scientific reports

OPEN

Augmented drug combination dataset to improve the performance of machine learning models predicting synergistic anticancer effects

Mengmeng Liu[1,4], Gopal Srivastava[2,4], J. Ramanujam[1,3] & Michal Brylinski[2,3]✉

Combination therapy has gained popularity in cancer treatment as it enhances the treatment efficacy and overcomes drug resistance. Although machine learning (ML) techniques have become an indispensable tool for discovering new drug combinations, the data on drug combination therapy currently available may be insufficient to build high-precision models. We developed a data augmentation protocol to unbiasedly scale up the existing anti-cancer drug synergy dataset. Using a new drug similarity metric, we augmented the synergy data by substituting a compound in a drug combination instance with another molecule that exhibits highly similar pharmacological effects. Using this protocol, we were able to upscale the AZ-DREAM Challenges dataset from 8798 to 6,016,697 drug combinations. Comprehensive performance evaluations show that ML models trained on the augmented data consistently achieve higher accuracy than those trained solely on the original dataset. Our data augmentation protocol provides a systematic and unbiased approach to generating more diverse and larger-scale drug combination datasets, enabling the development of more precise and effective ML models. The protocol presented in this study could serve as a foundation for future research aimed at discovering novel and effective drug combinations for cancer treatment.

Developing effective anticancer therapies is an important yet challenging task. Most currently available treatments employ a monotherapy, i.e., using a single drug to treat a particular disease[1,2]. Although widely used, monotherapies are known to suffer from certain problems, such as the acquired drug resistance and prominent side effects[1,3]. In contrast, combination therapies utilizing multiple pharmaceuticals to simultaneously target several biological processes generally have greater chances of overcoming these issues[4]. Not surprisingly, combination therapies against complex diseases, such as cancer, are attracting a significant attention. Nonetheless, exploring all possible drug combinations within a vast pharmacological space is a major obstacle to find those drug combinations exhibiting synergistic effects. Accurate computational methods to select the most promising therapeutic candidates for experimental testing can greatly facilitate the discovery of effective drug combinations.

Approaches utilizing machine learning (ML) are well suited to predict drug synergistic effects. Supervised learning techniques require large-scale experimental data to train models predicting effective drug combinations. These datasets differ with respect to the number of drugs and cell lines. For instance, A Large Matrix of Antineoplastic Agent Combinations from the National Cancer Institute (NCI-ALMANAC) contains 5232 drug pairs tested against 60 cancer cell lines[5]. Another resource provides drug responses measured for a panel of 39 cancer cell lines and 22 experimental drugs in all possible pairwise combinations and in combination with 16 approved drugs, totaling 583 compound pairs[6]. Other datasets are focused on a specific cell line, for example, 1833 bioactive drugs at 5 μM were tested in combination with temozolomide at 400 μM against a human glioblastoma cell line T98G[N7]. Furthermore, 1327 drug combinations from the CeMM library of unique drugs (CLOUD) dataset containing 308 prodrugs and active drugs[8] were found effective against a human chronic myeloid leukemia cell line KBM-7[9].

[1]Division of Electrical and Computer Engineering, Louisiana State University, Baton Rouge, LA 70803, USA. [2]Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA. [3]Center for Computation and Technology, Louisiana State University, Baton Rouge, LA 70803, USA. [4]These authors contributed equally: Mengmeng Liu and Gopal Srivastava. ✉email: michal@brylinski.org

Meta-datasets collect and standardize the results of individual drug combination screening studies in order to enable a more efficient utilization of these data resources. For instance, DrugComb is an open-access data portal to 739,964 combinations of 8397 drugs tested on 2320 cell lines from 33 tissues[10,11]. It quantifies the degree of drug-drug interactions over the full dose–response matrix with several synergy scores, Bliss independence (BLISS), Highest single agent (HSA), Loewe additivity (LOEWE), and Zero interaction potency (ZIP)[12–14]. SYNERGxDB is a comprehensive dataset compiled from nine individual datasets containing 22,507 pairwise combinations of 1977 drugs tested on 151 cell lines from 15 tissues[15]. Similar to DrugComb, SYNERGxDB also provides standardized synergy scores, BLISS and ZIP. Finally, Dialog for Reverse Engineering Assessments and Methods (DREAM) Challenges partnered with AstraZeneca and the Sanger Institute to compile a dataset of 20,483 synergy scores for 910 drug combinations involving 118 anticancer drugs tested against 85 cancer cell lines[16]. This dataset also provides a quality assessment score for each combination, ranging from − 3 to 1, where 1 indicates a synergy between drugs in the combination. Along with the synergy data for drug combinations, the AZ-DREAM Challenges data comprise various molecular data, such as mutations, copy number variation, gene expression, and the tissue of origin. These datasets offer unparalleled opportunities to develop highly accurate ML models to predict drug synergistic effects.

Since the performance of supervised ML strongly depends on the quality, quantity, and the contextual subject of training data, the data scarcity problem is one of the most common challenges to develop robust ML models. To overcome this difficulty, data augmentation techniques are widely employed to expand the volume of available data. For instance, classical augmentation methods, such as image flipping, image rotation, noise injection, kernel filters, random erasing, and image mixing, are frequently used in the medical image analysis domain[17–22]. Data augmentation techniques gaining attention in the medical time series analysis domain[23] include the time domain augmentation[24], the time–frequency domain augmentation[25], decomposition-based methods[26,27], statistical generative models[28,29], and learning-based methods[30–33]. In addition, more advanced deep learning-based augmentation techniques, including the feature space augmentation[34,35], generative adversarial networks (GAN)-based augmentation[36–40], the neural style transfer[41,42], and meta-learning schemes[43–45], have been proposed.

To combat overfitting in a neural network architecture with 60 million parameters for image recognition, two types of data augmentation were employed, label-preserving transformations and altering the intensities of the RGB channels in training images using Principal Component Analysis[46]. Indeed, these data augmentation techniques significantly reduced overfitting and improved performance, leading to the reduction in the top-1 error rate by more than 1%. CutMix is an interesting augmentation technique that combines regions from different images to create augmented samples[47]. CutMix improves model generalization by encouraging localization, providing diverse training examples, and enhancing model robustness against input corruption, as well as out-of-distribution detection performances. Augmenting training data with bilingual lexicon information was demonstrated to improve the performance of machine translation models on low-resource and unsupervised languages[48]. Three main types of lexical augmentation employed are codeswitching, lexical prompting, and raw token-pair training. Extensive experimentation results show that applying any of these augmentations to monolingual data yields substantial improvements, and that they can be combined for even greater effect.

Although image, language, and sequential data augmentation methods are well established, these approaches are, in principle, unsuitable to generate the heterogeneous data of cellular and molecular features for drug synergy prediction with supervised ML. On that account, a variety of domain-specific techniques have been developed. For instance, the fact that multiple simplified molecular-input line-entry system (SMILES) strings represent the same molecule was used to augment a molecular dataset of chemical species[49] using the SMILES enumeration[50]. Further, data augmentation utilizing multiple SMILES representations for a single compound was demonstrated to enhance the prediction accuracy of various molecular properties, such as solubility, lipophilicity, and bioactivity, irrespective of the specific machine learning model employed or the size of the dataset[51]. Another study doubled the size of a training dataset to predict anticancer drug synergism based on NCI-ALMANAC by generating duplicates with the reverse order of drugs[52]. Data up-sampling was also applied to increase the number of minor class instances for phenotype-based virtual screening of anticancer drug combinations[53]. Finally, an example of a deep learning-based data augmentation technique is the uniform graph convolutional network (UGCN)[54]. It employs a drug representation based on atomic interactions within organic compounds rather than hand-crafted features, such as molecular fingerprints, and string-based features, such as SMILES. UGCN can be used to augment chemical data by randomly sampling multiple complementary graphs for a single drug.

Despite the encouraging results reported for the abovementioned data augmentation techniques for drug synergy prediction, many of existing methods either are too general (up-sampling) or consider only drug structural information (SMILES enumeration and UGCN). To address these issues, we devised a new augmentation approach combining the drug chemical similarity with the system-level information on drug-target interactions. This approach employs a novel similarity metric, the drug action/chemical similarity (DACS) score, taking into account not only the chemical characteristics of drugs, but also their molecular targets. Applying the DACS score to augment the AZ-DREAM Challenges data with new compounds from PubChem[55] significantly increased the size and diversity of the training dataset for drug synergy prediction. To the best of our knowledge, this methodology represents the first systematic and effective protocol to augment a synergy dataset simultaneously utilizing the information on drug chemical structures and their protein targets. As a proof of concept, the augmented dataset was used to train several ML models demonstrating a higher accuracy of drug synergy prediction compared to those models trained on the original AZ-DREAM Challenges data.

