SCIENTIFIC REPORTS

OPEN

In-silico Prediction of Synergistic Anti-Cancer Drug Combinations Using Multi-omics Data

Received: 8 January 2019
Accepted: 29 May 2019
Published online: 20 June 2019

Remzi Celebi[1], Oliver Bear Don't Walk IV[2], Rajiv Movva[3], Semih Alpsoy[4] & Michel Dumontier[1]

Chemotherapy is a routine treatment approach for early-stage cancers, but the effectiveness of such treatments is often limited by drug resistance, toxicity, and tumor heterogeneity. Combination chemotherapy, in which two or more drugs are applied simultaneously, offers one promising approach to address these concerns, since two single-target drugs may synergize with one another through interconnected biological processes. However, the identification of effective dual therapies has been particularly challenging; because the search space is large, combination success rates are low. Here, we present our method for DREAM AstraZeneca-Sanger Drug Combination Prediction Challenge to predict synergistic drug combinations. Our approach involves using biologically relevant drug and cell line features with machine learning. Our machine learning model obtained the primary metric = 0.36 and the tie-breaker metric = 0.37 in the extension round of the challenge which was ranked in top 15 out of 76 submissions. Our approach also achieves a mean primary metric of 0.39 with ten repetitions of 10-fold cross-validation. Further, we analyzed our model's predictions to better understand the molecular processes underlying synergy and discovered that key regulators of tumorigenesis such as TNFA and BRAF are often targets in synergistic interactions, while MYC is often duplicated. Through further analysis of our predictions, we were also ble to gain insight into mechanisms and potential biomarkers of synergistic drug pairs.

The last decade has seen a revolution in the discovery of small molecule cancer drugs[1,2]. Drug development has trended away from the one-drug-fits-all paradigm towards a diverse array of targeted agents that exploit specific knowledge of individual tumors[3]. While this approach can provide success, the confinement of drugs to a single target fails to take into account the complex etiologies of many cancers[4]. Specifically, the single target model is highly susceptible to the genetic diversity of tumors; one cell with a resistance-conferring mutation can cause complete evolution of the tumor in a few months[5]. Thus, under the current system of drug development, acquired resistance and intratumor heterogeneity will continue to hinder effective and permanent cancer treatment.

Theoretically, combination drug therapy can address many of the limitations that single target agents cannot. The underlying rationale is that drugs targeting different components of an interconnected network (either a single pathway or two related pathways) can more effectively suppress a certain biological process[6]. Several model studies have supported this hypothesis: Simultaneous drug treatments are far more robust to mutation, since two unlikely independent events must happen instead of one (*i.e.*, $p_1 \approx 10^{-6}$, so $p_2 \approx p_1^2 \approx 10^{-12}$)[4]. Further, even in the presence of cross-resistance mutations, combination therapy still offers potential for treatment[7,8].

However, tangible development of drug combinations has lagged behind theoretical discussion, primarily because identifying successful combinations is a difficult problem. More often than not, simultaneous administration results in no interaction between drugs and thus no net beneficial effect (termed *additivity*), or adverse interactions leading to decreased efficiency and possible toxicity (*antagonism*). *Synergistic* combinations are drugs that amplify each other's activity, leading to elevated effects at low concentrations and, thus, reduced toxicity[9]. Picking out these synergistic combinations from the millions of possibilities requires meticulous experimentation and prohibitive levels of time and money[10].

To aid in the identification and development of combination therapies, a few *in silico* methods have been proposed to predict successful drug pairs for further experimental tests in recent years. DrugComboRanker,

[1]Maastricht University, Institute of Data Science, Maastricht, Netherlands. [2]Columbia University, Department of Biomedical Informatics, New York City, USA. [3]Stanford University, Department of Genetics, Palo Alto, USA. [4]Turkish-German University, Department of Molecular Biotechnology, Istanbul, Turkey. Correspondence and requests for materials should be addressed to R.C. (email: remzi.celebi@maastrichtuniversity.nl)

the method by[11], identifies synergistic drugs that target different signaling modules of a given disease network. However, this approach is limited to identification only of combinations that have known disease pathway interactions and is also highly susceptible to false positive pathway cross-talk. A method called DIGRE, proposed by[12], works by identifying secondary drugs that are more effective on cells post-treatment with the first drug. However, DIGRE relies on knowledge of differentially expressed genes post-drug treatment, for which data is not widely available or practical to obtain in a clinical setting; perhaps because it considers synergy for sequential drug treatment, which has been shown to be ineffective at overcoming tumor resistance[7]. Another approach, RACS, identifies labelled drug combinations that are most similar to unlabelled combinations in the context of seven target-related features, and then incorporates overlap of differentially expressed gene signatures to predict synergy[13]. Like DIGRE, RACS also relies on elusive post-treatment data, but its feature set also limits its predictions to direct drug-protein interactions; our work on compensatory pathway analysis shows that these first-order synergistic effects are far from exhaustive. Huang *et al.*[14] developed a computational model to predict drug combinations by using clinical side effects (SE) from post-marketing surveillance and the drug label. A database including 349 approved drug combinations was constructed with integration of drug information from SIDER, TWOSIDES, and DCDB sources. Logistic regression prediction model with 10-fold cross validation was utilized to determine predictive power of drug-drug combinations (DDC) relying on top 3 SE features identified by decision tree: pneumonia, haemorrhage rectum, and retinal bleeding. This approach does not use gene expression, pathway, and protein-domains information. They only look for marketed drugs in combination. Li *et al.*[15] aimed to predict synergistic drug combinations with various features including drug chemical structure similarity, target distance in protein-protein network, and targeted pathway similarity. They also used fifteen pharmacogenomics features using drug treated gene expression profiles and built a prediction model for synergistic drug combination using the Random Forest method. They only used gene expression profile data of MCF7 cell line following drug treatment on the cell line from CMap. Zhao *et al.*[16] developed a computational method for predicting synergistic activity of drugs used in combinations by integrating molecular and pharmacological data. They used STITCH, Drugbank and TTD databases to obtain compound-protein interactions. As a result of their analyses, they predicted 16 possible drug combinations. They reported that 11 out of their 16 predictions had already been identified as effective in the literature.

