, 10, 100\}$ (Table 4.7).

Table 4.4: Hyperparameters for DeepSynergy. The input data is normalized either by *tanh_norm* or *tanh* as given at Preuer *et al.*. *tanh* first standardizes and then applies the hyperbolic tangent. *tanh_norm* first standardizes, applies hyperbolic tangent, and standardizes again.

Hyperparameter	Tuned values
preprocessing	tanh_norm, tanh
number of hidden layer	2, 3
learning rate	512, 1024, 2048, 4096
dropout	0.5
input dropout	0.2
learning rate	0.0001

Table 4.5: Hyperparameters for TreeCombo. The input data is normalized with *tanh_norm*. The number of estimator parameters is set to 1000 as given at Janizek *et al.*. Maximum tree depth and learning rate ranges used the same as Janizek *et al.*

Hyperparameter	Tuned values
preprocessing	tanh_norm
maximum tree depth	4, 6, 8, 10, 12
learning rate	0.05, 0.10, 0.15
The number of estimators	1000

MatchMaker is trained using an early stopping mechanism, as described in [80]. Specifically, while minimizing the training loss, we monitor the validation loss and halt training if the validation loss does not improve after p epochs ($p = 100$). The model with the best validation loss is then saved. The maximum number of epochs is set to 1000. We employ the Adam optimizer [81] with a learning rate of

Table 4.6: Hyperparameters for Random Forest. The input data is normalized with *tanh_norm*.

Hyperparameter	Tuned values
preprocessing	tanh_norm
maximum tree depth	4, 6, 8, 10, 12
the number of estimator	10, 100, 500, 1000

10^{-4} and a batch size of 128 for the loss function. Xavier initialization [82] is used for weight initialization. To mitigate order bias and ensure each drug is trained within the DSN, we feed the sample for a drug pair $\langle i, j \rangle$ into the network in both $\langle i, j \rangle$ and $\langle j, i \rangle$ orders. During testing, predictions are obtained for both orders, $\langle i, j \rangle$ and $\langle j, i \rangle$, and the final Loewe score is computed by averaging the predicted scores from both configurations.

Table 4.7: Hyperparameters for ElasticNet. The input data is normalized with *tanh_norm*.

Hyperparameter	Tuned values
preprocessing	tanh_norm
a	0.001, 0.01, 0.1, 1.0, 10, 100
L1 ratio	0.20, 0.40, 0.60, 0.80

All models are trained on a SuperMicro SuperServer 4029GB-TRT, equipped with two Intel Xeon Gold 6140 processors (2.3 GHz, 24.75M cache) and 251 GB of RAM. MatchMaker and DeepSynergy are trained using a single NVIDIA TITAN RTX GPU (24 GB, 384-bit). The training durations were approximately 5 hours for MatchMaker and 4 hours for DeepSynergy. For model development, we utilize Keras [83] with a TensorFlow [84] backend for both MatchMaker and DeepSynergy. Scikit-learn [85] is used for the Random Forest and ElasticNet implementations, while TreeCombo is implemented using the XGBoost Python package [77]. TreeCombo, ElasticNet, and Random Forest are trained on the CPU, with training times of approximately 1 hour, 30 minutes, and 5 hours, respectively.

4.2.2 Performance Comparison

4.2.2.1 Evaluation Metrics

To quantitatively evaluate predictive performance, we employ several complementary metrics that capture different aspects of regression and classification ability:

- **Mean Squared Error (MSE):** MSE measures the average of the squared differences between predicted synergy scores and the true Loewe scores. It is defined as $MSE = \frac{1}{N} \sum_{i=1}^N (y_{i,j,k} - \hat{y}_{i,j,k})^2$, where y_i is the actual Loewe score and $\hat{y}_{i,j,k}$ is the model’s prediction. Lower MSE indicates more accurate predictions overall. We emphasize MSE because a lower error in a drug discovery context means the model better approximates the true quantitative synergy, which is critical for reliably ranking candidate combinations. Additionally, we used a weighted MSE during training to emphasize highly synergistic examples (with larger positive Loewe scores). However, all models are evaluated on standard unweighted MSE for fairness.
- **Pearson Correlation Coefficient:** Pearson’s r assesses the linear correlation between predicted and actual synergy values. It ranges from -1 to 1 , with 1 indicating a perfect linear relationship. A higher Pearson r means the model’s predictions vary in direct proportion to the actual scores. This metric is relevant because we want minor errors and the model to correctly capture the trend – e.g., combinations with higher true Loewe scores should consistently receive higher predicted scores. In drug synergy prediction, a strong Pearson correlation implies the model can successfully rank drug pairs from least to most synergistic, an important property for prioritization tasks.
- **Spearman Rank Correlation:** Spearman’s ρ measures the rank-order correlation between predictions and ground truth. It evaluates whether the predictions preserve the ordering of synergy scores, regardless of the exact numeric differences. This is useful in contexts where the relative ranking of

combinations is more important than their exact score. A high Spearman coefficient (close to 1) indicates the model assigns higher scores to truly more synergistic pairs, even if the scale is imperfect, capturing monotonic relationships. Spearman correlation is less sensitive to outliers and score distribution shape than Pearson correlation, so we report it to confirm that our models’ performance is robust in terms of ordering synergistic pairs.

- **Area Under the ROC Curve (AUC):** We use the area under the Receiver Operating Characteristic curve to evaluate the model’s ability to classify synergistic vs. non-synergistic pairs. The ROC curve plots the true positive rate against the false positive rate at various classification thresholds. The AUC (ranging from 0.5 for random guessing to 1.0 for perfect classification) summarizes this curve as a single number. A higher AUC means the model can better distinguish truly synergistic combinations from non-synergistic/antagonistic ones across all thresholds. We compute AUC by binarizing synergy outcomes based on a chosen Loewe score cutoff (discussed below) and treating above-threshold pairs as positives. This metric is relevant because, in practice, researchers often care about identifying the most synergistic drug pairs (positives) while avoiding false positives; AUC captures this discriminative power irrespective of a specific cutoff.
- **Partial AUC (pAUC) at Low FPR:** Because experimental follow-up of predicted synergies is resource-intensive, it’s particularly important that the top-ranked predictions have high precision (i.e., few false positives). Therefore, we consider the partial AUC up to a low false positive rate (we use 10% FPR). This pAUC focuses on the leftmost portion of the ROC curve, essentially measuring how well the model performs when we restrict attention to the most confident predictions. A higher pAUC at 0–0.1 FPR means that among the top candidates, the model flags as synergistic, and a large fraction are true positives. In drug discovery terms, a superior pAUC indicates the model is more effective at retrieving truly synergistic pairs with high confidence, which is crucial for prioritizing a short list of combinations for experimental testing.
- **Area Under the Precision-Recall Curve (AUPR):** We also report the

area under the precision–recall curve, a valuable metric for imbalanced data. In our scenario, truly synergistic combinations (positives) are relatively rare compared to non-synergistic ones. AUPR directly captures the trade-off between precision (positive predictive value) and recall (sensitivity). A high AUPR means that when the model calls a pair synergistic, it is often correct (high precision), and it also manages to identify a large portion of all truly synergistic pairs (high recall). This metric is especially insightful because it weights false positives more heavily; a model that labels everything as synergistic would achieve high recall but very low precision, yielding a poor AUPR. By examining AUPR alongside AUC, we ensure that our model not only separates classes well in theory but also yields practical precision in predicting synergy.

Together, these metrics provide a comprehensive evaluation: MSE measures absolute error magnitude; Pearson and Spearman correlations assess the correlation and ranking quality; and AUC/pAUC/AUPR assess classification-like performance in identifying synergistic pairs. Next, we present the results for MatchMaker and baseline methods using these metrics.

4.2.2.2 Regression Performance

We compare the models’ predictive performances on the held-out drug combinations. MatchMaker achieves the state-of-the-art performance in predicting continuous Loewe synergy scores, outperforming all baseline methods across all regression metrics. Figure 4.3 summarizes the results for different algorithms across various performance metrics. MatchMaker attains the lowest MSE and highest correlation on the test sets. Specifically, MatchMaker’s MSE is 79.49, substantially better (lower) than the next-best model, DeepSynergy (MSE = 105.9). This is a 33% reduction in MSE compared to DeepSynergy, and a 67% reduction compared to TreeCombo (MSE = 132.7). In practical terms, this means MatchMaker’s predicted Loewe scores are much closer to the true values on average,

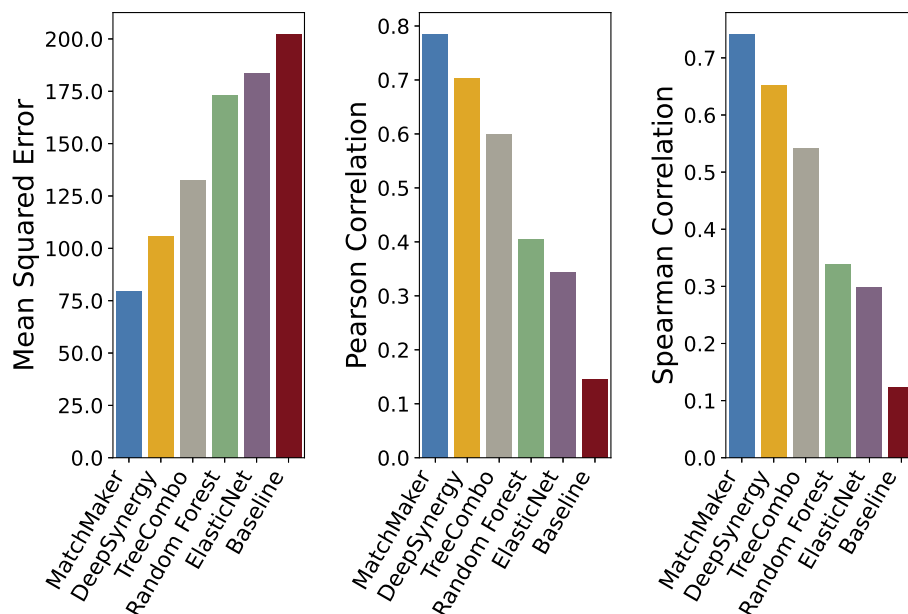


Figure 4.3: Comparison of model performance using mean squared error (MSE), Pearson correlation, and Spearman correlation coefficients. Among the evaluated models, MatchMaker achieved the best results with an MSE of 79.49, a Pearson correlation of 0.79, and a Spearman correlation of 0.74.

indicating more accurate quantitative predictions. Traditional models like Random Forest (MSE = 173.2) and Elastic Net (183.4) perform worst with errors more than double that of MatchMaker. A simple cell line mean baseline (predicting the average Loewe score per cell line) yielded an even higher error (202.5 MSE), underlining that the problem is non-trivial and cannot be solved by naive predictors.

Correlation-based metrics tell a consistent story. MatchMaker achieves a Pearson correlation of 0.79 and a Spearman rank correlation of 0.74 between its predictions and the true synergy scores. This means that MatchMaker’s predictions have a strong linear agreement with actual values and preserve the correct rank ordering of synergies to a high degree. In contrast, in the same test setting, the best baseline (DeepSynergy) reaches a Pearson correlation of 0.68 and Spearman correlation of 0.62. All methods follow the same relative ranking on both correlation and error metrics – models with lower MSE also have higher correlations –

but MatchMaker is consistently the top performer on all three metrics. In practical terms, a researcher using MatchMaker can trust that if the model assigns a higher synergy score to one drug pair than another, the prediction will likely hold experimentally, which is critical for prioritizing experiments.

To illustrate the model’s performance, we note that MatchMaker’s predictions for many drug pairs are very accurate. For instance, if a combination has an actual Loewe score of 20 (moderately synergistic), MatchMaker tends to predict a value close to this, often within a few points. Baseline models, however, might predict the same pair’s score much lower or higher, missing the magnitude of synergy. These differences accumulate across thousands of examples, hence the much lower MSE of MatchMaker. The high Spearman correlation (> 0.7) further implies that if one were to rank all tested combinations by predicted synergy, this ranking would be highly aligned with the true ranking by experimental Loewe score – a key property for screening tasks.

We also conducted an ablation experiment to assess the importance of the genomic context of the cell line. When we train MatchMaker without using cell line gene expression features, the performance drops dramatically: MSE increases to 131.9 (vs 79.5 with gene expression), Pearson correlation drops to 0.60, and Spearman correlation to 0.53. This is a substantial degradation in accuracy. It demonstrates that incorporating cell line-specific biology is crucial for synergy prediction – the model needs the gene expression context to correctly modulate how a given pair of drugs will interact in different cellular environments. Without these features, the model essentially reduces to a drug-only predictor and cannot explain why a drug pair might be synergistic in one cell line but not in another. The fact that performance declines to near the level of simpler models confirms that cell line features provide a significant predictive signal, and MatchMaker’s architecture effectively leverages this signal for improved performance.

In summary, MatchMaker’s novel dual-drug embedding architecture and the large training dataset lead to substantially improved regression accuracy over all baselines. It achieves the lowest prediction error and highest correlations, indicating that it can both precisely predict Loewe synergy scores and correctly

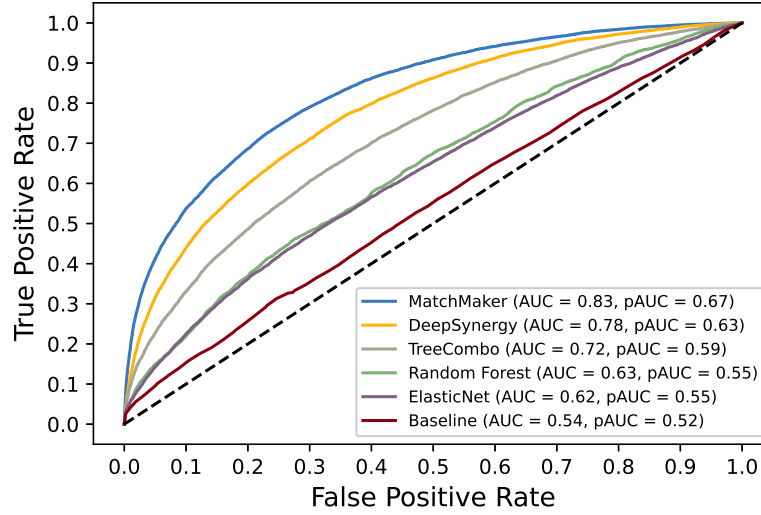
rank drug pairs by their synergistic effect. These results establish MatchMaker as the new state-of-the-art for continuous drug synergy prediction.

4.2.2.3 Classification Performance

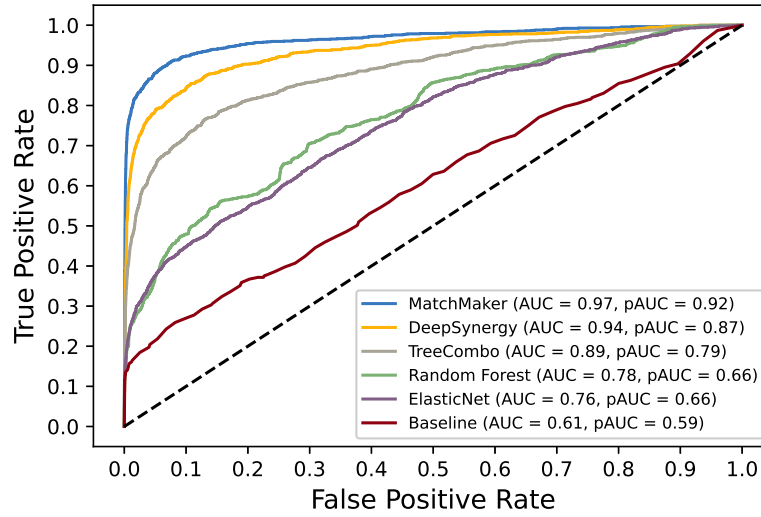
While predicting exact synergy scores is useful, one is often interested in identifying which drug combinations are significantly synergistic. To evaluate this, we tested MatchMaker’s ability to classify drug pairs as "synergistic" or "antagonistic" based on their Loewe scores. We compared it to other models in this classification-type task. We followed a procedure similar to prior work [22]: We binarized the outcomes using a synergy threshold t . If a drug pair’s Loewe score is greater than $+t$, it is labeled synergistic ($\bar{y} = 1$); if the score is less than $-t$, it is labeled antagonistic ($\bar{y} = 0$). Pairs with scores in between $-t$ and $+t$ are considered approximately additive and were not used as positives or negatives for this evaluation to avoid ambiguous cases. Based on guidelines from the SynergyFinder tool, we chose $t = 10$, meaning Loewe scores > 10 are treated as synergistic and < -10 as antagonistic. This threshold is biologically motivated: SynergyFinder suggests that scores in the range -10 to 10 are not confidently synergistic or antagonistic, whereas scores beyond ± 10 likely indicate real synergy or antagonism. In our dataset, this cutoff identifies a set of strongly synergistic combinations (positive class) and strongly antagonistic ones (negative class), allowing a meaningful classification analysis.

Figure 4.4b presents the ROC curves for MatchMaker and all baseline methods in this classification scenario. The ROC curves clearly illustrate the performance differences: MatchMaker’s curve rises towards the top-left corner, dominating those of the other models, which lie closer to the diagonal (the diagonal line would represent random guessing). MatchMaker achieves an AUC of 0.97, indicating near-perfect classification of synergistic vs. antagonistic pairs. By contrast, the next-best model, DeepSynergy, has a slightly lower AUC of 0.94. TreeCombo follows with an AUC of 0.89, and the classical models perform substantially worse (Random Forest 0.78, Elastic Net 0.76). The trivial baseline (using the cell line’s mean score) essentially fails this task (AUC 0.61, only slightly above random),

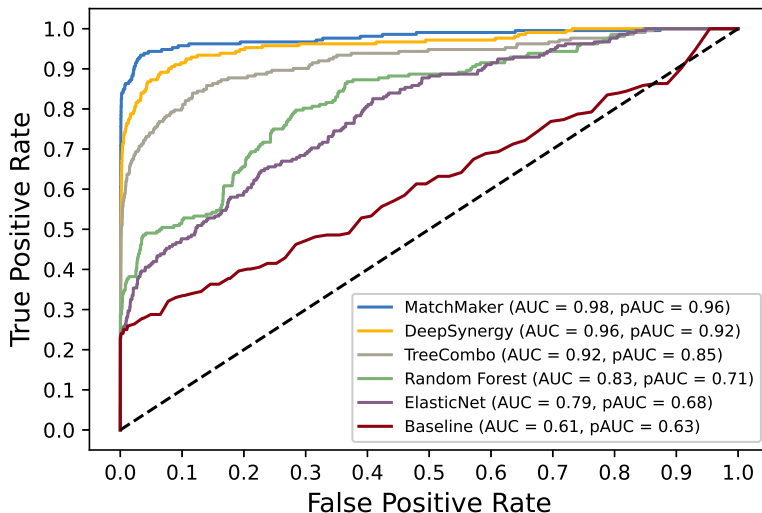
underscoring that identifying synergy is non-trivial. These AUC results show that MatchMaker is precise in regression and excels at identifying the truly synergistic cases.



(a) $t = 0$



(b) $t = 10$



(c) $t = 20$

Figure 4.4: Receiver Operating Characteristic (ROC) curves for all evaluated methods in the classification task. Combinations with Loewe scores greater than t are classified as positive (synergistic), while those with scores less than $-t$ are classified as negative (antagonistic). The figure also reports the corresponding Area Under the Curve (AUC) and partial AUC (pAUC) values computed up to a false positive rate (FPR) of 0.1.

Because synergistic pairs are rare, precision at the top predictions is vital. We therefore computed the partial AUC (pAUC) up to a 10% false positive rate to focus on early retrieval performance. In practical terms, pAUC asks: If we look at the model’s top predictions (allowing at most 10% of negatives to slip in as false positives), how well are we capturing true synergistic pairs? MatchMaker’s pAUC in this low-FPR regime is 0.92, significantly higher than DeepSynergy’s pAUC of 0.87. This indicates that MatchMaker retrieves more true synergistic combinations among the highest-scoring predictions before encountering any false alarms. In a drug screening context, this means if researchers tested, say, the top 5% of combinations ranked by MatchMaker, they would find a higher fraction of real synergies than if they used any other model’s ranking.

We also report the AUPR (Area Under Precision-Recall Curve) for this task since precision-recall is more informative when there are imbalanced classes. MatchMaker attains an AUPR of 0.85, outperforming DeepSynergy (0.74) by

a wide margin. This means that if one were to interpret the model’s output as a score for "synergy likelihood" and pick a threshold to call positives, MatchMaker would achieve a much better balance of high precision and high recall. In fact, MatchMaker’s precision remains high even as recall grows, whereas baseline methods see a drop-off in precision more quickly (reflected in their lower AUPR). For instance, the Random Forest and Elastic Net models have very low AUPR (0.37 and 0.36), showing that they miss most true synergies unless the threshold is set extremely leniently (and then precision suffers). Figure 4.4b and Table 4.8a summarize these classification metrics. Notably, MatchMaker’s advantage becomes even more pronounced when focusing on this low false-positive region, which is often the operating regime of interest.

Overall, MatchMaker’s classification performance is superior across all metrics (AUC, pAUC, AUPR), confirming that its predictions can be reliably used to identify top synergistic pairs. In practical application, MatchMaker can serve as an effective virtual screen: If one defines a synergy cutoff like $\text{Loewe} > 10$, MatchMaker can pick out those combinations with high accuracy. For example, if we treat $\text{Loewe} > 10$ as "actionable synergy," MatchMaker recovers such hits with very few false positives (as evidenced by the high pAUC and AUPR). This ability to flag truly effective combinations while minimizing false leads is crucial for experimental follow-up, where each false positive could lead to wasted resources.

Importantly, we examined whether these trends hold under different synergy thresholds. While our main analysis uses the ± 10 cutoff, one could choose a more lenient threshold ($t = 0$, treating any positive Loewe score as synergistic) or a stricter one ($t = 20$ for only very strong synergies). We found that MatchMaker maintained its lead in these scenarios as well. For instance, when we set $t = 0$ (classifying any positive Loewe as synergy vs. any negative as antagonism), or $t = 20$ (very high synergy vs. strong antagonism), the relative ordering of model performance does not change, MatchMaker continued to achieve the highest AUC and pAUC followed by DeepSynergy (see Figure 4.4a and 4.4c; and Table 4.8b and 4.8c). The absolute values of metrics shift (as the classification task difficulty changes with threshold), but the same trend holds. This consistency across various synergy definitions underscores the robustness of MatchMaker: Whether

one is interested in moderate synergies or only the most extreme cases, the model provides reliable predictions superior to other methods.

4.2.3 Robustness and Stability of Performance

Since machine learning models can sometimes be sensitive to how data is split or to random initialization, we assessed the confidence and stability of MatchMaker’s performance by running repeated cross-validations. We repeated the 5-fold cross-validation procedure 10 times with different random splits of the data as described in section 4.2.1 (shuffling the assignment of combinations to training and test folds each time). Figure 4.5 shows the average ROC curve for MatchMaker over these 10 runs, with a shaded region representing ± 1 standard deviation. The ROC curves across runs are nearly indistinguishable (the shaded band is very narrow), indicating that there is minimal variance in classification performance due to different train/test splits. In other words, no matter how we partition the data, MatchMaker reliably achieves almost the same high AUC (0.97) and maintains the same shape ROC curve. This gives us confidence that the model’s performance is not an artifact of a lucky split but generalizes to new data.

We observed similarly low variability in the regression metrics. Across the 10 random-split runs, MatchMaker’s average MSE on test sets was 82.3, with a standard deviation of only ± 2.6 . The mean Pearson correlation is 0.78 ± 0.009 and Spearman correlation is 0.75 ± 0.009 . These standard deviations are minimal, under 1.3% relative variation. Such consistency demonstrates that MatchMaker’s training process is stable and not overly dependent on initialization or data order. The results in each fold and each repetition remained close to those reported in Section 4.2.2.3 above, reinforcing that our earlier performance estimates are reliable. In practical terms, if an independent researcher were to train MatchMaker on a similar large dataset, they can expect to achieve very close to the reported performance, with little randomness in the outcome.

