PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Bioactivity assessment of natural compounds using machine learning models trained on target similarity between drugs

Vinita Periwal[1,2], Stefan Bassler[1,3], Sergej Andrejev[1], Natalia Gabrielli[1], Kaustubh Raosaheb Patil[4,5], Athanasios Typas[1], Kiran Raosaheb Patil[1,2]*

1 European Molecular Biology Laboratory, Heidelberg, Germany, **2** Medical Research Council Toxicology Unit, University of Cambridge, Cambridge, United Kingdom, **3** Faculty of Biosciences, Heidelberg University, Heidelberg, Germany, **4** Institute of Neuroscience and Medicine (INM-7), Jülich, Germany, **5** Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University, Düsseldorf, Germany

* kp533@cam.ac.uk

OPEN ACCESS

Citation: Periwal V, Bassler S, Andrejev S, Gabrielli N, Patil KR, Typas A, et al. (2022) Bioactivity assessment of natural compounds using machine learning models trained on target similarity between drugs. PLoS Comput Biol 18(4): e1010029. https://doi.org/10.1371/journal.pcbi.1010029

Editor: Costas D. Maranas, The Pennsylvania State University, UNITED STATES

Received: September 6, 2021

Accepted: March 17, 2022

Published: April 25, 2022

Data Availability Statement: The input data files can be accessed at (https://data.mendeley.com/datasets/7ft539gwf3/3) and codes for analysis and figure generation are available at https://github.com/periwal45/periwaletal2020.

Funding: VP and NG were supported by the EMBL Interdisciplinary Postdoc (EI3POD) program under Marie Skłodowska-Curie actions COFUND (grant number 664726). SB was supported by the Joachim Hertz Foundation (fellowship for

Abstract

Natural compounds constitute a rich resource of potential small molecule therapeutics. While experimental access to this resource is limited due to its vast diversity and difficulties in systematic purification, computational assessment of structural similarity with known therapeutic molecules offers a scalable approach. Here, we assessed functional similarity between natural compounds and approved drugs by combining multiple chemical similarity metrics and physicochemical properties using a machine-learning approach. We computed pairwise similarities between 1410 drugs for training classification models and used the drugs shared protein targets as class labels. The best performing models were random forest which gave an average area under the ROC of 0.9, Matthews correlation coefficient of 0.35, and F1 score of 0.33, suggesting that it captured the structure-activity relation well. The models were then used to predict protein targets of circa 11k natural compounds by comparing them with the drugs. This revealed therapeutic potential of several natural compounds, including those with support from previously published sources as well as those hitherto unexplored. We experimentally validated one of the predicted pair's activities, viz., Cox-1 inhibition by 5-methoxysalicylic acid, a molecule commonly found in tea, herbs and spices. In contrast, another natural compound, 4-isopropylbenzoic acid, with the highest similarity score when considering most weighted similarity metric but not picked by our models, did not inhibit Cox-1. Our results demonstrate the utility of a machine-learning approach combining multiple chemical features for uncovering protein binding potential of natural compounds.

Author summary

A large fraction of small-molecule drugs has originated from natural compounds making them an attractive resource for search of potential lead compounds. Yet, this resource is not extensively explored because of their vast number and technical barriers to obtaining

Interdisciplinary Science). KRP and VP received support from the UK Medical Research Council (grant number MC_UU_00025/11). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

them in pure form. Computational approaches can expedite exploration of natural compounds and their derivatives at a much larger scale. Towards this, we took advantage of the known protein targets of drugs to mine natural compounds with similarity to known small-molecule drugs. The underlying hypothesis is that two compounds binding to the same protein target are similar from a bioactivity viewpoint. To identify high-dimensional structural features of the compounds underlying their bioactivity, we computed various structural features of paired drugs (i.e., drugs sharing a common protein target) and used these to train machine learning classifiers. The trained classification models were then used to predict similarity between drugs and natural compounds. We assessed the resulting predictions–protein target binding by natural compounds—through an extensive literature survey, and experimental validated a novel prediction. Together, our results outline a workflow and provide a resource to explore therapeutic potential of natural compounds.