Predicting synergistic combinations using a wide range of cancer cell lines and drugs is much more challenging due to heterogeneity at molecular, chemical and biological level. Prior approaches have been limited by small dataset size and low data variety that could not reflect the extent of the standard prediction challenge. The DREAM AstraZeneca-Sanger Drug Combination Prediction Challenge offered one of the largest combinatorial cell line screening datasets, which also includes molecular data and chemical/biological data[17]. The dataset quantifies drug synergy with the Loewe model, defined as calculating the excess cell kill rate over the expected additive kill rate when the drug combination, is administered to cancer cell lines. The molecular information contained somatic mutations, copy-number alterations, DNA methylation, and gene expression profiles measured before drug treatment; and the compound information included putative drug targets, and where available, chemical properties. Here, we present our machine learning model developed to predict synergistic drug combinations for the DREAM AstraZeneca-Sanger Drug Combination Prediction Challenge. In order to best encapsulate the biological patterns underlying this synergy, we explored the most predictive and biologically relevant features for the prediction of drug synergies and trained a machine learning model using the features that characterise drugs and cell lines.

Our submission for Subchallenge 1 A of the DREAM AstraZeneca-Sanger Drug Combination Prediction Challenge was ranked in top 15 out of 76 submissions according to the primary metric used by the challenge organizers. Our machine learning model obtained the primary metric = 0.36 and the tie-breaker metric = 0.37 in the extension round of the challenge {https://www.synapse.org/#!Synapse:syn4231880/wiki/411305}. Our approach also achieves a mean primary metric of 0.39 with ten repetitions of 10-fold cross-validation. Through further analysis of our predictions, we were also able to gain insight into mechanisms and potential biomarkers of synergistic drug pairs. By automatically combining single-target drugs for synergistic therapy, our work paves the way towards efficient and widespread combinatorial cancer treatment.

Results

Model performance. *Machine learning models accurately predict synergy.* We began by comparing model performance training on the complete feature set (111,168 features) that contains all expression, copy number, and mutation data, and training on the abridged feature set (2121 features), to see if the latter completely encoded the relevant information. The abridged feature set includes the drug and cell line features that are biologically informative for drug synergy which have been extracted to train the machine learning models. Indeed, the abridged set performs equivalently for regression tasks across all five models (Fig. 1a), supporting our hypothesis that biologically informed feature curation can reduce overfitting and improve predictions. Thus, subsequent training was performed using the abridged set, both for its better performance and its faster training time.

Our next goal was to identify the most accurate model. We performed ten trials of 10-fold CV, and XGBoost and Random Forest stood out significantly from the others for regression ($P < 5 \times 10^{-4}$, two-sample z-test). With post-tuning, XGBoost achieved a weighted average Pearson correlation of $WAPCC = 0.39$; random forest was the next best model with $WAPCC = 0.36$ (Fig. 1b). Since XGboost was significantly better than Random Forest ($P < 0.01$, two-sample z-test), we used the XGBoost model for all downstream analyses.

Tuning XGBoost parameters. XGBoost performance and training time is heavily affected by choice of parameters[18]. We optimized four of these variables that cause most deviation: number of trees used (*n_estimators*), the maximum number of decisions (*max_depth*) for each tree, subsample ratio of observations and features (*subsample* and *colsample_bytree*) used to build each tree. Holding *n_estimators* constant, we varied the other three

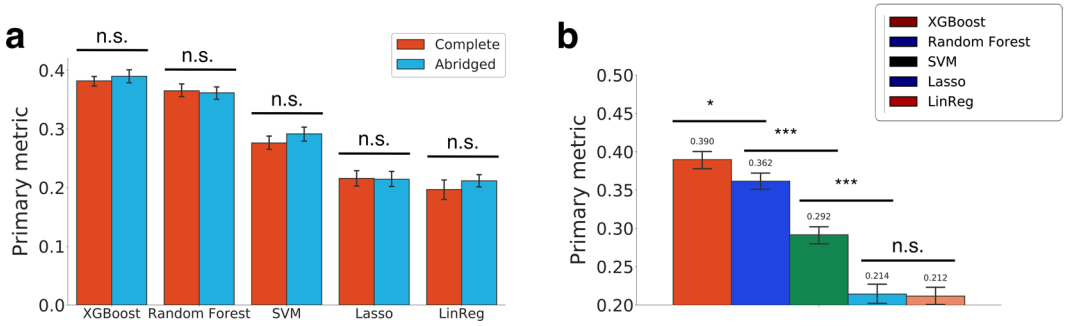

Figure 1. (**a**) Comparison of Primary metrics (weighted average Pearson correlations - WAPCCs) with the full and abridged feature sets. (**b**) Comparison of Primary metrics of the five models using the abridged feature set. $*P < 0.01$, $***P < 10^{-4}$, two-sample *z*-test. All error bars denote bootstrapped 95% confidence intervals. LinReg, Linear Regression; SVM, Support Vector Machine; n.s., not significant.

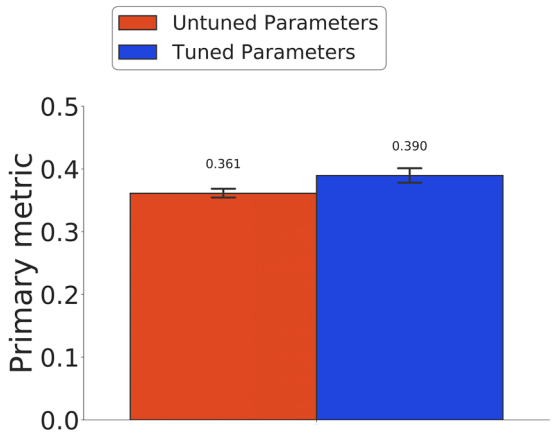

Figure 2. Parameter tuning of XGBoost. Comparison of ten repetitions of 10-fold cross validation weighted average Pearson correlations with the tuned parameters (*n_estimators* = 500, *max_depth* = 8, *subsample* = 0.75 and *colsample_bytree* = 1.0) and untuned default parameters (*n_estimators* = 250, *max_depth* = 8, *subsample* = 1.0 and *colsample_bytree* = 1.0) was obtained. Error bars denote bootstrapped 95% confidence intervals.

parameters and calculated cross-validation error at each step. We observed the minimum error with *max_depth* = 8. After setting the max depth to 8, we repeated the same process and varied the other three parameters. Error converged asymptotically for these iterations, so we took the best parameter values (*n_estimators* = 500, *max_depth* = 8, *subsample* = 0.75 and *colsample_bytree* = 1.0) that reached minimum error. Figure 2 shows the differences in regression performances (evaluated by ten trials of 10-fold CV) of the tuned vs. untuned models ($P < 0.01$).