This robustness also implies that the model likely did not overfit to peculiarities of one particular fold. Even when we randomize splits multiple times, we see the

Table 4.8: Comparison of model performance in the classification task using three evaluation metrics: Area Under the ROC Curve (AUC), partial AUC up to a false positive rate of 0.1 (pAUC), and Area Under the Precision-Recall Curve (AUPR). Combinations with Loewe scores above the threshold t are labeled as positive (synergistic), while those with scores below $-t$ are labeled as negative (antagonistic).

(a) $t=10$

Model	AUC	pAUC	AUC PR
MatchMaker	0.97	0.92	0.85
DeepSynergy	0.94	0.87	0.74
TreeCombo	0.89	0.79	0.61
Random Forest	0.78	0.66	0.37
ElasticNet	0.76	0.66	0.36
Baseline	0.61	0.59	0.23

(b) $t=0$

Model	AUC	pAUC	AUC PR
MatchMaker	0.83	0.67	0.65
DeepSynergy	0.76	0.61	0.52
TreeCombo	0.72	0.59	0.48
Random Forest	0.63	0.55	0.38
ElasticNet	0.62	0.55	0.37
Baseline	0.54	0.52	0.30

(c) $t=20$

Model	AUC	pAUC	AUC PR
MatchMaker	0.98	0.96	0.85
DeepSynergy	0.95	0.90	0.75
TreeCombo	0.92	0.85	0.65
Random Forest	0.83	0.71	0.41
ElasticNet	0.79	0.68	0.33
Baseline	0.61	0.63	0.28

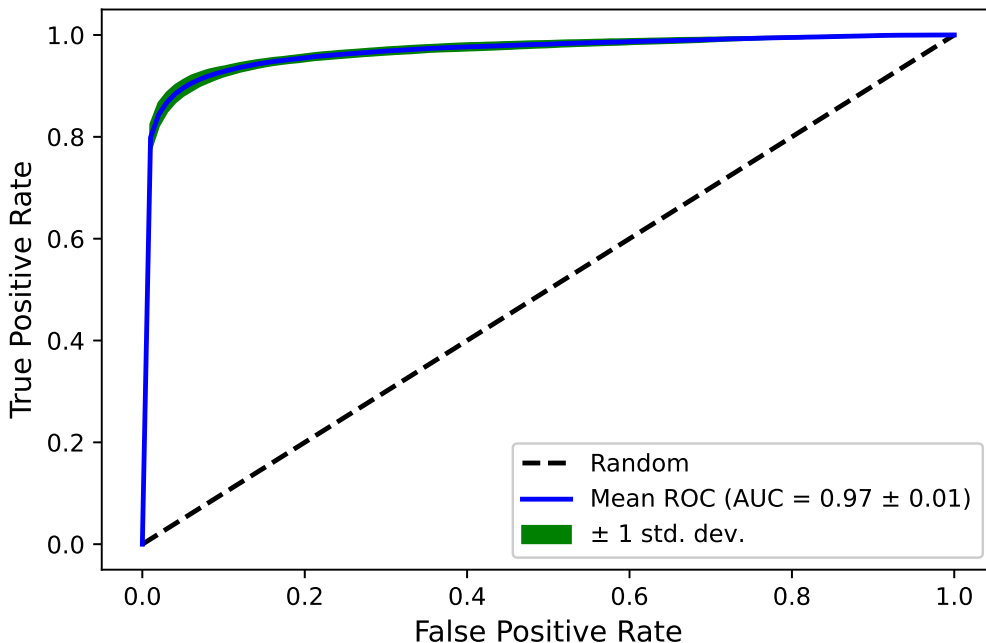


Figure 4.5: Area Under the Receiver Operating Characteristic (ROC) Curve for MatchMaker, along with the confidence interval, in the classification task using a threshold 10.

same accuracy and correlation, the model has truly learned generalizable patterns about drug synergy. It also suggests that the training dataset is sufficiently large and diverse such that no single subset dominates the learning; the model sees a wide variety of examples in any training fold. These observations prove that our performance gains are statistically significant and dependable. We can deploy MatchMaker with confidence that its predictions will be consistently accurate, and small changes in training conditions will not degrade its effectiveness.

4.2.4 MatchMaker Performance Across Cell Lines and Drug Pairs

An important question for translational use is whether the model performs uniformly well across different cell lines (biological contexts) and different drug pairs,

or if a subset of easier cases drives its accuracy. We analyzed MatchMaker’s results to identify any patterns in performance variation.

Cell line-specific performance: We computed the prediction error and correlations for each of the 81 cell lines separately – essentially treating all combinations tested on a given cell line k as a subgroup and evaluating how well the model predicts those. Figure 4.6 illustrates the MSE for each cell line. We found that for the vast majority of cell lines, performance is strong.

On the other hand, a few cell lines proved challenging. A small number (e.g., T98G, L-1236, NCI-H2122) exhibit higher prediction errors. These cell lines had the highest MSE values, indicating the model struggled more with combinations tested in those contexts. We investigated possible reasons: We suspect that poor performance could be due to limited training examples for a cell line or some biological outlier effect. Interestingly, we did not find a clear correlation with the number of training samples or with tissue type for these outlier lines. For instance, T98G (a brain cancer line) had 1,036 training examples, L-1236 (hematopoietic) had 1,391, and NCI-H2122 (lung) had 290 – so while NCI-H2122 had relatively fewer examples (which might explain some difficulty), T98G and L-1236 had plenty of data yet still yielded higher error. This suggests that the high errors might stem from intrinsic experimental noise or variability in those particular cell lines’ combination screenings. For example, the model would naturally have trouble if a cell line’s replicate experiments were inconsistent or the assay conditions were less reliable. Another possibility is that those cell lines exhibit biological behaviors (e.g., drug metabolism or mutation profiles) that were hard for the model to capture with just gene expression and chemical structure features. Interestingly, an analysis of the consistently predicted false positives (i.e., novel predictions) reveals that many experiments involving the T98G cell line are predicted as false positives, with a significant mean squared error (MSE) gap (see Section 4.2.5). One possible explanation for this observation is that the actual labels for these cell lines may be noisy, leading to a systematic error in the predictions for experiments involving these cell lines.

Despite these few outliers, it is reassuring that nearly all cell lines achieve

high predictive correlations, indicating that MatchMaker generalizes well across diverse cancer types. This is important for practical use: One can expect reliable performance in most cell contexts, with only occasional exceptions that might require further investigation. In fact, for 75 out of 81 cell lines (93%), the Pearson correlation between predicted and true scores is above 0.7, indicating excellent predictive power in most cellular contexts. Similarly, 71 of 81 cell lines have Spearman correlation greater than 0.7. The median cell line MSE is also low; the top 6 best-predicted cell lines have MSEs below 30, which is significantly better than the overall average. The best-performing cell lines include RD, A2780, and A427, where the model achieves particularly high accuracy (MSE well under 30 and correlations approaching 0.9). We provide these cell-line specific evaluations in Figure 4.7.

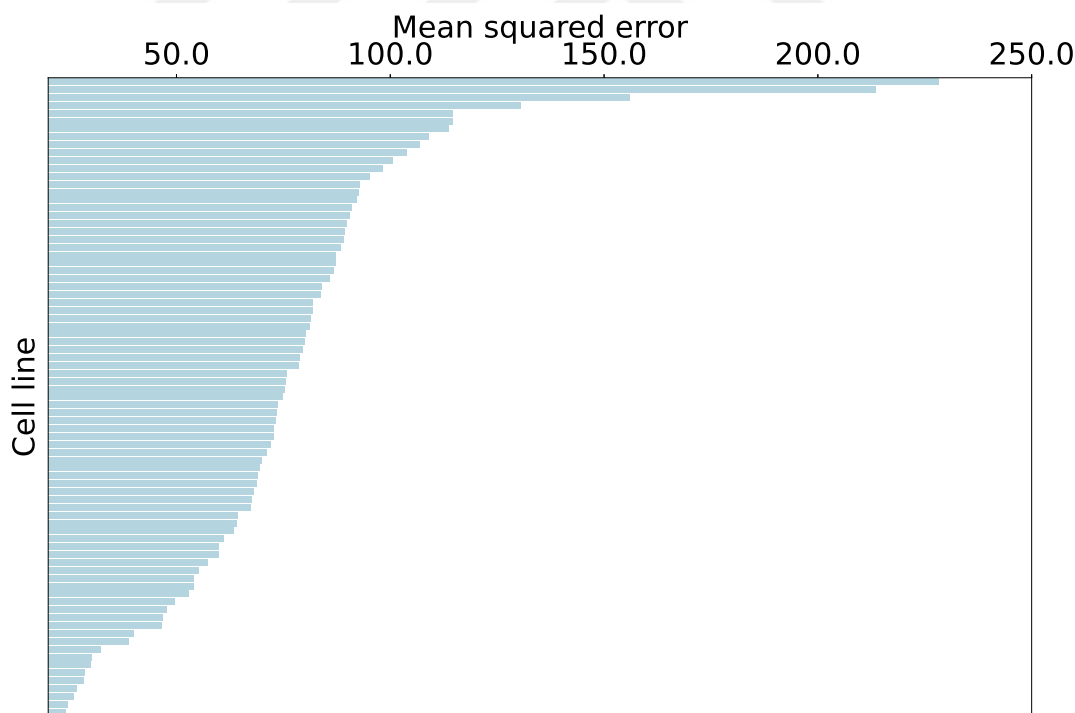


Figure 4.6: Cell line-specific MSE performance of MatchMaker is shown.

Drug pair-specific performance: We also evaluated how well the model predicts each individual drug pair across cell lines. Many drug pairs in the test set were tested on multiple cell lines (on average 25 different cell lines per drug pair), allowing us to see if the model consistently predicts a given pair’s synergy in

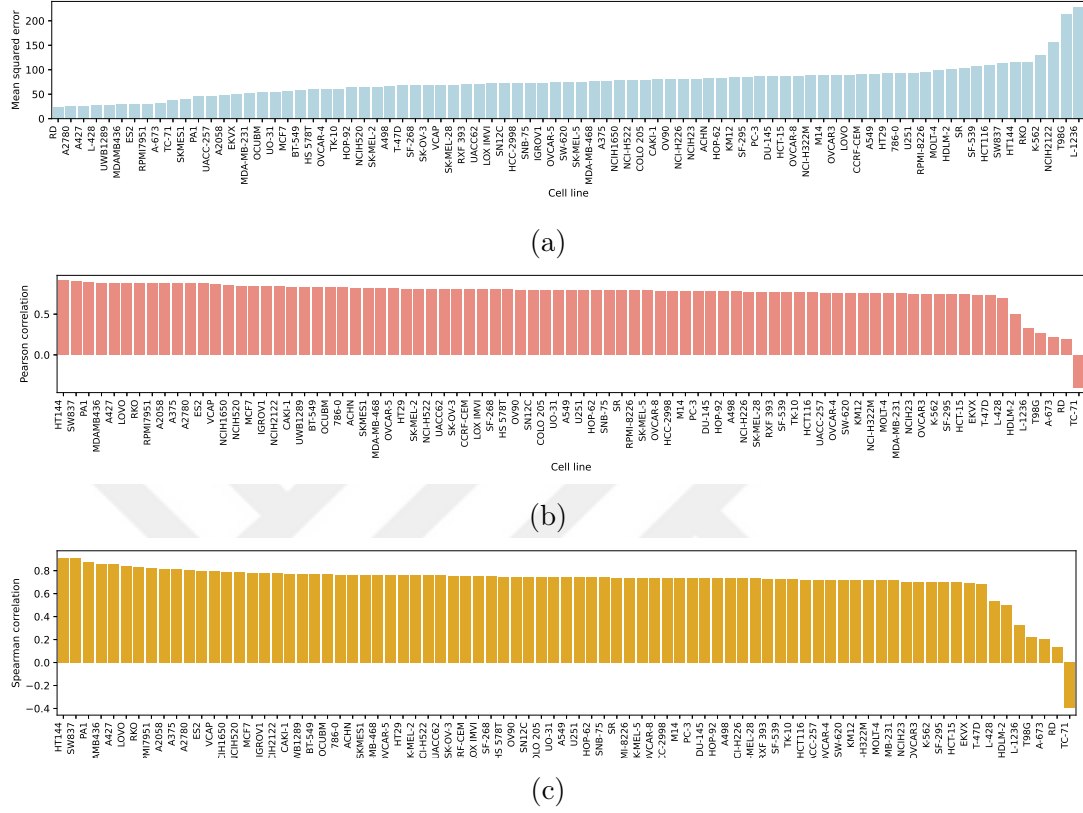


Figure 4.7: Cell line specific performances. The figures show the (a) MSE, (b) Pearson, and (c) Spearman correlation performances for each cell line.

various contexts. Out of 1,082 unique drug pairs in our test folds, we found that 74% of pairs have an MSE below the overall MSE (79.5), meaning the majority of drug pairs are predicted with error smaller than the average. This indicates that not just a few easy pairs drive the performance; instead, the model does well on a broad swath of drug combinations. Regarding correlation per drug pair, 46% of pairs have Pearson correlation > 0.7 and 35% have Spearman correlation > 0.7 when aggregating their predictions across cell lines. While a bit lower than the cell line-centric view, these fractions are still substantial – for nearly half the drug pairs, the model’s predictions of their synergy across multiple cell lines align closely with experimental measurements (Pearson correlation > 0.7). The somewhat lower rank correlations at the drug pair level suggest that certain drug pairs have more variable outcomes across cell types that the model doesn’t perfectly capture in rank order. This is perhaps expected: A drug pair might be highly synergistic in one cell line but only additive in another, and predicting

those context-dependent shifts is challenging. Still, the fact that 35–46% of pairs maintain high correlation indicates that the model often correctly predicts how a pair’s synergy will trend across different cellular backgrounds.

Table 4.9: The top predicted drug pairs based on MSE.

Drug1	Drug2	MSE	Pair count
LENALIDOMIDE	IMIQUMOD	4.93	54
BENDAMUSTINE HYDROCHLORIDE	PROCARBAZINE HYDROCHLORIDE	5.26	52
CELECOXIB	ERLOTINIB HYDROCHLORIDE	5.52	52
TENIPOSIDE	BORTEZOMIB	5.94	54
DEXRAZOXANE	PROCARBAZINE HYDROCHLORIDE	6.04	54
LOMUSTINE	ANASTROZOLE	6.07	53
LOMUSTINE	FULVESTRANT	6.13	54
BEZ-235	SN-38	6.37	27
VISMODEGIB	1-(5-DEOXPENTOFURANOSYL)-5-FLUORO-4- {[(PENTYLOXY)CARBONYL]AMINO}PYRIMIDIN-2(1H)-ONE	6.46	54
MITOXANTRONE	NSC733504	6.61	54

We also identified the top 10 best-predicted drug pairs – those with the lowest MSE across their appearances in test data. These serve as examples of combinations the model handles exceedingly well. Table 4.9 displays the top 10 drug pairs that yield the minimum MSEs. Interestingly, on analyzing these top pairs, we found that all but one of them had over 50 experiments in the dataset (meaning they were tested on more than 50 cell lines). In general, drug pairs that were tested in many cell lines (hence providing many training examples across contexts) tend to be predicted more accurately. The average number of cell lines per drug pair in our test set is 25, but these top-performing pairs far exceed that. This trend indicates that the model benefits from having multiple examples of a drug pair in training: Each cell line provides a new sample of that pair’s behavior, helping the model learn the pair’s interaction pattern more robustly. Essentially, the more data the model has seen for a specific pair (across different conditions), the better it can predict its synergy in a new context. This is intuitive and aligns with machine learning expectations; it also suggests a practical strategy that if a particular drug pair is of interest, obtaining training data on it in a few additional cell lines can substantially improve prediction confidence for that pair. Overall, the performance spread across drug pairs is wider than across cell lines (some pairs are inherently harder to predict consistently). However, most pairs still achieve good accuracy, and the best results occur for those with rich training data. This underscores the value of large datasets. Even though the total combination space is enormous, having many measurements for at least a subset

of pairs allows models like MatchMaker to shine on those. It sets a direction for future data collection to focus on depth (multiple contexts per pair) in addition to breadth.

Table 4.10: List of high-confidence false positives identified as novel synergistic drug pair predictions. These combinations exhibit an actual Loewe score below -10 and a MatchMaker-predicted Loewe score above +10 across all 10 random model initializations. The following five columns report additional sensitivity and synergy metrics: CSS, S, ZIP, BLISS, and HSA, with the latter four sourced from the DrugComb database. The final column summarizes the level of support from these alternative metrics, where the number of stars reflects how many of them also indicate a synergistic interaction.

Drug1	Drug2	Cell line	Actual Loewe	Predicted Loewe	CSS	S	ZIP	BLISS	HSA	Support
VEMURAFENIB	PLICAMYCIN	RPMI-8226	-47.72	18.45	30.78	20.37	3.76	8.00	0.51	*****
VEMURAFENIB	ACTINOMYCIN D	RPMI-8226	-24.6	24.67	58.83	22.70	10.29	12.72	9.72	*****
HYDROXYUREA	PRALATREXATE	K-562	-37.8	24.44	10.37	5.02	6.22	4.32	-5.08	****
VINBLASTINE SULFATE	PRALATREXATE	K-562	-10.84	35.41	34.80	18.08	-2.59	5.02	1.34	****
WNT-C59 (C59)	TEMOZOLOMIDE	T98G	-14.22	18.27	3.17	9.00	4.96	4.96	-14.22	****
SESAMIN	TEMOZOLOMIDE	T98G	-10.65	19.62	6.75	8.44	5.12	5.12	-10.65	****
PENFLURIDOL	TEMOZOLOMIDE	T98G	-12.97	15.54	4.43	18.75	13.24	13.24	-12.97	****
RUPATADINE FUMARATE	TEMOZOLOMIDE	T98G	-11.17	15.58	6.23	5.18	2.33	2.33	-11.17	****
SORAFENIB	PRALATREXATE	K-562	-86.9	24.64	29.61	22.02	-0.63	3.85	-5.38	***
HYDROXYUREA	PRALATREXATE	786-0	-36.42	18.37	33.89	24.84	-0.69	0.51	-1.10	***
VINBLASTINE SULFATE	PRALATREXATE	786-0	-11.96	39.38	7.64	7.53	-0.94	0.52	-1.16	***
ANASTROZOLE	PRALATREXATE	786-0	-12.23	25.00	7.21	7.23	-1.73	0.61	-0.84	***
BLEOMYCIN	PRALATREXATE	PC-3	-22.23	19.17	36.72	16.26	-4.39	-1.54	-0.59	**
VINBLASTINE SULFATE	PRALATREXATE	PC-3	-48.9	38.30	17.56	-7.60	-0.47	-2.53	-3.07	*
MITOTANE	PRALATREXATE	PC-3	-52.63	24.07	14.19	-10.75	-0.65	-4.63	-6.65	*
BENDAMUSTINE	PRALATREXATE	HCT116	-24.94	37.18	0.62	-0.80	-0.77	-0.73	-7.08	*
HYDROCHLORIDE	PRALATREXATE	HCT116	-24.94	37.18	0.62	-0.80	-0.77	-0.73	-7.08	*
JTC-801	TEMOZOLOMIDE	T98G	-30.02	14.30	54.64	-47.33	-32.62	-32.62	-29.94	*
TG100-115	TEMOZOLOMIDE	T98G	-13.09	24.01	18.05	-30.46	-25.04	-25.04	-13.06	*
CHLORPROPAMIDE	TEMOZOLOMIDE	T98G	-12.37	21.22	5.02	-13.91	-13.64	-13.64	-12.37	*
DOXOFYLLINE	TEMOZOLOMIDE	T98G	-16	16.74	13.49	-33.37	-28.24	-28.24	-15.97	*
GDC-0879	TEMOZOLOMIDE	T98G	-15.25	15.98	2.14	-9.92	-10.85	-10.85	-15.25	*
AZD2014	TEMOZOLOMIDE	T98G	-13.29	17.66	59.96	-30.62	-17.89	-17.89	-13.23	*
LY2109761	TEMOZOLOMIDE	T98G	-10.39	20.00	7.00	-3.83	-4.97	-4.97	-10.39	*
WZ811	TEMOZOLOMIDE	T98G	-10.82	12.53	6.57	-20.69	-18.97	-18.97	-10.82	*
FULVESTRANT	PRALATREXATE	DU-145	-12.35	27.61	-2.17	-8.59	-1.42	-4.0	-5.06	*

4.2.5 Novel Predictions

The reliability of the reference annotations provided by the DrugComb resource may be compromised by random noise or systematic biases. To interrogate this possibility, we examined those drug-cell-line combinations that MatchMaker persistently classifies as synergistic, although the benchmark labels designate them as antagonistic. In other words, we focus on **hard** false-positive instances in which the model predicts a positive Loewe interaction score yet the ground truth assigns a negative value.

Across 10 independently seeded MatchMaker runs, we identified every trio—two drugs plus one cell line—that was misclassified in all ten models. Because we use stratified 5-fold cross-validation (Section 4.2.1), each experiment appears in the held-out fold five times, meaning that we can truly detect systematic errors. Among the 286,421 total combinations, 2,540 distinct drug–cell-line trios are consistently labelled as false positives. To isolate only the most confident counter-examples, we retained those cases with an observed Loewe score below -10 coupled with a predicted Loewe score exceeding $+10$. This stringent filtering yields 25 high-confidence trios, listed in Table 4.10.

For these 25 instances, we additionally consulted complementary indicators of drug synergy: The combination sensitivity score (CSS), Bliss independence (BLISS) [86], highest single agent (HSA) [87], zero-interaction potency (ZIP) [88], and the S-synergy metric [89]. In two of the 25 trios, every one of these alternative measures corroborates a synergistic interaction; for the remaining trios, support ranges from partial to strong. The supplementary evidence is reported alongside the Loewe discrepancies in Table 4.10.

The top two entries— $\langle \text{VEMURAFENIB, PLICAMYCIN} \rangle$ and $\langle \text{VEMURAFENIB, ACTINOMYCIN D} \rangle$ in the RPMI-8226 (haematopoietic/lymphoid tissue) exemplify these tensions. MatchMaker infers robust synergy, the Loewe ground truth signals antagonism, yet all auxiliary metrics favour synergism. Literature inspection partially validates the model’s stance: Wishart et al. (2018) [90] documents that VEMURAFENIB elevates the serum concentration of ACTINOMYCIN D, indicating a pharmacokinetic interaction consistent with synergy.

The subsequent six combinations are endorsed by at least three of the four synergy indices and by a positive CSS. Particularly noteworthy is the PENFLURIDOL and TEMOZOLOMIDE (TMZ) pair, which has been shown to (i) retard glioblastoma tumour growth [91] and (ii) amplify apoptosis in glioblastoma multiforme (GBM) models [92] by down-regulating MGMT and GLI1 in T98G and other GBM cell lines. MatchMaker likewise contradicts the reference Loewe scores for TMZ in combination with (i) TG100-115, (ii) CHLORPROPAMIDE, and (iii) LY2109761. Independent studies support the therapeutic promise of these pairs:

TMZ potentiates TG100-115 against GBM cells and enhances murine survival [93, 94]; the LY2109761–TMZ duo suppresses tumour progression in T98G cultures [95]; and CHLORPROPAMIDE increases systemic TMZ exposure [90].

Table 4.10 further reveals that prediction errors are disproportionately large for combinations involving PRALATREXATE and, again, TMZ. Many of these experiments were performed in the T98G line, one of the poorest-performing cell lines for MatchMaker overall (Figure 4.6). As argued in Section 4.2.4, systematic inaccuracies in the T98G subset likely contribute to these recurrent discrepancies.

This analysis suggests that at least a subset of MatchMaker’s systematic "false positives" may, in fact, flag biologically plausible—or even experimentally supported—synergies that are mislabelled in the benchmark. These findings underscore the importance of complementary metrics and external evidence when assessing model performance on noisy pharmacogenomic datasets.