Introduction

Around 65% of the small-molecule drugs in use today have originated from natural compounds or their derivatives [1]. Therapeutic effects of natural compounds are thus central to drug discovery [2–5]. Further, identification of bioactive compounds present in the diet and their effect on health has been an active area of research since long [6,7]. A number of recent studies have reported that dietary natural compounds (such as polyphenols, alkaloids) can reduce the risk of many chronic diseases [8–10], lead to drug and food interactions (occurs when your food and medicine interfere with one another) [11–13], and significantly alter or diversify the composition of the human gut microbiome [14–17]. While natural compounds possess rich structural diversity, often have selective biological actions, and are prevalidated on various biological targets by evolutionary selection [18–21], they are generally less accessible in pure form than synthetic compounds. This is primarily due to their low abundance in natural sources and complex purification methods [22,23]. Recent technological advances in analytical methods such as metabolomics, metabolic engineering, and synthetic biology, as well as those in functional assays and phenotypic screens are opening new opportunities for natural compound-based drug discovery [2,22,24]. Increasing number of computational tools [25,26], techniques [27,28], and databases [29] are providing more accessible and powerful alternatives to explore the therapeutic potential of natural compounds.

An attractive approach to assess the bioactivity potential of a compound is comparing its chemical and structural similarity with that of the molecules with known activity [27,30,31]. Chemical structures encode complex atom and bond connectivity information which can be computationally exploited to predict their potential biological interactions. The chemical similarities between drug and natural compounds, especially dietary compounds, and their association with drug targets have been studied previously [12]. However, similarity assessments can vary considerably depending upon which structural fingerprint encoding is used [32,33]. Indeed, the structural similarity between two molecules is a subjective concept [34] and no single similarity measure can likely capture the complex structure-activity relationships (SAR). Thus, owing to the vast structural diversity of natural compounds, it would be advantageous to include more extensive similarity measure encodings to establish structure-activity relationships more accurately and to predict bioactivities [27,30,35,36].

Machine learning (ML) is being increasingly used to tackle complex structure-activity relations that are otherwise difficult to deconvolute [25,37–39]. ML has been effectively used to predict, among others, molecular targets [40,41], bioactivities [42], shared molecular interactions [43,44], toxicity [45,46], and drug-likeness of molecules [47]. Although ML models greatly facilitate predictions, they often lack interpretability which translates to the acceptance of their predictions in pharmaceutical or clinical settings. Thus, experimental validation assessing model prediction is crucial to build trust in a method [26].

Chemical proteomics (high-throughput methods for exploring drug-target-phenotype relationships) exists as a key and powerful method for target identification and elucidating mode of action of natural compounds [48,49]. We propose a powerful approach where we combine computational target prediction of natural compounds with an in-vitro validation of the binding partner.

In this study, we identify potential of natural compounds (especially ingested dietary molecules) to bind human proteins that are known drug targets. We trained binary machine learning classifiers (**Fig 1A**) using chemical similarity scores from multiple fingerprints and physicochemical properties of paired small-molecule drugs with their known protein targets. The resulting models are then used to predict the molecular targets of hundreds of natural compounds through assessing their similarity with the drugs. We validate the models by demonstrating a predicted link's Cox-1 binding activity by 5-methoxysalicylic acid (found in tea and herbs).

Results

Dataset of drugs with known targets

We utilized mappings between 1,410 FDA approved drugs (S1A **Table**) and their known, curated, targets (S1B **Table**) as our gold-standard dataset. The drugs were categorized according to their ATC (Anatomical Therapeutic Chemical) class, and into 16 chemical Superclasses (a hierarchy in chemical taxonomy with general structural identifiers such as organic acids and derivatives, organometallic compounds) [50] based on their chemical structures (Materials and methods). Many of the drugs target the nervous system (264), followed by cardiovascular (180), anti-infectives (148), multiple ATC (131) and anti-neoplastic (127) (S1A **Fig**). Among the 16 structural classes, benzenoids and organoheterocyclics constitute the major super-classes of drugs (840) encompassing all therapeutic classes except the nutraceuticals (S1A **Fig**).