Biological interpretation. *Feature importance analysis identifies biomarkers of synergy.* To determine biological factors underlying drug synergy, we computed an importance metric that represents percent contribution to the XGBoost model's prediction for each of the 2121 features (computed as accuracy improvement when that feature is included). We first looked at the total group scores for the 15 types of features (Fig. 3). As expected, trivial information (drug combination ID, cell line ID, tissue, disease, and sex) did not contribute much (8% total), justifying our creation of a more sophisticated feature set. The monotherapy features are the most informative features which accounted for 31% of predictive power. Notably, genomic context (expression, CNV, mutations) accounted for 32% of predictive power. Our three novel drug synergy network features had a net score of 3%, indicating their promise for future research towards any type of drug interaction prediction. Target protein domains did not help much (0.5%), perhaps indicating that most drugs were not promiscuous (and thus, the putative targets alone held most relevant information).

Although monotherapy and gene module expression are the two most important feature sets, we looked at the performance of a model which excluded these data in order to mimic current clinical settings. Also the competition organizers divided the first challenge (SC1) into two parts; in SC1A the competitors were asked to make prediction using all available data, whereas in SC1B the use of molecular data was limited to mutation and copy number variation. Figure 4 shows differences between the XGBoost models trained with/without the monotherapy data and gene expression. We also plotted the area under the receiver operating characteristic curve (ROC-AUC) to evaluate the performance of the binary predictions. In generating ROCs, we trained the models

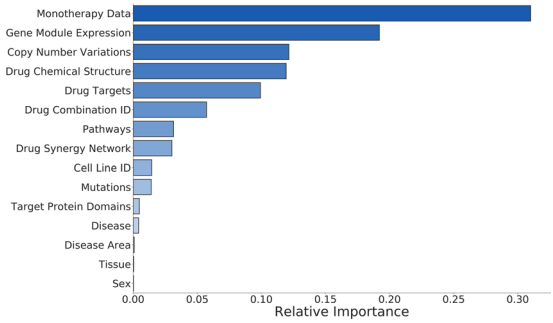

Figure 3. Bar plot of XGBoost feature group importances. For each training variable, an importance score is calculated as the improvement in predictive accuracy when that variable is included. Constituent scores (*e.g.*, the 53 individual importance scores for gene expression modules) are summed to determine the net importance of each feature class. The *x*-axis units represent fractional contribution (sum of bar lengths = 1).

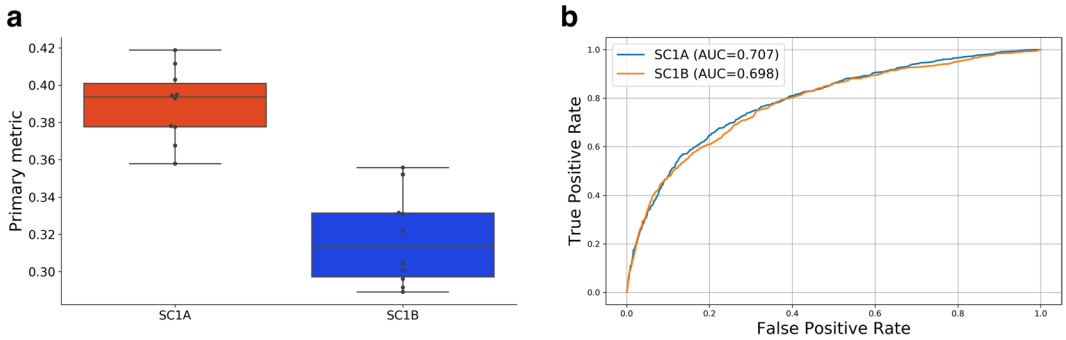

Figure 4. (**a**) Comparison of Primary metrics (weighted average Pearson correlations - WAPCCs) using all available molecular data (SC1A) and the molecular data excluding the monotherapy and gene expression feature sets (SC1B). (**b**) Comparison of the area under the receiver operating characteristic curve (ROC-AUC) with/without the monotherapy and gene expression feature sets.

for classification by binarizing the target values. The threshold was set at 20.0 as suggested by the challenge organizers (any score above 20.0 is represented by a 1, other scores map to 0). When gene expression and monotherapy information are excluded, the performance of XGBoost model drops significantly for real synergy value prediction ($WAPCC = 0.32$) but this difference is not significant in terms of binary predictions ($AUC = 0.70$).

Co-expression network created by the WGCNA approach identified cohesive modules having distinct gene expression patterns. Cohesiveness is a measure of how tightly a particular gene fits into its module. The more cohesive the module, the more similar the co-expression relationships are across the module. Module enrichment analysis performed after module identification showed that a majority of the modules have indeed biological functionality. When focused on these biologically important modules, the WGCNA detected several centrally located intramodular hub genes within the modules. These hub genes are highly connected to the rest of the genes in the module, so they can be regarded as major components of the module. Indeed, expression profiles of these hub genes are highly correlated to module eigengene values of the modules they belong to. In this respect, it seems that they are the most important genes in the modules, so effective drugs likely to attack to these hub genes. Thereby, they can be considered as biological targets or biomarker candidates for drug sensitivity.