Table 4.11: O’Neil and NCI ALMANAC datasets details

	O’Neil Dataset			NCI ALMANAC		
	Combinations	Drugs	Cell lines	Combinations	Drugs	Cell lines
Initial dataset	23,062	38	39	311,604	108	60
After feature filtering	18,346	38	30	256,711	99	55

Table 4.12: Performance comparison of the methods in the O’Neil and NCI ALMANAC datasets based on MSE, Pearson, and Spearman correlation coefficients.

	O’Neil dataset			NCI ALMANAC dataset		
	MSE	Pearson	Spearman	MSE	Pearson	Spearman
MatchMaker	267.9	0.69	0.69	3014.5	0.59	0.55
DeepSynergy	290.0	0.66	0.67	3881.6	0.40	0.36
TreeCombo	283.1	0.67	0.67	3651.3	0.47	0.47
Random Forest	342.4	0.57	0.58	3932.3	0.38	0.39
ElasticNet	305.3	0.62	0.63	4176.5	0.30	0.33

4.2.6 Evaluation on Other Datasets

To rigorously assess the generalizability of MatchMaker, we conducted independent external validation experiments on two completely separate synergy datasets:

The O’Neil dataset [27] and the NCI-ALMANAC dataset [21]. In each case, the MatchMaker model was retrained from scratch using only data from that dataset (with no carry-over of the DrugComb-trained model), and evaluated on held-out test data from the same source. This approach – training and testing afresh on each new dataset – treats each as an independent experiment, ensuring that performance is evaluated on an entirely new data distribution. All baseline models were likewise trained de novo on each dataset, permitting a fair, like-for-like performance comparison without any influence from the original DrugComb data.

O’Neil Dataset Results: The O’Neil dataset comprises 23,062 drug combination samples, spanning 39 cancer cell lines and 38 distinct compounds. We excluded 9 of the 39 cell lines to ensure consistency in gene expression features, as their untreated gene expression profiles are not available in the E-MTAB-3610 dataset [71]. Following this filtering step, the dataset is reduced to 18,346 samples, encompassing 30 cell lines and the full set of 38 drugs. The corresponding Loewe synergy scores for these retained samples are obtained from the version of the O’Neil dataset curated by [22]. A MatchMaker model trained from scratch achieved the best performance among all methods on this dataset. Specifically, MatchMaker obtained the lowest mean squared error (MSE 267.9) and the highest Pearson correlation (0.69) on the O’Neil test set, outperforming deep learning and machine learning baselines. The following best method (e.g., DeepSynergy or the tree-based TreeCombo) reached a Pearson correlation of only ~ 0.66 - 0.67 . In other words, MatchMaker improved correlation by ~ 2 - 3 points and reduced prediction error relative to these baseline models on the O’Neil data. It confirms that the performance gains observed on the DrugComb dataset carry over when the model is redeveloped on a new cohort, indicating robust effectiveness rather than overfitting the original training distribution.

NCI-ALMANAC Dataset: We further validated MatchMaker on the much larger NCI-ALMANAC dataset [21].

The NCI-ALMANAC dataset initially comprises 311,604 drug-drug combination experiments involving 108 unique compounds tested across 60 cancer cell lines. To ensure consistency with transcriptomic data, we excluded five cell lines

that lack corresponding gene expression profiles in the E-MTAB-3610 reference panel [71]. In addition, we removed 9 compounds either due to structural data being unavailable in the PubChem database or because multiple compounds mapped to the same PubChem compound identifier (CID) [34], introducing ambiguity. Following these filtering steps, the curated dataset consists of 256,711 drug–cell line pairs involving 99 drugs and 55 cell lines.

Unlike other synergy benchmarks that rely on Loewe scores, the NCI-ALMANAC evaluates drug interactions using the ComboScore metric [21]. ComboScore quantifies synergy by aggregating the deviations between observed and expected cell growth across all tested dose combinations. Specifically, the expected growth for a drug pair—denoted drugs A and B—is computed based on non-interaction assumptions, and the ComboScore sums the differences between these expectations and the observed effects to characterize the overall degree of synergy or antagonism. The expected growth of two drugs A and B is calculated as follows:

$$E^{A_p B_q} = \begin{cases} \min(Y^{A_p}, Y^{B_q}) & \text{if } Y^{A_p} \leq 0 \quad \text{or} \quad Y^{B_q} \leq 0 \\ \frac{Y^{A_p} * Y^{B_q}}{100} & \text{otherwise} \end{cases} \quad (4.5)$$

where Y^{A_p} and Y^{B_q} are the true (observed) growth rates of drugs A and B with concentrations p and q . The number of samples, drugs, and cell lines is summarized in Table 4.11.

Once again, MatchMaker was retrained *ab-initio* on this dataset and evaluated against similarly retrained models. On NCI-ALMANAC, MatchMaker significantly outperformed all baselines, demonstrating excellent generalization to a new distribution and synergy definition. For instance, MatchMaker achieved an MSE of 3014.5, substantially lower than the best baseline’s MSE (3651.3 by TreeCombo), and attained a higher Pearson correlation (0.59) compared to the best baseline’s Pearson correlation of 0.47. MatchMaker also led in Spearman rank correlation (0.55 vs. 0.47 by the closest competitor) on this dataset. These improvements amount to roughly a 17% reduction in error and a notable increase

in correlation over the strongest alternative method on the NCI-ALMANAC test set. The consistent superiority of MatchMaker across both external datasets is summarized in Table 4.12.

In summary, using O’Neil and NCI-ALMANAC as independent validation experiments underscores MatchMaker’s effectiveness on entirely new data distributions. Rather than relying on a pretrained DrugComb model, retraining MatchMaker on each dataset from scratch yielded the state-of-the-art results in both cases. MatchMaker outperforms all baseline models when trained and tested separately on each dataset. These findings confirm that MatchMaker’s modeling approach is inherently robust and generalizes well to diverse combination screens (even with different synergy metrics and vastly different data scales), thereby reinforcing its potential for broad applicability in drug synergy prediction.

Chapter 5

Personalized Drug Synergy Prediction (PDSP): Towards Patient-Specific Models

PDSP extends the MatchMaker framework by introducing a multi-task learning and transfer learning approach tailored for personalised medicine. To address inter-patient heterogeneity, PDSP is trained to jointly predict drug combination synergy and single-agent sensitivity. The model is first pre-trained on large-scale cell line data and then fine-tuned using patient-specific *ex-vivo* monotherapy response data, allowing it to adapt its predictions to individual tumor profiles.

5.1 Methodology

5.1.1 Problem Definition and Formulation

Similarly to MatchMaker, we design the drug combination synergy prediction as a supervised regression task, but we also predict the sensitivities of the drugs in the given cell line. Formally, for a given a drug pair $\langle i, j \rangle$ and a specific cancer

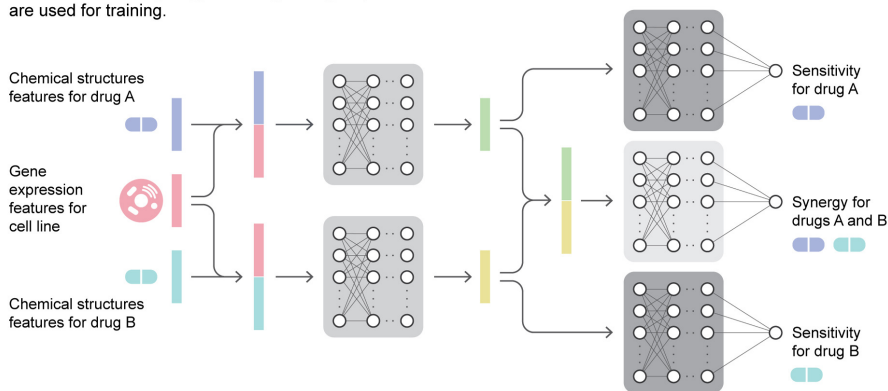
cell line k , the objective is to simultaneously predict the synergistic interaction score $y_{i,j,k} \in \mathbb{R}$ for the drug combination, as well as the individual sensitivity profiles of each drug when applied independently to the same cell line. In order to standardize the single-agent sensitivity measurements for supervised classification, we discretize the continuous drug response scores, as described in detail in Section 5.1.2.1. Accordingly, we define binary sensitivity labels $s_{i,k} \in \{0, 1\}$ and $s_{j,k} \in \{0, 1\}$, where a value of 1 indicates that the drug is considered effective or sensitive in cell line k , and a value of 0 denotes resistance or lack of significant response.

Each drug is characterized by a high-dimensional feature vector d_i or d_j , which encodes its physicochemical and structural properties derived from chemical descriptors. The biological context of the cell line k is captured by its gene expression profile, represented by the vector c_k , which reflects the underlying transcriptional state of the cell. These feature vectors collectively serve as the model’s input.

The predictive model is trained in a multi-task learning framework to jointly optimize two distinct but related objectives: (i) classifying the binary sensitivity status of each individual drug in the context of the cell line, and (ii) estimating the continuous synergy score associated with the combination of the two drugs in that same biological environment. To accomplish this, the model learns complex non-linear mappings from the input feature triplet (d_i, d_j, c_k) to both the sensitivity class probabilities $(\hat{s}_{i,k}, \hat{s}_{j,k})$ and the predicted synergy score $\hat{y}_{i,j,k}$, thereby enabling it to make biologically contextualized and therapeutically actionable predictions.

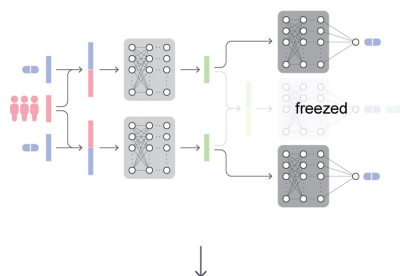
1. Training with cell line data

Both combination synergy and single-drug responses are used for training.

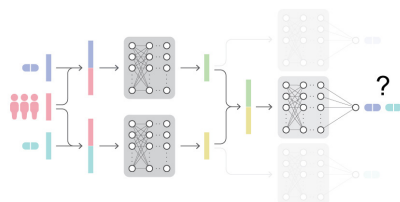


2.a. Fine-tuning with data from all patients

Only single-drug responses are used for fine-tuning.

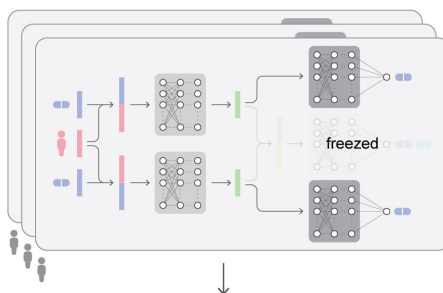


3.a. Use fine-tuned model for synergy prediction



2.b. Fine-tuning with data from each patient separately

Only single-drug responses are used for fine-tuning.



3.b. Use each fine-tuned model for synergy prediction for the patient

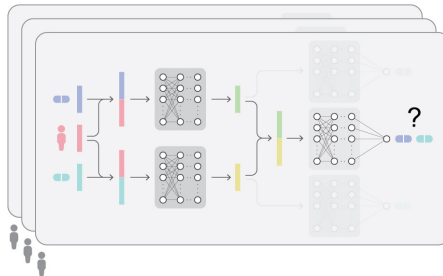


Figure 5.1: Overview of the PDSP framework pipeline: 1) The architecture and training procedure of the PDSP model, which is initially trained on drug combination data using cell line features. 2) Fine-tuning of the pretrained PDSP model. The pretrained PDSP model is adapted by updating the drug encoders using patient-specific single-drug response data and gene expression profiles, either collectively across all patients (Step 2.a) or individually for each patient (Step 2.b). 3) Synergy prediction is then performed using the fine-tuned models, enabling personalized evaluation of drug combinations.

5.1.2 Dataset Construction and Preprocessing

5.1.2.1 Cell Line Synergy Data

We first compile a large-scale dataset of drug pair synergies and half-maximal inhibitory concentration (IC_{50}) sensitivity values measured on cancer cell lines. Synergy scores are obtained from public high-throughput combination screens, aggregated via DrugComb (version v1.5, downloaded in Feb 2022). We specifically use the Bliss independence score [86] as the synergy metric, where positive values indicate synergistic effects and negative values indicate antagonism.

The IC_{50} value denotes the concentration of a drug required to inhibit a specific biological or biochemical process, such as cellular proliferation by 50%. It serves as a widely used proxy for drug efficacy. Lower IC_{50} values imply greater potency, indicating that the corresponding cell line exhibits higher sensitivity to the drug. In contrast, higher IC_{50} values suggest drug resistance, reflecting a diminished pharmacological response.

For downstream analyses, we transform continuous IC_{50} measurements into binary sensitivity labels. Consistent with the approach described in [96], we classify a cell line as sensitive to a particular drug if the natural logarithm of its IC_{50} value is less than -2 . Otherwise, the cell line is designated as resistant. This thresholding strategy enables categorical interpretation of drug response profiles while aligning with established practices in pharmacogenomic research.

To ensure robust model training, we filtered the raw data to include only combinations with available gene expression profiles for the corresponding cell line and chemical descriptors for each drug. Gene expression data for cell lines were taken from [71] (ArrayExpress [73] accession number is E-MTAB-3610), and small-molecule structures were retrieved from PubChem for descriptor calculation. Since DrugComb includes data from multiple sources, it sometimes contains replicates for the same drug pair. In these cases, we average Bliss scores across replicate measurements. After filtering, our final dataset contains 330,103

samples with 3,068 drugs and 81 cancer cell lines. This extensive dataset was the basis for pretraining the synergy prediction model and benchmarking against baseline methods.

Table 5.1: Summary of the DrugComb dataset and patient drug combination datasets following preprocessing. Each data instance represents a combination of two drugs and a single cell line (or patient sample).

	Source	Combinations	Drugs	Cell lines/Patients
Cell line	DrugComb [25]	330,103	3,068	81
Patient-derived	He et al., 2018	20	16	3

5.1.2.2 Patient-Derived Data

Since the patient-specific combination data are scarce, we leverage a focused dataset from [58] that provides both single-drug response and limited combination testing on patient samples. This dataset includes 654 single-drug sensitivity measurements on ex vivo leukemia patient samples, covering 218 unique compounds tested across three patients. In addition, 20 drug pair combinations were experimentally evaluated on these patient-derived cells, with 10 synergies and 10 antagonistic outcomes identified. These 20 patient-specific combination results (involving 16 total drugs) are held out as a validation set for personalized synergy prediction. Since patient molecular profiles can differ substantially from cell lines, baseline gene expression profiles of the patient samples (provided by [58]) were incorporated as patient-specific features for our model. However, the patient gene expression data included only 634 genes, a small panel compared to the 17,737 gene expression features available for cell lines. To make the model’s input feature space consistent, we intersected the gene sets, obtaining a common set $\mathcal{G}$ of 542 genes measured in both cell lines and patients. We use the expression values of these 542 genes to represent any cell line or patient sample in our model.

5.1.2.3 Input Feature Representation

Each drug in a combination is represented by a rich set of chemical features, and each cell context (cell line or patient) is represented by its gene expression profile. Following MatchMaker, we represent drugs with their chemical structure features, which we generate using the ChemoPy Python library [36] (more information provided in Table 4.3 and in the previous chapter). The cell lines and patients are represented with a vector of 542 gene expression levels for the shared gene set $\mathcal{G}$. To address systematic differences between gene expression platforms and conditions (cell line cultures vs. ex vivo patient cells), we apply quantile normalization [97] to the expression profiles. This normalization ensures the distribution of gene expression values is comparable between cell lines and patient samples, enabling the model to transfer learning across these domains.

For a given experiment (drug pair tested on a particular cell line or patient), we construct the model input by concatenating each drug’s descriptor vectors $d_i, d_j \in \mathbb{R}^{541}$ with the cell gene expression vector $c_k \in \mathbb{R}^{542}$. This results in two composite feature vectors (one per drug) that capture both the drug’s inherent properties and the cellular context in which the drug is applied. These features serve as inputs to our model to predict $y_{i,j,k} \in \mathbb{R}$, $s_{i,k} \in \{0, 1\}$, and $s_{j,k} \in \{0, 1\}$ simultaneously. By using identical feature types for both cell lines and patients (chemical descriptors + gene expression), we ensure the model can be trained on cell line data and later fine-tuned on patient data without changing its input format.

5.1.3 The Model Architecture

We develop Personalized Deep Synergy Predictor (PDSP) as a multi-task neural network that extends our earlier MatchMaker architecture to support patient customization. Figure 5.1 illustrates the overall design. The model is designed to jointly predict a drug pair’s synergy and each drug’s single-agent response in a

given biological context. The PDSP model extends our earlier framework, MatchMaker, by incorporating both monotherapy sensitivity and combination synergy prediction within a unified architecture. The model takes as input the vector representations of two drugs (d_i, d_j) and a cancer cell line (c_k), and processes them in two sequential stages.

In the first stage, PDSP employs Drug Encoder Networks (DENs) to generate conditional latent representations for each drug. Each DEN takes a drug–cell line pair as input and learns a representation that captures how a drug’s effect may vary depending on the cellular context. Both DENs are designed to share identical architectures and model weights to ensure symmetry and order invariance with the input drug pair. The second stage of PDSP leverages the encoded representations to predict drug sensitivity and synergy. It comprises two components: Drug Decoder Networks (DDNs) and a Synergy Aggregation Network (SAN). Each DDN receives a drug’s conditional encoding and decodes it to predict its sensitivity classification (e.g., sensitive or resistant) in the corresponding cell line. As with the encoders, the DDNs are symmetric and parameter-shared to preserve model consistency. While the DENs produce drug-specific latent representations, they do not capture the interactive effects between the two drugs. To address this, SAN concatenates the latent vectors from the DENs, thereby integrating information from both drugs within the cellular context. This aggregated representation is then used to predict the drug combination’s synergy score. Through this two-stage design, PDSP effectively models both individual drug responses and their combined therapeutic potential in a cell-line-specific manner.

This architecture is a novel extension of MatchMaker. MatchMaker is originally designed to predict combination responses by encoding drugs with cell contexts, but it does not include a mechanism to incorporate patient-specific calibration because it lacks the single-drug prediction component. PDSP’s addition of the DDN outputs for single-drug response is crucial – it enables the model to be fine-tuned on patient-specific drug sensitivity data, thereby personalizing the drug embeddings for a given patient. Notably, the design ensures that this additional task does not compromise the synergy prediction performance. The two tasks are related (both depend on how a drug affects the cell), so the

shared encoder can learn a valuable representation for both. We hypothesized that by training on the large cell line combination dataset, the DENs would capture generalizable drug–gene interactions. Then the DDN outputs could be used to adjust these representations to a particular patient’s biology without needing direct combination data from the patient.

The details of the DEN, DDN and SAN subnetworks are explained in the following sections.

5.1.3.1 Drug Encoder Network (DEN)

The PDSP architecture incorporates two identical Drug Encoder Networks, denoted as DEN_i and DEN_j , which process the input features of a cancer cell line (c_k) and two drugs (d_i, d_j) to produce conditional latent representations. These encoders are designed to capture how the biological context of a cell line modulates the effect of each drug individually.

Each DEN receives the gene expression profile of the cell line concatenated with the feature vector of one of the drugs in the pair as input. This concatenated input vector is passed through a series of fully connected (FC) layers. Non-linear activation functions—specifically, Rectified Linear Units (ReLU)—are applied after each FC layer, except the final layer. Dropout layers are also interleaved between FC layers to promote generalization and reduce overfitting. The formal representation of the DEN operation for a drug–cell line pair is given as:

$$\begin{aligned}
 den_1(d_m, c_k) &= \text{Dropout}(\text{ReLU}(\text{FC}(\text{Concat}(d_m, c_k)))) \\
 den_2(d_m, c_k) &= \text{Dropout}(\text{ReLU}(\text{FC}(den_1(d_m, c_k)))) \\
 DEN(d_m, c_k) &= \text{Dropout}(\text{Linear}(\text{FC}(den_2(d_m, c_k))))
 \end{aligned} \tag{5.1}$$

where d and c denote the input feature vectors for a drug m and a cancer cell line k , respectively. Each Drug Encoder Network ($DEN(d_i, c_k)$, $DEN(d_j, c_k)$)

processes one drug–cell line pair, transforming it into a conditional latent representation. These encoders consist of a sequence of intermediate layers, denoted den_1 and den_2 , typically composed of fully connected layers followed by non-linear activation functions and dropout. Importantly, $DEN(d_i, c_k)$ and $DEN(d_j, c_k)$ are constructed with identical architectures and share all model parameters and weights. This design ensures order-invariance, such that swapping the input order of the two drugs does not affect the final synergy prediction.

This parameter sharing enforces consistent learning of the drug–cell line interaction, independent of input order, and enables the model to generalize across drug pairs more effectively. The outputs of the encoders are denoted as $h_{i,k}$ and $h_{j,k} \in \mathbb{R}^n$, representing the conditional latent embeddings of drugs i and j in the context of cell line k .

Formally, the encoding process is defined as:

$$\begin{aligned} h_{i,k} &= DEN(d_i, c_k) \\ h_{j,k} &= DEN(d_j, c_k) \end{aligned} \tag{5.2}$$

5.1.3.2 Drug Decoder Network (DDN)

Building upon the latent drug representations $h_{i,k}$ and $h_{j,k}$, the PDSP model incorporates two identical Drug Decoder Networks (denoted as DDN_i and DDN_j) to predict the drug sensitivity statuses $\hat{s}_{i,k}$ and $\hat{s}_{j,k}$ for each drug i and j in the context of a given cancer cell line k . Both decoder networks share the same architecture and model weights, ensuring a consistent and order-invariant treatment of drug–cell line pairs.

Each DDN is implemented as a feed-forward neural network composed of three fully connected (FC) layers. The first two layers apply ReLU activation followed by dropout to promote non-linearity and prevent overfitting. The final FC layer outputs a continuous value using a linear activation function, corresponding to the predicted drug sensitivity.

The DDN architecture is formally defined as follows:

$$\begin{aligned}
ddn_1(h) &= \text{Dropout}(\text{ReLU}(\text{FC}(h))) \\
ddn_2(h) &= \text{Dropout}(\text{ReLU}(\text{FC}(ddn_1(h)))) \\
DDN(h) &= \text{Linear}(\text{FC}(ddn_2(h)))
\end{aligned} \tag{5.3}$$

where $h \in \{h_{i,k}, h_{j,k}\}$ is the conditional drug embedding obtained from the corresponding encoder network, and ddn_1 and ddn_2 represent the intermediate transformation layers of the decoder.

The final sensitivity predictions for drugs i and j on cell line k are computed as:

$$\begin{aligned}
\hat{s}_{i,k} &= DDN(h_{i,k}) \\
\hat{s}_{j,k} &= DDN(h_{j,k})
\end{aligned} \tag{5.4}$$

These predicted values $\hat{s}_{i,k}$ and $\hat{s}_{j,k}$ serve as estimations of how sensitive the cell line k is to each drug individually, and can be evaluated using regression or classification loss functions depending on the sensitivity label format.

5.1.3.3 Synergy Aggregation Network (SAN)

Analogous to the structure of DDN, the Synergy Aggregation Network (SAN) employs a feed-forward neural network architecture to predict drug combination synergy. SAN takes as input the conditional latent representations of two drugs, $h_{i,k}$ and $h_{j,k}$, derived from the shared drug encoder networks, and outputs a continuous synergy score $\hat{y}_{i,j,k}$.

Since SAN operates on drug pairs rather than individual drugs, it begins by concatenating the two latent vectors to form a unified representation that captures their joint effect within the cellular context. This concatenated vector is

then passed through a sequence of fully connected layers, each followed by ReLU activation and dropout, similar to the design of DDN. A final FC layer with a linear activation function produces the synergy prediction. The linear output is appropriate as the synergy score is continuous and, in this case, corresponds to the Bliss independence [29] synergy metric.

The SAN architecture is formally defined as:

$$\begin{aligned} san_1(h_1, h_2) &= \text{Dropout}(\text{ReLU}(\text{FC}(\text{Concat}(h_1, h_2)))) \\ san_2(h_1, h_2) &= \text{Dropout}(\text{ReLU}(\text{FC}(san_1(h_1, h_2)))) \\ SAN(h_1, h_2) &= \text{Linear}(\text{FC}(san_2(h_1, h_2))) \end{aligned} \quad (5.5)$$

where h_1 and h_2 represent the conditional latent encodings of drugs i and j , respectively, and san_1 and san_2 denote the intermediate transformation layers.