For the 1,410 drugs used, there were 1,262 known curated targets [51]. The number of drug targets ranged from 1 to 86 (i.e., some drugs have up to 86 known targets) (S1B **Table**), highlighting the fact that some drugs are well studied in terms of their target space. The most frequent targets were the different units and subunits of GABA receptors and GPCRs (adrenergic, muscarinic, histamine and dopamine receptors). The abundance of GABA receptors is consistent with the fact that many drugs (264) are targeting the nervous system.

Predictor variables for representing a chemical pair

To deconvolute the complex structure-activity relation between drugs and their targets, we created datasets for supervised binary classification. The underlying hypothesis is that a pair of compounds sharing at least one common protein target will be close in a high-dimensional structural space. Models are built as a binary classifier (one or more shared targets, or no shared target) with various structural similarity metrics as predictor variables. Such models can then be applied to predict the targets of natural compounds when compared to the drugs using the same predictors.

The first step towards building the classifier was to identify predictor variables. For this, similarity scores between all drug pairs were calculated using seven molecular fingerprints,

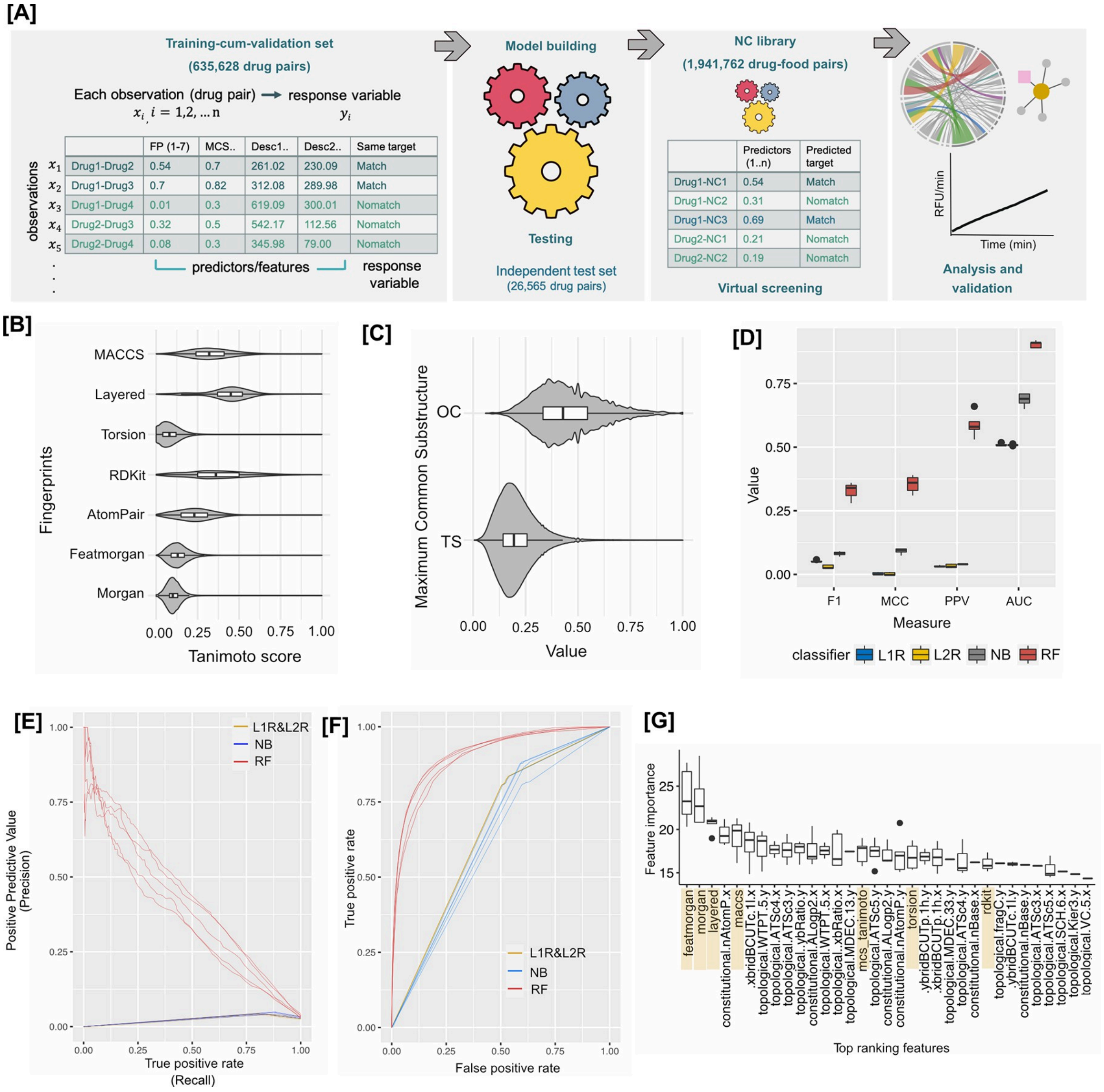


Fig 1. (A) **Overview of workflow deployed**. A training-cum-validation set comprising of drug pairs was created using various predictor variables (fingerprints, MCS and physicochemical properties). The model was trained for response variable (Match or Nomatch) and tested on an independent test set for performance evaluation. The natural compound library paired with drugs was virtually screened to obtain hit pairs, followed by analysis and in-vitro validation. (B-C)—**Similarity metrics (ML dataset)**. (B) Molecular fingerprints—the 7 fingerprints generate a different similarity score for the pairs of drug molecules compared. The median value of each is represented in the box plot (in the center) and the spread shows the density of the drug pairs around that score. (C) MCS—there are two types of scores reported by the MCS algorithm, one is the Tanimoto score and the other is the Overlap coefficient (OC). The violin plots were smoothed for density by an adjustment factor of 3. (D-F)—**Performance on the test set.