Journal Title Here, 2022, pp. 1–9

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

One-Hot News: Drug Synergy Models Take a Shortcut

Emine Beyza Çandır,[1] Halil İbrahim Kuru,[2] Magnus Rattray,[3] A. Ercument Cicek[2] and Oznur Tastan[1,*]

[1]Facutly of Engineering and Natural Sciences, Sabanci University, Istanbul,Turkey, [2]Department of Computer Engineering, Bilkent University, Ankara,Turkey and [3]Division of Informatics, Imaging and Data Sciences, University of Manchester, Manchester, UK
[*]Corresponding author. otastan@sabanciuniv.edu
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Combinatorial drug therapy holds great promise for tackling complex diseases, but the vast number of possible drug combinations makes exhaustive experimental testing infeasible. Computational models have been developed to guide experimental screens by assigning synergy scores to drug pair–cell line combinations, where they take input structural and chemical information on drugs and molecular features of cell lines. The premise of these models is that they leverage this biological and chemical information to predict synergy measurements. In this study, we demonstrate that replacing drug and cell line representations with simple one-hot encodings results in comparable or even slightly improved performance across diverse published drug combination models. This unexpected finding indicates that current models take a 'shortcut'; rather than learning the intended chemical or biological information, they exploit these representations as mere identifiers and capture general co-variation patterns of synergy measurements. While the current synergy prediction models can still aid in prioritizing drug pairs within a panel of tested drug and cell lines, our results highlight the need for better strategies to learn from intended features and generalize to unseen drugs and cell lines.

Key words: Drug Synergy, Drug Combinations, Generalization, Shortcut Learning

Introduction

Synergistic drug combinations enable lower dosing of each agent, reducing adverse side effects, which is especially important in treating complex diseases such as cancer (Mokhtari *et al.*, 2017; Al-Lazikani *et al.*, 2012; Tamma *et al.*, 2012; Möttönen *et al.*, 1999; Gradman *et al.*, 2010). However, the vast number of possible drug pair and cell line combinations makes exhaustive clinical evaluations of drug combinations infeasible. To address this limitation, several computational models have been developed to guide experimental efforts (Abbasi and Rousu, 2024). By predicting synergistic scores for drug pairs on cell lines, these models help prioritize which combinations merit further testing.

In a typical synergy prediction model, each input example is composed of a triplet: a drug pair and the cell line on which the two drugs's combined effect is measured. Models take numeral representations of these triplets using various descriptors for drugs and cell lines. For example, to describe drugs, many methods employ chemical and structural fingerprints (Xu *et al.*, 2023; Wang *et al.*, 2023; Preuer *et al.*, 2017; Kuru *et al.*, 2022; Güvenç Paltun *et al.*, 2021). Graph-based representations have also been used, such as in DeepDDS (Wang *et al.*, 2021). There are also multi-modal approaches that use different representations jointly, such as JointSyn (Li *et al.*, 2024), which uses fingerprints and molecular graphs. Alternative strategies have been proposed; for example, Marsy captures drug characteristics through differential expression signatures induced in two specific cell lines (El Khili *et al.*, 2023). The biological context of cell lines is often represented by the untreated gene expression profiles of the cell line. The premise of all these models is that they intend to leverage the chemical and biological information of drugs and cell lines to predict synergy.

Previously, models have been reported to perform well on new combinations of drugs and cell lines that are exclusively seen during training but fail on new drugs or cell lines during testing. This is revealed in the performances obtained in evaluation setups where the test data is split in a stratified way (Li *et al.*, 2024; Abbasi and Rousu, 2024; Wang *et al.*, 2024). These split scenarios for drug-drug-cell line combinations are illustrated in Figure 1 and are detailed in the Methods. Models perform very well on unseen triplet examples (leave-triplet-out) as this setup is relatively easy. High performance is also reported for the unseen drug pairs (leave-drug-pair-out), where both drugs are observed in training but only within triplets where they are paired with other drugs and cell lines. Model performance deteriorates when models are evaluated in triplets of an unseen cell line (leave-cell-line-out) (Preuer *et al.*, 2017; Xu *et al.*, 2023; Li *et al.*, 2024; Zhang and Tu, 2023). Model predictive performances also decrease when the test dataset is formed of triplets where one of the tested drugs is not observed