Using this formulation, the final predicted synergy score for the drug pair (i, j) in cell line k is computed as:

$$\hat{y}_{i,j,k} = SAN(h_{i,k}, h_{j,k}) \quad (5.6)$$

This predicted synergy score $\hat{y}_{i,j,k}$ is used to assess the potential therapeutic benefit of combining drugs i and j in the biological context defined by cell line k .

5.1.4 Model Training and Fine-Tuning Strategy

5.1.4.1 Pretraining on Cell Line Data

We trained the PDSP model on the cell line synergy dataset using a combined loss function that accounts for both synergy and single-drug predictions. The synergy prediction was treated as a regression problem, for which we used the

mean squared error (MSE) between the predicted Bliss score $\hat{y}_{i,j,k}$ and the ground truth $y_{i,j,k}$ as the loss component L_{reg} .

$$L_{reg} = \frac{1}{N} \sum_{i,j,k \in D} (y_{i,j,k} - \hat{y}_{i,j,k})^2 \quad (5.7)$$

The single-drug sensitivity predictions were trained using a *weighted binary cross entropy* (wBCE), reflecting the classification of each drug t on each cell k as sensitive or resistant. In particular, for each drug in a pair we compute a binary cross-entropy (BCE) between the true class (e.g., 1 for sensitive, 0 for resistant) and the predicted probability from the DDN. Because the dataset may be imbalanced (e.g., more drugs are classified as resistant than sensitive in the training data), we applied class-specific weights λ_1 and λ_2 to the positive and negative instances to form a weighted BCE loss. The formula for wBCE:

$$\begin{aligned} L_{cls} &= wBCE(i, k) + wBCE(j, k) \\ wBCE_{t,k} &= -\frac{1}{N} \sum_{t,k \in D} \lambda_1 s_{t,k} \log(\hat{s}_{t,k}) + \lambda_2 (1 - s_{t,k}) \log(1 - \hat{s}_{t,k}) \end{aligned} \quad (5.8)$$

where N is the number of drug combinations in the training dataset D . $y_{i,j,k}$ is the ground-truth Bliss score of drug pair $\langle i, j \rangle$ on cell line k . $s_{t,k}$ is the ground-truth sensitivity class of drug t on cell line k . $\hat{y}_{i,j,k}$ and $\hat{s}_{t,k}$ are the corresponding predictions. λ_1 and λ_2 are the loss weights for sensitive and resistant drugs on cell line k .

The total training objective for a given training sample (drug i , drug j , cell k) is a sum of the synergy regression loss and the two single-drug classification losses:

$$L = L_{reg} + \theta * L_{cls} \quad (5.9)$$

where θ is a hyperparameter that adjusts the weight between regression and

classification losses. We set θ to 10 in our experiments.

5.1.4.2 Transfer Learning for Patient Customization

A key innovation of PDSP is the ability to fine-tune (see Figure 5.1 Step 2) the model using patient-specific data, despite having very few combination examples for a patient. We devised a two-step fine-tuning strategy to customize the pretrained model for each new patient, leveraging the model’s single-drug prediction outputs. In essence, we use the patient’s single-drug response measurements (which are far more abundant than combination measurements) to adjust the drug encoder weights so that the model’s latent representation aligns with the patient’s biology.

We begin by pretraining the PDSP architecture using drug combination data derived from cancer cell line experiments, enabling the model to learn generalizable representations of drug–cell line interactions. In the subsequent fine-tuning phase, we adapt this pretrained model to patient-specific data to facilitate personalized predictions. During fine-tuning, we freeze the weights of the Synergy Aggregation Network (SAN) to preserve its synergy inference capability learned from the cell line data, and selectively update the parameters of the Drug Encoder Networks (DENs). This strategy allows the model to tailor the encoding of drug–patient interactions based on patient-specific gene expression profiles.

We explore two distinct fine-tuning strategies using patient single-drug response data and evaluate their impact on predictive performance. This process enables the model to learn individual variability while retaining its ability to generalize synergy predictions.

Let Φ_c denote the set of pretrained model weights obtained from cell line-based training. Each patient’s untreated gene expression profile is represented as $c_{p_t} \in \mathbb{R}^{542}$, where $t \in \{1, 2, 3\}$, and each drug’s chemical descriptor vector is denoted by $d_i \in \mathbb{R}^{541}$. The fine-tuning objective involves adapting Φ_c using patient-specific monotherapy data as follows:

Global Patient-Adaptation (Strategy 1, PDSP^{st1}): In the first fine-tuning strategy, we aggregate all available patient single-drug data (from the few patients we have) to adapt the model from the cell line domain to the patient domain. Starting from the cell-line pretrained PDSP model Φ_c , we continue training the model using only the single-drug sensitivity prediction losses for drugs tested on patient samples, while excluding any combination data (since we want to reserve the 20 patient drug pairs for final evaluation). In this step, the drug encoder (DEN) and decoder (DDN) weights are updated using patient data, while the synergy prediction branch (SAN) is frozen (it does not receive training signals because no patient combination labels are used in fine-tuning). We apply this adaptation using all patient samples together (treating the small set of patients as additional "cell lines" in a sense) to learn general shifts needed for patient gene expression distributions and drug response patterns. This global fine-tuning (which we call PDSP^{st1}) results in a single adapted model $\Phi_{c,all}$ that is not yet personalized to one specific individual, but rather tuned to patient-derived cells in general. It aims to adjust the model’s internal representations from those optimal for cell lines to those better aligned with primary patient cells.

Per-Patient Personalization (Strategy 2, PDSP^{st2}): In the second fine-tuning strategy, we personalize the model for each individual patient. Starting from the cell-line trained global model Φ_c , we fine-tune the model further on each patient’s own data. For a given patient, we take that patient’s set of single-drug sensitivity measurements (e.g., 218 drug sensitivity labels for that patient) and perform additional gradient updates to the DEN and DDN weights using only that patient’s data. This yields a unique model instance $\Phi_{c,t}$ ($t \in \{1, \dots, n\}$) for each patient (n is the number of patients). We evaluate each patient’s drug combination synergy score with the corresponding model’s weights $\Phi_{c,t}$. We refer to the models generated using this strategy PDSP^{st2}.

We found that this individualized training (even within an order of a few hundred data points) is feasible due to the strong initialization from PDSP. In both fine-tuning stages, we prevent catastrophic forgetting of the synergy task: Although the SAN weights were not updated (due to a lack of combo labels), the encoder adjustments can shift the inputs to SAN. We assume that as long as

the single-drug predictions drive the encoders to a better representation of the patient’s biology, the synergy predictions will indirectly become more accurate for that patient as well. By employing these transfer learning approaches, PDSP can incorporate patient-specific information (gene expression and drug response profile) to personalize synergy predictions. Notably, this personalization is accomplished without requiring any patient combination trials in the training process – an important property because, in practice, we often cannot test many combinations on a patient due to limited sample availability. Instead, PDSP leverages the relatively accessible single-drug testing data to bridge the gap between the patient and the preclinical cell line data.

5.1.5 Compared Methods and Baselines

To assess the effectiveness of PDSP, we compared it against several baseline methods and evaluated all models on both cell line and patient tasks. The baseline methods include both classical machine learning and recent deep learning approaches for drug synergy prediction. We compare PDSP’s performance with base model MatchMaker and the other state-of-the-art models: DeepSynergy [22], TreeCombo [23], and DeepDDS [49]. We also use a standard machine learning algorithm, Random Forest, as the baseline model. DeepDDS is a graph neural network-based model for drug synergy, which encodes each drug’s molecular graph structure with a GNN and combines them with cell line features. DeepDDS exemplifies using more advanced drug representations (graphs) and was reported to perform well on synergy tasks (other models are already explained in the previous chapter).

5.2 Results

5.2.1 Experimental Setup

By using the new drug combination scenario (see Section 2.4 for details), the entire dataset of drug combinations is partitioned into training, validation, and test sets with a ratio of 60%, 20%, and 20%, respectively. All model hyperparameters are optimized using the validation set, and final model performance is reported based on the held-out test set consisting of unseen drug combinations in cancer cell lines.

Additionally, the best-performing model and weights obtained from the training phase, characterized by Φ_c , are used as the initialization for fine-tuning on patient-specific single-drug response data, as described in Section 5.1.4.2. This transfer learning step allows us to leverage the knowledge from large-scale cell line data to inform personalized synergy prediction in a patient-specific context.

5.2.2 Performance Benchmark

5.2.2.1 Evaluation Metrics

We evaluated model performance on two levels: The regression quality of synergy prediction and classification accuracy in identifying synergistic vs. antagonistic pairs. For the regression task (predicting Bliss scores), we report the MSE between predicted and actual synergy, as well as the Pearson correlation coefficient and Spearman rank correlation between predictions and ground truths across the test set. These metrics measure how well the model captures the continuous variation in synergy scores and rank-orders drug pairs by efficacy. For the classification view, we define positive cases as those drug pairs with $\text{Bliss} > 0$ (synergy) and negative cases as $\text{Bliss} < 0$ (antagonism). We then evaluate the models in a post hoc manner: Without retraining, we take the synergy score predictions and consider their sign (or a threshold on the score) to classify each pair. We report

balanced accuracy (BACC) – the average of true positive and true negative rates, which accounts for class imbalance – and Cohen’s Kappa coefficient [98] to measure agreement with the true labels beyond chance. Additionally, we computed the partial Area Under the ROC Curve (pAUC) up to a false positive rate of 0.1. This pAUC focuses on performance in the low-false-positive regime, reflecting the model’s ability to confidently identify some synergistic pairs with very few false alarms (a desirable trait in drug discovery, where we want to flag a subset of top candidate pairs with high precision). No model was specifically trained for classification; this evaluation simply thresholded the regression outputs, which provides insight into the models’ practical utility in picking out synergistic vs. antagonistic outcomes.

For patient-specific evaluation, we emphasize classification-oriented metrics (precision, recall, accuracy) because the number of tested combinations per patient (only 20 in total across 3 patients) is too small to compute correlation coefficients reliably. In this setting, correctly identifying a synergistic or antagonistic pair is more meaningful than a slightly better numeric prediction. Therefore, we compare models based on how many 10 truly synergistic pairs and 10 truly antagonistic pairs they predict correctly for the patients, reporting metrics like precision (positive predictive value), recall (sensitivity) for each class, and overall accuracy. We also examine the confusion matrix of predictions (true vs. predicted synergistic/antagonistic) to understand the types of errors each model makes. This analysis is especially important for personalized models (PDSP^{st1}, PDSP^{st2}) versus non-personalized ones, highlighting the impact of fine-tuning.

All results are reported on hold-out test sets: For cell lines, on a reserved test set of combinations not seen during training; for patients, on the 20 patient combination outcomes which were never used in training or fine-tuning. Throughout training and evaluation, we ensured that no information from the test sets leaked into model development, thus the reported performance reflects true generalization.

5.2.2.2 Regression Performance on Cancer Cell Lines

Table 5.2: Comparison of model performance in synergy prediction using mean squared error (MSE), Pearson correlation, and Spearman correlation coefficients.

	MSE	Pearson	Spearman
PDSP	41.95	0.91	0.65
MatchMaker	41.60	0.91	0.64
DeepSynergy	50.65	0.89	0.59
TreeCombo	61.89	0.87	0.54
DeepDDS	177.33	0.54	0.26
Random Forest	165.72	0.60	0.31

On the cell line test set, PDSP achieved excellent accuracy in predicting combination synergy scores, which is on par with the best existing method. Table 5.2 summarizes the regression performance of PDSP against the baseline models in terms of MSE and correlation metrics. PDSP attained an MSE of 41.95, which is essentially tied with MatchMaker (41.60) and substantially lower (better) than the other methods. In practical terms, PDSP’s error was about 20% lower than the next best model beyond MatchMaker (DeepSynergy had MSE 50.65). Moreover, PDSP reached a Pearson correlation of 0.91 with the true Bliss scores on held-out cell line pairs. This indicates a strong linear agreement between predictions and actual outcomes – the model can explain a large fraction of the variance in synergy responses across diverse drugs and cell contexts. The Spearman rank correlation was 0.65 for PDSP, slightly better than MatchMaker (0.64). This suggests that PDSP preserved the ability to correctly rank-order combinations by synergy even after incorporating the extra task of single-drug prediction. Notably, the fact that PDSP and MatchMaker are almost indistinguishable in performance across all these metrics demonstrates an important point: Adding the drug response prediction branch (DDNs) did not degrade the core synergy prediction capability on cell line data. In other words, the multi-task learning approach successfully maintained synergy prediction accuracy (PDSP performed on par with MatchMaker, the previous state-of-the-art that lacks any personalization component). All other baselines lagged behind significantly. For instance, DeepSynergy reached a Pearson correlation of 0.89 (2% lower) and Spearman correlation of 0.59 (6% points lower), indicating that its predictions were less correlated with true

outcomes. TreeCombo and the Random Forest had even poorer correlations. In fact, the Random Forest’s error was extremely high (MSE 165.72), suggesting it essentially failed to learn meaningful patterns and mostly predicted near the mean response. Interestingly, DeepDDS underperformed even the simpler TreeCombo and Random Forest on this task. This could be due to the complexity of graph neural networks overfitting the limited data per cell line or difficulties in training. Overall, these results establish that PDSP’s base model is as accurate as the best existing method for predicting synergy in cell lines, and considerably better than other deep learning or machine learning approaches. This strong baseline is critical before moving to personalized predictions. It means we are not sacrificing any predictive power when building personalization capability.

We also analyzed the model performance on a per-cell-line basis to ensure PDSP is robust across different genetic backgrounds. For each cell line in the test set, we computed the correlations and MSE separately and examined whether performance correlates with how many training examples that cell line had in the training data. Interestingly, we observed no strong dependency between sample count and accuracy – some cell lines with relatively few training combinations were still predicted with high accuracy, and vice versa (detailed in Table 5.3). This suggests the model captures global patterns of drug response that are generalized even to cell contexts that it sees less frequently. It also implies that the training data was sufficiently diverse and the model did not simply memorize cell-line-specific effects, but learned transferable representations.

5.2.2.3 Classification Performance on Cancer Cell Lines

Beyond continuous prediction, we evaluated how well PDSP and other models can classify a drug pair as synergistic or antagonistic. This perspective is essential in practical settings: Researchers often want to flag which combinations are likely to work (synergy) and which are likely to be ineffective or counteractive (antagonism). Using the predicted Bliss scores from each model, we labeled pairs as positive (synergy) if $\hat{y}_{i,j,k} > 0$ and negative if $\hat{y}_{i,j,k} < 0$, and compared to the true labels (since the test set ground truth includes both positive and negative Bliss

Table 5.3: Top 10 cell line performance of different models

(a) PDSP

Cell line	# of occurrences in training data	MSE	Pearson	Spearman
OVCAR-4	2934	10.56	0.59	0.48
SNB-75	2822	11.56	0.65	0.49
NCI-H226	2991	13.37	0.60	0.48
LOVO	365	13.56	0.40	0.44
A427	365	13.70	0.68	0.66
TK-10	2962	14.10	0.67	0.50
UWB1289	365	14.23	0.56	0.54
OV90	365	14.30	0.57	0.65
EKVX	2876	14.72	0.60	0.51
OVCAR-5	2970	14.92	0.66	0.53

(b) MatchMaker

Cell line	# of occurrences in training data	MSE	Pearson	Spearman
OVCAR-4	2934	10.66	0.59	0.46
SNB-75	2822	12.03	0.63	0.47
LOVO	365	13.25	0.38	0.43
UWB1289	365	13.51	0.56	0.56
NCI-H226	2991	13.79	0.58	0.46
TK-10	2962	14.52	0.66	0.49
OV90	365	14.75	0.54	0.60
A427	365	15.04	0.64	0.66
SK-OV-3	3301	15.68	0.62	0.52
OVCAR-5	2970	15.78	0.63	0.53

(c) DeepSynergy

Cell line	# of occurrences in training data	MSE	Pearson	Spearman
OVCAR-4	2934	12.30	0.52	0.36
OV90	365	14.46	0.58	0.60
LOVO	365	14.67	0.28	0.31
SNB-75	2822	14.75	0.53	0.41
UWB1289	365	15.47	0.52	0.53
NCI-H226	2991	16.70	0.48	0.37
PC-3	2944	17.35	0.46	0.40
A498	3059	17.50	0.54	0.44
EKVX	2876	18.33	0.51	0.41
OVCAR-5	2970	18.51	0.59	0.49

values). Note that we do not perform any training for this task, and we use the discretized predictions and ground truth to measure classification performance.

Table 5.4 reports metrics for this binary classification on the cell line test set. Consistent with the regression results, PDSP and MatchMaker outperform other methods in classification performance. They exhibit similar performances, with only slight differences in different metrics. PDSP achieves slightly higher scores in AUPR (0.81 vs. 0.80) and Precision (0.70 vs. 0.69), while MatchMaker archives slightly higher F1 (0.73 vs. 0.72) and Recall (0.76 vs. 0.75) scores. PDSP and MatchMaker achieved the highest balanced accuracy (BACC), substantially outperforming the other models. Both models obtained a higher BACC (0.70), indicating they effectively distinguish synergistic from antagonistic outcomes despite the imbalance (typically, more combinations are non-synergistic). We also found that PDSP and MatchMaker had the best Cohen’s Kappa (0.40), reflecting strong agreement with ground truth labels beyond chance. In contrast, some baselines struggled: For example, the Random Forest and TreeCombo models tended to predict most pairs as synergistic, likely because predicting near the mean Bliss score (which is slightly above zero in many datasets) minimizes MSE. This led to a high false-positive rate for synergy and thus a low balanced accuracy. Their recall for antagonism was low. DeepSynergy performed better than those classical models, but still a notch below PDSP, while DeepDDS again underperformed, possibly outputting erratic predictions that did not align well with the binary labels.

Table 5.4: Comparison of classification performance across models using various evaluation metrics. Drug combination experiments with a Bliss score greater than 0 are labeled as positive (synergistic), while those below 0 are considered negative (antagonistic).

	AUC	pAUC	AUPR	F1	Precision	Recall	Kappa	BACC
PDSP	0.79	0.64	0.81	0.72	0.70	0.75	0.40	0.70
MatchMaker	0.79	0.64	0.80	0.73	0.69	0.76	0.40	0.70
DeepSynergy	0.76	0.62	0.78	0.70	0.68	0.73	0.35	0.68
DeepDDS	0.61	0.53	0.63	0.62	0.59	0.65	0.16	0.58
TreeCombo	0.73	0.60	0.75	0.68	0.67	0.79	0.32	0.66
Random Forest	0.63	0.55	0.66	0.54	0.63	0.48	0.17	0.59

One metric of particular interest is the partial AUC at 10% false positive rate

(pAUC). PDSP achieved a larger partial AUC in this low-FP region compared to baselines, meaning that if we allow only a 10% false alarm rate (a strict criterion useful for experimental prioritization), PDSP captures more true synergies within that allowance. In practical terms, this means PDSP can identify a set of top candidate combinations with high confidence (very few false hits). The strong performance in pAUC showcases PDSP’s potential in scenarios where only a handful of predictions can be followed up experimentally – it will likely suggest truly synergistic pairs near the top of its ranking. Altogether, the cell line classification results demonstrate that PDSP fits the quantitative synergy values well and makes qualitatively correct decisions about which drug pairs are synergistic, reinforcing its reliability for the next stage of patient-specific prediction.

5.2.2.4 Patient Personalized Classification Performance

Having established PDSP’s solid performance on cell line data, its ultimate goal is to perform personalized predictions for cancer patients. Therefore, we evaluated how well the model can predict drug pair synergy for individual patients after fine-tuning. As described earlier, we focus on the 20 drug pair tests from three leukemia patients (from [58]). These 20 pairs (each with a known synergy or antagonism outcome) represent a very challenging evaluation set, as they are completely outside the model’s original training distribution, and we have very limited patient-level ground truth data to train a complex machine learning model. For this reason, we employ a transfer learning approach in the PDSP design and use patients’ single drug responses to fine-tune the model as explained in Section 5.1.4.2.

We compare three scenarios: (i) the non-personalized PDSP (i.e., the base model trained on cell lines, applied directly to patient data without any fine-tuning), and the two personalized versions (ii) PDSP^{st1} (after global patient-data fine-tuning) and (iii) PDSP^{st2} (after subsequent per-patient fine-tuning). For additional context, we also applied the baseline models (MatchMaker, DeepSynergy, TreeCombo, DeepDDS and Random Forest) to the patient pairs. Those baseline models were trained on cell line data only and cannot be further personalized since

they do not have a mechanism for it. However, we can still feed the patient’s gene expression profile as input and obtain a prediction for the combination. For models that require gene expression, we use the patient’s profile just as we would a cell line’s profile. Figure 5.2 shows the number of *True Positives* (TPs), *True Negatives* (TNs), *False Positives* (FPs) and *False Negatives* (FNs).

Non-personalized Performance: The base PDSP model, without any patient-specific tuning, did not perform very well on the patient combination predictions – underscoring the necessity of our transfer learning approach. In fact, its performance was comparable to other cell-line-trained models: All suffered a drop in accuracy when faced with patient data. Qualitatively, we observed that non-personalized models often mislabeled antagonistic drug pairs as synergistic for the patients. For instance, the Random Forest baseline was an extreme case, it predicted every single pair as synergistic for all patients. This resulted in zero true negatives (it failed to recognize any actual antagonistic combinations) and many false positives. MatchMaker and TreeCombo, while slightly better, still predicted an overwhelming majority of pairs as synergistic (17/20 and 18/20 predicted positive, respectively). These models acted almost like "*one-class*" predictors. Consequently, their precision in identifying true synergistic pairs was low; many pairs they labeled as "synergy" were antagonistic in reality. Specifically, MatchMaker achieved only 0.47 precision for synergy predictions on the patient data, and TreeCombo about 0.56 precision. This means that over half of the combinations they flagged as synergistic were actually mispredictions (false alarms). The unpersonalized PDSP exhibited a similar issue: It tended to over-predict synergy. It achieved a precision of only 0.43 and an overall accuracy of 40% on the 20 patient pairs. Regarding the confusion matrix, PDSP (non-tuned) identified some antagonistic cases correctly, but still misclassified the majority of antagonistic pairs as synergistic. The accuracy of these non-personalized models was accordingly poor (around 40–50% at best, which is only slightly better than random guessing 50% in this balanced dataset of 10 synergistic/10 antagonistic). These findings highlight that models trained solely on cell lines do not directly translate to patients, likely due to the significant biological differences. It also sets a baseline: Any improvement with personalization should manifest as higher

Pat. ID	Compound 1	Compound 2	Random Forest	TreeCombo	DeepSynergy	MatchMaker	DeepDDS	PDSP	PDSP ^{st1}	PDSP ^{st2}
1263	FIMM003747-003	FIMM003787-001	+	+	-	+	+	+	+	-
1263	FIMM000618-004	FIMM109450-001	+	+	+	+	+	-	-	-
1263	FIMM100381-001	FIMM003747-003	+	+	+	-	+	-	-	-
1263	FIMM109441-001	FIMM003747-003	+	+	+	+	+	+	+	+
1263	FIMM100388-001	FIMM115484-001	+	+	-	+	-	-	-	-
1263	FIMM001244-002	FIMM115484-001	+	+	+	+	-	+	+	-
1409	FIMM109442-001	FIMM003735-001	+	+	+	+	-	+	-	+
1409	FIMM115616-002	FIMM003735-001	+	+	-	+	-	-	+	+
1409	FIMM109442-001	FIMM115616-002	+	+	+	+	-	+	+	+
1409	FIMM000296-002	FIMM000618-004	+	+	+	+	-	+	-	-
1409	FIMM115616-002	FIMM100346-001	+	+	+	+	-	-	-	+
1409	FIMM100381-001	FIMM003747-003	+	+	+	-	+	-	-	-
1409	FIMM109441-001	FIMM003747-003	+	-	+	+	+	+	+	+
1508	FIMM003790-001	FIMM003767-001	+	+	-	+	+	+	-	-
1508	FIMM000296-002	FIMM000618-004	+	+	-	+	+	+	-	-
1508	FIMM000618-004	FIMM109450-001	+	+	+	+	+	+	-	-
1508	FIMM000296-002	FIMM109450-001	+	+	+	-	+	+	-	-
1508	FIMM100381-001	FIMM003747-003	+	-	+	+	+	+	-	-
1508	FIMM109441-001	FIMM003747-003	+	+	+	+	+	+	+	-
1508	FIMM001244-002	FIMM115484-001	+	+	+	+	-	+	+	-
TRUE +			10	10	7	9	4	6	5	5
TRUE -			0	2	2	2	2	2	7	9
FALSE +			10	8	8	8	8	8	3	1
FALSE -			0	0	3	1	6	4	5	5
ACCURACY			$\frac{10}{20}$	$\frac{12}{20}$	$\frac{9}{20}$	$\frac{11}{20}$	$\frac{6}{20}$	$\frac{8}{20}$	$\frac{12}{20}$	$\frac{14}{20}$

Figure 5.2: Classification results on patient drug combination synergy data. All models are pretrained on DrugComb using cell line gene expression profiles. PDSP^{st1} and PDSP^{st2} represent personalized variants fine-tuned using Strategy 1 and Strategy 2, respectively. In the figure, correct predictions are shown with blue rectangles, while incorrect predictions are highlighted in red. Synergistic outcomes are marked with a plus (+) symbol, and antagonistic outcomes with a minus (-) symbol.

precision (fewer false positives) and higher recall of antagonism.