We next conducted a finer analysis, looking at the most significant of the 2121 variables individually (using the same percent contribution importance metric) to extract more specific biological information regarding synergy. For drug targets, a master cancer signaling protein[19], tumor necrosis factor alpha (TNFA), ranked highest. B-raf V600E, a mutant of the oncogene *BRAF* that determines drug sensitivity via signaling[20], and ATR, a kinase protein regulating DNA repair[21], were the next most significant. Using two-sample Kolmogorov-Smirnov test, TFNA ($P < 2.3^{-3}$) and B-raf V600E ($P < 2.0^{-15}$) can significantly favour synergistic combinations, while the drug combinations targeting AKT1 ($P < 0.05$) and ATR ($P < 2.2^{-14}$) proteins are likely to be antagonistic. For copy number variation, repetitions or deletions in the *MYC* and *NFKBIA* genes were most significant. MYC is a transcription factor whose copy number has been strongly correlated to colon cancer in the past[22], whereas NFKBIA is involved in several cancer pathways but only has a tenuous link between CNV and cancer[23]. Pathway analysis revealed that cell differentiation, apoptosis, and cancer signaling processes were most important. The membrane active transport pathway also ranked highly, perhaps for its role in regulating drug influx and efflux[24]. We also analyzed potential synergy mechanisms of highly ranked mutations, summarized in Table 1. Five of these mutations have been previously shown to be cancer risk factors. Thus, feature importance analysis combined with results from existing literature implicates the aforementioned variables as novel potential biomarkers of synergistic drug effects.

Mutation Pos.	Gene	Pathology	Hypothesized Influence on Synergy	References
3:179218294	*PIK3CA*	Breast	MSM in PIK3a domain changes drug sensitivity	[44]
11:36489991	*TRAF6*	Bladder	Affects MAPK apoptotic signaling pathway	[45]
9:130862983	*ABL1*	Colon	Confers resistance to tyrosine-kinase inhibitor drugs	[46]
4:102613584	*NFKB1*	Colon	MSM in binding domain affects transcription regulation	—
22:41166649	*EP300*	Lung	MSM disrupts transcriptional co-activation	—
10:121565526	*FGFR2*	Colon, Lung	Increased expression, affecting FGF signaling pathway	[47]
12:25245351	*KRAS*	Colon, Lung	Disrupts Akt/mTOR pathway through PPI networks	[48]

Table 1. Most predictive mutations of synergy, identified by XGBoost. The location of the mutation is given as its chromosome followed by its genomic coordinate. Brief hypotheses for the influence of each mutation on synergy are proposed. Pos., Position; PPI, Protein-Protein Interaction; MSM, Missense Mutation (results in different amino acid).

Discussion

Cancers are complex diseases that are regulated by multiple complementary or redundant pathways. As a result, acquired drug resistance is an issue plaguing the vast majority of current single-agent chemotherapy regimens. The design and development of targeted drug combinations that disrupt multiple modes of metastasis is thus becoming increasingly necessary. Here, we establish the first comprehensive machine learning framework that successfully predicts synergistic drug combinations and presents opportunity for further exploratory biological analysis.

In this study, IC50 is used as a sensitivity measure for predicting drug synergy since GDSC and DREAM studies reports the sensitivity of all the screened anti-cancer drugs with IC50. This measure is not a powerful indicator of drug activity as IC50 could not be measured when maximum drug concentration is not sufficient for killing the cells/cell lines. Indeed, we noticed that most of the screened cell lines in the GDSC and DREAM studies does not reach an IC50 point within screening concentration interval. In addition, we identified that there are substantial deviations in IC50 values reported for the cell lines screened by the same drugs in DREAM study. It shows either assay used in experimental procedure does not measure correct IC50 values or cell lines are genetically heterogenic, i.e, they consist resistant and sensitive sub-populations of cells. So using IC50 as a sensitivity measure might lead to underperformance of our *in-silico* models generated for predicting synergistic drug combinations. Instead of IC50, using alternative sensitivity measures such as Activity Area and Amax, which are regarded to be more reliable indicatiors of drug sensitivity, would improve the predictive power of our models and give us a more reliable picture of synergistic drug pairs.

This work takes a data-driven approach to drug synergy prediction, integrating comprehensive pharmacological data with molecular information to train powerful machine learning models. Importantly, we combine several different biological data types to build a comprehensive, novel feature set and thus optimize performance. We show that XGBoost is the most well-suited learning algorithm to synergy identification. Ultimately, our model's high correlation, generalizability to external data, and *de novo* discovery of drug combinations currently undergoing clinical trials alongside novel synergistic pairs all support its predictive success over previous methods.

We provide the workflow that generates the feature set and the results so that other labs can easily use or extend this methodology. The workflow can also handle additional features or missing features and can be run on a standard desktop machine. Note that the performance and training time of the XGBoost model are greatly influenced by the hyperparameters that need to be tuned. The overall performance of the method may be improved with the addition of gene expression and monotherapy data, however such data are challenging to obtain in clinical settings owing to financial, logistic and technological reasons. We note some possible limitations in our model's predictions. Using a synergy score as the output metric may not be ideal, since it is an integral over a wide range of concentrations (whereas in practice, treatments at lower concentrations are generally more clinically feasible). Additionally, computational models may report false positives, so our newly discovered combinations must be validated.

In the future, we hope to further explore undiscovered mechanisms of drug synergy. Specifically, drugs activating and repressing shared transcription factors via downstream effects has been recently suggested as a synergy mechanism[25]. We also plan to conduct experimental trials of predicted synergistic drug combinations on cancer-specific cell lines and patient organoid models to further support our *in silico* approach. Regardless, we expect our current framework to aid in rapid identification and development of synergistic drug combinations towards specific and comprehensive cancer treatment for all.

Methods

To better understand drug-drug interactions and suggest viable synergistic pairs, we approached the problem by aggregating as much open data as possible to build an accurate predictive model. We subsequently analyzed our predictions to identify novel and plausible mechanistic synergy hypotheses. Our workflow spanned three stages: feature compilation, building and evaluation of machine learning models, and biological interpretation of our results (Fig. 5).