** (D) performance of the four models, viz., regularized logistic regression (L1R and L2R), naïve bayes (NB) and random forest (RF) on independent test set for all 5 split-sets. Performance was evaluated using balanced measures: F1 score, matthews correlation coefficient (MCC), positive predictive value (PPV) and area under the curve (AUC). RF clearly had higher performance as compared to the logistic regression and naïve bayes models under all metrics and data splits. The performance of all models was also evaluated using (E) precision-recall and (F) ROC curve–the RF models achieved an AUC of 0.90 averaged on the all 5 test-split sets whereas NB and LRs performed relatively poor on all split-sets (average: NB: 0.68, L1R: 0.51 and L2R: 0.50). (G) High ranking features of RF models on the 5 split-sets–top features are displayed, showing most of the distance-based features provided maximum information gain with 'Featmorgan' performing best.

https://doi.org/10.1371/journal.pcbi.1010029.g001

viz., Morgan, Featmorgan, AtomPair, RDKit, Torsion, Layered and MACCS. The fingerprint similarity was scored using the Tanimoto Score, which is measured on a scale of '0–1'; higher the score, more similar are the molecules. Consistent with each molecular fingerprint assessing different features of the compound, the tanimoto score distribution for drug pairs differed across all the seven fingerprints (Fig 1B). For most of the drug pairs, the Morgan, Featmorgan and Torsion fingerprints consistently yielded a lower score (tanimoto score$_{median}$ = 0.11, 0.13, 0.08 respectively) as compared to AtomPair, RDKit, Layered, and MACCS (tanimoto score$_{median}$ = 0.23, 0.36, 0.45, 0.32 respectively). The drug pairs also showed broader tanimoto score distribution in AtomPair, RDKit, Layered and MACCS. A rank comparison of drug-pairs (S2 Fig) showed a low concordance amongst the different fingerprints supporting that each fingerprint captures different aspects of the structural similarity. This variance, together with the previous observations that none of the fingerprints alone is universally suited [32,52], we decided to utilize the similarity scores from all the seven fingerprints in training the ML classifier.