Impact of Fine-Tuning (PDSP^{st1} and PDSP^{st2}): Once PDSP was fine-tuned with patient data, we observed a dramatic improvement in its ability to predict patient-specific combination outcomes correctly. Even the first strategy of using all patients’ data (PDSP^{st1}) provided a noticeable boost. PDSP^{st1} managed to reduce the spurious synergy predictions and started to correctly identify more antagonistic pairs. According to the results, PDSP^{st1} achieved about 0.63 precision in its synergy predictions and an accuracy of 60%. In concrete terms, this model brought down the number of false positives (misclassified antagonisms) to 3 (out of 10 antagonistic pairs) and correctly predicted 7 out of 10 antagonistic cases (true negatives). Compared to the non-personalized PDSP, which had only 2 true negatives, this is a significant improvement in specificity. The recall for synergistic pairs for PDSP^{st1} is still modest. It missed a few of the actual synergistic pairs, which is expected, given that it became more conservative in calling synergy. Overall, PDSP^{st1}’s accuracy was around 60%, a 20% increase over the base model. This showcases the benefit of even a coarse-grained personalization: Using just a small pool of patient drug-response data, the model shifts its decision boundary to discriminate synergy from antagonism in patient contexts better.

The most striking results came with per-patient fine-tuning (PDSP^{st2}). By customizing the model to each individual, PDSP^{st2} achieved the best performance on all metrics for the patient data. Aggregating across the three patients’ outcomes, the per-patient models reached a precision of 0.83 – a huge jump from the 0.43 without fine-tuning. This high precision means that the majority of drug pairs PDSP^{st2} predicts as synergistic are truly synergistic, greatly reducing false positives. In terms of catching antagonistic combinations (which none of the non-personalized methods handled well), PDSP^{st2} was nearly perfect: It correctly recognized 9 out of the 10 actual antagonistic pairs (0.90 specificity), misclassifying only one antagonistic pair as synergistic (1 false positive). The cost of this increased antagonism detection was minimal. PDSP^{st2} still correctly identifies a significant fraction of truly synergistic pairs while becoming nearly perfect in antagonistic detection. Although PDSP^{st2} might have missed one or two of

the actual synergistic pairs (in exchange for minimizing false positives), it overall achieved the highest overall accuracy (0.70) across all models. These results demonstrate the practical clinical significance of PDSP’s personalization. In a clinical or *ex-vivo* testing scenario, predicting an antagonistic (or non-synergistic) drug combination is just as critical as predicting a synergistic one. Administering two drugs that antagonize each other could lead to wasted effort or even adverse effects, without providing patient benefit. The fine-tuned PDSP model’s strength lies in its ability to accurately identify and avoid antagonistic combinations for a patient.

While an ideal model would perfectly identify both synergy and antagonism, our results show that the fine-tuned models skewed towards high precision on antagonism. This bias can actually be desirable in practice: It is arguably more critical to avoid recommending a harmful or ineffective combo (false positive for synergy) than to miss a potentially useful combo (false negative for synergy), since the latter could be discovered in a later trial. The PDSP^{st2} model’s near-perfect antagonism detection gives clinicians a reliable filter to rule out dangerous combinations, thereby narrowing the search space to a smaller set of candidate synergistic pairs with high confidence. Indeed, our fine-tuned model boosts the overall accuracy of predicting patient-specific outcomes to about 0.70, a sizable improvement over the 0.30 accuracy of non-personalized approaches. This 30% gain means that personalized PDSP can make correct calls in roughly 6 more cases out of 20 than the original model, which in a real-world scenario could translate to sparing a patient from multiple ineffective treatments and guiding them towards effective ones.

In summary, the results confirm that PDSP’s transfer learning approach is effective. It retains top-tier performance on standard cell line benchmarks and, more importantly, delivers a substantial improvement in predicting drug synergy for individual patients. By fine-tuning on patient-specific drug response data, PDSP overcomes the translational gap that left other models ineffective, improving prediction accuracy and yielding much better predictions (fewer false positives for synergy). The personalized models particularly excel at identifying antagonistic drug pairs, which is crucial for clinical decision-making to avoid

detrimental combinations. These findings support the feasibility of in silico personalized medicine in oncology: Even with minimal patient data, a deep learning model like PDSP can be tailored to predict how a patient will respond to a new drug combination, thus aiding in the design of optimal combination therapies.



Chapter 6

DeepSynBa: A deep learning method to predict dose-response matrices for drug combinations

Previous drug synergy prediction models, including both classical machine learning approaches and deep learning frameworks like MatchMaker and PDSP, typically output a single scalar synergy score for each drug pair–cell line combination. While effective for ranking combinations, these models cannot capture dose-dependent interaction dynamics, offering limited guidance on optimal dosing strategies or mechanistic interpretation. To address this limitation, DeepSynBa introduces a novel framework that predicts the entire dose–response matrix for each drug combination in a given biological context. To parameterize the full dose–response surface and to disentangle key pharmacodynamic properties such as synergistic potency and efficacy, DeepSynBa adopts a methodology inspired by SynBa [1], incorporating a simplified variant of the two-dimensional generalized Hill equation derived from the MuSyC model [99]. A novel intermediate layer within the DeepSynBa architecture is implicitly trained to learn the parameters of this generalized Hill equation. These parameters allow for the reconstruction of the entire dose–response matrix for any given drug combination, as illustrated in panels (a) and (e) of Figure 6.2. Trained on the large-scale NCI-ALMANAC

dataset [21], the model significantly improves prediction accuracy over existing baselines and provides a mechanistically grounded foundation for optimizing combination therapies in precision oncology.

This modeling strategy supports continuous dose-response surface generation and permits the post-hoc computation of widely used synergy metrics, including Loewe additivity and Bliss independence. By predicting mechanistically interpretable pharmacodynamic parameters, DeepSynBa maintains compatibility with established drug synergy benchmarks while preserving architectural flexibility, avoiding constraints imposed by any single synergy definition.

6.1 Methodology

6.1.1 Problem Formulation

We address drug combination response prediction as a dose-response matrix regression task rather than reducing outcomes to a single synergy score. In this setting, each data point corresponds to a triplet of two drugs $\langle i, j \rangle$ and a cell line k , with the target output being the complete matrix of combination responses (e.g., percent viability or growth inhibition) across a grid of dose levels for the two drugs. Formally, given a dosage grid (e.g., concentrations x_i for drug i and x_j for drug j), the model is trained to predict all entries of the response matrix $y(i, j, k) \in \mathbb{R}^{N_i \times N_j}$, where N_i and N_j indicate the number of dosages of drugs i and j , respectively, used in the grid. Each drug i and j is represented by feature vectors d_i and d_j , while the cell line k is represented by a feature vector c_k . The goal is to predict $y_{x_i, x_j}(i, j, k)$ by leveraging the drug-specific features (d_i and d_j), cell line-specific features c_k , and the dosages (x_i, x_j) used in the experiment. By formulating the problem this way, we capture detailed dose-response surfaces instead of a single aggregated synergy value.

This approach retains information about potency (how rapidly efficacy is achieved as dose increases) and efficacy (the maximal achievable effect), which

are lost when condensing outcomes into one synergy score. Predicting the full response matrix enables post-hoc calculation of any synergy metric (like Loewe or Bliss scores) while avoiding the oversimplification inherent in single-score models. It also reflects experimental reality more closely, since high-throughput screens often measure viability at multiple dose combinations for each drug pair. Thus, we train a model to output the entire dose-response matrix for a given drug pair and cell line, treating it as a multi-output regression problem over the dose grid.

6.1.2 Dataset and Processing

To conduct a comprehensive comparative analysis of predictive models, we utilized the publicly available NCI-ALMANAC dataset [21]. This large-scale pharmacogenomic resource provides extensive measurements of drug combination effects across cancer cell lines. Our dataset comprises 274,516 dose-response matrices, representing the combinatorial effects of 101 unique small-molecule compounds across 60 genetically diverse cancer cell lines. Each matrix is structured as a 4×4 grid, where the first row and first column capture the individual responses to single-drug treatments. In subsequent modeling, these monotherapy responses are crucial for distinguishing between additive, synergistic, and antagonistic interactions.

When multiple replicates exist for the same drug-drug-cell line triplet at identical dosage levels, we resolve potential ambiguities by aggregating the values using the median response. This strategy prevents one-to-many mappings in the input-output relationship and ensures that the model receives a deterministic training target, enhancing reproducibility and interpretability.

For the definition of drug response values y , we adopt the **PercentGrowthNoTZ** column from the NCI-ALMANAC annotations, rather than the more commonly used **PercentGrowth** metric. Both metrics quantify normalized cellular proliferation, with 100 corresponding to the growth observed in untreated control wells. However, they differ in how they scale responses relative to baseline conditions. The **PercentGrowthNoTZ** column sets 0 as the point of complete cell death, while

the `PercentGrowth` column allows negative values, with -100 denoting complete cytotoxicity and 0 indicating no net growth.

We chose `PercentGrowthNoTZ` for two key reasons. First, its strictly non-negative range aligns with the domain assumptions of the generalized Hill equation, which underpins the parametric formulation used in our dose-response modeling. Second, the asymmetric scaling of growth and death responses in `PercentGrowth` introduces discontinuities and non-linearities that complicate the fitting of smooth, biologically meaningful functions (such as Equation 6.1). Using `PercentGrowthNoTZ` thus facilitates stable optimization and interpretable surface fitting, enhancing the model’s ability to learn from complex combination response patterns across a wide range of drug and cell line conditions.

6.1.3 DeepSynBa Architecture

DeepSynBa is a deep learning framework designed to predict the complete dose-response matrix for a drug pair on a specific cell line by explicitly disentangling combination potency and efficacy. The architecture (illustrated in Figure 6.1) integrates two key components: (i) cell line-conditioned drug representation networks that learn informative embeddings for each drug in the context of a given cell line’s biology, and (ii) an intermediate parameter prediction layer that outputs the coefficients of a parameterized dose-response surface model. Using these intermediate parameters, the model reconstructs the full dose-response surface over the continuous dose space. This design allows the model to predict the full response matrix and also yields interpretable parameters linked to pharmacological concepts of potency and efficacy, addressing a major limitation of prior approaches.

The model leverages feature representations for both drugs, denoted as d_i and d_j , and cell lines, represented as c_k . Drug features are derived from vector embeddings generated using the MoLFormer model [100], which captures each compound’s molecular structure and properties. In parallel, cell line features are encoded as vectors comprising untreated gene expression profiles, providing

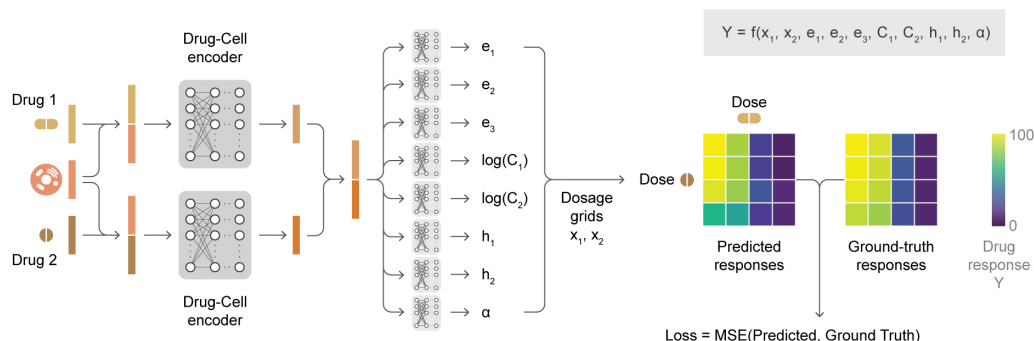


Figure 6.1: Overview of the DeepSynBa architecture. Drug and cell line features are first concatenated and passed through a series of fully connected layers to generate a unified latent representation. This representation is then processed by multiple prediction heads that output eight parameters defining the dose-response surface. During training, DeepSynBa uses the mean function from the SynBa likelihood model [1] to reconstruct the full dose-response matrices based on these predicted parameters. Model optimization is performed by minimizing the mean squared error (MSE) between the predicted and experimentally measured dose-response matrices.

a baseline characterization of the cellular state. To model drug-cell interactions, the feature vectors corresponding to each drug and the target cell line are concatenated, forming joint representations of the form $(d_i \oplus c_k)$ and $(d_j \oplus c_k)$. These concatenated vectors are subsequently passed through a series of fully connected layers. This neural transformation enables the extraction of higher-level abstractions, facilitating a more nuanced understanding of the interplay between molecular compounds and cellular environments.

The output of these layers constitutes drug-induced gene expression feature maps. These representations aim to capture the molecular-level impact of each drug on the gene expression profile of the given cell line. The model generates a detailed characterization of drug-specific cellular responses by encoding how individual drugs perturb the cellular transcriptome. Finally, the feature maps obtained from each drug-cell encoder are concatenated to form a unified representation. This aggregated vector serves as a comprehensive summary of the joint effect of both drugs on the cell line, enabling downstream tasks such as synergy

prediction or toxicity assessment. The model effectively captures the complex biological interactions underpinning drug responses in varying cellular contexts by integrating structurally-informed drug embeddings with transcriptomic profiles.

The concatenated drug-pair representation is then passed to eight prediction heads that predict parameters of a dose-response surface model. DeepSynBa follows the approach of SynBa [1] and uses a simplified form of the two-dimensional generalized Hill equation (derived from the MuSyC [101] model). This parametric model describes the combined effect of two drugs with separate terms for each drug’s single-agent efficacy and potency and additional terms capturing synergistic effects on potency and efficacy. Specifically, the prediction heads are $e_1, e_2, e_3, C_1, C_2, h_1, h_2$ and α from the likelihood function defined in SynBa, which are the parameters that characterise the expected value of the drug responses y .

Given the dosages x_1 for Drug 1 and x_2 for Drug 2, the expected value for a response y is defined as follows:

$$\begin{aligned} & \mathbb{E}[Y \mid e_0, e_1, e_2, e_3, C_1, C_2, h_1, h_2, \alpha, x_1, x_2] \\ &= \frac{C_1^{h_1} C_2^{h_2} e_0 + x_1^{h_1} C_2^{h_2} e_1 + C_1^{h_1} x_2^{h_2} e_2 + \alpha x_1^{h_1} x_2^{h_2} e_3}{C_1^{h_1} C_2^{h_2} + x_1^{h_1} C_2^{h_2} + C_1^{h_1} x_2^{h_2} + \alpha x_1^{h_1} x_2^{h_2}}, \end{aligned} \quad (6.1)$$

where e_0 represents the baseline drug effect observed without any pharmacological intervention. The parameters e_1 , C_1 , and h_1 are specific to Drug 1 and denote its maximal monotherapy effect, the concentration at which half-maximal effect is achieved (EC50), and the Hill coefficient reflecting the steepness of the dose–response curve, respectively. Analogously, e_2 , C_2 , and h_2 are the corresponding parameters characterizing the response to Drug 2 when administered alone. The term e_3 signifies the maximum effect achievable when both drugs are combined.

An additional parameter, denoted as α , works as an interaction coefficient that quantifies the degree to which the presence of one drug modulates the potency of the other. A value of $\alpha > 1$ indicates synergistic potency, implying that the combination of the two agents achieves a given pharmacological effect at lower

concentrations than would be required if the drugs acted independently. This synergy suggests a potentiation of efficacy when both agents are co-administered. For training stability within the DeepSynBa framework, the baseline effect e_0 is held constant and is dataset-dependent. For instance, within the NCI-ALMANAC dataset, e_0 is standardized to 100, ensuring consistency in model initialization and evaluation across experimental conditions.

Our approach incorporates the mean function of the Gaussian likelihood for drug responses as introduced in SynBa [1], and formally expressed in Equation 6.1. This mean function is derived from the two-dimensional generalized Hill equation proposed in MuSyC [101], which extends the classical Hill equation for monotherapy:

$$y = e_0 + \frac{e_1 - e_0}{1 + \left(\frac{C}{x}\right)^h} \quad (6.2)$$

to accommodate the interactive effects of two drugs in combination. Here, e_0 and e_1 represent baseline and maximal effects, C is the half-maximal effective concentration, x is the drug dose, and h is the Hill coefficient controlling the steepness of the response curve.

Among the five levels of complexity model variants in MuSyC’s two-dimensional formulation, we adopt the most complex model tier that still adheres to detailed balance, consistent with the approach taken in SynBa. This choice provides a principled trade-off: It allows for rich and biologically meaningful modeling of dose-response surfaces while avoiding the risk of overparameterization, which can lead to non-identifiability and instability in deep learning frameworks.

Within our architecture, the outputs of eight distinct prediction heads correspond to parameter estimates of the generalized Hill function. These parameters are then passed through the embedded SynBa Gaussian likelihood function to compute the full dose-response matrices. The model can be trained end-to-end by integrating this likelihood directly into the prediction pipeline. This design enables the accurate reconstruction of dose-response surfaces that capture the

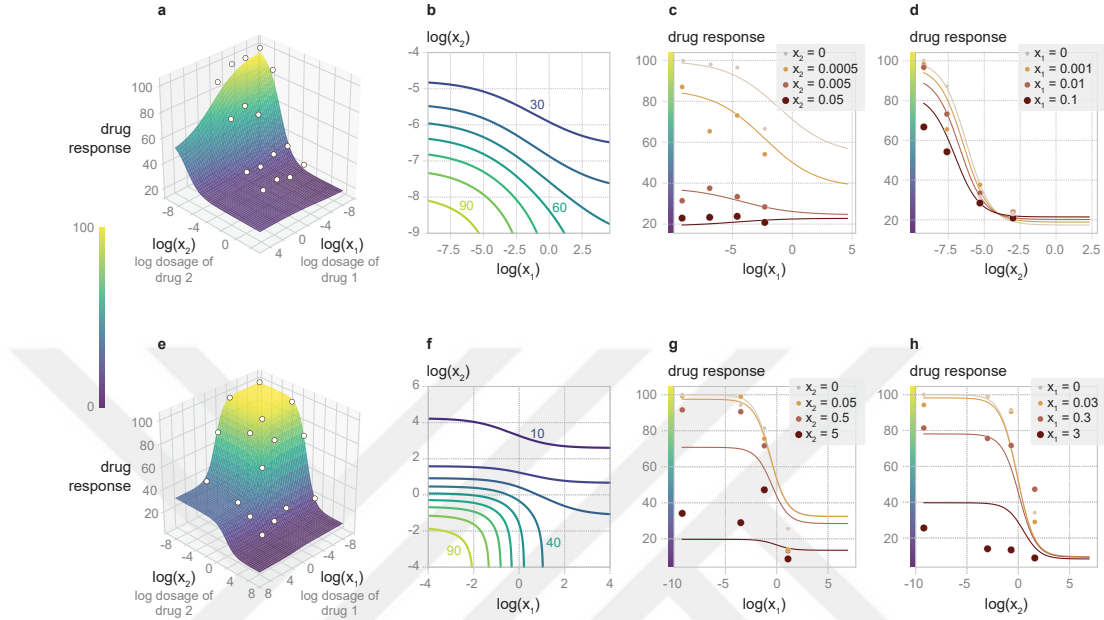


Figure 6.2: Predicted dose-response outputs from DeepSynBa for two distinct drug-cell line combinations, alongside comparisons with corresponding ground-truth experimental measurements. Panels (a)–(d) illustrate the combination of Drug 1 (Bleomycin Sulfate) and Drug 2 (Cabazitaxel) on cell line 786-0, while panels (e)–(h) depict the combination of Drug 1 (Azacitidine) and Drug 2 (Triethylenemelamine) on cell line MDA-MB-468. (a, e): Predicted dose-response surfaces with measured responses overlaid in red. (b, f): Contour plots showing predicted drug effects across varying dosage levels. (c, g): Dose-response curves at fixed concentrations of Drug 2 (in μM), comparing model predictions with experimental values. (d, h): Dose-response curves at fixed concentrations of Drug 1 (in μM), comparing predicted and measured responses.

nuanced effects of varying drug concentrations on cell line viability, supporting flexible and interpretable synergy modeling.

We can derive conventional synergy scores from the predicted surface as needed by plugging the surface into definitions of Loewe excess, Bliss independence, etc., thereby ensuring compatibility with existing synergy metrics. Significantly, the model is not constrained to optimize any synergy definition during training (avoiding bias toward one metric), but it remains flexible to compute any of them post hoc.

6.1.4 Training details

Model training was conducted for a maximum of 500 epochs using the AdamW optimizer [102], with a fixed learning rate of 10^{-4} , weight decay of 10^{-5} , and a mini-batch size of 128. Hyperparameters were tuned based on performance on the validation dataset to minimize the model’s loss function.

To optimize the performance of DeepSynBa, we systematically evaluated a range of hyperparameter configurations. The architecture of the Drug-Cell encoder comprises a single hidden layer, with its dimensionality selected from the set $\{256, 512, 1024, 2048, 4096\}$. This hidden layer is subsequently followed by batch normalization, a ReLU activation function, and dropout for regularization.

The model architecture also incorporates eight prediction heads, each structured as a sequence of three computational blocks. The dimensionality of the input layer to each prediction head is selected from $\{64, 128, 256\}$. Within each head, the successive blocks reduce the dimensionality by a factor of four at each stage. Each block is composed of a fully connected (dense) layer, followed by layer normalization, ReLU activation, and dropout, facilitating both stable training and effective generalization.

The final layer in each prediction head applies an activation function that depends on the expected range of the respective parameter of the SynBa likelihood function: Either a sigmoid activation for parameters constrained to the unit interval or a linear activation for unbounded parameters.

A comprehensive summary of all hyperparameter settings and tuning outcomes is provided in Table 6.1, which includes detailed justifications for selected configurations and their empirical performance during validation.

Table 6.1: The search space used for hyperparameter and training parameter optimization of the DeepSynBa architecture. This includes ranges and candidate values for network components, as well as optimization settings explored during model tuning.