Results

Similarity measure for cellular responses to drug treatment

During the data augmentation, new drug combinations are generated by replacing drugs with those molecules triggering similar pharmacological responses. The similarity of pharmacological effects of two drugs is quantified by the Kendall τ correlation coefficient between pIC_{50} values for the monotherapy treatments of multiple cancer cell lines. A positive value of Kendall τ indicates that two drugs have similar pharmacological effects in terms of the inhibition of the cancer growth, whereas a negative correlation and the lack of correlation point to different cellular responses to drug treatment. This concept is illustrated in Fig. 1 for crizotinib, a tyrosine kinase inhibitor used for the treatment of non-small cell lung carcinoma (NSCLC)[56], paired with six other anti-cancer drugs. Figures 1A–C are examples of a positive correlation between crizotinib and everolimus (Kendall τ of 0.50), entinostat (Kendall τ of 0.44), and perifosine (Kendall τ of 0.42), respectively. Everolimus, a derivative of sirolimus with cell proliferation and immunosuppressive properties, is used in combination with other anticancer agents for the treatment of kidney and breast cancer, and neuroendocrine tumors of gastrointestinal and lung origins[57]. Entinostat, a benzamide derivative with the antineoplastic activity, and perifosine, an allosteric AKT inhibitor with the antiglycolytic activity, are used for the treatment of NSCLC[58,59]. According to the analysis of pIC_{50} values against multiple cancer cell lines, these three drugs have similar profiles to that of crizotinib, i.e., they inhibit the growth of the same cancer cell lines and are ineffective against the same group of cell lines as well.

In contrast, cellular responses of crizotinib are uncorrelated with that of adavosertib (Fig. 1D, Kendall τ of − 0.06), vinorelbine (Fig. 1E, Kendall τ of − 0.03), and capivasertib (Fig. 1F, Kendall τ of − 0.01). Adavosertib is a tyrosine kinase WEE1 inhibitor used to improve the outcome in triple-negative breast cancer[60], vinorelbine is an agent to treat NSCLC and breast cancer[61], and capivasertib is AKT inhibitor used in the treatment of breast cancer[62]. Since these drugs have uncorrelated pharmacological effects, they cannot be used to replace crizotinib during the data augmentation process. The analysis of cellular responses with the Kendall τ is versatile and can be applied when two drugs have been tested on at least two common cell lines, otherwise the value of the Kendall τ is set to 0. The similarities of pharmacological effects between crizotinib and everolimus, entinostat, perifosine, adavosertib, vinorelbine, and capivasertib were calculated based on 7 + 2, 9 + 0, 7 + 0, 9 + 0, 0 + 13, and 9 + 10 common (breast + lung) cell lines, respectively.

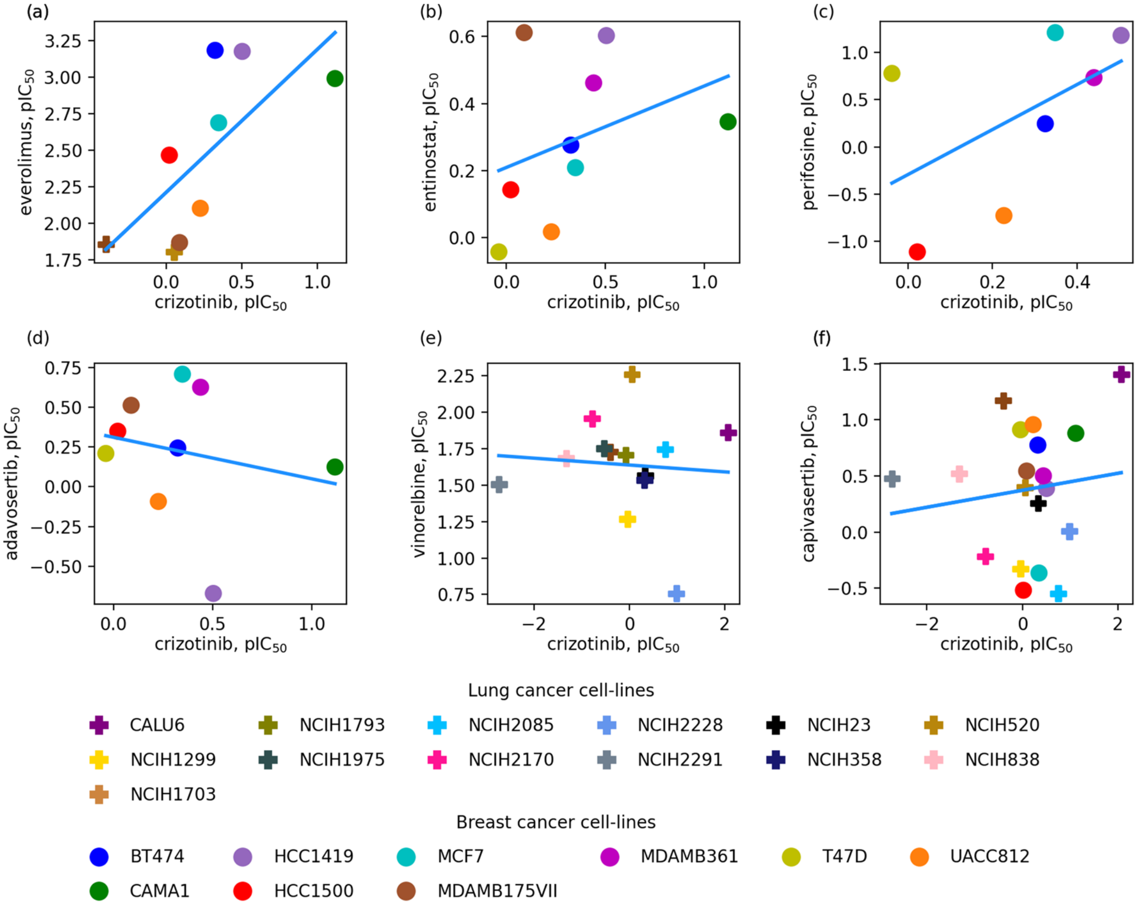

Figure 1. Similarity of pharmacological effects of two drugs quantified by the Kendall τ correlation coefficient. pIC_{50} values for the monotherapy treatments of multiple cancer cell lines with crizotinib are plotted against those for (**A**) everolimus, (**B**) entinostat, (**C**) perifosine, (**D**) adavosertib, (**E**) vinorelbine, and (**F**) capivasertib. (**A**, **B**, and **C)** are examples of the positive correlation, whereas (**D**, **E**, and **F)** represent the negative correlation. Individual breast cancer cell lines are shown as solid circles and lung cancer cell lines as solid plus signs.

Relation between drug similarity and pharmacological effects

Next, we investigate how similar two drugs need to be in order to trigger similar pharmacological effects. This analysis is performed for 4753 (98C2) possible pairs of 98 drugs in the AZ-DREAM Challenges dataset. Pharmacological responses are quantified with the Kendall τ correlation coefficient, whereas the drug similarity is measured with two metrics. The first score is the drug chemical similarity calculated as the Tanimoto coefficient (TC) between FP2 fingerprints[63]. Figure 2 (solid blue line) shows that, as expected, the fraction of drug pairs with the positive Kendall τ increases with the increasing chemical similarity and reaches a value of 1.0 for the TC threshold of 0.6. The second metric is the drug action similarity computed as the Matthews correlation coefficient (MCC)[64] between target proteins in the protein–protein interaction (PPI) network from the IHP-PING dataset[65]. Similar to the TC, the fraction of drug pairs with the positive Kendall τ also increases with the increasing MCC reaching 1.0 for the MCC threshold of 0.6 (Fig. 2, dashed purple line). For comparison, increasing the threshold for a random similarity does not increase the fraction of drug pairs with the positive Kendall τ (Fig. 2, dotted black line).

Drug action/chemical similarity score

Analyses presented above demonstrate that both chemical and drug action similarities can be used for data augmentation. However, their combination could potentially cover a larger chemical space than individual similarities while ensuring that the pharmacological profiles of drugs selected for augmentation are highly similar to those of their parent molecules. Therefore, we combined TC and MCC into a new metric, the drug action/chemical similarity (DACS) score. Figure 3 shows the relation between the DACS score and the fraction of drug pairs with the positive Kendall τ as the spatial heatmap in two dimensions corresponding to the individual similarities. The dark blue section in the upper left corner of the heatmap corresponds to the area of a low positive correlation, whereas the light blue section shows the combination of individual similarities resulting in a high positive correlation. The DACS score can be represented as a quarter circle in Fig. 3 (dashed black line). For example, above a DACS threshold of 0.6, as many as 85.7% drug pairs have a positive Kendall τ correlation.