Training data. We used a dataset recently released by AstraZeneca and the Dialogue for Reverse Engineering Assessments and Methods (DREAM) consortium[17] as the core training data for our method. The data are composed of synergy scores for 2790 experiments (Subchallenge 1 Training set + leaderboard) across 167

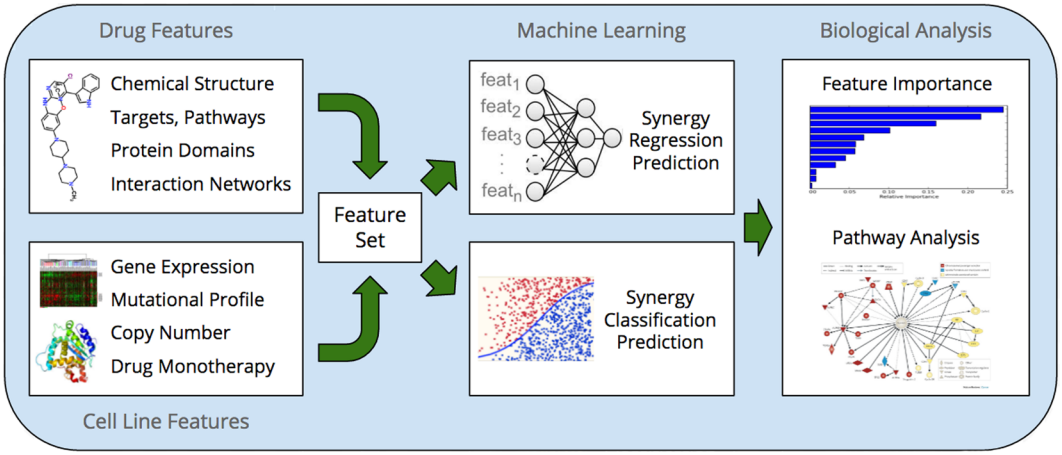

Figure 5. Our pipeline for modeling and analysis of drug synergy. We integrate features from two input streams: drug data and cell line data. We train our machine models on the compiled feature set and perform biological analysis of predictions to propose novel hypotheses explaining drug synergy.

drug combinations and 85 cell lines, representing a small fraction of the complete combinatorial space (2790/(167*85) = 19.6%) but still the largest combination drug screen to date. Synergy scores are defined by integrating experimental cell kill fraction minus the expected additive cell kill fraction as defined by the Loewe model[26].

Feature set. Pairs of drugs that are synergistic on one cell line are not necessarily synergistic on other cell lines[12]. Hence, we hypothesized that information on both the drugs and the tested cell line is predictive of synergy, making it necessary to incorporate both classes of features into our method. We extract the biologically relevant features, called abridged feature set. We detail the groups of features used to train our models below.

Chemical structure. Drug structure at the molecular level describes its binding activity. Chemical fingerprints are the most commonly used structural profile of drugs[27]. Fingerprints are bit vectors that indicate the presence (1) or absence (0) of various chemical features (*e.g.*, a C=N group, a six member ring, *etc.*). To integrate fingerprints into our pipeline, we used the Python OpenBabel 2.3 library[28] to take an input chemical formula (SMILES ID; given by AstraZeneca) and generate length 166 Molecular Access System (MACCS) binary structural feature lists[29]. For each drug combination, we used the sum of the two single drug bit vectors as features (*i.e.*, a 2 represents both drugs having the feature, a 1 represents one of the drugs having the feature, and a 0 represents neither of the drugs having the feature; this mapping worked best) preserving similarity resolution across each of the 166 structural elements. While individual elements may not be relevant, we expect our model to learn combinations of structures that are predictive.

Drug targets. Targets can shed light on the biological processes that the drug controls. We started by using summed bit vectors of the putative targets (given as part of the AstraZeneca synergy dataset), of which there were 185 across all the drugs; thus, a 2 represents a shared target, a 1 represents a target of one the drugs, and a 0 represents a target for neither of the drugs. However, this matrix was sparse across the training dataset, since the drugs have a median of one putative target each.

Target protein domains. To account for other drug-protein interactions, we generated structural protein domain features for the targets of each drug and mapped them to drug combinations in the same 2, 1, or 0 format. We used four databases (Pfam[30], Prosite[31], SMART[32], and SUPERFAMILY[33] with 131, 97, 67, and 75 features, respectively) resulting in 370 total domain features. These features may account for cases in which drugs are not known to interact specifically with a given target, but they still have some binding affinity (termed a *promiscuous* interaction).

Targeted pathways. We also generated 309 features for the biological pathways involving the drug targets using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. The rationale for these features was to provide a direct read on the specific metabolic, signaling, and regulatory processes that the drug combinations disrupt, which may inform synergistic effects.

Drug synergy network. We have explored the network based features to see if drug synergy is transferable between the cell lines and distinguish synergistic drug combinations. Previous studies have reported predictive success using network topology of drug-drug interaction networks[13,34]. We built an undirected synergy network, in which two synergistic drugs are connected by an edge. We identified a drug combination as synergistic if the majority of synergy scores for that drug combination across cell lines is greater than 20. Using the synergy

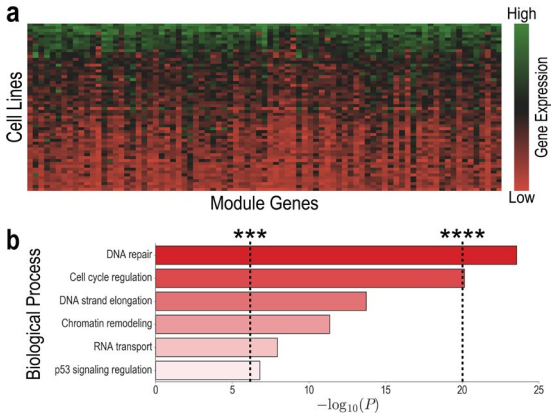

Figure 6. Weighted Gene Co-Expression Network Analysis identifies modules of correlated genes. (**a**) Expression heatmap for the 162 genes that form one of the modules. Note that in each cell line, the genes are either primarily highly expressed (green rows) or primarily lowly expressed (red rows), indicating that gene expression is correlated within modules. (**b**) Selected gene ontology biological term enrichments for genes in the cluster from (**a**) illustrate module-level biological function. $***P < 5 \times 10^{-7}$, $****P < 10^{-20}$ (Hypergeometric test with Bonferroni multiple testing correction).

network, we extracted three features frequently used in social network link prediction for each drug pair: number of common neighbors, Jaccard coefficient, Adamic-Adar coefficient[35].

