 doi:10.20944/preprints202104.0395.v1

Article

SynPred: Prediction of Drug Combination Effects in Cancer using Full-Agreement Synergy Metrics and Deep Learning

António J. Preto [1,2,3], Pedro Matos-Filipe [1,2], Joana Mourão [1,2], and Irina S. Moreira [4,1,*]

[1] University of Coimbra, Center for Neuroscience and Cell Biology, Center for Innovative Biomedicine and Biotechnology, 3004-504 Coimbra, Portugal; martinsgomes.jose@gmail.com, p.pintofilipe@gmail.com, jmourao@cnc.uc.pt, irina.moreira@cnc.uc.pt
[2] University of Coimbra, Institute for Interdisciplinary Research, 3030-789 Coimbra, Portugal; martinsgomes.jose@gmail.com, p.pintofilipe@gmail.com, jmourao@cnc.uc.pt
[3] PhD Programme in Experimental Biology and Biomedicine, Institute for Interdisciplinary Research (IIIUC), University of Coimbra, Casa Costa Alemão, 3030-789 Coimbra, Portugal; martinsgomes.jose@gmail.com
[4] Department of Life Sciences, University of Coimbra, Calçada Martim de Freitas, 3000-456 Coimbra, Portugal; irina.moreira@cnc.uc.pt
* Correspondence: irina.moreira@cnc.uc.pt; Tel.: +351 239 240 227

Simple Summary: Drug resistance in cancer is a multifactorial problem that can significantly compromise patient treatment, leading to higher mortality and morbidity. The administration of synergistic drug combinations can minimize this adaptation phenomenon; however, choosing the best combination for each cancer type is still difficult and complex. Our study aimed to develop a Machine Learning model that predicts the best synergistic anticancer drug combinations against specific cancer cell lines by integrating the four most used drug synergy metrics (Bliss, Loewe, HAS, and ZIP) with multi-omics features of cancer cell lines, phenotypic and biophysical data. Our model SynPred is an ensemble classifier that is fully integrated for the first time into a user-friendly webserver allowing users with different backgrounds, from scientists to clinicians, to use this model from drug synergy prediction, requiring only the upload of the two drug SMILES to be tested.

Abstract: High-throughput screening technologies continues to produce large amounts of multi-omics data from different populations and cell types for various diseases, such as cancer. However, analysis of such data encounters difficulties due to cancer heterogeneity, further exacerbated by human biological complexity and genomic variability. There is a need to redefine the drug discovery development pipeline, bringing an Artificial Intelligence (AI)-powered informational view that integrates relevant biological information and explores new ways to develop effective anticancer approaches. Here, we show SynPred, an interdisciplinary approach that leverages specifically designed ensembles of AI-algorithms, links omics and biophysical traits to predict synergistic anticancer drug synergy. SynPred exhibits state-of-the-art performance metrics: accuracy – 0.85, precision – 0.77, recall – 0.75, AUROC – 0.82, and F1-score - 0.76 in an independent test set. Moreover, data interpretability was achieved by deploying the most current and robust feature importance approaches. A simple web-based application available online at http://www.moreiralab.com/resources/synpred/ was constructed to predict synergistic anticancer drug combinations requiring only the upload of the two drug SMILES to be tested, allowing easy access by non-expert researchers.

 doi:10.20944/preprints202104.0395.v1

Keywords: Deep-Learning; Interpretability; Omics; Biophysics; Drug Synergy; Cancer

1. Introduction

Cancer, a heterogeneous group of diseases, is one of the leading causes of mortality and the most significant barrier to increase life expectancy worldwide. The International Agency for Research on Cancer estimates that, by 2040, approximately 29.5 million new cancer cases and 16.4 million deaths will be reported mainly due to the population's growth and ageing [1]. One of the significant contributors to this disease's global burden is the development of therapy resistance and, consequently, tumour relapse. Drug resistance in cancer is a multifactorial problem driven by the tumour microenvironment and genetic and nongenetic/epigenetic mechanisms that, along with cell plasticity, contribute to tumour heterogeneity [2]. In clinical settings, this problem is dealt with a combination of drugs administered together or in sequence, i.e., polytherapy. Targeting multiple components of different or interconnected cancer pathways is an efficient strategy to block vital biological processes [3].

Drug combinations with a synergistic effect, i.e., when the total therapeutic effect of both drugs is greater than the expected additive monotherapy effect [4], were successfully developed and applied in the treatment of different types of tumours, such as human epidermal growth factor receptor 2-positive breast cancer [5], chronic myeloid leukaemia [6], prostate cancer [7] or BRAF-mutant melanoma [8]. Nevertheless, this simultaneous administration can also result in a reduced therapeutic effect and possible toxicity (designated antagonism) or in the same beneficial effect when compared with the expected additive monotherapy effect (additivity) [4]. The experimental identification of successful synergistically effective combinations that amplify each other's activity is a well-known time-consuming, and expensive task. Therefore, there is still a significant need for efficient and user-friendly computational methods to complement and speed-up the traditional approaches by predicting the best synergistic drug combinations [9,10].

In the last years, the development and improvement of high-throughput technologies and computational tools boosted the use of large volumes of multi-omics data (e.g., genomic, transcriptomic, proteomic) essential to dissect and uncover the complex molecular signatures of cancer. Machine Learning (ML) algorithms, in particular, have attracted particular attention for their ability to learn new associations and extract useful insights from this type of data. A few ML models based on eXtreme Gradient Boosting, Random Forest, Elastic Nets, Support Vector Machine, and Naïve Bayes were already developed to predict the best combination of anticancer drugs by the integration of omics data with chemoinformatic properties of drugs or network information of their targets [11–14]. Likewise, Deep-Learning (DL) implemented via Deep-Neural Networks (DNNs) was particularly useful in dealing with the high multi-dimensionality of omics data in supervised and unsupervised contexts. DNNs classification and regression models such as DeepSynergy [15], AuDNNsynergy [16], MatchMaker [17] or DeepSignalingSynergy [18] were recently developed for drug combination prediction. However, a critical bottleneck was the use for class definition of a single drug synergy metric to access the degree of interaction, with very few using alternative approaches [14,19]. Besides, the vast majority require advanced knowledge in bioinformatics to be used and are only available through GitHub, posing several constraints for accessibility by the medical and scientific community. Most of the available web interfaces are mainly dedicated to drug combination response prediction using ML regression models, such as DECREASE [20] or DrugComb [21]. Notwithstanding, these require at least a set of laboratory experiments to upload a full or partial mandatory dose-response matrix, which difficult its use by the scientific community.

To overcome the current problems found in the field, we developed SynPred (SYNergy PREDiction), an *in-silico* ensemble classification model that considers several drug

Citation: Preto, J. P.; Matos-Filipe, P.; Mourão, J; Moreira, I. S. SynPred: Prediction of Drug Combination Effects in Cancer using Full-Agreement Synergy Metrics and Deep Learning. *Cancers* **2021**, *13*, x. https://doi.org/10.3390/xxxxx

Academic Editor: Firstname Lastname

Received: date
Accepted: date
Published: date

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2021 doi:10.20944/preprints202104.0395.v1

synergy metrics (Bliss, Loewe, HAS, and ZIP) for the prediction of effective drug combinations with accuracy, precision, recall, Area Under the Receiver Operator Curve (AUROC), and F1 scores of 0.85, 0.77, 0.75, 0.82, and 0.76 respectively. This model was developed by integrating not only multi-omics features of cell lines, phenotypic data but also biophysical data, in particular, physicochemical and structural features of drugs. SynPred was independently tested and validated in a new comprehensive database of synergistic drug combinations from the Ianevski study [20], achieving an accuracy of 0.98. We made available the stand-alone deployment at https://github.com/MoreiraLAB/synpred, which allows the user the opportunity to undergo bulk prediction with SynPred. Additionally, for the first time, a user-friendly web-based application was assembled and made freely available online at http://www.moreiralab.com/resources/synpred/ to predict drug combinations, requiring only the upload of the two drug SMILES to be tested. This interactive platform will allow users with different backgrounds, from scientists to clinicians, to test, reproduce and validate our models and data. The workflow used for the development of SynPred is depicted in **Figure 1**.

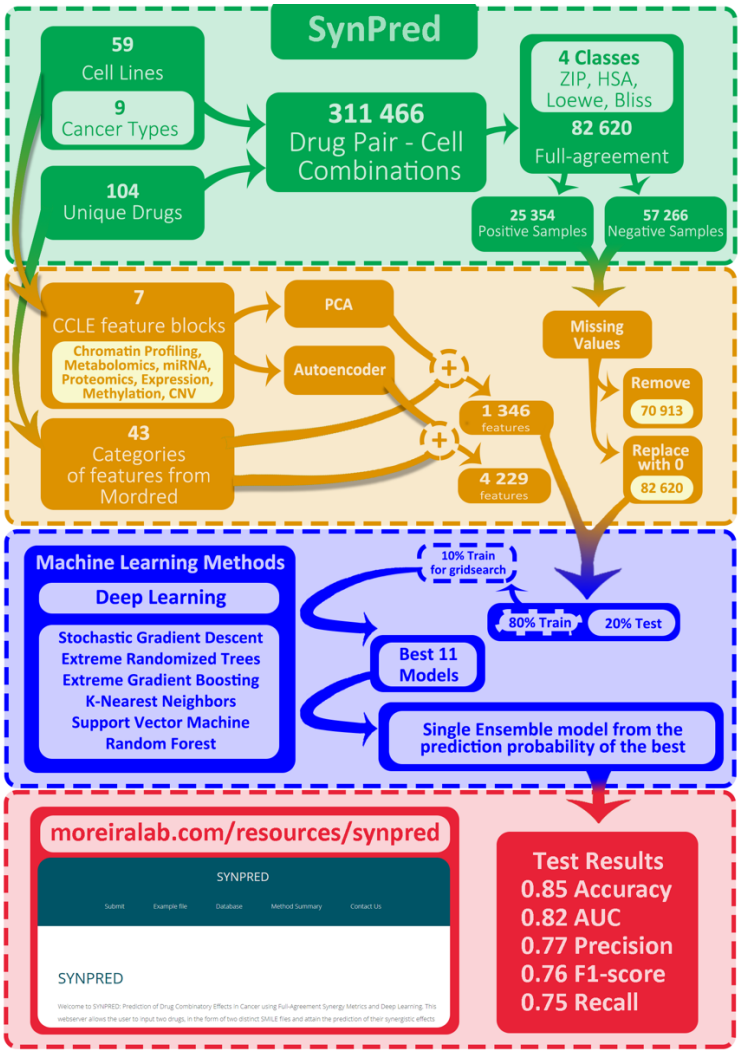


Figure 1. SynPred workflow summary. (Green) - Dataset construction. The National Cancer Institute - A Large Matrix of Anti-Neoplastic Agent Combinations database (phenotypic data) and the Cancer Cell Line Encyclopedia (CCLE) (multi-omics data) were used for this purpose. Four reference models (Zero Interaction Potency-ZIP, Highest Single Agent-HSA, Bliss, and Loewe) were used to quantify the degree of combination and retrieve a full agreement between all metrics. (Orange) – Feature extraction and data pre-processing. Included normalization and dimensionality reduction using auto-encoder or Principal Component Analysis (PCA). (Blue) - Prediction model development using a training set. This model was developed using different ML algorithms and a final ensemble model. (Red) – Model evaluation using an independent test set. The accuracy, precision, recall, Area Under the Receiver Operator Curve (AUROC), and F1-score, metrics were calculated.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2021 doi:10.20944/preprints202104.0395.v1