Dataset augmentation with DACS

The DACS metric is used as a guide to find the optimal number of new instances to be generated for the synergy dataset according to a procedure presented in Fig. 4. Each instance in the AZ-DREAM Challenges dataset consists of a pair of drugs targeting a cell line with a particular synergy score (Fig. 4A, drug pair *1*:*2*). During the augmentation procedure, candidate molecules to replace one drug in a pair are identified in the STITCH database[66] (Fig. 4B, drugs *3*, *4*, and *5*). Next, DACS scores against the drug to be replaced are calculated (Fig. 4C) and those molecules having scores larger than a cutoff are selected (Fig. 4D, drugs *3* and *5*). The original drug is then replaced by the selected molecules to create augmented pairs (Fig. 4E, drug pairs *3*:*2* and *5*:*2*). This procedure is repeated for the second drug in the original pair creating more augmented instances (Fig. 4F, drug pairs *1*:*6*).

The selection of a cutoff for DACS scores between the original drug to be replaced and the candidate substitute compounds is critical to create high-quality augmented instances. On that account, we conducted an analysis of the fraction of new drugs having similar pharmacological profiles to their parent molecules and the number of new instances that can be obtained from the STITCH database at different DACS similarity thresholds. Figure 5 shows that these two quantities are inversely related, i.e., increasing the DACS similarity threshold results in a higher chance of substitute compounds to trigger similar pharmacological responses (dashed purple line), however, at the same time, fewer molecules can be used to augment the dataset (solid blue line). The intersection point

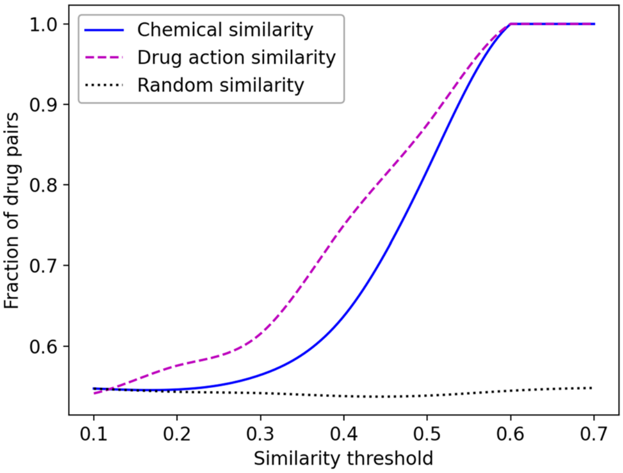

Figure 2. Fraction of drug pairs with positively correlated pharmacological effects as a function of their similarities. The chemical similarity (solid blue line) is measured with the Tanimoto coefficient between drug FP2 fingerprints. The drug action similarity (dashed purple line) is quantified with the Matthews correlation coefficient between target proteins in the IHP-PING protein–protein interaction network. Random similarity (dotted black line) is obtained by assigning a random number between 0 and 1.

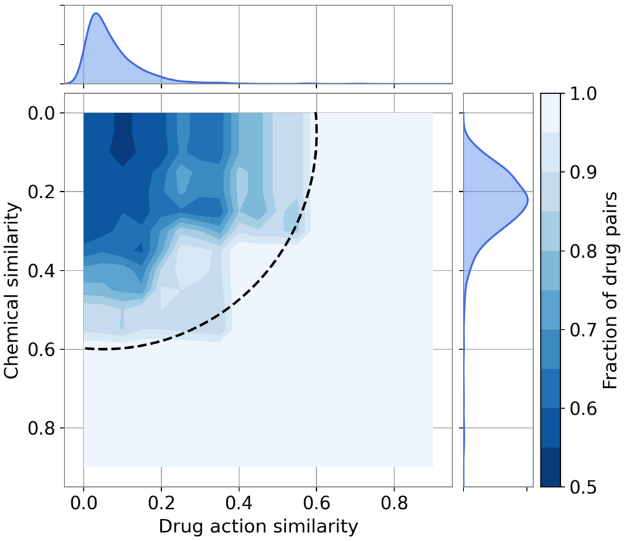

Figure 3. Heatmap of the fraction of drug pairs with positively correlated pharmacological effects. The fraction of drug pairs with the positive Kendall τ is displayed according to the color scale on the right. One-dimensional histograms show the distributions of the chemical similarity (a subplot on the right) and the drug action similarity (a subplot on the top). The dashed quarter circle represents a DACS threshold of 0.6.

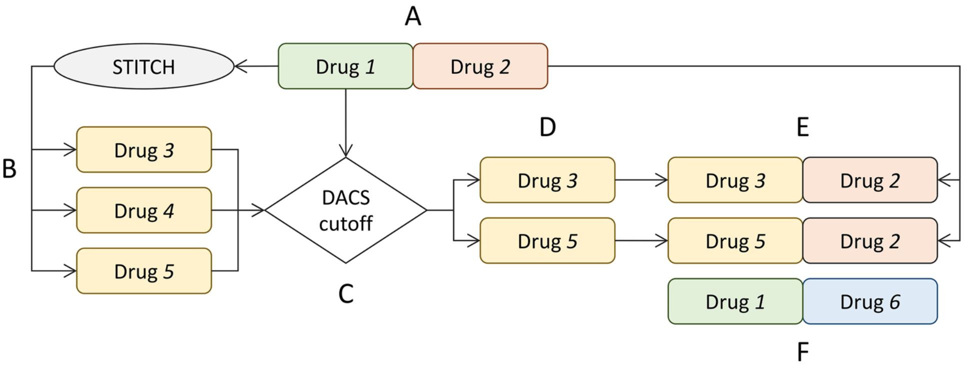

Figure 4. Flowchart of the augmentation procedure. The procedure starts with an original drug pair *1*:*2*, in which drug *1*, represented by a green rounded box, is to be replaced first (**A**). Candidate molecules *3*, *4*, and *5*, represented by yellow boxes, are selected from the STITCH database (**B**). DACS scores for compounds *3*, *4*, and *5* are calculated against drug *1* (**C**) and those molecules with scores larger than a cutoff are retained (**D**). These compounds are then combined with drug *2* creating augmented instances *3*:*2* and *5*:*2* (**E**). The same procedure is then applied to replace drug *2*, represented by an orange box. This generates more augmented instances containing drug *1*, such as an augmented pair *1*:*6*, in which molecule *6*, represented by a cyan box, is a substitute for drug *2* (**F**). The class of augmented instances (either synergistic or antagonistic) is transferred from the original drug pair *1*:*2*.

marked by a dotted black line in Fig. 5 represents the DACS cutoff of 0.53, at which the majority of substitute drugs (82%) have similar pharmacological profiles to their parent molecules and as many as 42,225 new drugs can be obtained from the STITCH database to augment the synergy dataset. Applying this threshold to replace one molecule in a drug pair in the AZ-DREAM Challenges dataset of 8798 instances produces an augmented dataset of 6,016,697 drug pairs annotated with synergy scores against various cancer cell lines.

Ideally, the distribution of synergy values across the augmented dataset should be the same as for the AZ-DREAM Challenges dataset. Figure 6 shows that these two distributions indeed are similar; the average synergy score ± standard deviation is 9.9 ± 26.1 for the AZ-DREAM Challenges dataset and 12.1 ± 28.5 for the augmented dataset. In addition, we compare various physicochemical properties of drugs present in the original and augmented dataset to those calculated for a set of 27,385 molecules selected randomly from the STITCH database[66]. Indeed, the original and augmented drugs have similar octanol–water partition coefficient (log*P*, 3.6 ± 2.0 and 3.8 ± 1.8), the number of hydrogen bond donors (HBD, 2.0 ± 1.2 and 2.0 ± 1.6) and acceptors (HBA, 6.8 ± 2.6 and 5.8 ± 2.4), and the Quantitative Estimate of Druglikeness[67] (QED, 0.48 ± 0.18 and 0.49 ± 0.20). For comparison, log*P*, HBD, HBA, and QED for random molecules are 3.2 ± 2.4, 1.9 ± 1.9, 5.0 ± 2.7, and 0.50 ± 0.22, respectively. These analyses demonstrate that the augmented dataset does not contain artifacts, such as molecules with certain

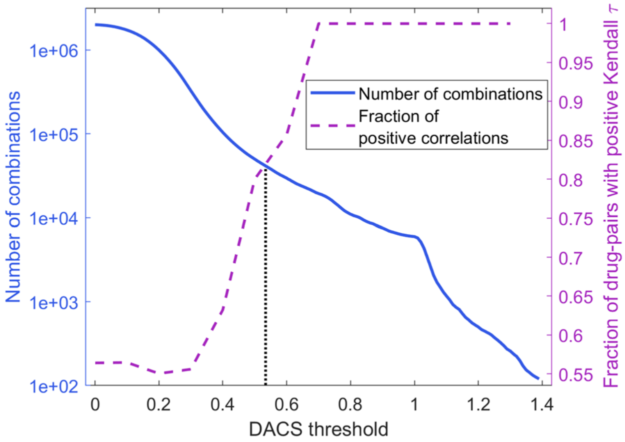

Figure 5. Selection of the optimal DACS threshold for data augmentation. The solid blue curve represents the number of potential substitutes for the original 98 drugs that can be found in the STITCH database as the DACS threshold is increased. The dashed purple line represents the change in the fraction of drug pairs with the positive Kendall τ as the DACS threshold is increased. The vertical dotted line marks the DACS threshold optimizing these two quantities.