Hyperparameter		Values
Drug-Cell Encoder	hidden layers	256, 512, 1024, 2048, 4096
	activations	ReLU
	input dropout	none, 0.2, 0.3, 0.5
	dropout	0.2, 0.5
	normalizations	batch normalization
Prediction heads	hidden layers	64, 128, 256
	activations	ReLU
	final layer activation	ReLU, sigmoid, linear
	normalizations	layer normalization
Training parameter	number of epochs	500
	loss	MSE, L1
	scheduler	LinearLR
	optimizer	AdamW
	batch size	(lr= 10^{-4} , wc= 10^{-5}) 128

6.2 Results

6.2.1 Case Studies

To demonstrate the enhanced interpretability and utility of DeepSynBa in predicting full dose–response surfaces, we analyze two illustrative cases from the NCI-ALMANAC dataset. These examples highlight the model’s capacity to capture nuanced interaction patterns between drug pairs, providing mechanistic insights beyond what scalar synergy scores can offer.

Case 1: Bleomycin Sulfate + Cabazitaxel on Renal Cancer Cell Line

786-0: The first case examines the combination of Bleomycin Sulfate (Drug 1), an antibiotic with chemotherapeutic properties, and Cabazitaxel (Drug 2), a potent chemotherapy agent used primarily in prostate cancer. The combination is evaluated on the 786-0 renal cancer cell line.

Figures 6.2(a)–(d) present the predicted and experimental results. Panel (a) shows the dose–response surface predicted by DeepSynBa, overlaid with actual measured responses (marked in white). The predicted surface exhibits a clear downward trend as the doses of both drugs increase, accurately reflecting the expected cytotoxic effects and matching the same pattern of measured responses. This decline along the y-axis represents increasing cell death.

Panel (b) provides a contour plot of the same surface, emphasizing dosage combinations that yield equivalent therapeutic effects. The contour lines for response levels of 90%, 80%, 70%, and 60% show inward curvature toward the origin. This pattern signals synergistic potency—i.e., when combined, lower doses of the two drugs produce the same level of efficacy as higher doses of either drug alone. In contrast, these contours would typically form concentric ovals or circles for non-synergistic interactions.

Interestingly, beyond the 60% response level, further increases in the dose of Drug 1 (right side of the x-axis) yield diminishing returns in therapeutic effect.

Meanwhile, Drug 2 continues to enhance efficacy, consistent with its role as a highly potent chemotherapeutic. This observation highlights Drug 2’s dominance in the combination at high doses and illustrates the potential benefit of synergistic potency at intermediate dosing ranges.

Panel (b) also demonstrates that synergy is dose-dependent, occurring only within specific dosage regions. This underlines the necessity of modeling full dose–response surfaces to uncover actionable dose windows that would remain undetected if synergy were collapsed into a single score.

To explore the dose-specific behavior in further detail, panels (c) and (d) show dose–response slices extracted from the full surface at fixed dosage levels. The predicted curves closely track the measured responses (represented as dots), validating the model’s fidelity. Both panels reveal that increasing the dose of one drug shifts the dose–response curve of the other downward, particularly in lower dose ranges. In panel (d), it is evident that Drug 1 does not improve maximal efficacy at higher doses, but reaches its IC50 at lower concentrations when combined with Drug 2. This illustrates the potential clinical benefit of using medium-to-low doses in combination to reduce off-target toxicity while maintaining therapeutic efficacy.

Case 2: Azacitidine + Triethylenemelamine on TNBC Cell Line MDA-MB-: The second example involves Azacitidine (Drug 1), a chemotherapy agent used primarily for leukemia, and Triethylenemelamine (Drug 2), which targets rapidly growing cells. The combination is tested on the MDA-MB-468 cell line, a model for triple-negative breast cancer (TNBC).

Panels (e)–(h) in Figure 6.2 present the corresponding results. The contour plot in panel (f) shows that the lines for response levels from 80% to 40% bend outward from the origin. This indicates a lack of synergistic interaction, as higher doses of both drugs are required to achieve the same therapeutic effect as monotherapies. Despite both drugs being individually effective, their combination does not offer added benefit in this context. This example illustrates how DeepSynBa can also identify non-beneficial combinations, offering critical insights

into when combination therapy may not be advantageous for a given tumor type.

These two case studies underscore DeepSynBa’s ability to disentangle and quantify synergistic potency and maximal efficacy across the dose space. Unlike conventional models that reduce interaction effects to a single synergy score, DeepSynBa provides a full landscape of drug interactions, revealing dose-specific behaviors, efficacy plateaus, and synergy windows.

As shown in Figure 6.2, such fine-grained analysis would not be possible if the interaction were summarized by a single metric. The rich representation of dose-response surfaces is essential for optimizing combination therapy design, enabling identification of safe and effective dosing strategies. Furthermore, this modeling framework supports better experimental planning and prioritization by highlighting dose ranges where synergistic benefits are most likely to occur, potentially reducing toxicity while maintaining or enhancing therapeutic outcomes.

6.2.2 Compared Methods

For a rigorous evaluation, we compared DeepSynBa against several state-of-the-art computational models for predicting dose-response outcomes in drug combination studies by modelling complex multi-dimensional interactions among drugs, dosages, and cell lines. In particular, we selected comboFM [51] and comboLTR [52] as representative baselines that, like DeepSynBa, predict the full dose-response matrix. comboFM uses higher-order tensor factorization techniques to model drug-dose-response interactions. It constructs a tensor encompassing drug features, dose levels, and responses, and factorizes this tensor using factorization machines to predict missing entries (responses). comboLTR takes a similar high-dimensional approach but replaces explicit factorization with a latent tensor reconstruction model using polynomial regression. These two methods leverage the same data types (drug chemical features, cell line info, dose) and output the same type of result (a full response surface), making them the most directly comparable to DeepSynBa. We also considered PIICM [cite], a Gaussian process model that predicts dose-response surfaces. However, PIICM does not utilize any

drug or cell line features and instead focuses on correlations among dose-response curves from experiments; as a result, it cannot generalize to new drugs or new cell lines that were not in the training data. Since our evaluation includes scenarios with novel drugs and cell lines, PIICM could not be meaningfully included. Therefore, we compare primarily against comboFM and comboLTR.

In addition to matrix prediction models, we compare DeepSynBa’s synergy score predictions against dedicated synergy scoring models. We include MatchMaker and DeepSynergy as benchmarks of methods that directly predict a single synergy value for a drug pair (see Chapter 4 for the details of these models). These models do not predict the full response matrix, only an aggregated metric of synergy. We emphasize that DeepSynBa is not trained on any particular synergy definition; nonetheless, generating dose-response surfaces can compute any synergy score post hoc, allowing a fair comparison to these algorithms on synergy prediction.

6.2.3 Performance Evaluation Setup

To assess the generalizability and robustness of our predictive models, we evaluate their performance under three distinct stratification scenarios: (i) new drug combinations, (ii) new cell lines, and (iii) new drugs (see Section 2.4 for details of each scenario). Each scenario is designed to mimic realistic challenges encountered in drug discovery and personalized medicine, providing a comprehensive evaluation framework.

For all three stratification strategies, the dataset is partitioned into training (60%), validation (20%), and test (20%) subsets. These proportions strike a balance between providing sufficient data for model optimization and ensuring robust evaluation.

These stratified evaluation protocols are essential for capturing the practical challenges associated with predictive modeling in drug discovery applications. By enforcing strict separations based on drug identity, drug pairing, and cellular

context, our evaluation design ensures a realistic and stringent test of model performance, ultimately advancing the development of computational tools for drug discovery and personalized therapeutic strategies.

6.2.4 Drug-Response Matrix Prediction Performance

To quantify performance, we use three complementary metrics computed over all entries of the dose-response matrices: Root Mean Squared Error (RMSE), Pearson correlation, and Spearman correlation. RMSE measures the average prediction error in percent growth, whereas Pearson and Spearman correlations measure the linear and rank-order agreement between the predicted and actual response values. These metrics are calculated over the pooled set of matrix entries from all test combinations in a given scenario. Table 6.2 summarises each model’s performance in terms of RMSE, Pearson, and Spearman correlations across these scenarios.

DeepSynBa achieves the best performance in all evaluation scenarios, outperforming both comboFM and comboLTR by a significant margin across all three metrics. Notably, in the new combination scenario (new drug pairs on known cell lines), DeepSynBa attained an RMSE of about 7.72 (Table 6.2a). This is substantially better than the next-best method, comboFM, which had an RMSE around 12.76 in the same setting. This difference corresponds to roughly a 0.65 reduction in prediction error. In practical terms, whereas comboFM’s predictions deviate from true viability by 12.76 points on average, DeepSynBa’s deviate by only 7.72 points, indicating a much closer fit to the experimental data. The improvement is similarly reflected in correlation metrics: DeepSynBa obtained higher Pearson and Spearman correlations than both baselines (for instance, the Pearson correlation for DeepSynBa is 0.95 in this scenario, compared to 0.87 for comboFM and 0.78 for comboLTR), demonstrating that our model not only reduces error magnitude but also captures the correct response trends across doses more faithfully. In the new cell line scenario, DeepSynBa’s RMSE (7.71) was again far better than comboFM (12.86), highlighting that the model generalizes well to unseen cellular

Table 6.2: The comparison of model performance across three evaluation scenarios: (a) new drug combination, (b) new cell line, and (c) new drug. Detailed descriptions of these scenarios are provided in Section 6.2.3.

(a) New drug combination scenario

	RMSE	Pearson	Spearman
comboFM	12.76	0.87	0.68
comboLTR	16.11	0.78	0.64
DeepSynBa	7.72	0.95	0.82

(b) New cell line scenario

	RMSE	Pearson	Spearman
comboFM	12.86	0.87	0.69
comboLTR	15.23	0.81	0.67
DeepSynBa	7.71	0.95	0.82

(c) New drug scenario

	RMSE	Pearson	Spearman
comboFM	42.39	0.35	0.24
comboLTR	40.22	0.40	0.40
DeepSynBa	18.69	0.71	0.64

contexts (see Table 6.2b). These results underscore DeepSynBa’s ability to learn the complex interactions between drug pairs and cellular features in a way that generalizes beyond the combinations it saw during training.

All models show some degradation in the most challenging new drug scenario, as expected when confronting entirely unseen chemical structures. However, even here, DeepSynBa maintained a clear advantage. DeepSynBa’s RMSE is 18.69 for combinations involving new drugs, whereas comboLTR and comboFM performed markedly worse with RMSE of 40.22 and 42.39, respectively (Table 6.2c. In fact, DeepSynBa roughly halved the prediction error compared to these baselines for new drugs. The Pearson and Spearman correlations likewise remained significantly higher for DeepSynBa in this scenario, indicating it still captured considerable signal despite the novelty of the compounds. This robustness may stem from the model’s parameter-based output: By learning fundamental patterns of dose-response behavior and tying them to learned drug representations, DeepSynBa can extrapolate how a new drug might behave in combination if its features are similar to known drugs. The baselines, in contrast, which rely on factorizing training data patterns, struggle when a drug with no prior interaction data is introduced. Overall, Table 6.2 and the above results provide strong evidence that DeepSynBa consistently outperforms the state-of-the-art methods in predicting full dose-response matrices, with especially large gains in scenarios requiring extrapolation to novel drugs or cell types.

6.2.4.1 Performance Across Cell Lines and Drugs

To provide a more detailed assessment of DeepSynBa’s predictive performance, we evaluated the model at the level of individual cell lines and drugs. This approach allows for a comprehensive analysis of the model’s generalization ability across diverse tumor types and drug agents, accounting for variations in both biological and pharmacokinetic properties. We calculated performance metrics for each cell line and drug, including root mean squared error (RMSE), Pearson correlation, and Spearman correlation. The results are visualized in Figure 6.3, which displays the distributions of these metrics using violin plots. To evaluate

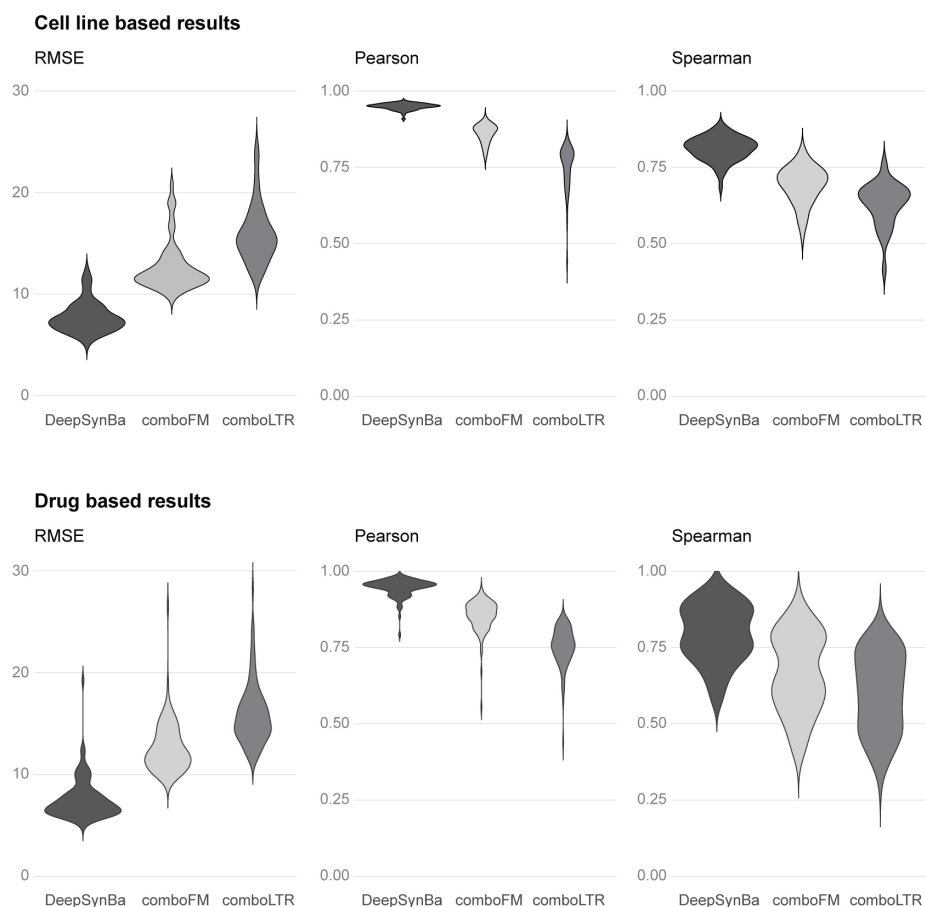


Figure 6.3: The distributions of RMSE, Pearson, and Spearman correlation metrics for dose-response surface predictions, shown across different cell lines (top panel) and drugs (bottom panel).

whether the differences in model performance between DeepSynBa and baseline methods were statistically significant, we conducted pairwise t -tests. All comparisons yield statistically significant results with p -values below 10^{-10} , indicating that DeepSynBa's improvements are unlikely to be due to chance (see Table 6.3).

As shown in the Figure 6.3, DeepSynBa's RMSE values are tightly concentrated at lower ranges, while its Pearson and Spearman correlations cluster toward higher values. These concentrated distributions demonstrate that DeepSynBa not only delivers higher accuracy but also exhibits greater consistency in performance. The narrower shape of the violin plots further suggests reduced

Table 6.3: The pairwise t-test p -value results for drugs and cell lines

(a) Cell line based results

	RMSE	Pearson	Spearman
DeepSynBa $\sim$ comboFM	2.91×10^{-51}	8.71×10^{-50}	6.58×10^{-9}
DeepSynBa $\sim$ comboLTR	1.57×10^{-43}	1.46×10^{-41}	1.14×10^{-38}
comboFM $\sim$ comboLTR	4×10^{-3}	6.58×10^{-9}	1.94×10^{-23}

(b) Drug based results

	RMSE	Pearson	Spearman
DeepSynBa $\sim$ comboFM	6.11×10^{-38}	4.17×10^{-33}	4.38×10^{-14}
DeepSynBa $\sim$ comboLTR	1.39×10^{-55}	4.63×10^{-58}	2.29×10^{-25}
comboFM $\sim$ comboLTR	2.26×10^{-15}	1.41×10^{-25}	4.52×10^{-4}

variance across cell lines and drugs, highlighting the model’s robustness across heterogeneous settings.

Overall, these findings confirm that DeepSynBa consistently outperforms alternative models’ accuracy and reliability, reinforcing its utility for large-scale drug combination prediction and translational applications.

6.2.4.2 Performance Across Tissues

We next examined model performance across the tissue of origin of the cell lines to see if certain cancer types were easier or harder for the models to predict. DeepSynBa continued to outperform the baselines in each of the nine tissue categories represented in NCI-ALMANAC (including breast, lung, colon, melanoma, ovarian, renal, leukaemia, etc.), as detailed in Table 6.4. Our model achieved the highest accuracy (lowest RMSE and highest correlations) in renal, breast, and ovarian cancers, while performance was relatively lower in leukemia cell lines. For instance, in renal cancer cell lines, DeepSynBa’s average RMSE was the smallest among all tissues, suggesting the model captured the response patterns in kidney-derived cells particularly well. We hypothesize that this may be due to well-defined molecular pathways in these cancers that the model can learn.

Table 6.4: Performance results across different tissue types under the new drug combination scenario.

	RMSE			Pearson			Spearman		
	DeepSynBa	comboFM	comboLTR	DeepSynBa	comboFM	comboLTR	DeepSynBa	comboFM	comboLTR
Breast	6.67	11.39	14.05	0.95	0.86	0.77	0.82	0.63	0.68
CNS	7.14	12.35	14.20	0.96	0.87	0.81	0.80	0.65	0.68
Colon	8.55	13.48	17.25	0.95	0.87	0.77	0.82	0.64	0.71
Leukemia	11.09	18.62	22.34	0.96	0.87	0.80	0.86	0.72	0.77
Melanoma	7.30	11.76	15.49	0.95	0.87	0.75	0.81	0.61	0.68
Non-Small Cell Lung	7.15	11.76	15.11	0.95	0.87	0.77	0.80	0.61	0.68
Ovarian	6.97	11.76	15.46	0.95	0.85	0.72	0.80	0.59	0.66
Prostate	7.09	11.73	15.09	0.96	0.88	0.78	0.82	0.65	0.68
Renal	6.93	11.16	14.87	0.95	0.87	0.75	0.82	0.64	0.70

Renal cancers often have characteristic alterations in signaling pathways like PI3K/AKT/mTOR, which are targets for many drugs [103, 104]. If our gene expression features capture these pathway activities, the model can more reliably predict how drugs affecting those pathways will work, leading to higher accuracy. Likewise, breast and ovarian cancers are frequently driven by recognizable genetic markers (e.g., BRCA1/2 mutations in many ovarian and breast tumors) [105, 106, 107]. Such markers produce clearer patterns of drug sensitivity – for example, BRCA-mutant cancers are very sensitive to DNA-damaging agents and PARP inhibitors, which likely made it easier for DeepSynBa to learn drug-response relationships. Indeed, the presence of strong molecular features can guide the model in associating specific drug mechanisms with strong effects in those tissues.

In contrast, the leukemia category proved most challenging, with DeepSynBa (and other models) showing the highest prediction error. Unlike solid tumors, leukemias are disseminated blood cancers that lack a tumor microenvironment and can have high genetic heterogeneity [108, 109, 110]. The diversity of leukemia cell lines (and possibly the different assay conditions for suspension cells) may make their responses harder to predict consistently. Additionally, many leukemia treatments involve targeted therapies (like tyrosine kinase inhibitors) with very cell-line-specific effects, which may not generalize as easily. Prior studies have noted that leukemias exhibit significant variability in drug sensitivity due to their heterogeneous genetic backgrounds. Our results align with those observations: Even though DeepSynBa performs best overall, its error in leukemia is

Table 6.5: Performance results across different drug combination categories under the new drug combination scenario.

	RMSE			Pearson			Spearman		
	DeepSynBa	comboFM	comboLTR	DeepSynBa	comboFM	comboLTR	DeepSynBa	comboFM	comboLTR
Chemotherapy-Chemotherapy	8.07	13.03	16.57	0.96	0.88	0.80	0.84	0.75	0.70
Chemotherapy-Targeted	8.33	12.39	16.16	0.94	0.86	0.73	0.80	0.67	0.59
Chemotherapy-Other	7.29	13.48	16.23	0.96	0.86	0.79	0.81	0.67	0.63
Targeted-Targeted	6.32	11.77	15.55	0.94	0.79	0.54	0.78	0.58	0.44
Targeted-Other	5.69	10.45	12.95	0.94	0.81	0.68	0.75	0.53	0.43
Other-Other	5.77	11.29	14.69	0.97	0.88	0.79	0.71	0.50	0.47

higher relative to other tissues, suggesting room for improvement, perhaps by incorporating more specific genomic features (mutations or cytogenetics) for blood cancers. Nevertheless, even in leukemia, DeepSynBa outperformed comboFM and comboLTR, indicating that the advantages of our approach carry over to this challenging setting. The overall trend of these tissue-specific results supports that DeepSynBa’s predictive power is robust across diverse cancer types, particularly strong in cancers with well-characterized drug response patterns. The insights also emphasize the importance of incorporating rich cell line features: By leveraging gene expression, our model can exploit known biology (like pathway addictions in specific tissues) to make better predictions.

6.2.4.3 Performance Across Drug Categories

We also analyzed performance by categorizing drug pairs based on the mechanistic class of each drug, as provided by NCI-ALMANAC’s annotations. Each drug is labeled as "chemotherapy", "targeted", or "other". This yields six possible pair types: chemo-chemo, chemo-targeted, chemo-other, targeted-targeted, targeted-other, and other-other. We computed prediction accuracy separately for each category of drug pairs (Table 6.5). Across all these categories, DeepSynBa again showed superior performance compared to the competing methods. Interestingly, the model was most accurate on combinations where at least one drug was in the "other" category, such as Targeted-Other and Other-Other pairs. These pair types had the lowest prediction errors (lowest RMSE and highest correlations). In contrast, combinations like Targeted-Targeted or Chemo-Targeted were relatively more difficult (though DeepSynBa still led in those cases as well).

The performance on "Other"-involving combinations may reflect the underlying distribution of response effects in those categories. We observed that drug pairs involving an "other" category drug often had more homogeneous and modest response profiles, typically inducing intermediate levels of effect. For example, many "other" drugs include metabolic modulators or hormonal agents that, by themselves, do not cause drastic cell death across the short timeframe of the screen. When combined with another drug, the overall effect tends to be moderate. Therefore, the variance in response for these combos is minor – they usually don't produce extreme synergy or antagonism, but relatively incremental changes. Indeed, in our data, the Targeted–Other and Other–Other pairs showed a narrower range of viabilities (see Figure 6.4). This narrow dynamic range makes the prediction task easier because the model doesn't need to account for very high or low outliers; a simpler relationship can fit most of the data. DeepSynBa appears to capitalize on this, achieving high accuracy for those pair types. In contrast, combinations like Targeted–Targeted sometimes yielded very high synergy or very context-specific effects (for instance, two targeted agents might have a dramatic effect only if the cell line has a particular mutation matching both drug targets).