2. Materials and Methods

2.1. Data acquisition and processing

2.1.1. Experimental drug combination phenotypic data

Drug combination phenotypic data was acquired via bulk-download from the largest-to-date dataset from National Cancer Institute - A Large Matrix of Anti-Neoplastic Agent Combinations (NCI-ALMANAC) through https://wiki.nci.nih.gov/display/NCIDTPdata/NCI-ALMANAC [22]. To this date, the dataset includes phenotypic data of tested cancer cell lines (growth percentage) of 104 unique FDA-approved drugs. These drugs were tested in combination against 59 cell lines from 9 cancer types currently included in the NCI [23,24], comprising a total of 311.466 drug pair/cell line combinations. Drug sensitivity assays included in NCI-ALMANAC were performed at the NCI's Frederick National Laboratory for Cancer Research, the Stanford Research Institute, and the University of Pittsburgh. Briefly, for each assay, cells were cultivated for 48 hours in a 3x3 or a 5x3 concentration matrix (different concentration values for each drug in combination) and the endpoint determined by Sulforhodamine B or CellTiter-Glo [22]. From these records, the authors retrieved the cell growth percentage at each drug concentration point, which corresponds to the percentage of growth of the cell lines in the presence of each combination, yielding a final viability assessment.

2.1.2. Combination scores and class definition

The phenotypic data from high-throughput drug combination screens were analysed using the "SynergyFinder" [25] R package (version 2.0.3). We used the percentage of cell growth included in the dataset to assess the degree of combination for each concentration matrix. For that, we applied four reference models: Bliss independence (**Equation 1**) [26], Loewe additivity (**Equation 2**) [27,28], Highest Single Agent (HSA) (**Equation 3**) [29], and Zero Interaction Potency (ZIP) (**Equation 4**) [30].

$$yBliss = y1 + y2 - y1y2 \quad (1)$$

Equation 1. Bliss independence model. yBliss – Bliss response; y1 – drug1 response; y2 – drug2 response.

$$yLoewe = \frac{Emin + Emax\left(\frac{x1 + x2}{m}\right)^{\lambda}}{1 + \left(\frac{x1 + x2}{m}\right)^{\lambda}} \quad (2)$$

Equation 2. Loewe additivity model. *yLoewe* – Loewe response; *Emin* – minimum drug response; *Emax* – maximum drug response; *m* – dose that produces midpoint effect between *Emin* and *Emax*; λ – shape parameter indicating the slope of the curve; *x1* – drug 1 dose; *x2* – drug 2 dose.

$$yHSA = \max(y1, y2) \quad (3)$$

Equation 3. Highest Single Agent (HSA) model; *yHSA* – HSA response; *y1* – drug 1 response; *y2* – drug 2 response.

$$yZIP = \frac{\left(\frac{x1}{m1}\right)^{\lambda 1}}{1 + \left(\frac{x1}{m1}\right)^{\lambda 1}} + \frac{\left(\frac{x2}{m2}\right)^{\lambda 2}}{1 + \left(\frac{x2}{m2}\right)^{\lambda 2}} - \left(\frac{\left(\frac{x1}{m1}\right)^{\lambda 1}}{1 + \left(\frac{x1}{m1}\right)^{\lambda 1}} * \frac{\left(\frac{x2}{m2}\right)^{\lambda 2}}{1 + \left(\frac{x2}{m2}\right)^{\lambda 2}}\right) \quad (4)$$

Equation 4. Zero Interaction Potency (ZIP) model; *yZIP* – ZIP response; *x1* – drug 1 dose; *x2* – drug 2 dose; *m1* – dose that produces midpoint effect for drug 1; *m2* – dose that produces midpoint effect for drug 2; $\lambda 1$ – shape parameter indicating the slope of the curve for drug 1; $\lambda 2$ – shape parameter indicating the slope of the curve for drug 2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2021 doi:10.20944/preprints202104.0395.v1

With this data, a binary classifier was developed to identify the type of combinatory effect present in each drug pair-cell line sample, where the 20% smaller values were classified as synergistic, and the remaining ones were classified as non-synergistic (**Figure A1**). The dataset used for training considered full-agreement combination assessment, i.e., we only kept the instances on which combination classification was the same across the four previous reference classifiers. For the dataset used, this process yielded 25.354 synergistic samples and 57.266 non-synergistic samples.

2.1.3. Drug molecular descriptors

Each drug included in NCI-ALMANAC was analysed to extract its physicochemical and structural features. A Simplified Molecular-Input Line-Entry System (SMILE) representation of the drugs was acquired from PubChem [31]. SMILEs were then used to mine molecular descriptors using the Python package "Mordred" (version 1.1.2) [32]. In total, was retrieved an array of 1.613 numeric features of 43 different categories making a two-dimensional molecular description of the drugs. Feature-arrays comprising non-numerical attributes or displaying zero variance were deleted. This pre-processing left 586 features describing each drug included in NCI-ALMANAC, distributed across 28 categories (**Table 1**). The resulting features were subjected to normalisation by removing the mean and scaling to unit variance with scikit-learn's StandardScaler [33].

Table 1. Number of features according to the molecular descriptor category of Mordred. Features are categorized as Energetic (E), Pharmacological (P), Structural (S) or Miscellaneous (M - in case of evaluating characteristics of multiple fields).

Number of Features per Descriptor Category					
E	Acidity/Basicity	2	S	Information Content	36
P	ADME	3	S	Molecular Complexity	1
S	Aromatics	2	P	Molecular Operating Environment	51
S	Atom Count	16	S	Molecule Graph	5
S	Atom-bond Connectivity	2	S	Path Count	21
M	Autocorrelation	180	E	Polarizability	2
S	Bond Count	9	S	Ring Count	66
E	Atomic Orbitals	10	S	Rotatable Bonds	1
S	Chirality	38	S	Topological Charges	21
S	Constitutional	14	S	Topological Index	7
E	Energy State	68	S	Topological Polar Surface Area	2
S	Fragment Complexity	1	S	Walk Counts	21
S	Framework	1	S	Weight	2
S	Hydrogen Bonds	2	M	Wildman-Crippen	2

2.1.4. Omics data of cancer cell lines

Omics data (expression, copy number variation, and methylation, global chromatin profiling, metabolomics, microRNA, proteomic profiling) describing the cancer cell lines were acquired via bulk download from the Cancer Cell Line Encyclopedia - CCLE (https://portals.broadinstitute.org/ccle/data) [34]. The number of cell lines included in CCLE varies depending on the type of omics data available at the time (**Table 2**). Correspondence of cell line IDs between NCI-ALMANAC and CCLE was performed according to data available at the Swiss Institute of Bioinformatics Cellosaurus Website [35]. According to the affected tissue, annotations acquired through Cellossaurus split the CCLE cell lines into 21 different cancer types. In agreement with the original publications [34,36], expression data were obtained through RNA-sequencing and processed to obtain level expression in transcripts per million by the expectation-maximization algorithm (file: CCLE_RNAseq_rsem_genes_tpm_20180929.txt.gz). Copy Number Variation (CNV) data were acquired from the Affymetrix SNP6.0 Arrays (file:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2021 doi:10.20944/preprints202104.0395.v1

CCLE_copynumber_byGene_2013-12-03.txt.gz). Copy numbers were normalized by the most similar HapMap normal samples [37]. Segmentation of normalized log2 (CN/2) ratios was achieved using the circular binary segmentation algorithm [34,38]. Methylation data were derived by quantifying CpG islands using Reduced Representation Bisulfite Sequencing (file: CCLE_RRBS_tss_CpG_clusters_20181022.txt.gz). Global chromatin profiling was attained using multiple reaction monitoring for 42 combinations of histone marks (file: CCLE_GlobalChromatinProfiling_20181130.csv). Metabolomics data were acquired in parallel with global chromatin profiling by reporting the abundance measures of 225 metabolites (file: CCLE_metabolomics_20190502.csv). MicroRNA associated with cancer dependencies was correlated, regarding 734 microRNAs, with the Achilles gene dependency dataset. Protein profiling was measured with Reverse Phase Protein Arrays for 213 antibodies (file: CCLE_RPPA_20181003.csv) [36].

Table 2. Number of features pertaining the omics data and the corresponding amount for both the autoencoder and the PCA processing.