Table 1. Summary of Drug and Cell Line Features Used by Each Model

Model	Drug Features	Cell Line Features
DeepSynergy (Preuer *et al.*, 2017)	ECFP6 molecular fingerprints (1309 features), physicochemical properties (802 features), and toxicophore features (2276 features).	Untreated gene expression values of 3984 genes.
MatchMaker(Kuru *et al.*, 2022)	Chemical descriptors of drugs (367 features), such as charge and connectivity descriptors (detailed in Table S3.	Untreated cell gene expression values of 972 landmark genes Subramanian *et al.* (2017).
DeepDDS(Wang *et al.*, 2021)	Molecular graphs: Derived from SMILES strings, where nodes represent atoms, edges represent bonds, and node features are binary vectors containing 5 atom-related properties.	Untreated gene expression profiles with 954 genes.
MARSY(El Khili *et al.*, 2023)	Landmark genes differential gene expression profiles measured in two cell lines (MCF7 and PC3) after drug treatment (total 1956 features per drug).	Untreated gene expression values of 4639 genes.
JointSyn(Li *et al.*, 2024)	Molecular graphs: Derived from SMILES strings with nodes representing atoms, edges representing bonds, and 78-dimensional atomic feature vectors. Morgan fingerprints: ECFP6 molecular fingerprints (1309 features).	Untreated gene expression values of 2087 genes.

in the training (leave-one-drug-out) (Preuer *et al.*, 2017; Xu *et al.*, 2023; Zhang and Tu, 2023). In particular, all model performances suffer when evaluated in triplets where both drugs in the triplets tested are not observed during training (Li *et al.*, 2024).

While a generalization problem has been reported in the literature, here we provide strong evidence that these models may not be learning from the intended chemical and biological information. For a set of diverse synergy prediction models, we replaced complex representations of drugs/cell lines with simple one-hot encodings, and the models' performances remained mostly on par and, in some cases, even improved slightly. These one-hot vectors simply serve as unique and orthogonal identifiers for each drug and cell line, bearing no biological content. To verify our findings, we experimented with a diverse set of published models that use different drug and cell line representations. We used source codes and data splits from the published studies, consistently observing similar results across all the models tested.

While previous work has documented generalization issues when the models are applied to unseen drugs and cell lines, our results reveal a more fundamental issue than previously acknowledged. The models take a shortcut and learn from the co-variation patterns of the synergy measurements rather than from the biologically relevant features, which could be the reason behind the generalization barrier. Below, we describe in detail the experiments we conducted.

Methods

Tested Models

In our analysis, we selected different models that are well-recognized or recent: DeepSynergy (Preuer *et al.*, 2017), DeepDDS (Wang *et al.*, 2021), MatchMaker (Kuru *et al.*, 2022), MARSY (El Khili *et al.*, 2023), and JointSyn (Li *et al.*, 2024). Each model uses different approaches and features for drug synergy prediction, while all use cell line gene expression profiles to represent the biological context. Below, we provide a summary of these approaches. The details about the number of features and the dataset descriptions are also provided in Table 1.

- **DeepSynergy** pioneered the use of deep learning for synergy prediction (Preuer *et al.*, 2017). It uses a fully connected neural network that concatenates feature vectors for two drugs and a cell line. Each drug is represented by three types of chemical descriptors: extended connectivity fingerprints (ECFP6) with 1309 features, physicochemical properties with 802 features, and 2276 toxicophore features. To characterize cell lines, DeepSynergy uses gene expression profiles of untreated cells (Iorio, 2016), filtered down to 3984 informative genes.
- **MatchMaker** employs two parallel drug-specific subnetworks alongside a Synergy Prediction Network (SPN), where each drug-specific subnetwork (DSN) processes the chemical features of one drug and the gene expression features of the corresponding cell line (Kuru *et al.*, 2022). The outputs of the DSNs are concatenated and fed into the SPN to predict synergy scores. Drugs are represented using 367 chemical descriptors calculated with the PyBioMed library. These descriptors encode the chemical structure of each drug. For cell lines, baseline gene expression values of untreated cells, consisting of 972 landmark genes, are used.
- **DeepDDS** is a GNN-based model designed to learn drug representations from molecular graphs created from SMILES strings. In these graphs, nodes represent atoms, and edges represent bonds. Each node is described by a binary vector containing five atomic properties, including atom type and aromaticity. For cell lines, DeepDDS uses baseline gene expression profiles of untreated cells, filtered to include 954 genes. These genes were selected by intersecting CCLE expression data with LINCS landmark genes and removing noncoding RNA transcripts (Wang *et al.*, 2021).
- **MARSY** generates representations for drug pairs and their interactions with cell lines through separate encoders (El Khili *et al.*, 2023). These representations are integrated within a multitask predictor to output both synergy scores and single-drug responses. MARSY represents drugs using their differential gene expression (DGE) signatures obtained from the LINCS database (Subramanian, 2017). These signatures were measured in two cell lines, MCF7 and PC3, 24 hours after drug treatment. Each drug was characterized by a concatenation of 978 landmark genes from both

Table 2. Summary of Synergy Scores and Dataset Statistics for Each Model. Note that while some models use datasets from the same sources, the sizes of these datasets vary according to the distinct filtering criteria employed in each study.