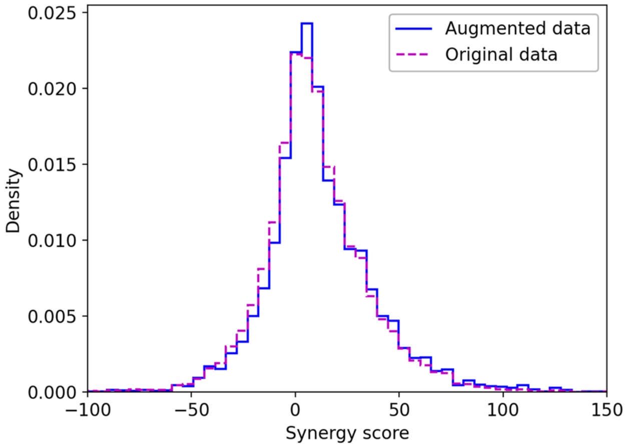

Figure 6. Distribution of synergy score across drug synergy datasets. The step histogram in purple dashed line shows the distribution of synergy scores in the original AZ-DREAM Challenges data, whereas the step histogram in blue solid line shows the distribution of synergy scores in the augmented dataset.

physicochemical properties, that could potentially bias the training of machine learning models toward a particular effect (either synergism or antagonism).

Drug synergy prediction with machine learning

Finally, we investigate whether training machine learning against the augmented data achieves a better classification performance than training against the original AZ-DREAM Challenges dataset. Four state-of-the-art machine learning methods are employed, Logistic Regression (LR)[68,69], Support Vector Machines (SVM)[70,71], Random Forest (RF)[72], and Gradient Boosting Trees (GBT)[73]. Following the original publication[16], drug pairs having synergy scores higher than 20 are labelled synergistic and those having synergy scores lower than − 20 are labelled antagonistic. First, we performed a fivefold cross-validation by randomly splitting the dataset into 5 subsets. Note that the augmented data are only used to train machine learning models, which are then validated against AZ-DREAM Challenges instances. Table 1 shows the classification performance evaluated with several metrics. Encouragingly, the performance of classifiers is improved when models are trained against the augmented data and the random-split validation is employed. For instance, the area under the receiver operating characteristic plot (AUC) increased from 0.802 to 0.809 for RF and from 0.859 to 0.863 for GBT classifiers.

Although a random-split cross-validation is often used to assess the performance of drug synergy predictors [16] , it leads to a significant overlap between training and validation subsets because those instances involving similar cell lines are present in both sets. Consequently, the trained model is going to have only a weak ability to generalize to unseen data, even though the validation accuracy may seem high. In order to mitigate this issue and more reliably evaluate the performance of machine learning trained on drug synergy data, we conducted a

Classifier	Validation protocol	Dataset	ACC	TPR	FPR	PPV	AUC	MCC	F1-score
LR	Random-split	Original	0.752	0.768	0.302	0.893	0.809	0.417	0.826
		Augmented	0.756	0.770	0.292	0.897	0.811	0.427	0.829
	Tissue-based	Original	0.631	0.718	0.509	0.769	0.659	0.200	0.728
		Augmented	0.637	0.711	0.475	0.777	0.661	0.221	0.729
SVM	Random-split	Original	0.745	0.755	0.287	0.897	0.803	0.414	0.819
		Augmented	0.750	0.757	0.271	0.902	0.803	0.429	0.823
	Tissue-based	Original	0.619	0.736	0.546	0.761	0.671	0.191	0.721
		Augmented	0.651	0.771	0.560	0.765	0.674	0.208	0.751
RF	Random-split	Original	0.754	0.787	0.358	0.879	0.802	0.392	0.831
		Augmented	0.757	0.788	0.347	0.882	0.809	0.402	0.832
	Tissue-based	Original	0.667	0.811	0.644	0.749	0.647	0.173	0.769
		Augmented	0.705	0.866	0.659	0.758	0.685	0.226	0.801
GBT	Random-split	Original	0.833	0.921	0.457	0.869	0.859	0.503	0.894
		Augmented	0.840	0.927	0.445	0.873	0.863	0.524	0.899
	Tissue-based	Original	0.716	0.930	0.803	0.734	0.688	0.176	0.815
		Augmented	0.736	0.940	0.743	0.750	0.734	0.260	0.828

Table 1. Performance of machine learning in the prediction of drug synergistic effects. *ACC* accuracy, *TPR* recall, *FPR* false positive rate, *PPV* precision, *AUC* area under the receiver operating characteristic plot, *MCC* Matthews correlation coefficient. Two protocols are employed utilizing the random-split of the data and the tissue-based cross-validation. The performance of Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Trees (GBT) classifiers is evaluated against the original AZ-DREAM Challenges data and the augmented dataset.

tissue-based cross-validation in which each fold comprises a particular tissue (or a group of tissues). This protocol has been shown to eliminate the overlap between training and validation subsets allowing for an unbiased assessment of the capabilities of machine learning to extract the information from input data [74].

Table 1 and receiver operating characteristic plots presented in Fig. 7 show that applying the more rigorous tissue-based validation protocol decreases the performance of machine learning predicting drug synergistic effects. However, this evaluation is more reliable because it better mimics a real scenario in which machine learning is applied to predict drug synergistic effects for unseen data, i.e., drug combinations against cell lines originating from tissues that have not been used to train the classifier. With this cross-validation protocol, machine learning trained on the augmented data yields even higher improvements in terms of the classification accuracy compared to models trained on the original AZ-DREAM Challenges dataset. For example, the AUC increased from 0.647 to 0.685 for RF and from 0.688 to 0.734 for GBT classifiers.

Table 2 shows AUC scores for each tissue fold and tree-based models trained on both the original and the augmented datasets. The comparison of AUC scores reveals that incorporating the augmented data into the training process systematically improves the classification performance regardless of the tissue type. In general, these findings indicate that incorporating augmented data can provide enhanced information for training machine learning models in a more effective manner.

Classification of instances with ambiguous synergy scores

The robustness of ML models stems from the foundation laid by the quality of the training data, ensuring that they can effectively handle diverse and complex scenarios with a high degree of accuracy. When a machine learning model encounters instances with ambiguous labels, it adapts by making predictions that are less confident for such cases. To illustrate this phenomenon, we evaluate the capability of the trained GBT model to handle instances with unclear class labels by assessing its performance across a spectrum of synergy scores. The GBT model was selected because its performance in fivefold cross-validation against instances with reliable synergy scores ≥ 20 (synergistic cases) and ≤ -20 (antagonistic cases) is better than those of LR, SVM, and RF. Figure 8 shows the distribution of prediction probabilities reported by the GBT model for drug combinations selected from the AZ-DREAM Challenges dataset with a varying degree of synergy scores with the corresponding statistics reported in Table 3.