Omics Data	Number of available cell lines	Number of features available	Number of features after autoencoder	Number of features after PCA	Explained variance (PCA)
Expression	1019	57820	1156	25	0.96
Copy Number Variation	1043	23316	466		0.75
Methylation	843	56146	1122		0.73
Global Chromatin Profiling	897	42	21		0.99
Metabolomics	928	225	112		0.88
microRNA	954	734	73		0.97
Proteomics	899	214	107		0.86

2.1.5. Dimensionality reduction of omics data

Data were normalized by removing the mean and scaling to unit variance with scikit-learn's StandardScaler [33]. Due to the omics data's high complexity, we performed dimensionality reduction to minimize the noise introduced in the dataset by highlighting the essential features. The datasets already described were used to build and train a Multi-Layer Perceptron (MLP) autoencoder, an unsupervised Artificial Neural Network (ANN) with a typical "hourglass" architecture, which is often used to perform dimensionality reduction in vast and high-dimensional datasets such as the ones observed with omics data [39–41]. This type of MLPs usually consists of three parts; an encoder that abstracts the input into hidden variables, i.e., a latent-space representation, a bottleneck layer that holds the smallest Hidden Layer (HL) (for purposes of dimensionality reduction, this is the layer that defines the size of the reduced dataset), and a decoder that reconstructs the original input data from the hidden data [42,43]. Seven autoencoders, one for each of the CCLE feature blocks, were developed by using Keras with a TensorFlow for Graphics Processing Units (GPU) (Version 2.3.1) backend [44]. Each of the autoencoders comprised seven layers, of which five were HLs. The input and output layers follow the number of available features in all cell lines, as displayed in **Table 2**. The number of nodes within the bottleneck layer of each of the seven autoencoders (used for extraction of the encoded features) corresponds to the autoencoder's final number of features. The two HL in each of the encoder and decoder sections vary in size according to the number of samples and features available (**Table A1**). In this stage, all models used Adam [45] as an optimizer function with a learning rate of 0.001. Rectified Linear Unit (ReLU) activation function was used in all layers. Mean Square Error (MSE) was used as a loss function. The models were trained for 1000, 250, or 100 epochs, depending on the dataset size (**Table A2**). After

 doi:10.20944/preprints202104.0395.v1

training, each autoencoder's bottleneck layer was used to perform dimensionality reduction of the omics data.

PCA, a commonly used method for dimensionality reduction [46], was also applied in the same datasets as the autoencoder, for which 25 Principal Components (PCs) were defined. It means that, by using PCA, each of the datasets was transformed to yield only 25 features, totalling 175 features to describe each unique cell line. As shown in **Table 2**, each feature block from CCLE had its variance explained in a range from 0.73 to 0.99. Since the seven blocks were used simultaneously for each sample, each cell line is thoroughly described by the components extracted with the PCA. Missing values (in both autoencoder and PCA) were processed by either dropping the sample entirely or replacing the missing values by zero. Both these approaches were performed after keeping the samples that complied with our full-agreement standard. Finally, we randomly split the data into training and test on an 80-20 ratio for model fitting (train) and evaluation (test) **(Table A3**).

2.2. Development of ML models

2.2.1. Multi-Layer Perceptron with Keras

The binary classification was fully developed using Keras with a TensorFlow (Version 2.3.1) backend [44]. Weights were updated using Adam optimizer [45] and a learning rate of 0.0001 along 250 epochs with binary cross-entropy as the loss function. All the HL were connected through ReLU activation, while the output layer was subject to sigmoid activation. As an initial approach, we performed a gridsearch for parameter optimization using 10% of the training set, fully detailed in the section "Parameter optimization". The best performing parameters were further selected and used to train the models using the full dataset.

2.2.2. ML algorithms with scikit-learn

The datasets presented in this work were also trained with the most commonly used algorithms for synergy prediction tasks, namely with MLPs [45,47–49], RF [50], ETC [51,52], Support Vector Machines (SVM) [53], Stochastic Gradient Descents (SGD) [54], KNN [55], and XGBoost [56]. The MLP, RF, ETC, SVM, SGD, and KNN models were built using the Python package "SciKit Learn" (Version 0.22.1) [33]. The XGBoost model was built using its dedicated package for Python (available at the Python Package Index as "xgboost") [56]. These six algorithms (except for scikit-learn's MLP) were also subject to gridsearch for parameter's optimization using 10% of the training set as described in the following section, with the best ones used to train the models with the full dataset.

2.2.3. Parameter optimization

In order to properly perform parameters' optimization in all the algorithms described, a grid search was performed using in-house built scripts for Keras DL models, and scikit-learn's GridSearchCV with 3-fold cross-validation (for ML algorithms with scikit-learn). We used 10% of the training set [57], a value that is in agreement with subset usage for parameter optimization [58], since using the full training dataset would increase exponentially an already long task. For Keras DL models, the gridsearch with 288 runs was performed with parameters covering the four available datasets, 30 different network architectures, and five different dropout rates (**Table A4**). In the case of the six ML algorithms (except for scikit-learn's MLP) a total of 820 runs including parameters and dataset combinations were used (**Table A5**).

2.2.4. Ensemble algorithms

After training both the DL and the remaining ML algorithms, the best-performing ones and their respective datasets were selected to generate an ensemble binary classifier trained with the full dataset. We measured the average probability prediction of all the selected algorithms for control. Furthermore, we deployed a new gridsearch for ensemble parameter optimization, taking the class probability of the selected algorithms as features, and developed a neural network that worked as an ensemble method (**Table A6**). This

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2021 doi:10.20944/preprints202104.0395.v1

neural network had a learning rate of 0.0001, trained for three epochs, used the Adam optimizer [45] and binary cross-entropy. Furthermore, all the seven algorithms.

2.2.5. Model evaluation and performance metrics

The binary classification models were evaluated through accuracy (acc), precision (prec), recall (rec), AUROC as well as F1-score as previously described [52].

2.2.6. Benchmark analysis using the DECREASE database

To ensure the final model's generalizability, we performed a literature review to search for databases that could be used to perform the benchmark and for which we could calculate/retrieve the necessary features. The O'Neil benchmark dataset could not be used as, after processing the dataset, there were not enough single dose responses associated with each concentration in the combination datasets [59]. Regarding the Forcina [60] and Mathews [61] datasets, the cell lines used were not available at CCLE, which renders cell line-associated feature extraction impossible. The chosen dataset that matched all the benchmark requirements was the DECREASE [20], comprising 210 unique drug-drug-cell line combinations corresponding to 34 drugs and 13 cell lines.

2.2.7. Assessing feature contribution among the prediction models

Due to the dimensionality reduction of cell lines, it was necessary to break the process into two stages to access feature contribution. Firstly, since the best performing dimensionality reduction approach was the PCA, the explained variance by each of the features in relation to the respective Principal Component was considered. This information was then extracted has an attribute from the PCA object using scikit-learn [33]. Secondly, the package eli5 [62], with Python deployment, was used to assess final feature weight by deploying Permutation Importance [50], a method that allows iterative exclusion of each of the features, to assess its contribution to the predictive model. The Permutation Importance was deployed on the test set because if the training set had been used, it would not be possible to assess the feature contribution under unbiased conditions. However, it is worth noting that this evaluation occurs after all model training; hence, not influencing the test results in any way.

2.3. Web-based application interface implementation

SynPred prediction models were implemented in a web-based application at http://www.moreiralab.com/resources/synpred/. The website's plots and front-end were constructed with plotly [63] and Flask [64], both freely available Python packages, on a framework that uses an in-house adaptation of Javascript, CSS, and HTML scripts. All the back-end hosting was mediated with Flask [64].

3. Results

3.1. Tuning and choosing the best ML parameters

ML performance and training time are deeply affected by specific model parameters, so an appropriate choice of the best ones should always be performed. With that in mind, we used a gridsearch approach to test a comprehensive array of parameters and dataset combinations, including several ML methods, a wide set of DL configurations and preprocessing setups. The choice of these parameters totalled 1.112 runs, of which were selected four best Keras based sets of parameters and the best configurations for each of the seven ML models trained with scikit-learn (**Table A7 and Table A8**). Regarding the preprocessing datasets, both PCA and PCA_drop perform very similarly for all metrics. As such, we chose to proceed with the PCA dataset, as the increased number of samples can lead to better results when using the full training set with the already tuned models. However, autoencoder datasets performed worse in the train sets and slightly worse for the test set. Due to the significantly larger size of the autoencoder datasets and their slight underperformance, they were excluded from further training. When considering the tested DL parameters, particularly Keras architectures, the best performing dropout rate

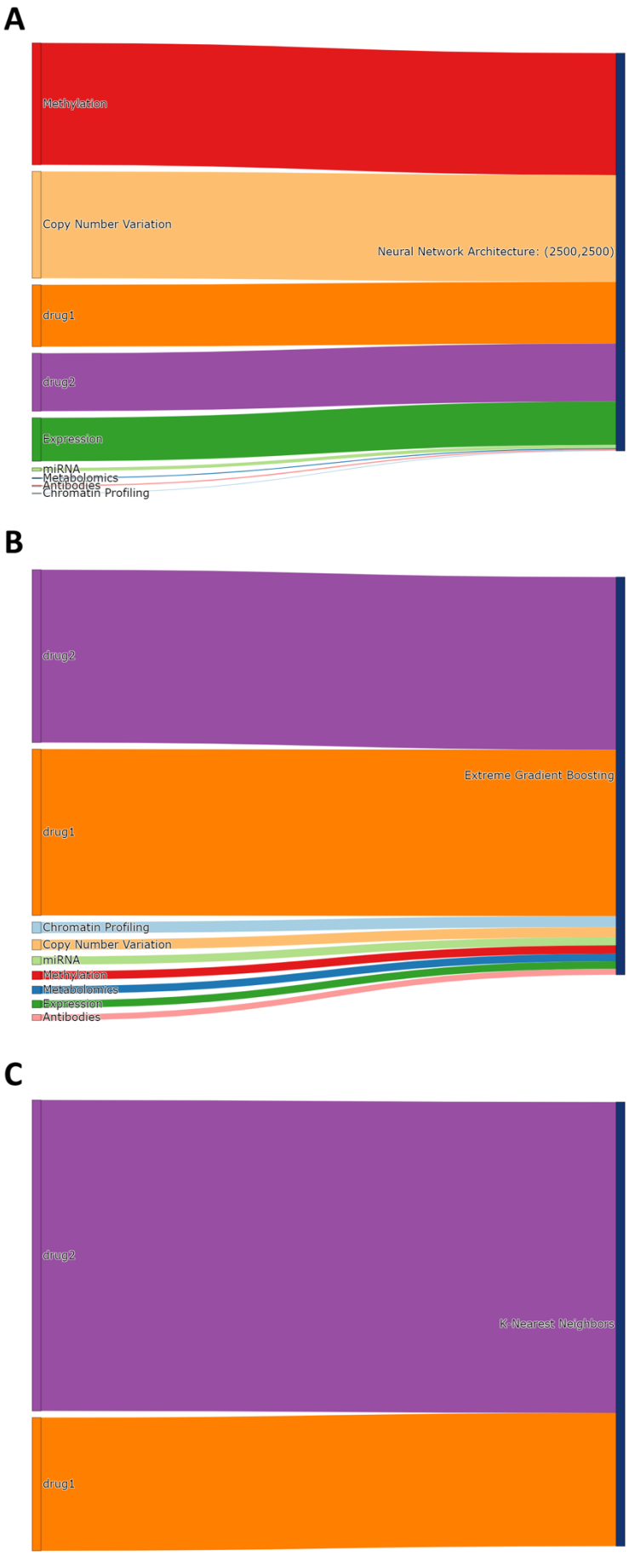
 doi:10.20944/preprints202104.0395.v1

was 0, and the best architectures, selected according to the best performance for each metric were: accuracy – (2114, 1057, 264, 16), precision – (2500, 2500), recall - (100, 100, 100, 100), AUROC – (2114, 1057) and F1-score (2500, 2500) (**Table A7**).