Models - Dataset	Synergy Score	# of Combinations	# of Drugs	# of Cell Lines
DeepSynergy - O'Neil	Loewe	23,052	38	39
MatchMaker - DrugComb	Loewe	426,239	3,057	167
MatchMaker - NCI-ALMANAC	ComboScore	264,424	99	54
DeepDDS - O'Neil	Binarized Loewe	12,415	36	31
MARSY - DrugComb	Zip	86,348	670	75
JointSyn - O'Neil	Loewe	12,033	38	34

cell lines, resulting in 1,956 features per drug. For cell lines, MARSY utilizes baseline gene expression profiles of untreated cells from the CCLE via the CellMiner database (Reinhold *et al.*, 2012). After filtering out lowly expressed and low-variance genes, the final cell line representation includes 4,639 genes.

- **JointSyn** is a recent model that reported competitive performance compared to other state-of-the-art methods (Li *et al.*, 2024). It integrates a joint graph encompassing drug combinations, drug features, and cell line representations within a dual-view architecture. These views generate embeddings for both the drug combination and the cell line, which are then passed to a prediction network to estimate synergy scores. JointSyn represents drugs using molecular graphs and Morgan fingerprints. Molecular graphs are derived from SMILES strings using RDKit (Landrum, 2016), where atoms are represented as nodes and bonds as edges. Each node is characterized by a 78-dimensional atomic feature vector computed with DeepChem (Ramsundar *et al.*, 2019). Additionally, Morgan fingerprints with 1,309 features capture further structural characteristics. For cell lines, JointSyn employs baseline gene expression profiles comprising 2,087 genes, which are filtered from the CCLE database based on drug sensitivity relevance.

Datasets and processing

Our objective is not to benchmark models on a single dataset but to compare each original model with its counterpart baseline model that is trained using one-hot-encoded features. To this end, we intentionally used the datasets reported in the original publications for each respective model except for Matchmaker, which we explain below. Synergy datasets comprise triplets: a drug pair and the cell lines, along with a synergy score derived from an experimental grid of measurements made at different drug dosages.

DeepSynergy was evaluated using the O'Neil dataset, available on the DeepSynergy website, which includes 23,052 drug–cell line combinations involving 38 drugs and 39 cell lines. Similarly, the DeepDDS model used another subset of the O'Neil dataset. This set included 12,415 combinations involving 36 drugs and 31 cell lines. JointSyn also uses a subset of the O'Neil dataset, as detailed in its publication, comprising 12,033 combinations of 38 drugs and 34 cell lines. MARSY was trained on a filtered version of the DrugComb dataset provided by its authors, containing 86,348 combinations involving 670 drugs and 75 cell lines.

For the MatchMaker model, we were able to conduct a more in-depth analysis. We used both the DrugComb and NCI-ALMANAC datasets. We employed an updated version of the DrugComb dataset, which was different from the original publication. We used this larger dataset because it is more suitable for the leave-drug-out split, wherein a subset of the data must be excluded. The DrugComb dataset comprises 739,964 drug-drug–cell line combinations involving 8,397 drugs and 2,320 cell lines. We filtered this dataset to include only drugs with available structural information in the PubChem database and cell lines with accessible gene expression data from the Genomics of Drug Sensitivity in Cancer (GDSC) database (Iorio, 2016). Following this filtering process, we obtained 426,239 combinations covering 3,057 drugs and 167 cell lines. The DrugComb dataset is imbalanced in the frequency with which individual drugs are represented across combinations. Approximately two-thirds of the drugs are involved in only one or two combinations within the dataset. In contrast, some drugs appear in more than ten thousand combinations. This imbalance extends to drug pairs as well; as illustrated in Figure S1, one drug in a pair may be present in a single combination, while another one is included in over 7,000 combinations. Given the imbalanced nature of the DrugComb dataset, which may contribute to the results, we extended our experiments to include the NCI-ALMANAC dataset, which is more balanced. The original NCI-ALMANAC dataset comprised 304,549 combinations of 104 drugs and 60 cell lines. After applying the same filtering criteria, based on drug structures and cell line gene expressions we used for DrugComb, there were 264,424 combinations involving 99 drugs and 54 cell lines. As illustrated in Figure S2, over 75% of drug pairs in the filtered dataset feature both drugs appearing between 5,000 and 6,000 times. Additionally, every drug in the dataset appears in at least 4,000 combinations, resulting in a more evenly distributed dataset.