6.2.5 Synergy Score Prediction Performance

While DeepSynBa is designed to predict full dose-response matrices, it is essential to verify that this comprehensive approach does not come at the expense of accurately predicting standard synergy scores. Therefore, as a final evaluation, we computed two widely used synergy metrics – Loewe additivity scores [30] and Bliss independence scores [29] for both the ground truth and the model-predicted dose-response matrices, and compared these to the predictions of dedicated synergy models. Based on the full response surface, a Loewe or Bliss synergy score summarizes whether a drug pair is synergistic or antagonistic. We calculated these scores for every drug pair in the test set using the true experimental data, dose-response models' predicted matrices, and obtaining the scores directly predicted by MatchMaker and DeepSynergy for those pairs. Table 6.6 presents the results

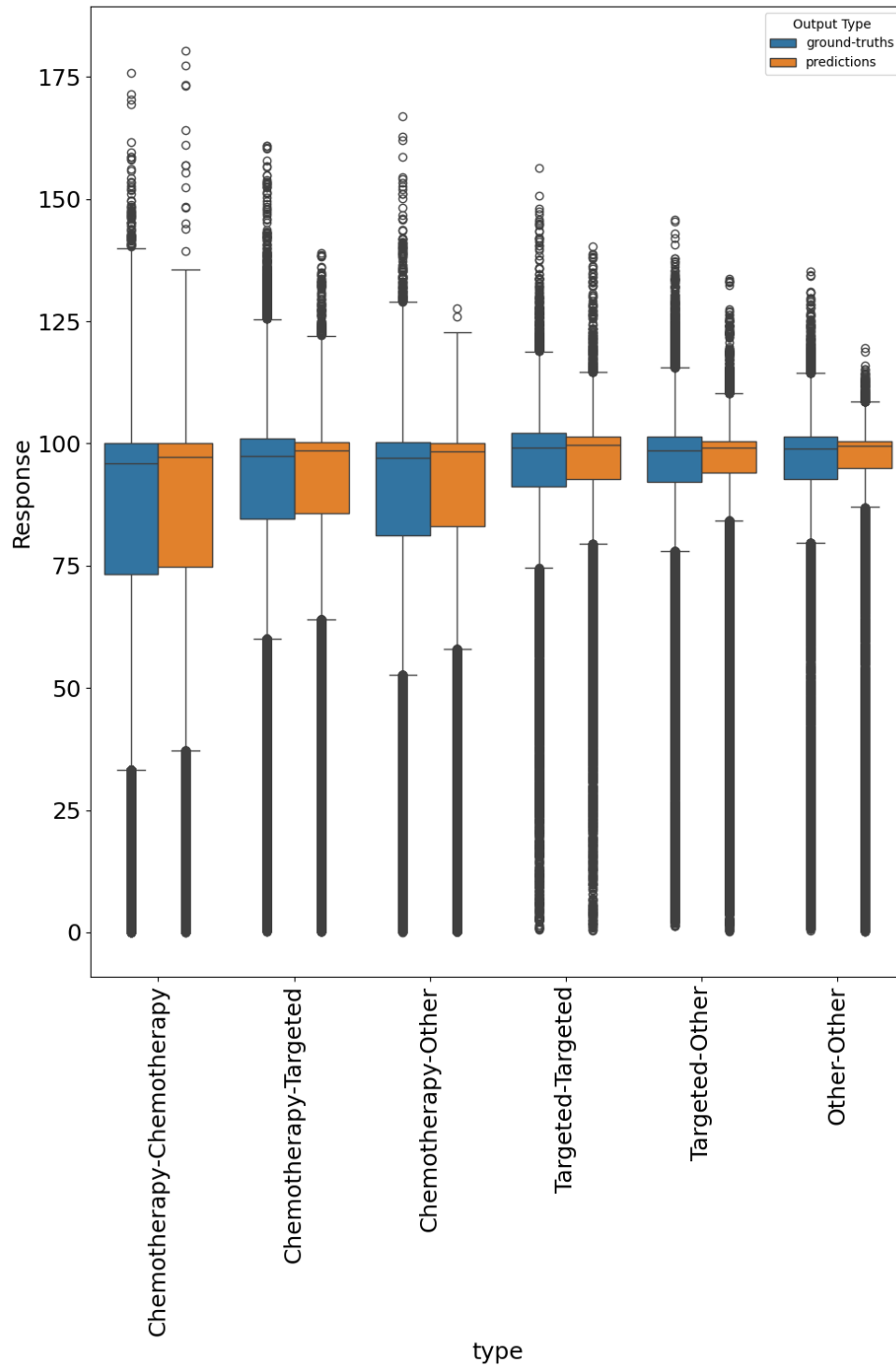


Figure 6.4: The ground-truth and the predicted response distributions for different drug type combinations.

Table 6.6: Performance comparison using Loewe and Bliss synergy scores. Ground-truth Loewe/Bliss values are computed from experimentally measured dose-response matrices. For DeepSynBa, ComboFM, and ComboLTR, predicted synergy scores are derived from their respective predicted dose-response matrices. In contrast, MatchMaker and DeepSynergy are trained directly on Loewe/Bliss scores as regression targets.

	Loewe			Bliss		
	RMSE	Pearson	Spearman	RMSE	Pearson	Spearman
DeepSynBa	6.37	0.93	0.93	7.06	0.52	0.35
comboFM	9.70	0.83	0.83	9.32	0.24	0.15
comboLTR	11.46	0.75	0.76	10.97	-0.01	-0.01
MatchMaker	5.69	0.94	0.95	5.22	0.78	0.73
DeepSynergy	11.10	0.85	0.85	6.77	0.60	0.53

in terms of prediction error and correlation for synergy scores. DeepSynBa consistently outperformed the other dose-response matrix prediction methods (comboFM, comboLTR) in predicting both Loewe and Bliss synergy values. This is expected given DeepSynBa’s higher accuracy on the underlying matrices. Still, it confirms that our model’s improvements carry through to the level of summary indices: It better captures the combinations’ overall synergy/antagonism.

Remarkably, although DeepSynBa is not explicitly trained to predict any synergy score, its performance is close to that of synergy score predictor models. In particular, DeepSynBa’s synergy predictions were comparable to those of MatchMaker and DeepSynergy. DeepSynBa achieved nearly the same accuracy for Loewe synergy scores as MatchMaker (which directly learns Loewe scores). In fact, DeepSynBa slightly outperformed DeepSynergy in several metrics for Loewe predictions – our model’s comprehensive learning may be capturing the essential factors that determine Loewe additivity. Bliss score predictions by DeepSynBa were also very close to those of the dedicated models. The key point is that DeepSynBa can recover traditional synergy assessments reliably: For a given pair, one can compute its Bliss or Loewe score from the predicted matrix and trust that it will be as accurate as if one had used a specialized synergy predictor. This is a significant advantage — DeepSynBa offers far more information (the full

matrix) without sacrificing the summary metrics that researchers and clinicians commonly use to prioritize combinations.

Overall, the synergy score evaluation provides further evidence of DeepSynBa’s strengths. It excels at the primary task of dose-response matrix prediction and indirectly excels at the secondary task of synergy scoring. By contrast, models like MatchMaker and DeepSynergy cannot provide the nuanced dose-dependent insights that DeepSynBa does, since they collapse everything into one number. DeepSynBa bridges that gap: It gives a rich output that can be reduced to conventional metrics while maintaining top-tier predictive performance. These results highlight that DeepSynBa improves upon prior models in both scope and accuracy – it goes beyond the limitations of single-score predictors and non-interpretable matrix factorization methods, delivering more comprehensive and actionable predictions for drug combinations.

Chapter 7

Conclusion

7.1 MatchMaker: Advancing Accuracy and Generalization in Synergy Prediction

MatchMaker is introduced as a deep learning framework to predict whether two drugs will act synergistically, leveraging drug chemical information and cell line-specific gene expression profiles. By training on the largest available drug combination dataset at the time (DrugComb), MatchMaker achieves the state-of-the-art predictive performance, outperforming previous models like DeepSynergy by up to 15% in correlation and 33% in MSE on synergy prediction benchmarks. These gains in predictive performance translate to more reliable identification of synergistic drug pairs, enabling the model to prune the otherwise unmanageable search space of possible drug combinations. In practice, this means MatchMaker can serve as an effective *in silico* prescreening tool, prioritizing promising drug pairs for experimental validation and thereby accelerating the discovery of effective combinations. The model’s design also emphasizes generalization by conditioning drug effects on cell line gene expression; it captured context-specific responses, allowing it to generalize across diverse cancer types and molecular contexts. This

cell line-conditioned representation approach helped ensure that predictions remained robust even when extrapolating to new cell lines or unseen drug pairs, addressing a key challenge in drug synergy prediction where performance often degrades outside the training distribution. In summary, MatchMaker’s primary contributions lie in its high predictive accuracy and robust generalization for drug synergy prediction in cancer cell lines, setting a strong foundation for subsequent advancements in the field.

However, despite its strengths, MatchMaker is limited by a notable shortcoming: A lack of patient personalization. The model relies on cancer cell line data as a proxy for human tumors, which does not capture individual patient heterogeneity. Genetic and molecular differences between patients mean that a combination highly synergistic in one context may underperform in another. Moreover, MatchMaker predicted synergy using a summary score (e.g., Loewe, Bliss, or similar metrics) at fixed conditions, without accounting for varying dose effects or individual patient biomarkers. While its focus on widely available features (drug structure and gene expression) enhanced general applicability, the absence of a mechanism to tailor predictions to a specific patient’s tumor reduced its direct clinical relevance. This recognition of personalization as the next frontier motivated the development of the Personalized Deep Synergy Predictor (PDSP) framework.

7.2 PDSP: Personalizing Drug Synergy Prediction via Transfer Learning

PDSP is developed to address MatchMaker’s personalization gap by introducing a patient-specific calibration mechanism built on transfer learning. The core idea of PDSP is to leverage the knowledge learned from large-scale cell line data and adapt it to individual patients, thus combining the best of both worlds: The statistical power of big data with the relevance of patient-specific information. To achieve this, PDSP first trains a base model on extensive cell line synergy

datasets (essentially an enhanced version of MatchMaker) that simultaneously learns to predict drug pair synergy and single-drug response for each cell line. This multi-task learning (predicting both combination synergy scores and individual drug sensitivities) imbues the model with a rich understanding of drug effects. In the second stage, PDSP performs fine-tuning to calibrate the model for a specific patient by using that patient’s gene expression profile and measured single-drug response data from ex vivo samples. In essence, the model "personalizes" the drug representation and prediction layers using the patient’s data, while retaining the general patterns learned from cell lines. This transfer learning approach proved highly effective when evaluated on data from leukemia patients. PDSP improved synergy prediction accuracy by roughly 27% compared to a non-personalized model trained only on cell line data. Such a substantial gain underscores the value of incorporating patient-specific calibration, by fine-tuning on limited but relevant patient data, PDSP significantly boosts prediction performance in clinical contexts.

Beyond improved accuracy, PDSP’s patient-specific modeling brings considerable clinical relevance. By customizing predictions to reflect an individual patient’s tumor biology, PDSP moves drug combination forecasts closer to practical decision support in precision oncology. For example, a clinician could use PDSP to predict which of several candidate drug pairs is most likely synergistic for a particular patient, based on that patient’s molecular profile and drug sensitivity measurements. This personalized prediction capability is especially crucial given the heterogeneity of cancer: Two patients with the same diagnosis might respond very differently to the same drug pair. PDSP addresses this by effectively creating a unique model for each patient, thus acknowledging inter-patient variability. Moreover, the framework demonstrates the power of deep transfer learning in a biomedical setting – knowledge learned from in vitro cell line experiments is repurposed to inform in silico predictions for real patients. This represents a meaningful step towards integrating AI-driven synergy prediction with personalized medicine. Rather than relying solely on generic models, PDSP shows that we can bridge preclinical data to patient-specific contexts, enhancing the translational potential of synergy modeling. However, it is worth noting that

PDSP’s approach requires that some patient data (gene expression and single-drug responses) be available, which may involve an extra experimental step (like high-throughput drug sensitivity screening on a patient’s biopsy). Despite this, given the relatively easier accession of single-drug response data compared to combination data, PDSP offers a pragmatic path forward. It successfully addresses the key limitation of MatchMaker by introducing personalization, all while maintaining strong performance through its two-stage learning strategy.

7.3 DeepSynBa: Predicting Dose–Response Surfaces and Disentangling Synergy Mechanisms

While MatchMaker and PDSP focused on predicting a summary synergy score (e.g., a Loewe or Bliss synergy value) for a drug pair, DeepSynBa expands the synergy prediction scope by modeling a drug combination’s entire dose–response matrix. Instead of outputting a single synergy metric, DeepSynBa predicts how a pair of drugs will jointly affect cell viability across a range of doses for each drug. This approach addresses a critical limitation of using a one-number summary for synergy: Synergy scores can be misleading if the tested dose range misses key effect levels. By learning the full dose–response surface, DeepSynBa provides a much richer characterization of drug interactions, capturing both the magnitude and nuances of combination effects across doses. This enables researchers to calculate any desired synergy metric post hoc (such as Loewe, Bliss, or others) from the predicted dose-response matrix. More importantly, it allows identification of optimal dosing regions; for instance, DeepSynBa can indicate at which concentrations two drugs exhibit the strongest synergy or if there’s a dose region where antagonism occurs, information that a single summary score might obscure.

A defining feature of DeepSynBa is its ability to disentangle synergistic potency and efficacy in drug combinations. Drawing inspiration from a two-dimensional generalised Hill equation framework, DeepSynBa’s architecture explicitly learns separate parameters for combination potency and efficacy. Synergistic potency

refers to how the presence of one drug affects the concentration–response curve of the other (e.g., shifting the EC50). In contrast, synergistic efficacy refers to how the maximal achievable effect is altered by the drug combination. By modeling these aspects distinctly, DeepSynBa provides interpretability that previous dose-matrix prediction models lacked. For example, suppose DeepSynBa predicts a particular drug pair that has high potency synergy but no efficacy gain. In that case, this suggests one drug makes the other work at lower doses, but the maximal effect remains unchanged – a nuanced insight that can inform mechanism-of-action studies. Technically, this is achieved via an intermediate layer in the network that outputs the parameters of the parameterized dose–response surface (such as Hill coefficients, maximal effect, potency ratios, etc.), which are then used to reconstruct the full response curve. This design means that DeepSynBa not only forecasts outcomes but also provides a set of interpretable coefficients for each drug pair, bridging the gap between black-box predictions and mechanistic understanding.

DeepSynBa builds upon the architectural strengths of MatchMaker while substantially extending its capabilities. It reuses the cell line–conditioned drug encoder modules from MatchMaker to incorporate cell context, ensuring that the dose–response predictions remain sensitive to the biological background (e.g., cancer-specific gene expression of the cell line or potentially patient). In doing so, it retains the generalization ability to different biological contexts. However, unlike its predecessors, DeepSynBa directly tackles the dose dimension of synergy. This advancement has several important implications: First, by predicting complete dose–response surfaces, DeepSynBa achieved better accuracy in fitting combination experimental data than models that predict only a single synergy score. Its comprehensive training on full matrices allows it to capture subtle combination effects that summary-score models might miss. Second, the model’s outputs (the learned parameters of the dose–response surface) make it possible to compute traditional synergy scores or any custom metric on demand, which means DeepSynBa remains compatible with existing synergy evaluation frameworks. Third, and perhaps most significantly, DeepSynBa’s interpretability and detail can guide experimental and clinical decision-making. Researchers can use

DeepSynBa to simulate combination responses at untested doses, identify promising dose regions to explore in follow-up experiments, and effectively suggest how to refine dose selection in synergy trials. In a translational sense, if applied to patient-derived cells (with appropriate fine-tuning as PDSP does), DeepSynBa could inform oncologists not just which drug pairs are synergistic for a patient, but at what dose levels the synergy is maximized, thus aiding in treatment optimization. By disentangling efficacy and potency effects, clinicians could discern whether a drug pair is likely to increase the overall cell kill (efficacy) or achieve the same effect at lower doses (potency) – valuable information for therapy planning.

In summary, DeepSynBa represents a significant expansion of deep learning-based synergy models by incorporating dose-response surface prediction and mechanistic interpretability. It complements the earlier frameworks, where MatchMaker provided accuracy, PDSP provided personalization, and DeepSynBa provided granularity and insight. Together, these models form a progressive sequence of advancements – from accurate prediction, to personalized adaptation, to mechanistic and dose-specific understanding – each addressing different facets of the drug synergy prediction challenge.

7.4 Integration with Personalized Medicine and Future Directions

The convergence of these three frameworks – MatchMaker, PDSP, and DeepSynBa – highlights a path forward for integrating drug synergy prediction into personalized medicine. Personalized oncology benefits immensely from AI models that can predict how a given patient might respond to combination therapies. In this thesis, PDSP already demonstrates the feasibility of tailoring synergy predictions to individual patients by incorporating their genomic and pharmacological profiles. A logical next step is to combine the strengths of PDSP and DeepSynBa: For instance, a model that predicts patient-specific dose-response surfaces for drug combinations. Such a framework would require fine-tuning DeepSynBa’s

dose-response predictions with patient-specific data (analogous to PDSP’s calibration) to yield truly personalized dose–response matrices. The outcome would be a tool that suggests which drug pairs a particular patient is most likely to benefit from and the optimal dosing strategies for those drugs. This integration would directly support precision medicine by informing combination treatment design on a patient-by-patient basis, potentially improving therapeutic efficacy while minimizing toxicity.

Looking ahead, several future research directions emerge from the insights of this work:

- **Integrating Multi-omics Data:** All three models thus far primarily utilize gene expression (transcriptomic) data as the cellular context. Future models could incorporate additional omics layers – such as genomic mutations, proteomic profiles, or epigenetic markers – to capture a more holistic view of tumor biology. For example, integrating mutational data might help the model account for pathway activations or deactivations that strongly influence drug synergy. Including such data could further personalize and refine predictions and potentially improve generalization to diverse patient subpopulations. However, we should note that not all of these multi-omics features are available for all cell lines. Therefore, dataset sizes get smaller when we try to use all such feature sets, which may lower the models’ performance, especially deep learning architectures. This is a trade-off that we need to consider.
- **Improving Model Interpretability:** While DeepSynBa takes a step toward interpretability by outputting pharmacologically meaningful parameters, a broader need remains to make deep learning models transparent and trustworthy for clinical use. Research could focus on explainable AI techniques applied to drug synergy models, such as attention mechanisms or feature attribution methods, to identify which genes or molecular features drive a predicted synergy. Another approach is to incorporate known biological pathways or network information into the model architecture (a hybrid mechanistic-statistical model) so that predictions can be traced to

pathway interactions. Enhanced interpretability will be crucial for clinical acceptance of model recommendations and for generating scientific insights (e.g., revealing novel mechanisms of synergy).

- **Expanding to Additional Therapeutic Contexts:** Thus far, the frameworks have been developed and validated mostly in cancer cell lines and ex vivo contexts. Future work could extend these models to other diseases where combination therapy is relevant (such as infectious diseases or neurological disorders) or in vivo preclinical models. Adapting the models to predict synergy in patient-derived organoids or mouse models could be a stepping stone toward clinical translation, helping to bridge the gap between lab predictions and actual patient outcomes.
- **Exploring Combination Complexity:** Another frontier is moving beyond pairwise combinations to model higher-order combinations (triplets or regimens involving more than two drugs). While the complexity grows combinatorially, advances in model efficiency and training data (e.g., high-throughput screens of triple-drug combos) could allow future frameworks to tackle this. Additionally, the dose-response surface approach of DeepSynBa could, in principle, be extended to higher dimensions for triple-drug dose cubes, though interpretability would likewise become more complex.
- **Clinical Integration and Validation:** Importantly, prospective clinical studies will be needed to truly measure these models' impact. A future direction is to design clinical trials where model-predicted synergistic combinations are tested in patients, evaluating whether AI-guided combination selection improves outcomes or reduces trial-and-error in finding effective therapies. Such studies would provide valuable feedback to further refine the models (e.g., by highlighting scenarios of false positives/negatives) and would mark a significant step toward routine clinical deployment.

Finally, reflecting on the broader significance of this work, it is evident that advanced computational models for drug synergy have a transformative potential in drug discovery and translational oncology. By intelligently narrowing the vast space of drug pairs, models like MatchMaker accelerate the identification of

promising combinations, which can save considerable time and resources in early-stage drug development. When these models incorporate patient-specific factors as PDSP does, they pave the way for *personalized combination therapy*, moving us closer to treatments tailored to the unique genetic and molecular makeup of each patient’s cancer. This is particularly crucial in oncology, where the heterogeneity of tumors often means that the efficacy of drug combinations can vary widely from patient to patient. Furthermore, the interpretability and detailed predictions offered by approaches like DeepSynBa can facilitate a deeper understanding of how drugs interact biologically. This not only benefits clinicians in making informed decisions but also aids researchers in uncovering new biological insights (for example, understanding why a particular drug pair is synergistic could reveal a previously unappreciated interaction between signaling pathways). In a broader drug discovery context, synergy models can help repurpose existing drugs by identifying novel synergistic partners, thereby opening up new therapeutic options without the need to develop entirely new compounds.

A recent study [111] raises critical concerns about a limitation in many existing drug synergy prediction models. It turns out that several models, including MatchMaker, perform nearly the same, or sometimes even better, when drug and cell line features are replaced with simple one-hot encodings. This finding suggests that these models may not be learning meaningful biological or chemical patterns, but instead rely on shortcuts hidden in the data, such as repeated combinations or dataset-specific artifacts. While such models can still be useful for ranking known combinations within a familiar dataset, their ability to generalize to new drugs or new biological contexts remains questionable. These results highlight the need for more robust model evaluation strategies, improved data representations, and inductive biases that enforce biological relevance.

In conclusion, the trio of frameworks presented – MatchMaker, PDSP, and DeepSynBa – collectively advances the state-of-the-art in computational drug synergy prediction. They address challenges spanning from accuracy and scalability to personalization and interpretability. Together, these contributions demonstrate a powerful paradigm: Using deep learning to bridge experimental data and clinical insight, we can more effectively chart the landscape of drug combinations

for cancer therapy. This work provides concrete tools and methods for synergy prediction and lays out a vision for integrating such models into the pipeline of precision oncology and drug development. As research continues to build on these foundations, we move closer to a future where combination treatments for cancer (and other complex diseases) are discovered and optimized with the assistance of AI, leading to more effective and individualized patient care.



Bibliography

- [1] H. Zhang, C. H. Ek, M. Rattray, and M. Milo, “Synba: improved estimation of drug combination synergies with uncertainty quantification,” *Bioinformatics*, vol. 39, no. Supplement_1, pp. i121–i130, 2023.
- [2] V. T. DeVita and P. S. Schein, “The use of drugs in combination for the treatment of cancer: rationale and results,” *New England Journal of Medicine*, vol. 288, no. 19, pp. 998–1006, 1973.
- [3] B. Al-Lazikani, U. Banerji, and P. Workman, “Combinatorial drug therapy for cancer in the post-genomic era,” *Nature biotechnology*, vol. 30, no. 7, p. 679, 2012.
- [4] P. Csermely, T. Korcsmáros, H. J. Kiss, G. London, and R. Nussinov, “Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review,” *Pharmacology & therapeutics*, vol. 138, no. 3, pp. 333–408, 2013.
- [5] Y. Sun, Z. Sheng, C. Ma, K. Tang, R. Zhu, Z. Wu, R. Shen, J. Feng, D. Wu, D. Huang, *et al.*, “Combining genomic and network characteristics for extended capability in predicting synergistic drugs for cancer,” *Nature communications*, vol. 6, no. 1, pp. 1–10, 2015.
- [6] E. D. Clercq, “The design of drugs for hiv and hcv,” *Nature reviews Drug discovery*, vol. 6, no. 12, pp. 1001–1018, 2007.
- [7] P. D. Home, S. J. Pocock, H. Beck-Nielsen, P. S. Curtis, R. Gomis, M. Hanefeld, N. P. Jones, M. Komajda, J. J. McMurray, R. S. Team, *et al.*, “Rosiglitazone evaluated for cardiovascular outcomes in oral agent combination

- therapy for type 2 diabetes (record): a multicentre, randomised, open-label trial,” *The Lancet*, vol. 373, no. 9681, pp. 2125–2135, 2009.
- [8] R. Wang, J. F.-W. Chan, S. Wang, H. Li, J. Zhao, T. K.-Y. Ip, Z. Zuo, K.-Y. Yuen, S. Yuan, and H. Sun, “Orally administered bismuth drug together with n-acetyl cysteine as a broad-spectrum anti-coronavirus cocktail therapy,” *Chemical science*, vol. 13, no. 8, pp. 2238–2248, 2022.
 - [9] J. Jia, F. Zhu, X. Ma, Z. W. Cao, Y. X. Li, and Y. Z. Chen, “Mechanisms of drug combinations: interaction and network perspectives,” *Nature reviews Drug discovery*, vol. 8, no. 2, p. 111, 2009.
 - [10] D. Sun, W. Gao, H. Hu, and S. Zhou, “Why 90% of clinical drug development fails and how to improve it?,” *Acta Pharmaceutica Sinica B*, vol. 12, no. 7, pp. 3049–3062, 2022.
 - [11] B. M. Altevogt, M. Davis, D. E. Pankevich, and S. M. P. Norris, “Improving and accelerating therapeutic development for nervous system disorders: workshop summary,” 2014.
 - [12] L. He, E. Kuleskiy, J. Saarela, L. Turunen, K. Wennerberg, T. Aittokallio, and J. Tang, “Methods for high-throughput drug combination screening and synergy scoring,” in *Cancer systems biology*, pp. 351–398, Springer, 2018.
 - [13] R. Macarron, M. N. Banks, D. Bojanic, D. J. Burns, D. A. Cirovic, T. Garyantes, D. V. Green, R. P. Hertzberg, W. P. Janzen, J. W. Paslay, *et al.*, “Impact of high-throughput screening in biomedical research,” *Nature reviews Drug discovery*, vol. 10, no. 3, pp. 188–195, 2011.
 - [14] D. Day and L. L. Siu, “Approaches to modernize the combination drug development paradigm,” *Genome medicine*, vol. 8, no. 1, p. 115, 2016.
 - [15] X.-M. Zhao, M. Iskar, G. Zeller, M. Kuhn, V. Van Noort, and P. Bork, “Prediction of drug combinations by integrating molecular and pharmacological data,” *PLoS computational biology*, vol. 7, no. 12, 2011.
 - [16] F. Cheng, I. A. Kovács, and A.-L. Barabási, “Network-based prediction of drug combinations,” *Nature communications*, vol. 10, no. 1, pp. 1–11, 2019.