3.2. Measuring feature contribution for model development

To understand the importance of each group of included features for the final model performance, and to attain a higher more interpretable model, we analysed each of the individual models with Permutation Importance. We perceived that more complex models, particularly DL-based models, tend to make a more extensive use of the omics-based features (over 70% of the total feature contribution) (**Figure 2A**), while simpler models, such as K-Nearest Neighbors (kNN), made exclusive use of the drug features (**Figure 2B**). Other non-DL based models, such as Extreme Gradient Boosting (XGboost), made residual (around 7%) usage of the omics features (**Figure 2C**). This observation proves to be critical to find drug pair-cell line combinations specific for each cancer tissue.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2021 doi:10.20944/preprints202104.0395.v1



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2021 doi:10.20944/preprints202104.0395.v1

Figure 2. Sankey diagram showing the results of permutation importance. Feature block contribution for a Keras DL model (A), kNN (B), and XGboost (C). The left and right axes represent the association between the multiomics data or drug features and the ML model, while the connection's width is proportional to each feature's contribution to the final predictive model. The colour scheme represents CCLE multiomics data (light blue - chromatin profiling, salmon – copy number variation, light green – miRNA; red – methylation, dark blue – metabolomics, green – expression, pink – antibodies) and drug descriptors (orange – drug1, purple – drug 2).

We then looked for a possible biological relevance of the Top 5 genes in each group of the most critical multiomics features to understand if genes contributing more were also implicated in tumorigenesis. Of the 15 ranked genes from expression, methylation and CNV variations, all of them, except for C11orf52 and DAZ2, the last one mainly associated with male infertility, are used as prognostic cancer markers or have a role in tumour progression and treatment (**Table 3**). These data suggest that our models, especially DNNs, are likely to capture the most relevant information for each group of multiomics features for synergistic drug combinations. The remaining ranked genes organised by each ML model's best-contributing features are presented in interactive Sankey diagrams on the website landing page (Database dropdown menu at webserver page).

Table 3. Permutation importance of the top5 proteins associated with expression, methylation and CNV features as well as their associated biological relevance.

Type of Feature	Gene Name	Protein Description	Biological relevance [1]
Expression	TMSB4X	Thymosin beta 4 X linked	Prognostic marker in renal cancer (unfavorable)
	MT-CO2	Mitochondrially encoded cytochrome c oxidase II	Prognostic marker in liver cancer (favorable) and pancreatic cancer (favorable)
	MT-RNR2	Mitochondrially Encoded 16S RRNA	Studies have suggested their role as a potential biomarker in bladder cancer [65]
	MT-CO3	Mitochondrially encoded cytochrome c oxidase III	Prognostic marker in pancreatic cancer (favorable) and liver cancer (favorable)
	COX6C	Cytochrome c oxidase subunit 6C	However, several studies have shown an abnormal level of COX6C in several types of cancer [66]
Methylation	C11orf52	Chromosome 11 open reading frame 52	No supported role in cancer prognostic or progression.
	NPY1R	Neuropeptide Y receptor Y1	Prognostic marker in breast cancer (favorable)
	TMBIM6	Transmembrane BAX inhibitor motif containing 6	Prognostic marker in renal cancer (favorable), head and neck cancer (unfavorable) and breast cancer (unfavorable)
	C2CD4D	C2 calcium dependent domain containing 4D	This gene was suggested as epigenetic markers in ovarian cancer [67]
	EDNRB	Endothelin receptor type B	Prognostic marker in renal cancer (favorable)
CNV	UTY	Ubiquitously transcribed tetratricopeptide repeat containing, Y-linked	Loss/mutation of this gene (together with UTX) associated with multiple cancers [68]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2021 doi:10.20944/preprints202104.0395.v1

Type of Feature	Gene Name	Protein Description	Biological relevance [1]
	MACROD2	Mono-ADP ribosylhydrolase 2	This gene was suggested to have a potential role in tumorigenesis [69]
	WWOX	WW domain containing oxidoreductase	Prognostic marker in renal cancer (favorable) and breast cancer (unfavorable)
	DAZ2	Deleted in azoospermia 2	No role in cancer prognostic. Gene associated with male infertility
	KANK1	KN motif and ankyrin repeat domains 1	Gene involved in progression of several solid tumours [70]

[1] The protein description and biological importance were retrieved from The Human Proteins Atlas (https://www.proteinatlas.org/) and The Human Gene Database (https://www.genecards.org/). When this information was not listed in these databases, we presented the study that supports the biological relevance. Favourable and unfavourable is related to gene contribution for cancer progression.

3.3. SynPred model for drug combination prediction

After selecting the best parameters for both DL with Keras and ML with scikit-learn, we trained eleven models (four from Keras and seven from scikit-learn) with the full training set. With a class imbalance of over two negative samples for each positive, it was crucial to consider several metrics on independent test set evaluation. Prior to ensemble deployment, the best scikit-learn ML model in the independent test set was XGBoost (acc=0.84, prec=0.78, rec=0.68, AUROC=0.80, F1=0.73) followed by decision-tree-based (acc=0.83-0.84, prec=0.77-0.81, rec=0.62-0.64, AUROC=0.78, F1=0.70) and DL (acc=0.80, prec=0.69, rec=0.67, AUROC=0.76, F1=0.68) (**Table A9 and Table A10**). ML models for synergy prediction with good performance values were already developed using Extreme Randomized Trees (ETC) (AUROC=0.89-0.95, Area Under the Precision-Recall Curve=0.51-0.71) algorithms, although without the incorporation of multiomics data [19]. Contrarily, Random Forests (RF) and XGBoost-based models trained with multiomics data such as cell lines expression, mutation, CNV and/or methylation presented slightly lower performance values (AUROC=0.68-0.75, Weighted Average Pearson Correlations=0.32-0.39) [12,13]. Besides methodological variations, we should highlight that differences in the database used (NCI-ALMANAC versus O'Neil) and the inclusion of multiomics data could justify the differences in models' performance. In fact, we demonstrated that chromatin profiling, metabolomics, miRNA, CNV, expression, methylation, and antibodies data tend to have a lower contribution in the development of non-DL-based models (**Figure 2**). Given the importance of multiomics features for cell lines characterization and treatment response [71], we believe that their inclusion is of utmost importance for developing accurate synergy prediction models.

The eleven models previously selected and trained with the full training dataset were used to make ensemble predictors by attaining each class probability. The ensemble models trained were then analysed by the best performing metrics (**Table 4**). All the best performing models were DL models trained with Keras with dropout rate different from 0. Our final SynPred model achieved an accuracy, precision, recall, AUROC, and F1-score of 0.85, 0.77, 0.75, 0.82 and 0.76, respectively, on an independent test set. This model is an ensemble model of four DL-based models and seven ML-based models, attained with a DL model with 4 hidden layers of size 50, a dropout rate of 0.60 and PCA pre-processing of the omics features with replacement of the missing values with 0. We then compare our model performance metrics with the current state-of-art classification models that used neural networks for synergy drug combinations prediction and the O'Neil *et al.* dataset. Our model, SynPred, achieved much better precision (0.77) than other relevant methods in the area such as DeepSynergy (0.56) [15], and AuDNNsynergy (0.72) [16], although it

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2021 doi:10.20944/preprints202104.0395.v1

performed slightly worse in terms of accuracy (0.85 versus 0.92-0.93) and AUROC (0.82 versus 0.90-0.91). We hypothesize that the slightly lower predictive performance in accuracy and AUROC arises from the fact that we used a stricter full agreement for synergy class definition, while DeepSynergy and AuDNNsynergy only used the Loewe additivity model. Besides, the initial balance of the dataset used in our study, albeit skewed slightly towards the negatives, is much more balanced than the O'Neil dataset. This can easily lead to lower values in some metrics although producing an overall superior performance predictor. Notwithstanding, considering all the individual metrics, our model shows a very high performance, with a great balance between true positive and negative predictions.

Table 4. Best performing ensemble methods, with the result for the mean probability (averaged class probability from each of the eleven models trained with the full training set) shown as a control.

Method	Architecture	Dropout Rate	Subset	Acc	Prec	Rec	AUROC	F1
Mean	NA	NA	Train	0.99	0.97	0.99	0.98	0.98
			Test	0.85	0.66	0.81	0.83	0.73
DL	(25, 25, 25, 25, 25, 25, 25)	0.4	Train	1.00	0.99	0.99	1.00	0.99
			Test	0.85	0.71	0.79	0.83	0.75
	(10, 10)	0.5	Train	0.97	1.00	0.91	0.95	0.95
			Test	0.78	0.89	0.59	0.76	0.71
	(500, 500, 500, 500)	0.9	Train	0.70	0.00	1.00	0.85	0.00
			Test	0.69	0.00	1.00	0.84	0.00
	(10, 10, 10)	0.7	Train	1.00	1.00	1.00	1.00	1.00
			Test	0.83	0.55	0.87	0.85	0.67
	(50, 50, 50, 50)	0.6	Train	1.00	1.00	0.99	1.00	1.00
			Test	0.85	0.77	0.75	0.82	0.76

Abbreviations: NA, not applicable.