In addition to the fingerprint metrics, maximum common substructure (MCS, which is based on overlap between the two molecules represented as chemical graphs) [53] and the physicochemical descriptors (numerical properties) were included as additional predictor variables based on their previously noted utility [25,54,55]. The MCS calculation reports several statistics, amongst which the MCS size (median = 8), tanimoto score (median = 0.19) and OC (overlapping coefficient) (median = 0.43) score are important measures to assess similarity. The tanimoto score and OC score distribution is shown in Fig 1C. OC, measured on a scale of 0–1, accounts for size difference amongst molecules and is a useful indicator when there is a significant size difference between molecules being compared. The MCS measures are more intuitive to interpret as the substructure graph shared between the two molecules can be visualized and can be mapped back to the underlying molecules to extract which are the common and unique features, while this is not possible with the fingerprints and the molecular descriptors.

For molecular descriptors, we used 5 different categories (constitutional, topological, geometrical, electronic and hybrid) to capture individual physicochemical properties of the drugs. S1D Table, reports the number of descriptors used in each category. Majority of the physical and chemical information comes from the constitutional and topological descriptors. In total 225 molecular descriptors (for example, molecular weight, logP, aromatic bonds, and ring blocks) were calculated for each molecule.

Data processing

For each drug pair the distance-based measures (tanimoto score of Morgan, Featmorgan, AtomPair, RDKit, Torsion, Layered and MACCS), MCS features (MCS size, MCS tanimoto score, MCS overlap coefficient), and the molecular descriptors (constitutional, topological, geometrical, electronic and hybrid) were concatenated to create a vector of 460 predictors with a binary response variable ('Match' or 'Nomatch').

Training and test set. Prior to training the classifiers, the data was split into a training-cum-validation set (80%) and a hold-out test set (20%). A naive splitting might result in overly similar examples (i.e., the physicochemical descriptors) in train and test set as the drugs are paired in turn resulting in unrealistically optimistic predictive performance. Thus, to avoid overlapping drug predictors, we adopted a systematic procedure to split the drug pairs to ensure that the train and test sets were independent. This would essentially mean that all the drug pairs present in the test set were exclusive to it and were not seen by the classifier during learning from the training set. This procedure resulted in 1,128 drugs in the train set and 282 drugs in the test set and this split was performed 5 times (referred now as 5 split-sets) (S2A Table). After the splits the drugs were paired resulting in 635,628 pairs in train set and 39,621 pairs in test set (S2A Table). The whole dataset was imbalanced with a class ratio of 0.03 and this imbalance was maintained in all 5 split-sets. We monitored the Superclass information in all 5 split-sets to obtain a balanced representation of the structural diversity of drugs in all sets (S2B **Table**).

Data pre-processing. All 5 sets of train and test data followed same processing and model building steps. They were preprocessed to remove non-informative features by removing constant variables such as with all 0's or all 1's ($n = 82$), and to filter out observations with missing values or null for any predictor variable. This resulted in 378 predictors. The classification models were then trained to classify the binary response variable for target referred as 'Match' and 'Nomatch' using 3 classifiers.

Machine learning

We employed three learning algorithms commonly used for binary classification tasks for large datasets, namely, logistic regression (LR) [56], naïve bayes [NB] [57], and random forests (RF) [58]. This selection was done to cover models that can capture linear and nonlinear relationships as both provide interpretability in terms of feature weights as well as probabilistic predictions and feasibility with computational overhead. The classifiers were run on all 5 split-sets to capture any variability arising because of the random train and test splitting. We used various commonly used performance measures for imbalanced data to report the results.