The models are trained to predict different synergy scores. MatchMaker uses the Loewe Additivity score (Loewe, 1953) when DrugComb is employed. When trained on NCI-ALMANAC MatchMaker, the ComboScore is employed instead of Loewe. MARSY uses the ZIP score from the DrugComb dataset. DeepDDS applies a threshold on the Loewe Additivity score to binarize. Combinations with a synergy score greater than 10 are classified as synergistic, while scores below zero are categorized as antagonistic.

Table 2 summarizes these datasets used in our experiments for each model, including the synergy score metrics, as well as the number of drug combinations, drugs, and cell lines.

Data Splitting Strategies for Evaluations

The drug combination model performances can be assessed in different evaluation sets, where the criteria the test data set aside differs. These scenarios affect the prediction task difficulty and are important to understand the model's capabilities and performance claims. We illustrate these scenarios in Figure 1.

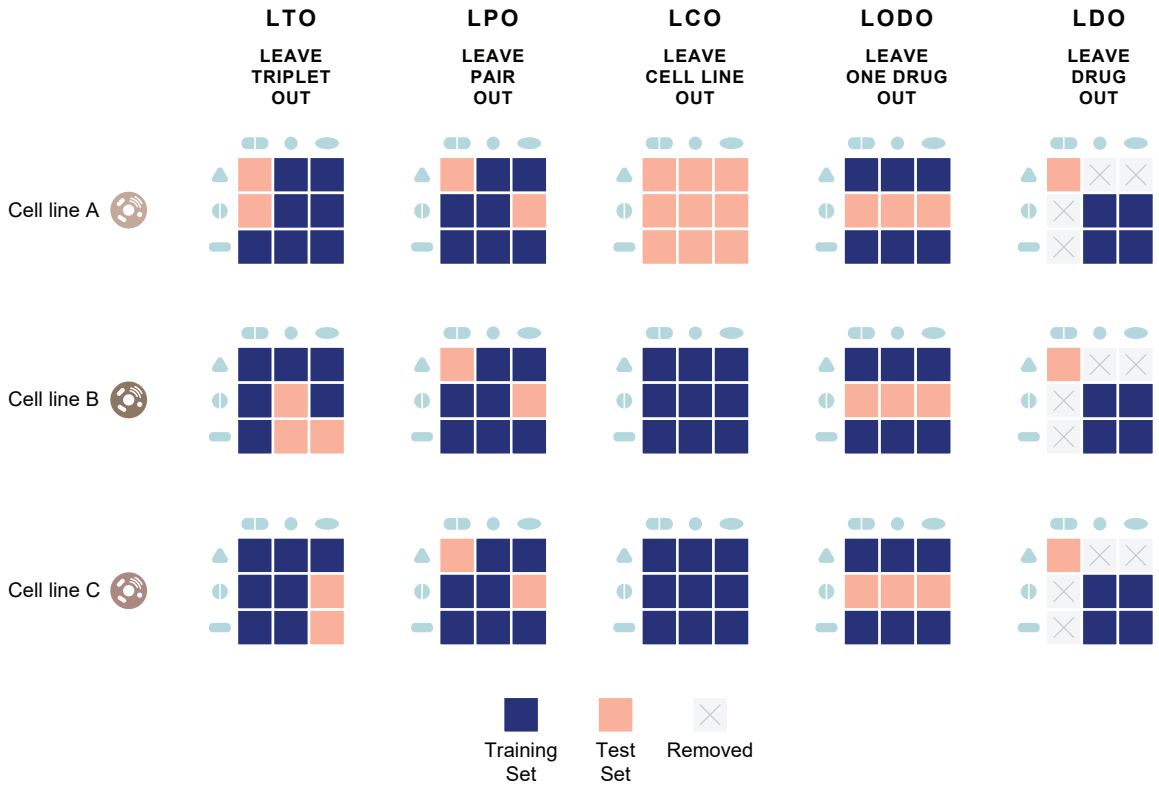

Fig. 1. Illustration of data split strategies, inspired by Preuer *et al.* (2017). Each evaluation strategy shows how drugs and cell lines are included or excluded in the training and test sets. The detailed description of these strategies is explained in Section 2.3.