The network proximity features for a drug pair (x, y) in the drug synergy network are defined as follows:

$$\textbf{Common Neighbors}(\mathbf{x}, \mathbf{y}) = |\Gamma(x) \cap \Gamma(y)| \tag{1}$$

$$\textbf{Jaccard}(\mathbf{x}, \mathbf{y}) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \tag{2}$$

$$\textbf{Adamic/Adar}(\mathbf{x}, \mathbf{y}) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log|\Gamma(z)|} \tag{3}$$

where $\Gamma(x)$ represents neighbors of node x, $\Gamma(y)$ represents neighbors of node y in the drug synergy network.

Monotherapy information. To calculate synergy scores defined by excess over the Loewe additivity model, the AstraZeneca study also conducted cell viability assays for the 69 individual drugs involved in the 167 combinations[36]. These monotherapy features included, for each drug in the combination, the maximum concentration used in the assay, the IC50 value (concentration where half of maximum kill is achieved), the Hill coefficient H (slope of the dose-response curve), the max kill percentage E_{inf} and data quality check information.

Gene expression profiles using weighted correlation network analysis. Microarray expression data of 17,419 genes were generated for the 85 cell lines by the Genomics for Drug Sensitivity in Cancer (GDSC) group[37]. However, using such a large number of gene expression values directly can be detrimental, since data for individual genes have noisy deviations across cell lines that are not biologically meaningful (we saw minimal improvement in model performance when we used raw expression). To overcome this issue and summarize biological processes that are otherwise difficult to learn, we leveraged Weighted Gene Co-Expression Network Analysis (WGCNA), a robust technique to identify systems-level gene modules[38]. Modules are determined by hierarchal clustering of the 17,419 × 17,419 gene expression correlation matrix[39]. As expected, genes within a given module have highly correlated expression profiles (Fig. 6a), but are also frequently enriched for Gene Ontology (GO) terms that indicate biological function (Fig. 6b). Thus, we used mean expression values of the 53 modules as cell line features.

Mutations and copy number variations. Genomic sequence features also provided important information for cell line-specific context. The Catalogue Of Somatic Mutations In Cancer (COSMIC) database performed whole-exome sequencing of the 85 cell lines to identify coding single nucleotide polymorphisms (SNPs) and copy number variations[40]. In total, there were 75,281 SNPs that occurred in at least one cell line, but the vast majority of these mutations were not predictively relevant. We filtered out all SNPs in genes that are not in the KEGG cancer pathways, resulting in 876 features represented in binary format; these included *BRAF*, *TP53*, and other canonical tumorigenesis mutations[41]. Copy number variations (CNVs) are long, repeated segments of genes that have been increasingly implicated in disease in recent years[42]. To filter CNVs, we correlated the copy number of each gene with its expression across the 85 cell lines. Genes in the cancer pathways with a statistically significant, above median correlation ($P < 0.01$, Fisher's correlation test; Spearman rank correlation >0.17) were hypothesized to

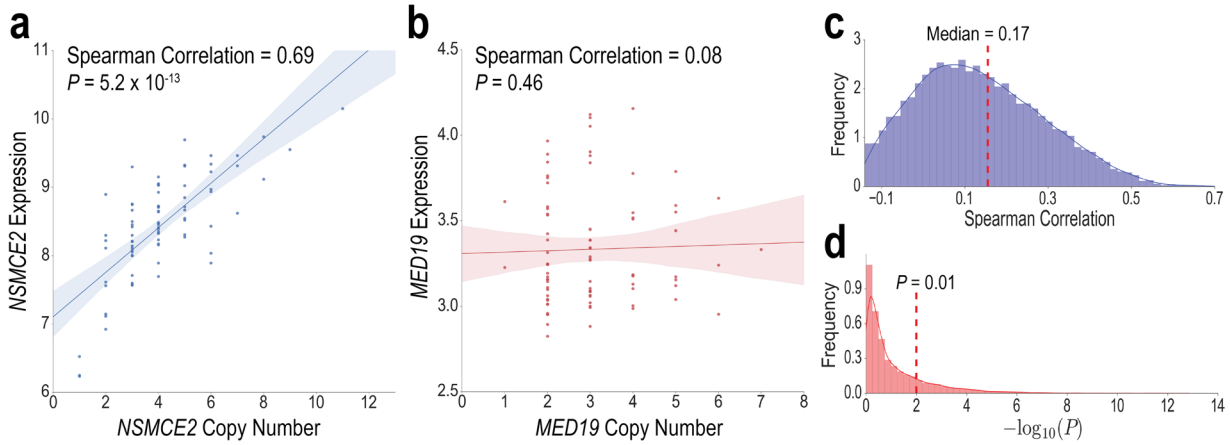

Figure 7. Expression and copy number variation (CNV) correlations differ across genes. (**a**) *NSMCE2* expression varies with CNV for the 85 cancer cell lines, while (**b**) *MED19* does not have a significant correlation. (**c**) Distribution (probability density) of Spearman rank correlations and (**d**) distribution of negative log *P*-values for all 17,419 genes. CNV of cancer genes with above median Spearman and a significant *P*-value were used as features. *P*-values are generated with Fisher's *r*-to-*z* transformation for correlation testing.

possibly have functionally relevant CNVs (Fig. 7), and their copy numbers were included in the feature set (143 genes).

Machine learning prediction. To learn patterns from the feature set and make accurate synergy score predictions, we trained multiple machine learning models with the AstraZeneca data to identify an optimal framework.