Including ambiguous labels represented by synergy scores close to 0 lowers the confidence, and the model attempts to reflect this uncertainty in its predictions. For instance, Fig. 8A shows that the median (Q_2) prediction probability is 0.981 when the most ambiguous positive cases with the synergy score of > 0 are included, while it is as high as 0.999 when the model is applied to only the most reliable positive cases with the synergy score of ≥ 20. This trend can also be observed for negative instances (Fig. 8B), for which the median prediction probability increases from 0.248 for the most ambiguous cases with the synergy score of < 0 to 0.687 for the most reliable cases with the synergy score of ≤ -20. Another indication of the lack of strong prediction confidence when instances having unclear labels are included is the increased spread of prediction probabilities. Indeed, wider interquartile ranges (Q_3-Q_1) are observed when ambiguous positive cases are considered compared to those obtained for the most reliable drug combinations only. For negative cases, Q_2 and Q_3 values decrease as

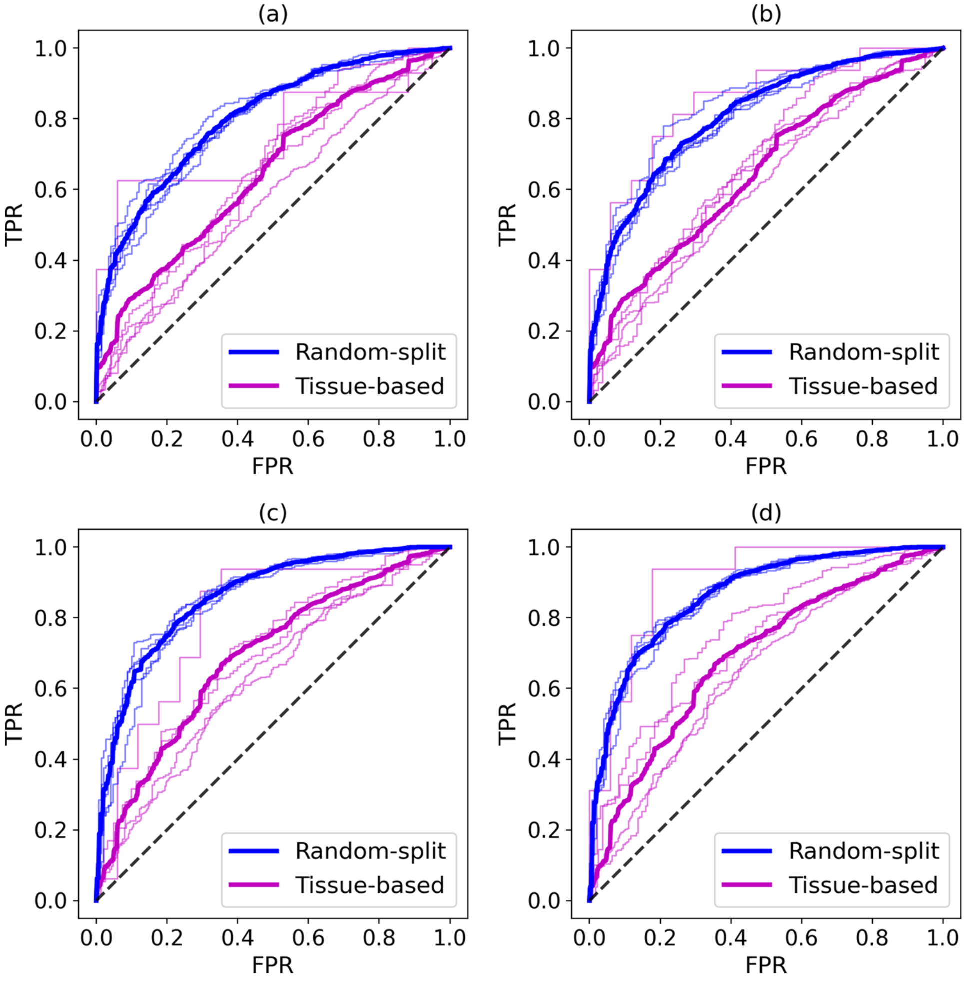

Figure 7. Performance of machine learning in the prediction of drug synergistic effects. Receiver operating characteristics plots for the Random Forest classifier against (**A**) the original AZ-DREAM Challenges data and (**B**) the augmented dataset, and for the Gradient Boosting Trees classifier against (**C**) the original AZ-DREAM Challenges data and (**D**) the augmented dataset. Blue lines were calculated for the random-split protocol, while purple lines were obtained for the tissue-based cross-validation. Thick lines show the mean performance averaged over individual folds represented by thin lines.

Classifier	Dataset	Breast tissue	Digestive system	Excretory system	Respiratory system	Other
RF	Original	0.574	0.628	0.650	0.636	0.746
	Augmented	0.613	0.640	0.664	0.658	0.849
GBT	Original	0.637	0.675	0.631	0.704	0.794
	Augmented	0.648	0.715	0.649	0.752	0.904

Table 2. Area under the receiver operating characteristic plot (AUC) scores for each fold in the tissue-based cross-validation. The performance of Random Forest (RF) and Gradient Boosting Trees (GBT) classifiers is reported for the original and the augmented AZ-DREAM Challenges datasets.

more unclear instances are included, meaning there is a concentration of prediction probability towards the lower values, which signifies the declined prediction confidence for those instances and a diminished level of assurance in the ability to assign accurate classifications by the model.

Evaluation against "unseen" data

To further evaluate the generalizability of a model trained on the AZ-DREAM Challenges augmented data, we conducted the performance evaluation against an independent dataset of 250 drug combinations selected from DrugCombDB[75]. It is important to note that since drugs in this set are chemically dissimilar to those in the

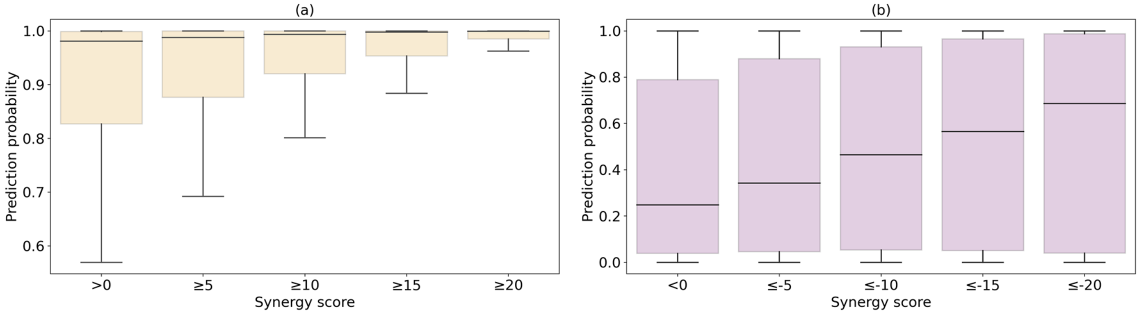

Figure 8. Distribution of prediction probabilities across varying degrees of drug synergy. Predictions are generated by the Gradient Boosting Trees classifier for (**A**, light yellow) positive instances with an increasing threshold for the synergy scores from > 0 to ≥ 20 and (**B**, light purple) negative instances with a decreasing threshold for the synergy scores from < 0 to ≤ -20. Boxes end at quartiles Q_1 and Q_3, Q_2 is the median. Whiskers extend from Q_1 and Q_3 to the most extreme data points within $Q_1 - 1.5 \times IQR$ and $Q_3 + 1.5 \times IQR$, respectively, where IQR is the inter-quartile range.

Quartile	Synergy score (positives)					Synergy score (negatives)				
	> 0	≥ 5	≥ 10	≥ 15	≥ 20	< 0	≤ -5	≤ -10	≤ -15	≤ -20
Q_1	0.827	0.877	0.921	0.954	0.985	0.039	0.046	0.053	0.050	0.041
Q_2	0.981	0.987	0.994	0.998	0.999	0.248	0.342	0.466	0.564	0.687
Q_3	0.999	0.999	1.000	1.000	1.000	0.789	0.879	0.931	0.965	0.989

Table 3. Statistics for the distribution of prediction probabilities across varying degrees of drug synergy. Predictions are generated by the Gradient Boosting Trees classifier for positive instances with an increasing threshold for the synergy scores from > 0 to ≥ 20 and negative instances with a decreasing threshold for the synergy scores from < 0 to ≤ -20. Q_2 is the median.

AZ-DREAM Challenges dataset, DrugCombDB instances can be regarded as "unseen" data. In this analysis, two GBT models were trained, one using the original AZ-DREAM Challenges data and the other using both the original and augmented instances. A GBT model trained solely on the original data correctly classified only 76/250 drug combinations (12 synergistic and 64 antagonistic) yielding the accuracy of 0.30 and a high false positive rate (FPR) of 0.73. In contrast, a GBT model that incorporated augmented data during training correctly predicted 141/250 drug combinations (11 synergistic and 130 antagonistic) achieving a much higher accuracy of 0.56 and a significantly lower FPR of 0.45. This improved performance by employing augmented instances highlights the importance of data augmentation techniques in enhancing the ability of machine learning models to generalize to new drug synergy data. Through exposure to a comprehensive and diverse dataset, the model acquired improved pattern recognition capabilities and achieved more accurate classifications, resulting in an enhanced reliability for drug synergy predictions in a real-world application scenario.

Discussion

In this study, we devised a data augmentation protocol to solve the data scarcity problem in predicting synergistic effects of anti-cancer drug combinations with machine learning models. The augmentation protocol expands the synergy dataset by replacing a compound in a drug combination instance with another molecule having highly similar pharmacological effects. This is achieved through the use of the DACS similarity metric between two drugs, which incorporates both chemical structure and drug action similarities. Compared to existing techniques used in synergy data augmentation, such as the upsampling[53], the SMILES enumeration[50], and the reverse order of drugs[52], which essentially duplicate the existing data points, our approach expands the dataset by including new, unbiased instances. As a results, this augmentation methodology not only enriches the available data points, but also enhances the diversity of the data, which is highly beneficial to improve the generalizability of machine learning models. Additionally, in contrast to other augmentation approaches involving a learning process[54], our method generates data points in a shorter amount of time.