3.4. Benchmarking with DECREASE

The DECREASE database [20] containing 34 drugs, 13 cell lines, and 210 combinations was processed to attain the features associated with its data using the same pipeline applied to SynPred. The samples corresponding to MCF-10A and HEK293 cell lines were disregarded, as these were not present in CCLE, making the acquisition of multiomics features impossible. From the two original published datasets, one with 192 and the other with 18 samples, a joint dataset was generated with 188 drug-drug-cell line unique combinations. When calculating the class with our strict full-agreement requirements, all the possible samples turned out to be negative, which indicates that the dataset, although appropriate, seems to be limited in the range of synergy values for both classification and regression tasks. When predicting the class, our SynPred ensemble achieved 0.98 of accuracy.

3.5. Web-based application description

The method for predicting the type of combinatory effect in each dug pair-cell line sample is available as a web-based application at http://www.moreiralab.com/resources/synpred/. All the eleven described single models are deployed on user submission, as well as the ensemble approach. The user needs to submit two drugs as input in the *.smile format and selects from a dropdown menu, the primary body site corresponding to the tested cancer cell lines. The drugs are then subject to feature extraction by Mordred and a standard pre-processing (feature elimination and normalization) as thoroughly described in the material and methods section. The output, displayed in a downloadable heatmap, is a binary drug combination prediction effect (synergistic and non-synergistic) for each of the individual cell lines and the average of both models based on

 doi:10.20944/preprints202104.0395.v1

the prediction values. The results are returned to the provided email and displayed on the submission web page. Additionally, users can assess, explore, and visualize through different plots as well as export a summary of the synergy scores (calculated using ZIP, Bliss, HSA and Loewe metrics) by cell line used to develop the original dataset of SynPred. To our knowledge, this is the first webserver that can predict new drug synergy combinations without the need of uploading a partial or full dose-response matrix. This feature is an advantage compared with regression models implemented in webservers that need these types of data for drug combination response prediction [20,21].

4. Discussion

Synergistic anticancer drug combinations are a powerful tool to help tackle cancer drug resistance since they can simultaneously target multiple key molecules or pathways. The rational design of combination therapies is warranted to improve the efficacy, although this is a well-known time-consuming and expensive task. In recent years, ML algorithms' applicability for drug-repurposing or novel drug design has been essential to demonstrate the importance of *in silico* methodologies to help overcome this problem. A few classification ML models for predicting drug synergy combinations were already developed [12,13,15,16,19], although the suitability of most of them is hindered by the use of only one reference model for the calculation of drug synergy (e.g., Bliss, Loewe, HSA, or ZIP). Given the different sensitivity observed between these reference models in evaluating the degree of combination, a more comprehensive and rigorous approach that leverage all metrics to predict drug synergy is still needed.

This study introduced a new synergy prediction model, SynPred, that combines comprehensive multiomics data of cancer cell lines with physicochemical and structural features of drugs. This work is one of the first that takes full agreement class between the four most used synergy metrics and uses one of the most comprehensive and equilibrated databases in terms of class balance, the NCI-ALMANAC. Our top-ranked model, an ensemble developed with the best machine learning models, achieved state-of-the-art performance to predict synergistic drug combinations in an independent dataset. Besides, we provide the complete workflow in our GitHub coupled with a freely available and easy-to-use webserver that only requires two drug SMILES as inputs, thus alleviating the need of uploading a conventional and laborious dose-response matrix. SynPred can be a valuable tool to the scientifical and medical community for drug repurposing or *in-silico* discovery of new anticancer drug combinations.

Additionally, given the importance of multiomics data in cell line classification and therapy response, we combined all the available multiomics features in the CCLE database to explore their individual contribution to model development. The knowledge mined from this analysis demonstrates the capacity of different ML models to deal with multiomics data, with DL algorithms being much more able to learn and leverage this complex type of features. We found that the most ranked proteins in each of the most contributing multiomics features are important cancer biomarkers or have a role in tumorigenesis, demonstrating DNN models' capacity to capture their significance and use this information for the final model development. In the future, we expect to include protein-protein interactions data and network analysis to improve the model performance, aiming to identify drug combinations with potential new targets across different cell lines.

5. Conclusions

In summary, our Machine Learning model can accurately predict synergistic anticancer drug combinations simply using as input the two SMILES of the drugs being tested, not relying on the need of a time-consuming and laborious dose-response matrix. This is a non-time-consuming approach that can accelerate discovering new anticancer drugs combinations with synergistic activity.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Figure S1: Distribution of synergy scores (x-axis) along with the NCI-ALMANAC dataset according to each

 doi:10.20944/preprints202104.0395.v1

one of the reference models used. (A) ZIP, (B) Bliss, (C) HSA, and (D) Loewe, Table S1: Conditions for dimensionality reduction with autoencoders. Hidden and bottleneck layers definition according to the Number of Features (NoF), Table S2: Conditions for dimensionality reduction with autoencoders. Number of epochs of the autoencoder training according to either the Number of Samples (NoS) or NoF, Table S3: Final datasets to be subjected to training, Table S4: Gridsearch combination parameters using 10% on the training set with DL algorithms, Table S5: Gridsearch combination parameters using 10% on the training set with non-DL algorithms, Table S6: Gridsearch combination parameters of the ensemble neural network, Table S7: Gridsearch results regarding the runs with 10% randomly selected entries of the training dataset for the DL models trained with Keras, Table S8: Gridsearch results regarding the runs with 10% randomly selected entries of the training dataset for the ML models trained with scikit-learn, Table S9: Results for the best scikit-learn ML models, MLP, RF, ETC, SVM, SGD, KNN and XGBoost, when evaluated in the training set and an independent test set, Table S10: The results for DL with Keras, when evaluated in the training set and an independent test set, after the gridsearch from the training random subset.

Author Contributions: Conceptualization, I.S.M. and J.M.; methodology, A.J.P., J.M. and P.M.F.; software, A.J.P. and P.M.F.; validation, A.J.P. and J.M.; formal analysis, A.J.P. and J.M.; investigation, A.J.P., P.M.F. and J.M.; resources, A.J.P., P.M.F., J.M. and I.S.M.; data curation, A.J.P., P.M.F. and J.M.; writing—original draft preparation, P.M.F., J.M. and A.J.P.; writing—review and editing, J.M., A.J.P. and I.S.M.; visualization, A.J.P. and P.M.F.; supervision, I.S.M. and J.M.; project administration, I.S.M and J.M.; funding acquisition, I.S.M. All authors have read and agreed to the published version of the manuscript.

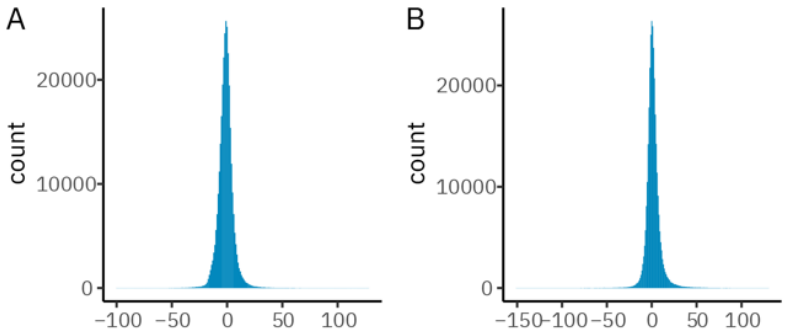
Funding: This research was funded by the European Regional Development Fund through the COMPETE 2020 - Operational Programme for Competitiveness and Internationalisation and Portuguese national funds via Fundação para a Ciência e a Tecnologia (FCT) [POCI-01-0145-FEDER-031356, UIDB/04539/2020, and LA/P/0058/2020]. FCT also supported A.J.P. with a PhD scholarship [SFRH/BD/144966/2019]. Funding for open access charge: Fundação para a Ciência e a Tecnologia [POCI-01-0145-FEDER-031356].

Data Availability Statement: SynPred is a free, open-source web-based application available for non-commercial use at http://www.moreiralab.com/resources/synpred/ without any login or registration requirements. The source code of the web-based application implementation is deposited in the GitHub repository (https://github.com/MoreiraLAB/synpred) to allow the stand-alone use of the application and further integration and comparison with other models.

Acknowledgments: Authors would like also to acknowledge STRATAGEM - New diagnostic and therapeutic tools against multidrug-resistant tumors, CA17104.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2021 doi:10.20944/preprints202104.0395.v1

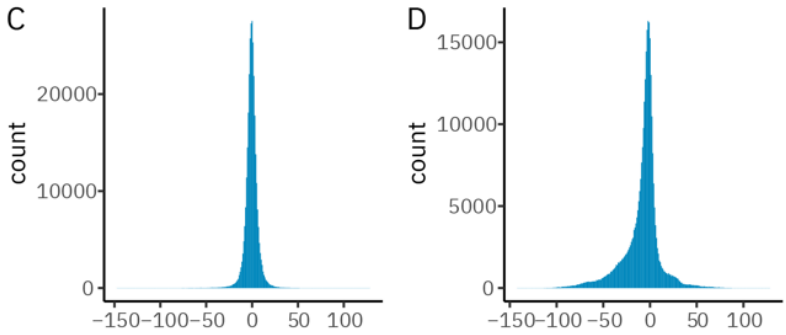


Figure A1. Distribution of synergy scores (x-axis) along with the NCI-ALMANAC dataset according to each one of the reference models used. (A) ZIP, (B) Bliss, (C) HSA, and (D) Loewe.

Table A1. Conditions for dimensionality reduction with autoencoders. Hidden and bottleneck layers definition according to the Number of Features (NoF).

Number of Features (NoF)	Hidden layers size	Bottleneck layer size
$NoF \leq 250$	HL1: $NoF * 2$ HL2: NoF HL4: NoF HL5: $NoF * 2$	$\frac{NoF}{2}$
$250 < NoF \leq 1000$	HL1: NoF HL2: $\frac{NoF}{2}$ HL4: $\frac{NoF}{2}$ HL5: NoF	$\frac{NoF}{10}$
$NoF \geq 1000$	HL1: $\frac{NoF}{2}$ HL2: $\frac{NoF}{4}$ HL4: $\frac{NoF}{4}$ HL5: $\frac{NoF}{2}$	$\frac{NoF}{50}$

Table A2. Conditions for dimensionality reduction with autoencoders. Number of epochs of the autoencoder training according to either the Number of Samples (NoS) or NoF.