Types of models trained on the feature set. We trained five machine learning models on the feature set for synergy score prediction: linear regression, Lasso, support vector machine (SVM), random forest, and XGBoost. The first four models were trained using the {sklearn} Python package[43], while XGBoost was trained using the {xgboost} Python package[18]. Linear regression fits a line to the training data by minimizing its *cost function*, which we chose to be the sum of the squared distances of the predictions from the actual synergy scores. Lasso is a linear regression with an L1-regularization term (*i.e.*, the sum of the absolute values of the coefficients) in the cost function to prioritize models with smaller coefficients and thus reduce overfitting. SVM works by building a hyperplane in an f-dimensional feature space such that all dimensions are within ε of the target value and the L2-norm (sum of squared coefficients) is minimized. Random forest stochastically assigns a set of features and training examples to each of its n decision trees that are independently trained and then averaged for the final prediction. XGBoost is similar but adds L2-regularization and boosting, a method to prioritize the weights learned from the most mispredicted examples.

Evaluating predictive performance using cross validation. For evaluation of model performance using only the AstraZeneca dataset, we did ten repetitions of 10-fold cross-validation (CV). This involves randomly splitting the data into 10 equally sized segments, iteratively training on 9 of the 10 folds, and testing on the remaining tenth. We used the weighted average Pearson correlation coefficient (WAPCC) of the experimental value vs. our prediction as the primary evaluation metric suggested by the challenge organizers. The primary metric is defined as follows:

$$WAPCC = \frac{\sum_{i=1}^{N} \sqrt{n_i - 1} \cdot \rho_i}{\sum_{i=1}^{N} \sqrt{n_i - 1}} \tag{4}$$

where $N = 167$ is the number of the tested drug combinations, ρ_i is the Pearson correlation for drug combinations i, n_i is the number of cell lines that drug combination i is applied to.

The reported metrics are not on the final test set, and are rather cross-validation scores. All the performances we achieve should thus be considered post-hoc analyses and may appear higher than our final performance on the challenge itself.

Biological interpretation with feature importance. To analyze the relative predictive power of the different biological classes of features and identify potential biomarkers, we calculated how much each feature increased the accuracy of the XGBoost model (termed 'gain')[18].

Data Availability

Challenge data is available through registration via https://openinnovation.astrazeneca.com/data-library.html and it is expected to be public soon. Code is accessible via https://www.synapse.org/#!Synapse:syn5605365/wiki/394725 and https://github.com/rcelebi/dream-drugcombo.