While random-split cross-validation is frequently utilized for data partitioning, it may lead to tissue-level overlap and elevate the possibility of model overfitting, particularly when dealing with data containing multiple cell lines from the same tissue. The reason for this is that those instances involving similar cell lines tend to have comparable feature representations, such as gene expression profiles and the gene-disease association. The overlap is likely going to occur when these instances are present in both the training and validation sets[76]. In such cases, the trained model may exhibit a strong performance due to the presence of overlapping data, but it will not perform well on novel, unseen data. Consequently, the model may be overestimated in terms of its true performance and fail to generalize to other datasets. On the other hand, a tissue-based cross-validation can effectively eliminate the data overlap issue. By excluding all instances originating from a validation tissue from the training set for each fold, the generalizability of a machine learning model can be properly evaluated.

Tree-based models (RF and GBT) employed in this study are robust, interpretable, and widely adopted by AZ-DREAM Challenges participants[16]. These models have the ability to deal with complex non-linear input–output relationships and can handle sizable datasets to a certain degree. Neither tree-based models nor other classifiers like LR and SVM are designed to exploit intricate relationships between features. This limitation is especially notable when dealing with heterogeneous features, including protein–protein interactions, gene expression levels, and drug-protein associations. In such cases, these models may struggle to find the optimal decision boundaries, generally leading to an unsatisfactory performance. Neural networks, on the other hand, are better equipped to handle diverse data types and can learn complex relationships between features with hidden layers and non-linear activation functions. This ability to integrate multiple heterogeneous data into a single model can often result in an improved performance compared to tree-based models. Our future research will concentrate on exploring this aspect.

The augmentation protocol devised in this study is not limited to anti-cancer drug data can be used to expand other synergy datasets as well; it has the potential to become a universal tactic for enhancing datasets in drug discovery and related fields. This could result in a greater amount of data being accessible and ultimately lead to better research results. Furthermore, the developed new drug similarity measure, the DACS score, improves the way drug similarity is assessed. By integrating both structural and target similarities, DACS provides a more exhaustive and inclusive perspective on drug similarity compared to traditional methods that only examine a single aspect, such as the chemical similarity. By offering a more holistic approach to analyzing and evaluating the similarities between drugs, DACS can help improve the accuracy and efficiency of the drug discovery process.

Deep learning, with its ability to dissect complex data and reveal underlying patterns and relationships, has become a pivotal tool in the field of pharmacology and drug development[77,78]. The varied and comprehensive synergy dataset created in this study has the potential to significantly aid deep learning models by offering a diverse range of data for training purposes. The utilization of sufficient data enables deep learning algorithms to recognize intricate relationships and connections among cellular, molecular, and biological system-level features, thereby elevating the precision and efficacy of synergistic effect predictions. Moreover, an extensive and varied dataset reduces the risk of overfitting, a common issue where models become too reliant on limited training data and struggle to generalize to new data. Thus, the utilization of a comprehensive synergy dataset can lead to more robust and dependable deep learning models and ultimately, more advanced outcomes in drug discovery and related fields.

In addition to being used in deep learning-based drug discovery, the proposed anti-cancer drug synergy dataset has the potential to facilitate other applications, such as drug repositioning, drug target identification, toxicity analysis, the modeling of drug interactions, systems pharmacology, and precision medicine. By providing valuable insights into the interactions between drugs, targets, and biological systems, the synergy data can contribute to the development of more effective and safer pharmaceutics. Overall, the wide-ranging possibilities arising from this study may have significant implications for the drug discovery and development field. Ultimately, this could result in the creation of novel therapeutic approaches for a range of diseases.

Methods

Similarity of drug pharmacological effects

The Kendall τ rank correlation coefficient is employed to measure the ordinal association between the pharmacological effects of two drugs against a set of cell lines. First, common cell lines targeted by both drugs are identified and two lists ranked by pIC_{50} values for monotherapy treatments are calculated. Next, the value of the Kendall τ accounting for ties (τ_b)[79,80] is computed:

$$\tau_b = \frac{n_c - n_d}{\sqrt{(n_c + n_d + n_1)(n_c + n_d + n_2)}} \tag{1}$$

where n_c is the number of concordant cell line pairs (having the same order in both drug lists), n_d is the number of discordant cell line pairs (having different order in both drug lists), n_1 is the number of pairs tied only in the first list, and n_2 is the number of pairs tied only in the second list. τ_b of $+1$ indicates a perfectly positive association, i.e., the two drugs having the same pharmacological effects in terms of the inhibition of the cancer growth across multiple common cell lines. A value of -1 indicates a perfectly negative association, i.e., the opposite pharmacological effects, and a value of 0 indicates the lack of any association. The Kendall τ coefficient is calculated when pIC_{50} values are available for monotherapy treatments of at least two common cell lines, otherwise it is set to 0.

Similarity of drug molecular mechanism of action

Similarity of the mechanism of action of two drugs is quantified with the MCC[64] computed for 19,968 proteins in the IHP-PING dataset[65] according to chemical-protein associations obtained from the STITCH database[66]:

$$MCC = \frac{(T \times N) - (A \times B)}{\sqrt{(T + A)(T + B)(N + A)(N + B)}} \tag{2}$$

where T is the number of proteins targeted by both drugs, N is the number of proteins not targeted by any drug, A is the number of proteins only targeted by the first drug, and B is the number of proteins only targeted by the second drug. MCC ranges from -1 to $+1$ with high positive values indicating a significant overlap between the molecular targets of two drugs, thus a similar mechanism of action. The MCC for a pair of drugs having different mechanisms of action is going to be around 0.

Drug action/chemical similarity score

The DACS measure provides a convenient and informative way to combine the drug structure similarity with the similarity of the molecular mechanisms of action. It is calculated as:

$$DACS = \sqrt{TC^2 + MCC^2} \tag{3}$$

where *TC* is the Tanimoto coefficient between drug FP2 fingerprints[63] and *MCC* is the similarity of drug mechanism of action defined in Eq. (1). When one of the component metrics, either TC or MCC, is sufficiently high, then the other metric does not need to be as high for the DACS score to be over a predefined threshold. In rare cases of negative MCC values, the MCC component of the DACS score is set to 0.

Classification datasets

Following the original paper on the AZ-DREAM Challenges dataset[16], we compiled the primary dataset by excluding those instances having ambiguous synergy scores between − 20 and 20 to create a classification dataset of 3210 drug combinations comprising 2461 synergistic (a synergy score $\geq$ 20) and 749 antagonistic (a synergy score ≤ -20) cases. The corresponding augmented dataset contains 1,850,037 synergistic and 465,288 antagonistic combinations totaling 2,315,325 labeled instances. Further, the following four datasets were constructed at varying degrees of drug synergy to evaluate the performance of ML against instances having ambiguous labels, 8817 combinations comprising 5839 synergistic (a synergy score $>$ 0) and 2978 antagonistic (a synergy score $<$ 0) cases, 6974 combinations comprising 4882 synergistic (a synergy score $\geq$ 5) and 2092 antagonistic (a synergy score ≤ -5) cases, 5408 combinations comprising 3913 synergistic (a synergy score $\geq$ 10) and 1495 antagonistic (a synergy score ≤ -10) cases, and 4180 combinations comprising 3119 synergistic (a synergy score $\geq$ 15) and 1061 antagonistic (a synergy score ≤ -15) cases.

In addition to the primary dataset, an independent validation set was created based on DrugCombDB[75]. Applying the same synergy score criteria and excluding molecules with the TC of $\geq$ 0.4 to any compound in the AZ-DREAM Challenges dataset resulted in 250 drug combinations with 14 synergistic and 236 antagonistic effects, referred to as "unseen" data.

Feature vectors

Input data for machine learning consist of drug and cell features. The former are computed with Mol2vec[81] by encoding a drug chemical structure to a 300-dimensional vector. The latter features are calculated by embedding 17,419 gene expression values for a cell line obtained from the AZ-DREAM Challenges dataset with an adversarial deconfounding autoencoder[82]. Similar to drug embeddings, the gene expression profile is encoded to a 300-dimensional vector. The final, 900-dimensional feature vector is generated by concatenating two drug feature vectors and a cell feature vector.