Number of Samples (NoS)	Number of Features (NoF)	Number of epochs
$NoS \leq 100$	$NoF \leq 100$	1000
$100 < NoS < 1000$	$100 < NoF \leq 250$	250
$NoS \leq 1000$	$NoF \geq 250$	100

Table A3. Final datasets to be subjected to training.

Dataset name	Feature pre-processing	Sample prepro-cessing	Feature's size	Sample's size	Random sample split
PCA	PCA	Replace missing val-ues with 0	1.347	82.620	**Train:** 66.095 **Test:** 16.525
PCA_drop	PCA	Drop missing values	1.347	70.913	**Train:** 56.706 **Test:** 14.207
Autoencoder	Autoencoder	Replace missing val-ues with 0	4.229	82.620	**Train:** 66.095 **Test:** 16.525
Autoen-coder_drop	Autoencoder	Drop missing values	4.229	70.913	**Train:** 56.706 **Test:** 14.207

Table A4. Gridsearch combination parameters using 10% on the training set with DL algorithms.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2021 doi:10.20944/preprints202104.0395.v1

Datasets	Architecture	Dropout rate
PCA, PCA_drop, autoencoder, autoencoder_drop	(100,100), (100,100,100), (100,100,100,100), (500,500), (500,500,500), (500,500,500,500), (1000,1000), (1000,1000,1000), (1000,1000,1000,1000), (2500,2500), (2500,2500,2500), (2500,2500,2500,2500), (673,336), (673,336,84), (673,336,84,5), (2114,1057), (2114,1057,264), (2114,1057,264,16)	0.00, 0.25, 0.50, 0.75

Table A5. Gridsearch combination parameters using 10% on the training set with non-DL algorithms.

Method (package name)	Datasets	Parameters
RF (RandomForestClassifier)	PCA, PCA_drop, autoencoder, autoencoder_drop	n_estimators: (10,100,1000) max_depth: (None, 1, 5) min_samples_split: (2, 5,10) min_samples_leaf: (1, 2, 4)
ETC (ExtraTreesClassifier)		n_estimators: (10,100,1000) min_samples_split: (2, 5,10) max_depth: (None, 1, 5) min_samples_leaf: (1, 2, 4)
SVM (LinearSVC)		C: (0.1,0.5,1.0)
SGD (SGDClassifier)		penalty: (l2, l1, elasticnet) alpha: (0.00001, 0.0001, 0.001)
KNN (KNeighborsClassifier)		n_neighbors: (2, 5, 10, 25)
XGB (XGBClassifier)		max_depth: (2, 6, 10) alpha: (0, 0.25, 0.50) n_estimators: (10, 100, 1000)

Table A6. Gridsearch combination parameters of the ensemble neural network.

Architecture	Dropout rate
(10,10), (10,10,10), (10,10,10,10), (10,10,10,10,10), (10,10,10,10,10,10,10) (25,25), (25,25,25), (25,25,25,25), (25,25,25,25,25), (25,25,25,25,25,25,25), (50,50), (50,50,50), (50,50,50,50), (50,50,50,50,50), (50,50,50,50,50,50,50), (100,100), (100,100,100), (100,100,100,100), (100,100,100,100,100), (100,100,100,100,100,100,100), (500,500), (500,500,500), (500,500,500,500), (500,500,500,500,500), (500,500,500,500,500,500,500)	0.00, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90

Table A7. Gridsearch results regarding the runs with 10% randomly selected entries of the training dataset for the DL models trained with Keras.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2021 doi:10.20944/preprints202104.0395.v1

Variable under scope	Value under scope	Subset	Acc	Prec	Rec	AUC	F1
Dataset	PCA	Train	0.92	0.88	0.90	0.92	0.88
		Test	0.70	0.53	0.52	0.65	0.51
	PCA_drop	Train	0.94	0.90	0.91	0.93	0.90
		Test	0.71	0.54	0.49	0.65	0.50
	autoencoder	Train	0.73	0.54	0.22	0.59	0.29
		Test	0.69	0.39	0.15	0.54	0.20
	autoencoder_drop	Train	0.73	0.48	0.17	0.57	0.23
		Test	0.69	0.34	0.11	0.53	0.16
Dropout rate	0.00	Train	0.88	0.85	0.71	0.83	0.77
		Test	0.69	0.51	0.39	0.61	0.44
	0.25	Train	0.85	0.84	0.59	0.78	0.65
		Test	0.70	0.53	0.33	0.60	0.38
	0.5	Train	0.83	0.70	0.50	0.74	0.52
		Test	0.70	0.46	0.29	0.59	0.31
	0.75	Train	0.77	0.41	0.41	0.67	0.38
		Test	0.69	0.30	0.26	0.57	0.24
Architecture	(100,100)	Train	0.82	0.78	0.46	0.72	0.52
		Test	0.70	0.50	0.22	0.57	0.28
	(100,100,100)	Train	0.79	0.62	0.55	0.72	0.52
		Test	0.68	0.41	0.37	0.59	0.34
	(100,100,100,100)	Train	0.76	0.55	0.55	0.71	0.50
		Test	0.66	0.36	0.39	0.58	0.33
	(500,500)	Train	0.85	0.80	0.58	0.78	0.62
		Test	0.70	0.52	0.31	0.60	0.36
	(500,500,500)	Train	0.84	0.72	0.57	0.76	0.59
		Test	0.70	0.46	0.34	0.60	0.36
	(500,500,500,500)	Train	0.82	0.65	0.58	0.75	0.57
		Test	0.68	0.42	0.38	0.6	0.36
	(1000,1000)	Train	0.86	0.75	0.59	0.78	0.63
		Test	0.70	0.46	0.33	0.6	0.36
	(1000,1000,1000)	Train	0.84	0.70	0.55	0.76	0.59
		Test	0.70	0.43	0.31	0.59	0.34
	(1000,1000,1000,1000)	Train	0.84	0.68	0.57	0.76	0.59
		Test	0.70	0.43	0.32	0.60	0.35
	(2500,2500)	Train	0.85	0.73	0.60	0.78	0.63
		Test	0.70	0.52	0.34	0.60	0.37
	(2500,2500,2500)	Train	0.84	0.74	0.55	0.76	0.59
		Test	0.70	0.48	0.31	0.60	0.35
	(2500,2500,2500,2500)	Train	0.84	0.68	0.56	0.76	0.60
		Test	0.70	0.44	0.32	0.60	0.36
	(673,336)	Train	0.85	0.79	0.59	0.78	0.62
		Test	0.70	0.50	0.32	0.60	0.36
	(673,336,84)	Train	0.84	0.65	0.54	0.76	0.57
		Test	0.70	0.40	0.30	0.59	0.33
	(673,336,84,5)	Train	0.81	0.52	0.41	0.70	0.44
		Test	0.70	0.33	0.23	0.57	0.25
	(2114,1057)	Train	0.86	0.80	0.60	0.78	0.64
		Test	0.70	0.50	0.34	0.60	0.37
	(2114,1057,264)	Train	0.84	0.74	0.55	0.76	0.59
		Test	0.70	0.47	0.30	0.59	0.35

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2021 doi:10.20944/preprints202104.0395.v1

Variable under scope	Value under scope	Subset	Acc	Prec	Rec	AUC	F1
	(2114,1057,264,16)	Train	0.84	0.70	0.53	0.75	0.57
		Test	0.70	0.45	0.30	0.59	0.34

Table A8. Gridsearch results regarding the runs with 10% randomly selected entries of the training dataset for the ML models trained with scikit-learn.

Subset	Method	Parameters	Subset	Acc	Prec	Rec	AUC	F1
PCA	RF	max_depth:None, min_samples_leaf:2, min_samples_split:2, n_estimators:1000	Train	1.00	1.00	0.99	1.00	1.00
			Test	0.76	0.69	0.39	0.66	0.50
	ETC	max_depth:None, min_samples_leaf:2, min_samples_split:5, n_estimators:1000	Train	1.00	1.00	1.00	1.00	1.00
			Test	0.77	0.67	0.48	0.69	0.56
	SVM	C:0.5	Train	0.68	0.48	0.39	0.60	0.43
			Test	0.66	0.44	0.37	0.58	0.40
	SGD	alpha:1e-05, penalty:l1	Train	0.67	0.46	0.49	0.62	0.47
			Test	0.62	0.40	0.44	0.57	0.42
	KNN	n_neighbors:25	Train	0.71	0.60	0.17	0.56	0.27
			Test	0.69	0.49	0.14	0.54	0.22
	XGB	alpha:0, max_depth:6, n_estimators:1000	Train	1.00	1.00	1.00	1.00	1.00
			Test	0.75	0.64	0.50	0.68	0.56
PCA_drop	RF	max_depth:None, min_samples_leaf:1, min_samples_split:5, n_estimators = 1000	Train	1.00	1.00	1.00	1.00	1.00
			Test	0.76	0.69	0.41	0.66	0.51
	ETC	max_depth:None, min_samples_leaf:1, min_samples_split:10, n_estimators: 100	Train	0.99	1.00	0.97	0.98	0.98
			Test	0.76	0.68	0.42	0.67	0.52
	SVM	C:0.1	Train	0.61	0.40	0.55	0.60	0.46
			Test	0.60	0.40	0.54	0.58	0.46
	SGD	alpha:0.001	Train	0.68	0.48	0.43	0.61	0.45
			Test	0.66	0.45	0.41	0.59	0.43
	KNN	n_neighbors:10	Train	0.73	0.69	0.23	0.59	0.35
			Test	0.69	0.53	0.17	0.55	0.26
	XGB	alpha:0.0, max_depth:6, n_estimators:100	Train	1.00	1.00	1.00	1.00	1.00
			Test	0.75	0.63	0.45	0.66	0.52
autoencoder	RF	max_depth:None, min_samples_leaf:1, min_samples_split:2, n_estimators:1000	Train	1.00	1.00	1.00	1.00	1.00
			Test	0.74	0.67	0.34	0.63	0.45
	ETC	max_depth:None, min_samples_leaf:1, min_samples_split:10, n_estimators:100	Train	1.00	1.00	0.99	0.99	0.99
			Test	0.74	0.64	0.35	0.63	0.46
	SVM	C:0.1	Train	0.62	0.38	0.39	0.56	0.38
			Test	0.62	0.38	0.37	0.55	0.37