- **Leave-Triple-Out (LTO):** The drug-drug–cell line triplets are randomly split without any stratification. The test set includes unseen drug-drug–cell line triplets. However, individual drugs, drug pairs or cell lines within these combinations may still appear with another partner in the training data. This is the easiest scenario.
- **Leave-Pair-Out (LPO):** Drug pairs present in the test set are not included in the training set. However, individual drugs within these pairs can still appear in the training data, paired with other drugs. There is no restriction on cell lines; the test set may include cell lines seen during training. LPO evaluates the model's ability to predict interactions between drug pairs it has not encountered before. This is the most commonly used strategy in the literature.
- **Leave-Cell-Line-Out (LCO):** This strategy sets aside test data based on cell lines. Consequently, the train and test triplets do not share common cell lines. LCO tests the model's ability to generalize predictions to new biological contexts represented by unseen cell lines. This is crucial for evaluating how well the model can adapt to different cellular contexts, which is important for applications like personalized medicine, where patient-specific cell responses are considered.
- **Leave-One-Drug-Out (LODO):** In stratification, at least one drug in each drug pair within the test set is completely absent in the training triplets. The other drug in the pair may still appear in the training set, interacting with different drugs. There is no restriction on cell lines; the test set may contain cell lines seen during training. The purpose of LODO is to evaluate the model's ability to predict interactions when it has incomplete information—specifically when one drug is entirely new to the model. This scenario reflects real-world situations where a new drug is introduced.
- **Leave-Drug-Out (LDO):** If a drug is seen in the training data, this drug and all of its drug interactions are excluded from the test set. As a result, the model does not see any of the test drugs during training. Implementing this split requires first splitting the drugs and then adding their associated triplet; this reduces the number of training examples (Figure 1). There is no restriction on cell lines; the test set can include cell lines in the training data. LDO evaluates how well it can predict interactions involving drugs it has never seen in the training.

Experimental Setup

In our experiments, we adopted the hyperparameters specified in the original papers for each model. For dataset splits, we utilized the original splits provided by the authors. Regarding DeepSynergy, we conducted LPO 5-fold nested cross-validation, consistent with the procedure described in the original paper, using the folds provided by the authors. For DeepDDS, we followed the LTO split procedure outlined in the original work, applying 5-fold cross-validation. In the MARSY experiments, the model employed the LPO split method along with 5-fold cross-validation. For JointSyn, the LTO split was generated

using the authors' provided code, resulting in 10 replicates per fold.

While the MatchMaker paper initially utilized the LPO split, we applied additional split strategies to study generalization and the behaviour of one-hot encoded features under different conditions. These included LPO, LCO, LODO, and LDO. In each case, 60% of the triplets are set for training, 20% for validation, and 20% for testing. For each split strategy, we generated 10 different train/test splits by varying the random split state. Results averaged across these splits are reported. Due to the nature of the LDO split, where the test set contains only unseen drugs, not all data could be utilized. Consequently, the DrugComb and NCI-ALMANAC datasets were reduced by approximately 68%

All models were trained using both their original drug and cell line features as well as one-hot-encoded features constructed based on their datasets. For JointSyn, we could not train the models with their original drug/cell line features due to a compatibility issue in the required libraries and the CUDA version on the cluster. Therefore, we used the results reported in their paper for these features. For DeepDDS and JointSyn, when training one-hot-encoded feature models, the graph components of the original architectures were removed, and one-hot-encoded features were directly provided as input to the models. The original architecture of DeepDDS, including the graph components, was fully maintained using its drug and cell line features.

Computational Environment

We performed DeepSynergy, MatchMaker, MARSY, JointSyn, and DeepDDS experiments on systems equipped with Tesla V100-PCIE-32GB and Tesla V100S-PCIE-32GB GPUs. The Tesla V100-PCIE-32GB featured 31.74 GiB of memory, 80 cores, and a memory bandwidth of 836.37 GiB/s, while the Tesla V100S-PCIE-32GB offered the same memory and cores but with a bandwidth of 1.03 TiB/s. These systems included CPUs with frequencies ranging from 2.29 GHz to 3.59 GHz. TensorFlow with CUDA 10.1 and cuDNN 7 was used for training.