- [17] J. Yang, H. Tang, Y. Li, R. Zhong, T. Wang, S. Wong, G. Xiao, and Y. Xie, “Digre: Drug-induced genomic residual effect model for successful prediction of multidrug effects,” *CPT: pharmacometrics & systems pharmacology*, vol. 4, no. 2, pp. 91–97, 2015.
- [18] M. Bansal, J. Yang, C. Karan, M. P. Menden, J. C. Costello, H. Tang, G. Xiao, Y. Li, J. Allen, R. Zhong, *et al.*, “A community computational challenge to predict the activity of pairs of compounds,” *Nature biotechnology*, vol. 32, no. 12, p. 1213, 2014.
- [19] K. Pang, Y.-W. Wan, W. T. Choi, L. A. Donehower, J. Sun, D. Pant, and Z. Liu, “Combinatorial therapy discovery using mixed integer linear programming,” *Bioinformatics*, vol. 30, no. 10, pp. 1456–1463, 2014.
- [20] T. Yu, “Astrazeneca-sanger drug combination prediction dream challenge,” 2015.
- [21] S. L. Holbeck, R. Camalier, J. A. Crowell, J. P. Govindharajulu, M. Hollingshead, L. W. Anderson, E. Polley, L. Rubinstein, A. Srivastava, D. Wilsker, *et al.*, “The national cancer institute almanac: a comprehensive screening resource for the detection of anticancer drug pairs with enhanced therapeutic activity,” *Cancer research*, vol. 77, no. 13, pp. 3564–3576, 2017.
- [22] K. Preuer, R. P. Lewis, S. Hochreiter, A. Bender, K. C. Bulusu, and G. Klambauer, “Deepsynergy: predicting anti-cancer drug synergy with deep learning,” *Bioinformatics*, vol. 34, no. 9, pp. 1538–1546, 2017.
- [23] J. D. Janizek, S. Celik, and S.-I. Lee, “Explainable machine learning prediction of synergistic drug combinations for precision cancer medicine,” *bioRxiv*, p. 331769, 2018.
- [24] H. I. Kuru, O. Tastan, and A. E. Cicek, “Matchmaker: a deep learning framework for drug synergy prediction,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 4, pp. 2334–2344, 2021.

- [25] B. Zagidullin, J. Aldahdooh, S. Zheng, W. Wang, Y. Wang, J. Saad, A. Malyutina, M. Jafari, Z. Tanoli, A. Pessia, *et al.*, “Drugcomb: an integrative cancer drug combination data portal,” *Nucleic acids research*, vol. 47, no. W1, pp. W43–W51, 2019.
- [26] H. I. Kuru, A. E. Cicek, and O. Tastan, “From cell lines to cancer patients: personalized drug synergy prediction,” *Bioinformatics*, vol. 40, no. 5, p. btae134, 2024.
- [27] J. O’Neil, Y. Benita, I. Feldman, M. Chenard, B. Roberts, Y. Liu, J. Li, A. Kral, S. Lejnine, A. Loboda, *et al.*, “An unbiased oncology compound screen to identify novel combination strategies,” *Molecular cancer therapeutics*, vol. 15, no. 6, pp. 1155–1162, 2016.
- [28] M. P. Menden, D. Wang, M. J. Mason, B. Szalai, K. C. Bulusu, Y. Guan, T. Yu, J. Kang, M. Jeon, R. Wolfinger, *et al.*, “Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen,” *Nature communications*, vol. 10, no. 1, pp. 1–17, 2019.
- [29] C. Bliss, “The calculation of microbial assays,” *Bacteriological reviews*, vol. 20, no. 4, pp. 243–258, 1956.
- [30] S. Loewe, “The problem of synergism and antagonism of combined drugs,” *Arzneimittelforschung*, vol. 3, pp. 285–290, 1953.
- [31] R. J. Tallarida, “Drug combinations: tests and analysis with isoboles,” *Current protocols in pharmacology*, vol. 72, no. 1, pp. 9–19, 2016.
- [32] D. Weininger, “Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules,” *Journal of chemical information and computer sciences*, vol. 28, no. 1, pp. 31–36, 1988.
- [33] H. L. Morgan, “The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service,” *Journal of chemical documentation*, vol. 5, no. 2, pp. 107–113, 1965.

- [34] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, *et al.*, “Pubchem substance and compound databases,” *Nucleic acids research*, vol. 44, no. D1, pp. D1202–D1213, 2016.
- [35] D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. De Veij, E. Félix, M. P. Magariños, J. F. Mosquera, P. Mutowo, M. Nowotka, *et al.*, “ChEMBL: towards direct deposition of bioassay data,” *Nucleic acids research*, vol. 47, no. D1, pp. D930–D940, 2019.
- [36] D.-S. Cao, Q.-S. Xu, Q.-N. Hu, and Y.-Z. Liang, “ChemoPy: freely available python package for computational biology and chemoinformatics,” *Bioinformatics*, vol. 29, pp. 1092–1094, 03 2013.
- [37] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin, *et al.*, “The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity,” *Nature*, vol. 483, no. 7391, pp. 603–607, 2012.
- [38] W. Yang, J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. A. Smith, I. R. Thompson, *et al.*, “Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells,” *Nucleic acids research*, vol. 41, no. D1, pp. D955–D961, 2012.
- [39] L. Chen, B.-Q. Li, M.-Y. Zheng, J. Zhang, K.-Y. Feng, and Y.-D. Cai, “Prediction of effective drug combinations by chemical interaction, protein interaction and target enrichment of kegg pathways,” *BioMed research international*, vol. 2013, 2013.
- [40] G. R. Zimmermann, J. Lehar, and C. T. Keith, “Multi-target therapeutics: when the whole is greater than the sum of the parts,” *Drug discovery today*, vol. 12, no. 1-2, pp. 34–42, 2007.
- [41] S. Li, B. Zhang, and N. Zhang, “Network target for screening synergistic drug combinations with application to traditional chinese medicine,” *BMC systems biology*, vol. 5, no. S1, p. S10, 2011.

- [42] L. Huang, F. Li, J. Sheng, X. Xia, J. Ma, M. Zhan, and S. T. Wong, “Drugcomboranker: drug combination discovery based on target network analysis,” *Bioinformatics*, vol. 30, no. 12, pp. i228–i236, 2014.
- [43] J. Zhao, X.-S. Zhang, and S. Zhang, “Predicting cooperative drug effects through the quantitative cellular profiling of response to individual drugs,” *CPT: pharmacometrics & systems pharmacology*, vol. 3, no. 2, pp. 1–7, 2014.
- [44] G. Jin, H. Zhao, X. Zhou, and S. T. Wong, “An enhanced petri-net model to predict synergistic effects of pairwise drug combinations from gene microarray data,” *Bioinformatics*, vol. 27, no. 13, pp. i310–i316, 2011.
- [45] P. Sidorov, S. Naulaerts, J. Arieu-Bonnet, E. Pasquier, and P. J. Ballester, “Predicting synergism of cancer drug combinations using nci-almanac data,” *Frontiers in chemistry*, vol. 7, p. 509, 2019.
- [46] F. Xia, M. Shukla, T. Brettin, C. Garcia-Cardona, J. Cohn, J. E. Allen, S. Maslov, S. L. Holbeck, J. H. Doroshov, Y. A. Evrard, *et al.*, “Predicting tumor cell line response to drug pairs with deep learning,” *BMC bioinformatics*, vol. 19, no. 18, pp. 71–79, 2018.
- [47] T. Zhang, L. Zhang, P. R. Payne, and F. Li, “Synergistic drug combination prediction by integrating multiomics data in deep learning models,” *Translational bioinformatics for therapeutic development*, pp. 223–238, 2021.
- [48] Q. Liu and L. Xie, “Transynergy: mechanism-driven interpretable deep neural network for the synergistic prediction and pathway deconvolution of drug combinations,” *PLoS computational biology*, vol. 17, no. 2, p. e1008653, 2021.
- [49] J. Wang, X. Liu, S. Shen, L. Deng, and H. Liu, “Deepdds: deep graph neural network with attention mechanism to predict synergistic drug combinations,” *Briefings in Bioinformatics*, vol. 23, no. 1, p. bbab390, 2022.
- [50] P. Zhang and S. Tu, “Mgae-dc: Predicting the synergistic effects of drug combinations through multi-channel graph autoencoders,” *PLoS computational biology*, vol. 19, no. 3, p. e1010951, 2023.

- [51] H. Julkunen *et al.*, “Leveraging multi-way interactions for systematic prediction of pre-clinical drug combination effects,” *Nature communications*, vol. 11, no. 1, pp. 1–11, 2020.
- [52] T. Wang *et al.*, “Modeling drug combination effects via latent tensor reconstruction,” *Bioinformatics*, vol. 37, no. Supplement_1, pp. i93–i101, 2021.
- [53] L. Rønneberg, P. D. Kirk, and M. Zucknick, “Dose–response prediction for in-vitro drug combination datasets: a probabilistic approach,” *BMC bioinformatics*, vol. 24, no. 1, p. 161, 2023.
- [54] K. M. Gayvert, O. Aly, J. Platt, M. W. Bosenberg, D. F. Stern, and O. Elemento, “A computational approach for identifying synergistic drug combinations,” *PLoS computational biology*, vol. 13, no. 1, p. e1005308, 2017.
- [55] R. Celebi, O. Bear Don’t Walk IV, R. Movva, S. Alpsoy, and M. Dumontier, “In-silico prediction of synergistic anti-cancer drug combinations using multi-omics data,” *Scientific Reports*, vol. 9, no. 1, p. 8949, 2019.
- [56] B. Zagidullin, Z. Wang, Y. Guan, E. Pitkänen, and J. Tang, “Comparative analysis of molecular fingerprints in prediction of drug combination effects,” *Briefings in bioinformatics*, vol. 22, no. 6, p. bbab291, 2021.
- [57] H. Li, T. Li, D. Quang, and Y. Guan, “Network propagation predicts drug synergy in cancers,” *Cancer research*, vol. 78, no. 18, pp. 5446–5457, 2018.
- [58] L. He, J. Tang, E. I. Andersson, S. Timonen, S. Koschmieder, K. Wennerberg, S. Mustjoki, and T. Aittokallio, “Patient-customized drug combination prediction and testing for t-cell prolymphocytic leukemia patients,” *Cancer research*, vol. 78, no. 9, pp. 2407–2418, 2018.
- [59] D. Baptista, P. G. Ferreira, and M. Rocha, “A systematic evaluation of deep learning methods for the prediction of drug synergy in cancer,” *PLOS Computational Biology*, vol. 19, no. 3, p. e1010200, 2023.
- [60] X.-W. Wang, L. Madeddu, K. Spirohn, L. Martini, A. Fazzone, L. Becchetti, T. P. Wytock, I. A. Kovács, O. M. Balogh, B. Benczik, *et al.*, “Assessment of

- community efforts to advance network-based prediction of protein–protein interactions,” *Nature communications*, vol. 14, no. 1, p. 1582, 2023.
- [61] X. Wang, H. Zhu, D. Chen, Y. Yu, Q. Liu, and Q. Liu, “A complete graph-based approach with multi-task learning for predicting synergistic drug combinations,” *Bioinformatics*, vol. 39, no. 6, p. btad351, 2023.
- [62] F. Zhang, X. Zhao, J. Wei, and L. Wu, “Pathsynergy: a deep learning model for predicting drug synergy in liver cancer,” *Briefings in Bioinformatics*, vol. 26, no. 2, p. bbaf192, 2025.
- [63] B. Sun, H. Du, S. Hou, Q. Hu, X. Pang, D. Wei, and X. Wang, “Enhancing drug synergy combination: Integrating graph transformers and bilstm for accurate drug synergy prediction,” *IEEE Journal of Biomedical and Health Informatics*, 2025.
- [64] F. Rafiei, H. Zeraati, K. Abbasi, J. B. Ghasemi, M. Parsaeian, and A. Masoudi-Nejad, “Deeptrasynergy: drug combinations using multimodal deep learning with transformers,” *Bioinformatics*, vol. 39, no. 8, p. btad438, 2023.
- [65] M. Xu, X. Zhao, J. Wang, W. Feng, N. Wen, C. Wang, J. Wang, Y. Liu, and L. Zhao, “Dffndds: prediction of synergistic drug combinations with dual feature fusion networks,” *Journal of Cheminformatics*, vol. 15, no. 1, p. 33, 2023.
- [66] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.
- [67] X. Li, B. Shen, F. Feng, K. Li, Z. Tang, L. Ma, and H. Li, “Dual-view jointly learning improves personalized drug synergy prediction,” *Bioinformatics*, vol. 40, no. 10, p. btae604, 2024.

- [68] Y. Liu, P. Zhang, D. Li, C. Che, and B. Jin, “Mgtnsyn: Molecular structure-aware graph transformer network with relational attention for drug synergy prediction,” *Expert Systems with Applications*, p. 127699, 2025.
- [69] J.-J. G. Gutierrez, E. Lau, S. Dharmapalan, M. Parker, Y. Chen, M. A. Álvarez, and D. Wang, “Multi-output prediction of dose–response curves enables drug repositioning and biomarker discovery,” *npj Precision Oncology*, vol. 8, no. 1, p. 209, 2024.
- [70] G. C. Forcina, M. Conlon, A. Wells, J. Y. Cao, and S. J. Dixon, “Systematic quantification of population cell death kinetics in mammalian cells,” *Cell systems*, vol. 4, no. 6, pp. 600–610, 2017.
- [71] F. Iorio, T. A. Knijnenburg, D. J. Vis, G. R. Bignell, M. P. Menden, M. Schubert, N. Aben, E. Gonçalves, S. Barthorpe, H. Lightfoot, *et al.*, “A landscape of pharmacogenomic interactions in cancer,” *Cell*, vol. 166, no. 3, pp. 740–754, 2016.
- [72] M. E. Yohe, B. E. Gryder, J. F. Shern, Y. K. Song, H.-C. Chou, S. Sindiri, A. Mendoza, R. Patidar, X. Zhang, R. Guha, *et al.*, “Mek inhibition induces myog and remodels super-enhancers in ras-driven rhabdomyosarcoma,” *Science translational medicine*, vol. 10, no. 448, 2018.
- [73] A. Athar, A. Füllgrabe, N. George, H. Iqbal, L. Huerta, A. Ali, C. Snow, N. A. Fonseca, R. Petryszak, I. Papatheodorou, *et al.*, “Arrayexpress update—from bulk to single-cell expression data,” *Nucleic acids research*, vol. 47, no. D1, pp. D711–D715, 2019.
- [74] A. Subramanian, R. Narayan, S. M. Corsello, D. D. Peck, T. E. Natoli, X. Lu, J. Gould, J. F. Davis, A. A. Tubelli, J. K. Asiedu, *et al.*, “A next generation connectivity map: L1000 platform and the first 1,000,000 profiles,” *Cell*, vol. 171, no. 6, pp. 1437–1452, 2017.
- [75] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

- [76] Y. Li and Y. Yuan, “Convergence analysis of two-layer neural networks with relu activation,” in *Advances in neural information processing systems*, pp. 597–607, 2017.
- [77] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, ACM, 2016.
- [78] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [79] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *Journal of the royal statistical society: series B (statistical methodology)*, vol. 67, no. 2, pp. 301–320, 2005.
- [80] Y. Yao, L. Rosasco, and A. Caponnetto, “On early stopping in gradient descent learning,” *Constructive Approximation*, vol. 26, no. 2, pp. 289–315, 2007.
- [81] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [82] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, 2010.
- [83] F. Chollet *et al.*, “Keras: Deep learning library for theano and tensorflow,” *URL: <https://keras.io/k>*, vol. 7, no. 8, p. T1, 2015.
- [84] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, “Tensorflow: A system for large-scale machine learning,” in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pp. 265–283, 2016.
- [85] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, “Scikit-learn: Machine learning in python,” *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

- [86] C. Bliss, “The toxicity of poisons applied jointly 1,” *Annals of applied biology*, vol. 26, no. 3, pp. 585–615, 1939.
- [87] M. Berenbaum, “What is synergy? pharmacol. rev. 41, 93–141. erratum published in 1989,” *Pharmacol. Rev.*, vol. 41, p. 422, 1989.
- [88] B. Yadav, K. Wennerberg, T. Aittokallio, and J. Tang, “Searching for drug synergy in complex dose–response landscapes using an interaction potency model,” *Computational and structural biotechnology journal*, vol. 13, pp. 504–513, 2015.
- [89] A. Malyutina *et al.*, “Drug combination sensitivity scoring facilitates the discovery of synergistic and efficacious drug combinations in cancer,” *PLoS computational biology*, vol. 15, no. 5, p. e1006752, 2019.
- [90] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, *et al.*, “Drugbank 5.0: a major update to the drugbank database for 2018,” *Nucleic acids research*, vol. 46, no. D1, pp. D1074–D1082, 2018.
- [91] H. Kim, K. Chong, B.-K. Ryu, K.-J. Park, M. O. Yu, J. Lee, S. Chung, S. Choi, M.-J. Park, Y.-G. Chung, *et al.*, “Repurposing penfluridol in combination with temozolomide for the treatment of glioblastoma,” *Cancers*, vol. 11, no. 9, p. 1310, 2019.
- [92] A. Ranjan and S. K. Srivastava, “Penfluridol suppresses glioblastoma tumor growth by inhibiting sonic hedgehog signaling,” 2016.
- [93] J. Li, M. Kaneda, K. Patel, J. Wang, J. Akers, V. Ramakrishnan, T. Jiang, B. Carter, J. Varner, and C. Chen, “Os06. 7 pi-3k γ inhibition suppresses glioblastoma tumorigenicity through disruption of an il11-stat3-myc signaling axis between microglia and glioblastoma,” *Neuro-oncology*, vol. 19, no. Suppl 3, p. iii12, 2017.
- [94] J. Li, M. M. Kaneda, J. Ma, M. Li, K. Patel, T. Koga, A. L. Sarver, F. Furnari, B. Xu, S. Dhawan, *et al.*, “Pi3k γ inhibition suppresses microglia/tam accumulation in glioblastoma microenvironment to promote exceptional temozolomide response,” *bioRxiv*, 2020.

- [95] M. Zhang, T. W. Herion, C. Timke, N. Han, K. Hauser, K. J. Weber, P. Peschke, U. Wirkner, M. Lahn, and P. E. Huber, “Trimodal glioblastoma treatment consisting of concurrent radiotherapy, temozolomide, and the novel $\text{tgf-}\beta$ receptor i kinase inhibitor ly2109761,” *Neoplasia (New York, NY)*, vol. 13, no. 6, p. 537, 2011.
- [96] Y. Chang, H. Park, H.-J. Yang, S. Lee, K.-Y. Lee, T. S. Kim, J. Jung, and J.-M. Shin, “Cancer drug response profile scan (cdrscan): a deep learning model that predicts drug effectiveness from cancer genomic signature,” *Scientific reports*, vol. 8, no. 1, pp. 1–11, 2018.
- [97] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed, “A comparison of normalization methods for high density oligonucleotide array data based on variance and bias,” *Bioinformatics*, vol. 19, no. 2, pp. 185–193, 2003.
- [98] J. Cohen, “A coefficient of agreement for nominal scales,” *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.
- [99] D. Wooten *et al.*, “MuSyC is a consensus framework that unifies multi-drug synergy metrics for combinatorial drug discovery,” *Nature communications*, vol. 12, no. 1, pp. 1–16, 2021.
- [100] J. Ross, B. Belgodere, V. Chenthamarakshan, I. Padhi, Y. Mroueh, and P. Das, “Large-scale chemical language representations capture molecular structure and properties,” *Nature Machine Intelligence*, vol. 4, no. 12, pp. 1256–1264, 2022.
- [101] C. T. Meyer *et al.*, “Quantifying drug combination synergy along potency and efficacy axes,” *Cell systems*, vol. 8, no. 2, pp. 97–108, 2019.
- [102] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.
- [103] D. Miricescu, D. G. Balan, A. Tulin, O. Stiru, I. A. Vacaroiu, D. A. Mihai, C. C. Popa, R. I. Papacoea, M. Enyedi, N. A. Sorin, *et al.*, “Pi3k/akt/mtor signalling pathway involvement in renal cell carcinoma pathogenesis,” *Experimental and Therapeutic Medicine*, vol. 21, no. 5, pp. 1–7, 2021.

- [104] H. Guo, P. German, S. Bai, S. Barnes, W. Guo, X. Qi, H. Lou, J. Liang, E. Jonasch, G. B. Mills, *et al.*, “The pi3k/akt pathway and renal cell carcinoma,” *Journal of genetics and genomics*, vol. 42, no. 7, pp. 343–353, 2015.
- [105] N. Petrucelli, M. B. Daly, and T. Pal, “Brca1-and brca2-associated hereditary breast and ovarian cancer,” 2022.
- [106] M. R. Radu, A. Prădatu, F. Duică, R. Micu, S. M. Crețoiu, N. Suciu, D. Crețoiu, V. N. Varlas, and V. E. Rădoi, “Ovarian cancer: biomarkers and targeted therapy,” *Biomedicines*, vol. 9, no. 6, p. 693, 2021.
- [107] V. Barili, E. Ambrosini, B. Bortesi, R. Minari, E. De Sensi, I. R. Cannizzaro, A. Taiani, M. Michiara, A. Sikokis, D. Boggiani, *et al.*, “Genetic basis of breast and ovarian cancer: Approaches and lessons learnt from three decades of inherited predisposition testing,” *Genes*, vol. 15, no. 2, p. 219, 2024.
- [108] L. Tang, Z. Huang, H. Mei, and Y. Hu, “Immunotherapy in hematologic malignancies: achievements, challenges and future prospects,” *Signal Transduction and Targeted Therapy*, vol. 8, no. 1, p. 306, 2023.
- [109] E. Estey, R. L. Levine, and B. Löwenberg, “Current challenges in clinical development of “targeted therapies”: the case of acute myeloid leukemia,” *Blood, The Journal of the American Society of Hematology*, vol. 125, no. 16, pp. 2461–2466, 2015.
- [110] A. Calderon, C. Han, S. Karma, and E. Wang, “Non-genetic mechanisms of drug resistance in acute leukemias,” *Trends in Cancer*, vol. 10, no. 1, pp. 38–51, 2024.
- [111] E. B. Candir, H. I. Kuru, M. Rattray, A. E. Cicek, and O. Tastan, “One-hot news: Drug synergy models take a shortcut,” *bioRxiv*, pp. 2025–04, 2025.

Appendix A

Code

The implementations of each model is available at their GitHub repositories. Each repository includes the code and the instructions to train corresponding models.

- GitHub page of MatchMaker: <https://github.com/hikuru/matchmaker>
- GitHub page of PDSP: <https://github.com/hikuru/pdsp>
- GitHub page of DeepSynBa: <https://github.com/hikuru/DeepSynBa>