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2021 doi:10.20944/preprints202104.0395.v1

Subset	Method	Parameters	Subset	Acc	Prec	Rec	AUC	F1
Autoencoder_drop	SGD	alpha:0.0001, penalty:l1	Train	0.67	0.46	0.55	0.64	0.50
			Test	0.65	0.45	0.53	0.62	0.49
	KNN	n_neighbors:2	Train	0.83	1.00	0.46	0.73	0.63
			Test	0.69	0.51	0.20	0.56	0.29
	XGB	alpha:0.5, max_depth:6, n_estimators:100	Train	1.00	1.00	1.00	1.00	1.00
			Test	0.75	0.64	0.44	0.67	0.52
	RF	max_depth:None	Train	1.00	1.00	1.00	1.00	1.00
			Test	0.74	0.65	0.35	0.63	0.45
	ETC	max_depth:None, min_samples_leaf:1, min_samples_split:10, n_estimators = 1000	Train	1.00	1.00	0.99	0.99	0.99
			Test	0.74	0.66	0.36	0.64	0.47
	SVM	C:0.1	Train	0.62	0.40	0.44	0.57	0.42
			Test	0.62	0.40	0.44	0.57	0.42
	SGD	alpha:0.001, penalty:l1	Train	0.68	0.48	0.51	0.63	0.50
			Test	0.67	0.47	0.48	0.62	0.47
	KNN	n_neighbors:10	Train	0.74	0.70	0.26	0.61	0.38
			Test	0.69	0.52	0.18	0.55	0.27
	XGB	alpha:0, max_depth:6, n_estimators:100	Train	1.00	1.00	1.00	1.00	1.00
			Test	0.74	0.62	0.44	0.66	0.51

Table A9. Results for the best scikit-learn ML models, MLP, RF, ETC, SVM, SGD, KNN and XGBoost, when evaluated in the training set and an independent test set.

Machine Learning Method	Subset	Acc	Prec	Rec	AUC	F1
MLP	Train	0.93	0.89	0.88	0.91	0.88
	Test	0.80	0.69	0.67	0.76	0.68
RF	Train	0.99	1.00	0.96	0.98	0.98
	Test	0.84	0.81	0.62	0.78	0.70
ETC	Train	0.99	0.99	0.99	0.99	0.99
	Test	0.83	0.77	0.64	0.78	0.70
SVM	Train	0.61	0.38	0.44	0.56	0.41
	Test	0.60	0.38	0.44	0.56	0.41
SGD	Train	0.64	0.41	0.40	0.57	0.40
	Test	0.63	0.40	0.39	0.57	0.40
KNN	Train	0.75	0.69	0.34	0.63	0.45
	Test	0.72	0.62	0.29	0.61	0.40
XGBoost	Train	1.00	1.00	1.00	1.00	1.00
	Test	0.84	0.78	0.68	0.80	0.73

Table A10. The results for DL with Keras, when evaluated in the training set and an independent test set, after the gridsearch from the training random subset.

Architecture	Subset	Acc	Prec	Rec	AUC	F1
(100,100,100,100)	Train	0.95	0.92	0.92	0.94	0.92
	Test	0.77	0.63	0.61	0.73	0.62
(2500,2500)	Train	0.94	0.92	0.89	0.93	0.91
	Test	0.80	0.69	0.65	0.76	0.67
(2114,1057)	Train	0.95	0.91	0.92	0.94	0.91
	Test	0.80	0.68	0.67	0.76	0.67

Architecture	Subset	Acc	Prec	Rec	AUC	F1
(2114,1057,264,16)	Train	0.94	0.91	0.90	0.93	0.91
	Test	0.80	0.68	0.65	0.76	0.66

References

1. International Agency for Research on Cancer. GLOBOCAN - Cancer Tomorrow via Global Cancer Observatory, **2020**, https://gco.iarc.fr/tomorrow/en.
2. Vasan, N.; Baselga, J.; Hyman, D.M. A View on Drug Resistance in Cancer. *Nature* **2019**, *575*, 299–309.
3. Chatterjee, N.; Bivona, T.G. Polytherapy and Targeted Cancer Drug Resistance. *Trends in Cancer* **2019**, *5*, 170–182.
4. Roell, K.R.; Reif, D.M.; Motsinger-Reif, A.A. An Introduction to Terminology and Methodology of Chemical Synergy—Perspectives from Across Disciplines. *Front. Pharmacol.* **2017**, *8*, 158.
5. Brandão, M.; Pondé, N.F.; Poggio, F.; Kotecki, N.; Salis, M.; Lambertini, M.; de Azambuja, E. Combination Therapies for the Treatment of HER2-Positive Breast Cancer: Current and Future Prospects. *Expert Review of Anticancer Therapy* **2018**, *18*, 629–649.
6. Westerweel, P.E.; te Boekhorst, P.A.W.; Levin, M.-D.; Cornelissen, J.J. New Approaches and Treatment Combinations for the Management of Chronic Myeloid Leukemia. *Front. Oncol.* **2019**, *9*, 665.
7. Xu, J.; Qiu, Y. Current Opinion and Mechanistic Interpretation of Combination Therapy for Castration-Resistant Prostate Cancer. *Asian J Androl* **2019**, *21*, 270.
8. Ribas, A.; Lawrence, D.; Atkinson, V.; Agarwal, S.; Miller, W.H.; Carlino, M.S.; Fisher, R.; Long, G.V.; Hodi, F.S.; Tsoi, J.; et al. Combined BRAF and MEK Inhibition with PD-1 Blockade Immunotherapy in BRAF-Mutant Melanoma. *Nat Med* **2019**, *25*, 936–940.
9. Wang, Z.; Deisboeck, T.S. Dynamic Targeting in Cancer Treatment. *Front. Physiol.* **2019**, *10*, 96.
10. Wang, Z.; Li, H.; Guan, Y. Machine Learning for Cancer Drug Combination. *Clin. Pharmacol. Ther.* **2020**, *107*, 749–752.
11. Janizek, J.D.; Celik, S.; Lee, S.-I. Explainable Machine Learning Prediction of Synergistic Drug Combinations for Precision Cancer Medicine; Cancer Biology, *biorXiv*, **2018**, doi: http://doi.org/10.1101/331769, 27 May 2018, pre-print: not peer-reviewed.
12. Li, H.; Li, T.; Quang, D.; Guan, Y. Network Propagation Predicts Drug Synergy in Cancers. *Cancer Res* **2018**, 5446-5457.
13. Celebi, R.; Bear Don't Walk, O.; Movva, R.; Alpsoy, S.; Dumontier, M. In-Silico Prediction of Synergistic Anti-Cancer Drug Combinations Using Multi-Omics Data. *Sci Rep* **2019**, *9*, 8949.
14. Malyutina, A.; Majumder, M.M.; Wang, W.; Pessia, A.; Heckman, C.A.; Tang, J. Drug Combination Sensitivity Scoring Facilitates the Discovery of Synergistic and Efficacious Drug Combinations in Cancer. *PLoS Comput Biol* **2019**, *15*, e1006752.
15. Preuer, K.; Lewis, R.P.I.; Hochreiter, S.; Bender, A.; Bulusu, K.C.; Klambauer, G. DeepSynergy: Predicting Anti-Cancer Drug Synergy with Deep Learning. *Bioinformatics* **2018**, *34*, 1538–1546.
16. Zhang, T.; Zhang, L.; Payne, P.R.O.; Li, F. Synergistic Drug Combination Prediction by Integrating Multi-Omics Data in Deep Learning Models. In: Markowitz,J. (eds) Translational Bioinformatics for Therapeutic Development. Methods in Molecular Biology, **2020**, vol 2194. Humana, New York, NY, Vol. 2194, pp. 223-238.
17. Kuru, H.I.; Tastan, O.; Cicek, A.E. MatchMaker: A Deep Learning Framework for Drug Synergy Prediction; Bioinformatics, *biorXiv*, **2020**, doi: http://doi.org/10.1101/2020.05.24.113241, 30 May 2020, pre-print: not peer-reviewed.
18. Zhang, H.; Feng, J.; Zeng, A.; Payne, P.; Li, F. Predicting Tumor Cell Response to Synergistic Drug Combinations Using a Novel Simplified Deep Learning Model; Bioinformatics, *biorXiv*, **2020**, doi: http://doi.org/10.1101/2020.04.10.036491, 13 April 2020, pre-print: not peer-reviewed.
19. Gilvary, C.; Dry, J.R.; Elemento, O. Multi-Task Learning Predicts Drug Combination Synergy in Cells and in the Clinic; Cancer Biology, *biorXiv*, **2019**, doi: http://doi.org/10.1101/576017, 13 March 2019, pre-print: not peer-reviewed.
20. Ianevski, A.; Giri, A.K.; Gautam, P.; Kononov, A.; Potdar, S.; Saarela, J.; Wennerberg, K.; Aittokallio, T. Prediction of Drug Combination Effects with a Minimal Set of Experiments. *Nat Mach Intell* **2019**, *1*, 568–577.
21. Zagidullin, B.; Aldahdooh, J.; Zheng, S.; Wang, W.; Wang, Y.; Saad, J.; Malyutina, A.; Jafari, M.; Tanoli, Z.; Pessia, A.; et al. DrugComb: An Integrative Cancer Drug Combination Data Portal. *Nucleic Acids Research* **2019**, *47*, W43–W51.
22. Holbeck, S.L.; Camalier, R.; Crowell, J.A.; Govindharajulu, J.P.; Hollingshead, M.; Anderson, L.W.; Polley, E.; Rubinstein, L.; Srivastava, A.; Wilsker, D.; et al. The National Cancer Institute ALMANAC: A Comprehensive Screening Resource for the Detection of Anticancer Drug Pairs with Enhanced Therapeutic Activity. *Cancer Research* **2017**, *77*, 3564–3576.
23. Shoemaker, R.H. The NCI60 Human Tumour Cell Line Anticancer Drug Screen. *Nature Reviews Cancer* **2006**, *6*, 813–823.
24. Division of Cancer Treatment & Diagnosis, DCTD Tumor Repository - A Catalog of in Vitro Cell Lines, Transplantable Animal and Human Tumors and Yeast; National Cancer Institute at Frederick, **2020**, https://dtp.cancer.gov/organization/btb/docs/DCTDTumorRepositoryCatalog.pdf.
25. He, L.; Kulesskiy, E.; Saarela, J.; Turunen, L.; Wennerberg, K.; Aittokallio, T.; Tang, J. Methods for High-throughput Drug Combination Screening and Synergy Scoring. In *Methods in Molecular Biology (Clifton, N.J.)*; 2018; Vol. 1711, pp. 351–398 ISBN 978-1-4939-7493-1.
26. Bliss, C.I. The Toxicity of Poisons Applied Jointly. *Annals of Applied Biology* **1939**, *26*, 585–615.
27. Loewe, S.; Muischnek, H. Über Kombinationswirkungen. *Archiv für Experimentelle Pathologie und Pharmakologie* **1926**, *114*, 313–26.