Results

To assess the utility of the biochemical information we compared for each model chosen, we trained it with its original biological and chemical information as reported in the publication and compared it with models, where we replaced these biochemical features with one-hot-encoded features. By stripping away the biological and chemical information, we aim to assess to what extent the models rely on this intended information. We evaluated each model with its original source, dataset, and train/test splits. The details of these setups are provided in Section Methods and the code reproducing the analysis are available at `https://github.com/tastanlab/ohe`.

Comparison of the Models with the OHE Baseline Models

Figure 2 shows that the performance differences between the original drug and cell line features and OHE representations are minimal across all the models tested. For MatchMaker on the DrugComb dataset under the LPO split, the MSE difference between the original features and OHE representations was just 3.11%, indicating that the model performance remains largely unaffected if the drug and cell line molecular features are removed. DeepSynergy exhibited a minimal deviation (-0.92%) on the O'Neil dataset with OHE representations, demonstrating comparable performance to results obtained using the original drug and cell line features. Notably, MARSY, tested on the DrugComb dataset, showed a marginal improvement (-8.14%) when OHE features were used instead of the original feature. It suggests that the model's predictive performance does not depend on the chemical and biological features. The performance gap was negligible for models like JointSyn and DeepDDS, which incorporate graph-based embedding techniques. JointSyn reported only a 1.07% deviation in MSE under the LTO split, while DeepDDS showed a mere 1.08% deviation in AUC. Thus, the results were persistent even for more complex graph-based architectures.

The performances of the models remain largely consistent regardless of whether original features or OHE representations were used. This consistent finding across a different set of synergy prediction models points out that the performance does not genuinely depend on the intended biological or chemical information provided to the model. Instead, the models take shortcuts by learning from the co-variation patterns of the measurements. This could be an instance of shortcut learning Geirhos *et al.* (2020).

MARSY, along with synergy scores, predicts single-drug responses. In our experiment, we also observed a similar result for single drug response prediction in MARSY; models trained with OHE, despite lacking any transcriptomic information, perform on par with the models trained with the biological features (Table S5). This finding suggests that this issue may extend beyond the drug synergy prediction domain and might be prevalent for single-drug response predictions as well. This result prompts a re-evaluation of drug representation and evaluation strategies in synergy prediction models, and an assessment of whether the same problem holds in other drug prediction tasks is needed.

A Deeper Analysis with MatchMaker

To investigate generalization further, we experimented with Matchmaker in other evaluation setups that use different split strategies (explained in Methods). We omit the LTO strategy as this is the most straightforward scenario, and the models generally perform very well. We compared Matchmaker's performance, which uses chemical descriptors and gene expression levels, with the model trained on the OHE representations for drugs and cell lines. We conducted these experiments both on DrugComb and NCI-ALMANAC datasets.

The model architectures correctly exploit underlying covariation patterns in the synergy measurements to make predictions for drug combinations within a tested drug panel and cell line, as indicated by the LPO split. However, the results degrade as the evaluation strategy sets a harder challenge. The model performance worsens on unseen cell lines, as shown in the LCO results. The results obtained on LODO, when one of the drugs is not seen, are even lower. The model fails to generalize to the case when both drugs in the test data are new, the LDO setup. These results show the models' inability to transfer this knowledge to new, unseen drugs and cell lines.

The models trained with original features and OHE features show comparable performance on the LPO and LTO splits. The models relying on OHE struggle significantly when predicting synergy for drug pairs not present in the training set, the LDO scenario. In OHE, each drug is treated as an

Model	Dataset	Split Method	Metric	Drug & Cell Line Features	One-Hot Encoding	Percent Deviation (%)
DeepSynergy	O'Neil	LPO	MSE	252.55 ± 12.46	250.23 ± 13.24	-0.92 ± 1.76
			PCC	0.72	0.73	1.39
			SCC	0.72	0.73	1.39
DeepDDS	O'Neil	LTO	ROC AUC	0.93	0.94	1.08
			PR AUC	0.93	0.94	1.08
			ACC	0.85	0.86	1.18
Matchmaker	DrugComb	LPO	MSE	97.23 ±1.14	100.25 ±1.15	3.11 ±0.54
			PCC	0.76	0.75	-1.32
			SCC	0.72	0.70	-2.78
MARSY	DrugComb	LPO	MSE	32.94 ± 1.20	30.26 ± 1.11	-8.14 ± 1.49
			PCC	0.87	0.88	1.15
			SCC	0.74	0.75	1.35
JointSyn	O'Neil	LTO	MSE	76.56 ±0.48*	77.38 ±2.21	1.07
			PCC	0.89	0.88	-1.12
			R^2	0.78	0.78	0.00