Cross-validation protocols

Two cross-validation procedures are employed utilizing a random and a tissue-based data split. In the random-split cross-validation, the classification dataset is randomly partitioned into five equal-size folds. In the tissue-based cross-validation, the dataset is assigned to five groups according to the tissue type of cell lines, the breast tissue, the digestive system, the excretory system, the respiratory system, and other tissues. Note that tissue types in the augmented dataset are the same as in the original dataset because the augmentation process does not affect cell lines. A fivefold cross-validation is conducted the usual way, i.e., in each round, the machine learning model is trained on the augmented data for 4 subsets and then validated against the original AZ-DREAM Challenges instances in the remaining subset. This protocol ensures that the augmented data is used only to train classifiers and the validation is performed on the original data and labels. Since the original dataset is imbalanced, comprising 76.7% synergistic and 23.3% antagonistic instances, a stratified split is used to preserve the percentage of samples for each class in each fold. When augmenting the training set, the ratio is preserved by proportionally adding instances of each class. In the tissue-based split, although the proportions of synergistic and antagonistic instances are different in each tissue, the training set is augmented in a way to preserve the ratio of synergistic/antagonistic instances in individual folds.

Machine learning

Four machine learning models are used to evaluate the performance of supervised learning algorithms on the original and the augmented datasets of drug combinations, Logistic Regression, Support Vector Machines, Random Forest, and Gradient Boosting Trees. LR is a supervised machine learning algorithm designed for binary classification tasks to predict the likelihood of an instance belonging to one of two classes (synergistic or antagonistic in our case). It employs the logistic function to transform a linear combination of input features into a probability score, allowing for intuitive interpretation[68,69]. Model training involves minimizing the logistic loss function through optimization techniques such as gradient descent. The coefficients of the linear equation are estimated during the training process to create a predictive model. The following parameters were used in the LR model: L2 penalty, the tolerance for stopping criteria of 0.0001, the inverse of regularization strength of 0.45, the maximum number of iterations of 500, and class weights set to "balanced" to deal with the imbalanced dataset.

SVM is a powerful supervised machine learning algorithm used for classification and regression tasks. In the classification context, it aims to find the optimal hyperplane in the feature space to maximize the margin between data points belonging to different classes[70,71]. SVM is effective in dealing with high-dimensional features and can handle non-linear relationships through the use of kernel functions implicitly mapping the input features into a higher-dimensional space. The following parameters were used in the SVM model: the regularization parameter of 0.42, a linear kernel type, the tolerance for stopping criterion of 0.001, a probability set to true to enable probability estimation, and class weights set to "balanced" to deal with the imbalanced dataset.

The RF classifier utilizes a collection of individual trees built independently to determine the final output by the majority vote[72]. In contrast, the GBT classifier builds trees additively to reduce the bias of the previous tree, and then combines the output of all trees scaled by the learning rate to calculate the final output[73]. Parameters of both classifiers were manually tuned to optimize their classification performance. The following parameters were used in RF: the number of trees in the forest of 300, the minimum number of samples per leaf node of 85, the number of features to consider for the best split equal to the square root of total number of features, and class weights set to: "balanced" in order to deal with the imbalanced dataset. The following parameters were used in GBT: the number of boosting stages of 650, the minimum number of samples per leaf node of 120, the number of features to consider for the best split equal to the square root of total number of features, the learning rate of 0.28, and the maximum depth of the individual regression estimators of 5. In validation calculations against "unseen" data, a GBT model is first trained on the AZ-DREAM Challenges dataset, utilizing either the original instances or the original and augmented data. The trained model is then employed to classify instances in the DrugCombDB dataset[75].

Data availability

All data are freely available at https://github.com/MengLiu90/Synergy-Data-Augmentation.

Received: 23 October 2023; Accepted: 11 January 2024
Published online: 18 January 2024

References

1. Liu, Y. & Zhao, H. Predicting synergistic effects between compounds through their structural similarity and effects on transcriptomes. *Bioinformatics* **32**(24), 3782–3789 (2016).
2. Vogel, C. L. *et al.* Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. *J. Clin. Oncol.* **20**(3), 719–726 (2002).
3. Bayat Mokhtari, R. *et al.* Combination therapy in combating cancer. *Oncotarget* **8**(23), 38022–38043 (2017).
4. Rafique, R., Islam, S. M. R. & Kazi, J. U. Machine learning in the prediction of cancer therapy. *Comput. Struct. Biotechnol. J.* **19**, 4003–4017 (2021).
5. Holbeck, S. L. *et al.* The National cancer institute ALMANAC: A comprehensive screening resource for the detection of anticancer drug pairs with enhanced therapeutic activity. *Cancer Res.* **77**(13), 3564–3576 (2017).
6. O'Neil, J. *et al.* An unbiased oncology compound screen to identify novel combination strategies. *Mol. Cancer Ther.* **15**(6), 1155–1162 (2016).
7. Forcina, G. C. *et al.* Systematic quantification of population cell death kinetics in mammalian cells. *Cell Syst.* **4**(6), 600–610 (2017).
8. Markt, P. *et al.* CLOUD – CeMM library of unique drugs. *J. Cheminform.* **4**, P23 (2012).
9. Licciardello, M. P. *et al.* A combinatorial screen of the CLOUD uncovers a synergy targeting the androgen receptor. *Nat. Chem. Biol.* **13**(7), 771–778 (2017).
10. Zheng, S. *et al.* DrugComb update: A more comprehensive drug sensitivity data repository and analysis portal. *Nucleic Acids Res.* **49**(W1), W174–W184 (2021).
11. Zagidullin, B. *et al.* DrugComb: An integrative cancer drug combination data portal. *Nucleic Acids Res.* **47**(W1), W43–W51 (2019).
12. Berenbaum, M. C. What is synergy?. *Pharmacol. Rev.* **41**(2), 93–141 (1989).
13. Loewe, S. The problem of synergism and antagonism of combined drugs. *Arzneimittelforschung* **3**(6), 285–290 (1953).
14. Yadav, B. *et al.* Searching for drug synergy in complex dose-response landscapes using an interaction potency model. *Comput. Struct. Biotechnol. J.* **13**, 504–513 (2015).
15. Seo, H. *et al.* SYNERGxDB: An integrative pharmacogenomic portal to identify synergistic drug combinations for precision oncology. *Nucleic Acids Res.* **48**(W1), W494–W501 (2020).
16. Menden, M. P. *et al.* Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nat. Commun.* **10**(1), 2674 (2019).
17. Shorten, C. & Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *J. Big Data* **6**(1), 1–48 (2019).
18. Taylor, L. & Nitschke, G. Improving deep learning with generic data augmentation. in *2018 IEEE Symposium Series on Computational Intelligence (SSCI)* (IEEE, 2018).
19. Moreno-Barea, F. J. *et al.* Forward noise adjustment scheme for data augmentation. in *2018 IEEE Symposium Series on Computational Intelligence (SSCI)* (IEEE, 2018).
20. Zhong, Z. *et al.* Random erasing data augmentation. in *Proceedings of the AAAI conference on artificial intelligence* (2020).
21. Inoue, H. *Data Augmentation by Pairing Samples for Images Classification.* arXiv preprint arXiv:1801.02929 (2018).
22. Summers, C. & Dinneen, M. J. Improved mixed-example data augmentation. in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV).* (IEEE, 2019).
23. Wen, Q. *et al. Time Series Data Augmentation for Deep Learning: A Survey.* arXiv preprint arXiv:2002.12478 (2020).
24. Le Guennec, A., Malinowski, S. & Tavenard, R. Data augmentation for time series classification using convolutional neural networks. in *ECML/PKDD Workshop on Advanced Analytics and Learning on Temporal Data* (2016).
25. Steven Eyobu, O. & Han, D. S. Feature representation and data augmentation for human activity classification based on wearable IMU sensor data using a deep LSTM neural network. *Sensors* **18**(9), 2892 (2018).
26. Gao, J. *et al. Robusttad: Robust Time Series Anomaly Detection Via Decomposition and Convolutional Neural Networks.* arXiv preprint arXiv:2002.09545 (2020).
27. Wen, Q. *et al.* RobustSTL: A robust seasonal-trend decomposition algorithm for long time series. in *Proceedings of the AAAI Conference on Artificial Intelligence* (2019).
28. Cao, H., Tan, V. Y. & Pang, J. Z. A parsimonious mixture of Gaussian trees model for oversampling in imbalanced and multimodal time-series classification. *IEEE Transact. Neural Netw. Learn. Syst.* **25**(12), 2226–2239 (2014).
29. Kang, Y., Hyndman, R. J. & Li, F. GRATIS: GeneRAting time series with diverse and controllable characteristics. *Stat. Anal. Data Min. ASA Data Sci. J.* **13**(4), 354–376 (2020).
30. Esteban, C., Hyland, S. L. & Rätsch, G. *Real-Valued (medical) Time Series Generation with Recurrent Conditional Gans.* arXiv preprint arXiv:1706.02633 (2017).
31. Ratner, A. J. *et al.* Learning to compose domain-specific transformations for data augmentation. *Adv. Neural Inf. Process. Syst.* **30** (2017).
32. Zhang, X. *et al. Adversarial Autoaugment.* arXiv preprint arXiv:1912.11188 (2019).