28. Chou, T.-C. Drug Combination Studies and Their Synergy Quantification Using the Chou-Talalay Method. *Cancer Research* **2010**, *70*, 440–446.
29. Foucquier, J.; Guedj, M. Analysis of Drug Combinations: Current Methodological Landscape. *Pharmacology research & perspectives* **2015**, *3*, e00149.
30. Yadav, B.; Wennerberg, K.; Aittokallio, T.; Tang, J. Searching for Drug Synergy in Complex Dose-Response Landscapes Using an Interaction Potency Model. *Computational and structural biotechnology journal* **2015**, *13*, 504–13.
31. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B.A.; Thiessen, P.A.; Yu, B.; et al. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Research* **2018**, *47*, D1102–D1109.
32. Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *Journal of Cheminformatics* **2018**, *10*, 4.
33. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.
34. Barretina, J.; Caponigro, G.; Stransky, N.; Venkatesan, K.; Margolin, A.A.; Kim, S.; Wilson, C.J.; Lehár, J.; Kryukov, G.V.; Sonkin, D.; et al. The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity. *Nature* **2012**, *483*, 603–607.
35. Bairoch, A. The Cellosaurus, a Cell-Line Knowledge Resource. *J Biomol Tech* **2018**, *29*, 25–38.
36. Ghandi, M.; Huang, F.W.; Jané-Valbuena, J.; Kryukov, G.V.; Lo, C.C.; McDonald, E.R.; Barretina, J.; Gelfand, E.T.; Bielski, C.M.; Li, H.; et al. Next-Generation Characterization of the Cancer Cell Line Encyclopedia. *Nature* **2019**, *569*, 503–508.
37. The International HapMap Consortium The International HapMap Project. *Nature* **2003**, *426*, 789–796.
38. Venkatraman, E.S.; Olshen, A.B. A Faster Circular Binary Segmentation Algorithm for the Analysis of Array CGH Data. *Bioinformatics* **2007**, *23*, 657–663.
39. Chaudhary, K.; Poirion, O.B.; Lu, L.; Garmire, L.X. Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. *Clin Cancer Res* **2018**, *24*, 1248–1259.
40. Zhang, L.; Lv, C.; Jin, Y.; Cheng, G.; Fu, Y.; Yuan, D.; Tao, Y.; Guo, Y.; Ni, X.; Shi, T. Deep Learning-Based Multi-Omics Data Integration Reveals Two Prognostic Subtypes in High-Risk Neuroblastoma. *Front. Genet.* **2018**, *9*, 477.
41. Simidjievski, N.; Bodnar, C.; Tariq, I.; Scherer, P.; Andres Terre, H.; Shams, Z.; Jamnik, M.; Liò, P. Variational Autoencoders for Cancer Data Integration: Design Principles and Computational Practice. *Front. Genet.* **2019**, *10*, 1205.
42. Hinton, G.E.; Salakhutdinov, R.R. Reducing the Dimensionality of Data with Neural Networks. *Science* **2006**, *313*, 504–507.
43. Wang, Y.; Yao, H.; Zhao, S. Auto-Encoder Based Dimensionality Reduction. *Neurocomputing* **2016**, *184*, 232–242.
44. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv*, **2015**, doi: http://arxiv.org/pdf/1603.04467, 16 March 2016, pre-print: not peer-reviewed.
45. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv*, **2017**, doi: http://arxiv.org/pdf/1412.6980, 16 March 2016, pre-print: not peer-reviewed.
46. Meng, C.; Oana A. Zeleznik; Gerhard G. Thallinger; Bernhard Kuster; Amin M. Gholami; Aedín C. Culhane Dimension Reduction Techniques for the Integrative Analysis of Multi-Omics Data. *Briefings in Bioinformatics* **2016**, *17*, 628–641.
47. Hinton, G.E. Connectionist Learning Procedures. *Artificial intelligence* **1989**, *40*, 185–234.
48. Glorot, X.; Bengio, Y. Understanding the Difficulty of Training Deep Feedforward Neural Networks. *International Conference on Artificial Intelligence and Statistics* **2010**, 9, 429-256.
49. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on Imagenet Classification. *arXiv*, **2015**, doi: http://arxiv.org/pdf/1502.01852, 06 February 2015, pre-print: not peer-reviewed.
50. Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32.
51. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely Randomized Trees. *Machine Learning* **2006**, *63*, 3–42.
52. Preto, A.J.; Moreira, I.S. SPOTONE: Hot Spots on Protein Complexes with Extremely Randomized Trees via Sequence-Only Features. *IJMS* **2020**, *21*, 7281.
53. Fan, R.E.; Chang, K.W.; Hsieh, C.J.; Wang, X.R.; Lin, C.J. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research* **2008**, *9*, 1871–1874.
54. Zadrozny, B.; Elkan, C. Transforming Classifier Scores into Accurate Multiclass Probability Estimates. *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* **2002**, 694–699.
55. Altman, N.S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician* **1992**, *46*, 175–185.
56. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **2016**, 785–794.
57. Noemí DeCastro-García; Ángel Luis Muñoz Castañeda; David Escudero García; Miguel V. Carriegos Effect of the Sampling of a Dataset in the Hyperparameter Optimization Phase over the Efficiency of a Machine Learning Algorithm. *Advances in Complex Systems and Their Applications to Cybersecurity* **2019**.
58. Swersky, K.; Snoek, J.; Adams, R.P. Multi-Task Bayesian Optimization. *NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems* **2013**, *2*, 2004–2012.
59. O'Neil, J.; Benita, Y.; Feldman, I.; Chenard, M.; Roberts, B.; Liu, Y.; Li, J.; Kral, A.; Lejnine, S.; Loboda, A.; et al. An Unbiased Oncology Compound Screen to Identify Novel Combination Strategies. *Mol Cancer Ther* **2016**, *15*, 1155–1162.

60. Forcina, G.C.; Conlon, M.; Wells, A.; Cao, J.Y.; Dixon, S.J. Systematic Quantification of Population Cell Death Kinetics in Mammalian Cells. *Cell Systems* **2017**, *4*, 600-610.e6.
61. Mathews Griner, L.A.; Guha, R.; Shinn, P.; Young, R.M.; Keller, J.M.; Liu, D.; Goldlust, I.S.; Yasgar, A.; McKnight, C.; Boxer, M.B.; et al. High-Throughput Combinatorial Screening Identifies Drugs That Cooperate with Ibrutinib to Kill Activated B-Cell–like Diffuse Large B-Cell Lymphoma Cells. *Proc Natl Acad Sci USA* **2014**, *111*, 2349–2354.
62. Korobov, M.; Lopuhin K. ELI5 Documentation. **2021**, https://readthedocs.org/projects/eli5/downloads/pdf/latest/.
63. Plotly Technologies Inc. Collaborative data science *Plotly*; Plotly Technologies Inc. Collaborative data science: Montréal, QC, **2015**.
64. Grinberg, M. Flask Web Development: Developing Web Applications with Python, 2nd Edition. Reilly Media, Inc, **2018**.
65. Omar, N.N.; Tash, R.F.; Shoukry, Y.; ElSaeed, K.O. Breaking the Ritual Metabolic Cycle in Order to Save Acetyl CoA: A Potential Role for Mitochondrial Humanin in T2 Bladder Cancer Aggressiveness. *Journal of the Egyptian National Cancer Institute* **2017**, *29*, 69–76.
66. Tian, B.-X.; Sun, W.; Wang, S.-H.; Liu, P.-J.; Wang, Y.-C. Differential Expression and Clinical Significance of COX6C in Human Diseases. *Am J Transl Res* **2021**, *13*, 1–10.
67. Widschwendter, M.; Zikan, M.; Wahl, B.; Lempiäinen, H.; Paprotka, T.; Evans, I.; Jones, A.; Ghazali, S.; Reisel, D.; Eichner, J.; et al. The Potential of Circulating Tumor DNA Methylation Analysis for the Early Detection and Management of Ovarian Cancer. *Genome Med* **2017**, *9*, 116.
68. Gozdecka, M.; Meduri, E.; Mazan, M.; Tzelepis, K.; Dudek, M.; Knights, A.J.; Pardo, M.; Yu, L.; Choudhary, J.S.; Metzakopian, E.; et al. UTX-Mediated Enhancer and Chromatin Remodeling Suppresses Myeloid Leukemogenesis through Noncatalytic Inverse Regulation of ETS and GATA Programs. *Nat Genet* **2018**, *50*, 883–894.
69. Feijs, K.; Cooper, C.; Žaja, R. The Controversial Roles of ADP-Ribosyl Hydrolases MACROD1, MACROD2 and TARG1 in Carcinogenesis. *Cancers* **2020**, *12*, 604.
70. Gu, Y.; Zhang, M. Upregulation of the Kank1 Gene Inhibits Human Lung Cancer Progression in Vitro and in Vivo. *Oncol Rep* **2018**, *40*, 1243-1250.
71. Berger, M.F.; Mardis, E.R. The Emerging Clinical Relevance of Genomics in Cancer Medicine. *Nat Rev Clin Oncol* **2018**, *15*, 353–365.