Fig. 2. Comparison of Drug & Cell Line Features vs OHE Representations Across Different Models and Datasets. Drug and cell line features refer to the original representations of drugs and cell lines used in their respective models. Performance metrics, percent deviation, and standard error (SE) are included in the table. The details of the SE calculations are provided in the Supplementary Section on Standard Error Calculations. For clarity, SE values less than 0.01 are omitted from the table but are available in the Supplementary Table S6. *Results for JointSyn using original features were sourced from the respective publication.

Table 3. Performance Comparison of MatchMaker Model Using Drug & Cell Line Features vs OHE Representations on DrugComb Dataset Across Different Split Methods. Chemical descriptors were originally used to represent drug features, while cell line gene expression levels represented cell line features in the MatchMaker model. These representations serve as the baseline for comparison with one-hot-encoded drug and cell line representations.

Split Method	Feature Type	MSE	PCC	SCC
LPO	Drug & Cell Feature	97.23 ± 1.14	0.76 ± 0.002	0.72 ± 0.001
	One-Hot-Encoding	100.25 ± 1.15	0.75 ± 0.002	0.70 ± 0.002
LCO	Drug & Cell Feature	161.68 ± 1.50	0.52 ± 0.013	0.45 ± 0.012
	One-Hot-Encoding	158.11 ± 1.45	0.53 ± 0.012	0.46 ± 0.012
LODO	Drug & Cell Feature	199.22 ± 1.75	0.39 ± 0.009	0.36 ± 0.007
	One-Hot-Encoding	202.16 ± 1.76	0.33 ± 0.009	0.28 ± 0.007
LDO	Drug & Cell Feature	237.42 ± 3.34	0.13 ± 0.012	0.14 ± 0.010
	One-Hot-Encoding	223.57 ± 3.18	0.10 ± 0.009	0.10 ± 0.008

Table 4. Performance Comparison of MatchMaker Model Using Drug & Cell Line Features vs OHE Representations on NCI-ALMANAC Dataset Across Different Split Methods. Chemical descriptors were originally used to represent drug features, while cell line gene expression levels represented cell line features in the MatchMaker model. These representations serve as the baseline for comparison with one-hot-encoded drug and cell line representations.

Split Method	Feature Type	MSE	PCC	SCC
LPO	Drug & Cell Feature	2912.40 ± 68.03	0.57 ± 0.012	0.54 ± 0.005
	One-Hot-Encoding	3041.02 ± 70.18	0.54 ± 0.010	0.52 ± 0.002
LCO	Drug & Cell Feature	2956.30 ± 43.27	0.57 ± 0.007	0.46 ± 0.005
	One-Hot-Encoding	2930.47 ± 45.97	0.57 ± 0.006	0.47 ± 0.005
LODO	Drug & Cell Feature	3903.41 ± 83.75	0.26 ± 0.014	0.27 ± 0.012
	One-Hot-Encoding	3943.30 ± 84.28	0.24 ± 0.010	0.25 ± 0.004
LDO	Drug & Cell Feature	4587.08 ± 152.46	0.16 ± 0.021	0.17 ± 0.018
	One-Hot-Encoding	4670.97 ± 154.58	0.09 ± 0.010	0.09 ± 0.007

independent category without shared features that facilitate extrapolation to unseen combinations. Models utilizing feature-rich representations potentially encode drug similarities and relationships, enabling better generalization. Nevertheless, the results are still poor for both models in the LDO case.

Discussion

Effective predictions can streamline the identification of effective drug combinations in different biological contexts, reducing the time and cost associated with experimental validation. While the current synergy prediction models report high predictive performance when tested on new combinations of drugs or cell lines that are observed in training, they exhibit significant limitations in generalizing to unseen drugs. In this work, by replacing the biological and chemical information on drugs and cell lines with simple identifiers, we show that the models learn from the co-variation patterns of the synergy measurements rather than from the domain-relevant features.