References

1. Mignani, S., Huber, S., Tomás, H., Rodrigues, J. & Majoral, J.-P. Why and how have drug discovery strategies in pharma changed? What are the new mindsets? *Drug Discov. Today* **21**, 239–249, https://doi.org/10.1016/j.drudis.2015.09.007 (2016).
2. Dias, M. H., Kitano, E. S., Zelanis, A. & Iwai, L. K. Proteomics and drug discovery in cancer. *Drug Discov. Today* **21**, 264–277, https://doi.org/10.1016/j.drudis.2015.10.004 (2016).
3. Hoelder, S., Clarke, P. A. & Workman, P. Discovery of small molecule cancer drugs: Successes, challenges and opportunities. *Mol Oncol* **6**, 155–176, https://doi.org/10.1016/j.molonc.2012.02.004 (2012).
4. Lavecchia, A. & Cerchia, C. *In silico* methods to address polypharmacology: current status, applications and future perspectives. *Drug Discov. Today* **21**, 288–298, https://doi.org/10.1016/j.drudis.2015.12.007 (2016).
5. McGranahan, N. & Swanton, C. Biological and Therapeutic Impact of Intratumor Heterogeneity in Cancer Evolution. *Cancer Cell* **27**, 15–26, https://doi.org/10.1016/j.ccell.2014.12.001 (2015).
6. Al-Lazikani, B., Banerji, U. & Workman, P. Combinatorial drug therapy for cancer in the post-genomic era. *Nat Biotech* **30**, 679–692, https://doi.org/10.1038/nbt.2284 (2012).
7. Bozic, I. *et al.* Evolutionary dynamics of cancer in response to targeted combination therapy. *eLife* **2**, e00747, https://doi.org/10.7554/eLife.00747 (2013).
8. Hu, C.-M. J. & Zhang, L. Nanoparticle-based combination therapy toward overcoming drug resistance in cancer. *Biochem. Pharmacol.* **83**, 1104–1111, https://doi.org/10.1016/j.bcp.2012.01.008 (2012).
9. Ma, Y. *et al.* High-Dose Parenteral Ascorbate Enhanced Chemosensitivity of Ovarian Cancer and Reduced Toxicity of Chemotherapy. *Science Translational Medicine* **6**, 222ra18–222ra18, https://doi.org/10.1126/scitranslmed.3007154 (2014).
10. Griner, L. A. M. *et al.* High-throughput combinatorial screening identifies drugs that cooperate with ibrutinib to kill activated B-cell—like diffuse large B-cell lymphoma cells. *PNAS* **111**, 2349–2354, https://doi.org/10.1073/pnas.1311846111 (2014).
11. Huang, L. *et al.* DrugComboRanker: drug combination discovery based on target network analysis. *Bioinformatics* **30**, i228–i236, https://doi.org/10.1093/bioinformatics/btu278 (2014).
12. Bansal, M. *et al.* A community computational challenge to predict the activity of pairs of compounds. *Nat Biotech* **32**, 1213–1222, https://doi.org/10.1038/nbt.3052 (2014).
13. Sun, Y. *et al.* Combining genomic and network characteristics for extended capability in predicting synergistic drugs for cancer. *Nat Commun* **6**, https://doi.org/10.1038/ncomms9481 (2015).
14. Huang, H., Zhang, P., Qu, X. A., Sanseau, P. & Yang, L. Systematic prediction of drug combinations based on clinical side-effects. *Sci. reports* **4** (2014).
15. Li, X. *et al.* Prediction of synergistic anti-cancer drug combinations based on drug target network and drug induced gene expression profiles. *Artif. Intell. Medicine* (2017).
16. Zhao, X.-M. *et al.* Prediction of Drug Combinations by Integrating Molecular and Pharmacological Data. *PLOS Comput Biol* **7**, e1002323, https://doi.org/10.1371/journal.pcbi.1002323 (2011).
17. Menden, M. P. *et al.* Community assessment of cancer drug combination screens identifies strategies for synergy prediction. bioRxiv 200451, https://doi.org/10.1101/200451 (2018).
18. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. arXiv:1603.02754 [cs] 785–794, https://doi.org/10.1145/2939672.2939785 (2016).
19. Chu, W.-M. Tumor necrosis factor. *Cancer Letters* **328**, 222–225, https://doi.org/10.1016/j.canlet.2012.10.014 (2013).
20. Andrulis, M. *et al.* Targeting the BRAF V600e mutation in multiple myeloma. *Cancer Discov* **3**, 862–869, https://doi.org/10.1158/2159-8290.CD-13-0014 (2013).
21. Toledo, L. I., Murga, M. & Fernandez-Capetillo, O. Targeting ATR and Chk1 kinases for cancer treatment: A new model for new (and old) drugs. *Molecular Oncology* **5**, 368–373, https://doi.org/10.1016/j.molonc.2011.07.002 (2011).
22. Tseng, Y.-Y. *et al.* PVT1 dependence in cancer with MYC copy-number increase. *Nature*. https://doi.org/10.1038/nature13311 (2014).
23. Patane, M. *et al.* Frequency of NFKBIA deletions is low in glioblastomas and skewed in glioblastoma neurospheres. *Mol. Cancer* **12**, 160, https://doi.org/10.1186/1476-4598-12-160 (2013).
24. Ruiz, N., Gronenberg, L. S., Kahne, D. & Silhavy, T. J. Identification of two inner-membrane proteins required for the transport of lipopolysaccharide to the outer membrane of Escherichia coli. *Proc. Natl. Acad. Sci. USA* **105**, 5537–5542, https://doi.org/10.1073/pnas.0801196105 (2008).
25. Mitrofanova, A. *et al.* Predicting Drug Response in Human Prostate Cancer from Preclinical Analysis of *In Vivo* Mouse Models. *Cell Reports* **12**, 2060–2071, https://doi.org/10.1016/j.celrep.2015.08.051 (2015).
26. Geary, N. Understanding synergy. *Am. J. Physiol. Endocrinol. Metab.* **304**, E237–253, https://doi.org/10.1152/ajpendo.00308.2012 (2013).
27. Melville, J. L. & Hirst, J. D. Tmacc interpretable correlation descriptors for quantitative structure activity relationships. *J. Chem. Inf. Model.* **47**, 626–634, https://doi.org/10.1021/ci6004178 (2007).
28. O'Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **3**, 33, https://doi.org/10.1186/1758-2946-3-33 (2011).
29. Durant, J. L., Leland, B. A., Henry, D. R. & Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **42**, 1273–1280, https://doi.org/10.1021/ci010132r (2002).
30. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucl. Acids Res.* **44**, D279–D285, https://doi.org/10.1093/nar/gkv1344 (2016).
31. Sigrist, C. J. A. *et al.* New and continuing developments at PROSITE. *Nucleic Acids Res.* **41**, D344–347, https://doi.org/10.1093/nar/gks1067 (2013).
32. Letunic, I., Doerks, T. & Bork, P. SMART: recent updates, new developments and status in 2015. *Nucl. Acids Res.* **43**, D257–D260, https://doi.org/10.1093/nar/gku949 (2015).
33. Wilson, D. *et al.* Superfamily sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucl. Acids Res.* **37**, D380–D386, https://doi.org/10.1093/nar/gkn762 (2009).
34. Xu, K.-J., Song, J. & Zhao, X.-M. The drug cocktail network. *BMC Syst Biol* **6**, S5, https://doi.org/10.1186/1752-0509-6-S1-S5 (2012).
35. Adamic, L. A. & Adar, E. Friends and neighbors on the Web. *Social Networks* **25**, 211–230, https://doi.org/10.1016/S0378-8733(03)00009-1 (2003).
36. Dry, J. *et al.* AstraZeneca-Sanger Drug Combination Prediction DREAM Challenge - syn4231880 (2015).
37. Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740–754, https://doi.org/10.1016/j.cell.2016.06.017 (2016).

38. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559, https://doi.org/10.1186/1471-2105-9-559 (2008).
39. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* **24**, 719–720, https://doi.org/10.1093/bioinformatics/btm563 (2008).
40. Forbes, S. A. *et al*. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43**, D805–811, https://doi.org/10.1093/nar/gku1075 (2015).
41. Seton-Rogers, S. T. Mutant relationships. *Nat Rev Cancer* **15**, 135–135, https://doi.org/10.1038/nrc3917 (2015).
42. Shlien, A. & Malkin, D. Copy number variations and cancer. *Genome Med* **1**, 62, https://doi.org/10.1186/gm62 (2009).
43. Pedregosa, F. *et al*. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
44. Nik-Zainal, S. *et al*. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54, https://doi.org/10.1038/nature17676 (2016).
45. Arthur, J. S. C. & Ley, S. C. Mitogen-activated protein kinases in innate immunity. *Nat Rev Immunol* **13**, 679–692, https://doi.org/10.1038/nri3495 (2013).
46. Greuber, E. K., Smith-Pearson, P., Wang, J. & Pendergast, A. M. Role of ABL Family Kinases in Cancer: from Leukemia to Solid Tumors. *Nat Rev Cancer* **13**, 559–571, https://doi.org/10.1038/nrc3563 (2013).
47. Fletcher, M. N. C. *et al*. Master regulators of FGFR2 signalling and breast cancer risk. *Nat Commun* **4**, 2464, https://doi.org/10.1038/ncomms3464 (2013).
48. Paplomata, E. & O'Regan, R. The PI3k/AKT/mTOR pathway in breast cancer: targets, trials and biomarkers. *Ther Adv Med Oncol* **6**, 154–166, https://doi.org/10.1177/1758834014530023 (2014).

Author Contributions

R.C., O.B., R.M., S.A. and M.D. conceived the experiment(s), R.C., O.B. and R.M. conducted the experiment(s), R.C., O.B., R.M., S.A. and M.D. analysed the results. All authors reviewed the manuscript.

Additional Information

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.