33. Dash, S. *et al.* Medical time-series data generation using generative adversarial networks. in *International Conference on Artificial Intelligence in Medicine* (Springer, 2020).
34. DeVries, T. & Taylor, G.W. *Dataset Augmentation in Feature Space.* arXiv preprint arXiv:1702.05538 (2017).
35. Wong, S. C. *et al.* Understanding data augmentation for classification: When to warp?. in *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)* (IEEE, 2016).
36. Frid-Adar, M. *et al. Gan-Based Data Augmentation for Improved Liver Lesion Classification.* (2018).
37. Calimeri, F. *et al.* Biomedical data augmentation using generative adversarial neural networks. in *International Conference on Artificial Neural Networks* (Springer, 2017).
38. Frid-Adar, M. *et al.* GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing* **321**, 321–331 (2018).
39. Han, C. *et al.* GAN-based synthetic brain MR image generation. in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (IEEE, 2018).
40. Madani, A. *et al.* Chest x-ray generation and data augmentation for cardiovascular abnormality classification. in *Medical Imaging 2018: Image Processing* (SPIE, 2018).
41. Gatys, L. A., Ecker, A. S. & Bethge, M. *A Neural Algorithm of Artistic Style.* arXiv preprint arXiv:1508.06576 (2015).
42. Jackson, P. T. *et al.* Style augmentation: Data augmentation via style randomization. in *CVPR Workshops*. (2019).
43. Wang, J. & Perez, L. The effectiveness of data augmentation in image classification using deep learning. *Convol. Neural Netw. Vis. Recogn.* **11**, 1–8 (2017).
44. Lemley, J., Bazrafkan, S. & Corcoran, P. Smart augmentation learning an optimal data augmentation strategy. *IEEE Access* **5**, 5858–5869 (2017).
45. Cubuk, E. D. *et al. Autoaugment: Learning Augmentation Policies from Data.* arXiv preprint arXiv:1805.09501 (2018).
46. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**(6), 84–90 (2017).
47. Yun, S. *et al.* Cutmix: Regularization strategy to train strong classifiers with localizable features. in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019).
48. Jones, A. *et al. Bilex Rx: Lexical Data Augmentation for Massively Multilingual Machine Translation*. arXiv preprint arXiv:2303.15265 (2023).
49. Sutherland, J. J., O'brien, L. A. & Weaver, D. F. Spline-fitting with a genetic algorithm: A method for developing classification structure– activity relationships. *J. Chem. Inf. Comput. Sci.* **43**(6), 1906–1915 (2003).
50. Bjerrum, E. J. *SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules.* arXiv preprint arXiv:1703.07076 (2017).
51. Kimber, T. B., Gagnebin, M. & Volkamer, A. Maxsmi: Maximizing molecular property prediction performance with confidence estimation using smiles augmentation and deep learning. *Artif. Intell. Life Sci.* **1**, 100014 (2021).
52. Sidorov, P. *et al.* Predicting synergism of cancer drug combinations using NCI-ALMANAC data. *Front. Chem.* **7**, 509 (2019).
53. Ye, Z. *et al.* ScaffComb: A phenotype-based framework for drug combination virtual screening in large-scale chemical datasets. *Adv. Sci.* **8**(24), 2102092 (2021).
54. Liu, Q. *et al.* DeepCDR: A hybrid graph convolutional network for predicting cancer drug response. *Bioinformatics* **36**, i911–i918 (2020).
55. Kim, S. *et al.* PubChem in 2021: New data content and improved web interfaces. *Nucleic Acids Res.* **49**(D1), D1388–D1395 (2021).
56. Chuang, J. C. & Neal, J. W. Crizotinib as first line therapy for advanced ALK-positive non-small cell lung cancers. *Transl. Lung Cancer Res.* **4**(5), 639–641 (2015).
57. Royce, M. E. & Osman, D. Everolimus in the treatment of metastatic breast cancer. *Breast Cancer (Auckl)* **9**, 73–79 (2015).
58. Ruiz, R., Raez, L. E. & Rolfo, C. Entinostat (SNDX-275) for the treatment of non-small cell lung cancer. *Expert Opin. Investig. Drugs* **24**(8), 1101–1109 (2015).
59. Le Grand, M. *et al.* Akt targeting as a strategy to boost chemotherapy efficacy in non-small cell lung cancer through metabolism suppression. *Sci. Rep.* **7**, 45136 (2017).
60. Keenan, T. E. *et al.* Clinical efficacy and molecular response correlates of the WEE1 inhibitor adavosertib combined with cisplatin in patients with metastatic triple-negative breast cancer. *Clin. Cancer Res.* **27**(4), 983–991 (2021).
61. Cazzaniga, M. E. *et al.* Metronomic oral vinorelbine in advanced breast cancer and non-small-cell lung cancer: Current status and future development. *Fut. Oncol.* **12**(3), 373–387 (2016).
62. Smyth, L. M. *et al.* Capivasertib, an AKT Kinase Inhibitor, as monotherapy or in combination with fulvestrant in patients with. *Clin. Cancer Res.* **26**(15), 3947–3957 (2020).
63. O'Boyle, N. M. *et al.* Open babel: An open chemical toolbox. *J. Cheminform.* **3**, 33 (2011).
64. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **405**(2), 442–451 (1975).
65. Mazandu, G. K. *et al.* IHP-PING—generating integrated human protein–protein interaction networks on-the-fly. *Brief. Bioinformat.* **22**(4), 277 (2021).
66. Szklarczyk, D. *et al.* STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res.* **44**(D1), D380–D384 (2016).
67. Keller, T. H., Pichota, A. & Yin, Z. A practical view of 'druggability'. *Curr. Opin. Chem. Biol.* **10**(4), 357–361 (2006).
68. Hosmer, D. & Lemeshow, S. *Applied Logistic Regression* 2nd edn. (Wiley, New York, 2000).
69. Tolles, J. & Meurer, W. J. Logistic regression: Relating patient characteristics to outcomes. *JAMA* **316**(5), 533–534 (2016).
70. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
71. Ben-Hur, A. *et al.* Support vector clustering. *J. Mach. Learn. Res.* **2**, 125–137 (2001).
72. Breiman, L. Random forests. *Mach. Learn.* **45**(1), 5–32 (2001).
73. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* 1189–1232 (2001).
74. Pu, L. *et al.* CancerOmicsNet: A multi-omics network-based approach to anti-cancer drug profiling. *Oncotarget* **13**, 695–706 (2022).
75. Liu, H. *et al.* DrugCombDB: A comprehensive database of drug combinations toward the discovery of combinatorial therapy. *Nucleic Acids Res.* **48**(D1), D871–D881 (2020).
76. Singha, M. *et al.* GraphGR: A graph neural network to predict the effect of pharmacotherapy on the cancer cell growth. *bioRxiv* (2020).
77. Nag, S. *et al.* Deep learning tools for advancing drug discovery and development. *3 Biotech* **12**(5), 110 (2022).
78. Vamathevan, J. *et al.* Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **18**(6), 463–477 (2019).
79. Kendall, M. G. *Rank Correlation Methods*. (1962).
80. Agresti, A. *Analysis of ordinal categorical data* (Wiley, 2010).
81. Jaeger, S., Fulle, S. & Turk, S. Mol2vec: Unsupervised machine learning approach with chemical intuition. *J. Chem. Inf. Model* **58**(1), 27–35 (2018).
82. Dincer, A. B., Janizek, J. D. & Lee, S.-I. Adversarial deconfounding autoencoder for learning robust gene expression embeddings. *Bioinformatics* **36**, i573–i582 (2020).

Acknowledgements

Portions of this research were conducted with computing resources provided by Louisiana State University.

Author contributions

M.L. processed the data, performed statistical analyses, conducted machine learning experiments, and analyzed the results. G.S. curated the data, performed preliminary statistical analyses, and augmented the data. J.R. contributed to the augmentation design and the interpretation of the results. J.R. and M.B. secured funding for the project. M.L. and G.S. drafted the paper. M.B. supervised the project and wrote the final version of the manuscript.

Funding

This work has been supported in part by the National Institute of General Medical Sciences of the National Institutes of Health award R35GM119524, the US National Science Foundation award CCF-1619303, the Louisiana Board of Regents contract LEQSF (2016–19)-RD-B-03, and the Center for Computation and Technology at Louisiana State University.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.B.