Shortcut learning instances have been reported in the literature for deep vision and natural language processing tasks (Geirhos *et al.*, 2020; Hermann *et al.*, 2023). For example, a deep neural network might seem to recognize cows well, but it struggles with images where cows are not set against a typical grassy background—exposing that it had inadvertently used the presence of grass as a shortcut to predict cows (Beery *et al.*, 2018). Other well-known examples include a model learning to use the presence of rulers in lesion images as a shortcut for predicting malignancy (Narla *et al.*, 2018) or learning based on the presence of hospital identifiers in X-ray scans (Zech *et al.*, 2018). These shortcuts are caused by confounding correlations between the inputs and outputs. In the present case, the correlations are of a different nature; the algorithm learns non-trivial correlations between the drug and cell-line input representations that are informative about the target. But rather than learning anything about these rich representations, the model learns to simply recognise these high-dimensional patterns associations with the target.

This study highlights once more that diagnosing the model behaviours with different strategies is imperative to understanding the capabilities and advancing these models. These results also caution us to have multiple baseline models to interrogate the predictions.

Despite the challenges in generalizing to novel drug pairs, models that capture co-variation patterns remain valuable for discovering new synergistic combinations within an existing drug panel. By leveraging the learned interactions and correlations, these models can efficiently identify promising combinations that warrant further experimental testing, thereby accelerating the drug discovery process within the scope of the tested drugs. However, broader generalization remains an area for improvement.

Competing interests

No competing interest is declared.

Author contributions statement

B.C. collected data, conducted the experiments, and wrote the first draft of the manuscript; H.İ.K. contributed to software development and reviewed the manuscript; A.E.C. and M.R. analyzed the results and reviewed the manuscript; O.T. conceived the experiments, analyzed results, and wrote and reviewed the manuscript.

References

Abbasi, F. and Rousu, J. (2024). New methods for drug synergy prediction: A mini-review. *Current Opinion in Structural Biology*, **86**, 102827.

Al-Lazikani, B. *et al.* (2012). Combinatorial drug therapy for cancer in the post-genomic era. *Nature Biotechnology*, **30**(7), 679–692.

Beery, S. *et al.* (2018). Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473.

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

Geirhos, R. *et al.* (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, **2**(11), 665–673.

Gradman, A. H. *et al.* (2010). Combination therapy in hypertension. *Journal of the American Society of Hypertension*, **4**(2), 90–98.

Güvenç Paltun, B. *et al.* (2021). Machine learning approaches for drug combination therapies. *Briefings in Bioinformatics*, **22**(6), bbab293.

Hermann, K. L. *et al.* (2023). On the foundations of shortcut learning. *CoRR*, **abs/2310.16228**.

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

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

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

Li, X. *et al.* (2024). Dual-view jointly learning improves personalized drug synergy prediction. *Bioinformatics*, **40**(10).

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

Mokhtari, R. B. *et al.* (2017). Combination therapy in combating cancer. *Oncotarget*, **8**(23), 38022–38043.

Möttönen, T. *et al.* (1999). Comparison of combination therapy with single-drug therapy in early rheumatoid arthritis: a randomised trial. *The Lancet*, **353**(9164), 1568–1573.

Narla, A. *et al.* (2018). Automated classification of skin lesions: from pixels to practice. *Journal of Investigative Dermatology*, **138**(10), 2108–2110.

Preuer, K. *et al.* (2017). Deepsynergy: predicting anti-cancer drug synergy with deep learning. *Bioinformatics*, **34**(9), 1538–1546.

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

Reinhold, W. C. *et al.* (2012). Cellminer: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the nci-60 cell line set. *Cancer research*, **72**(14), 3499–3511.

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

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

Tamma, P. D. *et al.* (2012). Combination therapy for treatment of infections with gram-negative bacteria. *Clinical Microbiology Reviews*, **25**(3), 450–470.

Wang, J. *et al.* (2021). Deepdds: deep graph neural network with attention mechanism to predict synergistic drug combinations. *Briefings in Bioinformatics*, **23**(1).

Wang, Y. *et al.* (2024). Deep learning for predicting synergistic drug combinations: State-of-the-arts and future directions. *Clinical and Translational Discovery*, **4**(3).

Wang, Z. *et al.* (2023). Deml: drug synergy and interaction prediction using ensemble-based multi-task learning. *Molecules*, **28**(2), 844.

Xu, M. *et al.* (2023). Dffndds: prediction of synergistic drug combinations with dual feature fusion networks. *Journal of Cheminformatics*, **15**(1).

Zech, J. R. *et al.* (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, **15**(11), e1002683.

Zhang, P. and Tu, S. (2023). Mgae-dc: Predicting the synergistic effects of drug combinations through multi-channel graph autoencoders. *PLoS computational biology*, **19**(3), e1010951.