Logistic regression. Logistic regression models apply the logistic function to weighted linear combinations of their input predictors to obtain predictions. We used two types of regularized logistic regression (L1R and L2R) so that the overall error (cost function) during training is minimized to optimize performance while controlling the complexity of the models leading to better generalizability [59]. The models were trained for optimized learning by setting the cost parameter 'C' and class weights because of data imbalance. The value for 'C' was determined using a heuristics on a balanced subset of data [60]. To account for the highly imbalanced nature of the training set we applied class weights (Match– 0.97, Nomatch– 0.03). Both regularization types, L1R and L2R, were used separately to train models and their performance on the hold-out test set was evaluated respectively on the 5 split-sets (**Fig** 1D). As can be seen from **Fig** 1D, both the linear models failed to adequately predict the response variable suggesting higher order of complexity that cannot be segregated by a linear model.

Naïve bayes. Naïve bayes is a simple yet powerful algorithm which computes the conditional probabilities of target variable assuming independent predictor variables using the Bayes theorem [57]. It is known to perform well on large datasets and has very fast processing times. The performance results of NB on the 5 split-sets are also reported in **Fig** 1E and 1F. As can be assessed from **Fig** 1D, NB performed slightly better than LR but was also not efficient to delineate the complexity of the input dataset.

Random forest. A random forest (RF) is a collection of decision trees, each of which is trained on a subsampled version of the original dataset. The predictions of individual trees are averaged to provide a final prediction for the forest. Classically, RFs are known to have strong performance on various computational chemistry tasks and have been state-of-the-art in various cheminformatics settings till date such as to predict chemical binding similarities [35], learning drug functions from their chemical structures [61], in-vitro toxicity prediction [62], and drug-target interactions [63].

Since the default parameters might not be optimal for complex learners, we used hyperparameter tuning for building an optimized RF model. Hyperparameters (ntree (number of trees), nodesize (number of observations at terminal nodes), mtry (number of variables to split at each node) and classwt (class weight) were randomly searched over 10-iterations in a 5-fold cross-validation (CV) setting. This generated 10 combinations of tuned parameters from which the final parameters were selected using the performance metric MCC [64]. Even with other commonly adopted metrics such as in (**Fig** 1D) or accuracy (S2C **Table**), the set of best performing hyperparameter combinations generated consistent results in the 10 iterations each with 5-fold CV. The hyperparameter tuned results (S2C **Table** showed all 10 iterations performed comparably indicating robustness of the RF hyperparameters on our dataset. The best performing hyperparameters combination for each split-set i.e.,–*ntree*, *nodesize*, *mtry*, *classwt* were selected by the best value of MCC which was used to train the final models. The models were tested using the respective independent test set and performance metrics are depicted in **Fig** 1D–1F. Data skewness can significantly impact performance metrics [65]. We therefore used a combination of metrics to evaluate the model's performance on the test set. We focused on balanced threshold metrics (**Fig** 1D) (Matthews correlation metric (MCC) and F1) in combination with rank metrics (**Fig** 1E and 1F) (precision-recall curve and ROC curves) to evaluate our model's performance (metrics definitions explained in S3 **Table**). Overall, the RF model performed robustly, achieving an AUC of 0.90, precision of 0.59 and recall of 0.23 averaged over all 5 split-sets.

Extracting feature importance. During training, RF models were configured to generate feature/variable importance measures. For each tree, the prediction error on the out-of-bag (estimating the prediction performance by evaluating predictions on those observations that were not used in building of the base learner) portion of the data is recorded (i.e., error rate for classification). Then the same is done after permuting each predictor variable. The difference between the two are then averaged over all trees and normalized by the standard deviation of the differences. If the standard deviation of the differences is equal to 0 for a variable, the division is not done (but the average is almost always equal to 0 in that case) [58]. We extracted these feature importance values from the final trained model to explore which predictors contained highly predictive information. The measure is the mean decrease in accuracy calculated on out-of-bag data. Higher the number, higher is the importance. The feature importance from high to low (only top ranking) are shown in **Fig** 1G. The 'Featmorgan' was the best predictor variable with the highest importance value s in our 5 trained models closely followed by other distance-based fingerprints.

Predicting natural compounds and drugs similarities

Source of natural compounds and creation of drug-food pairs. A catalogue of 11,788 natural compounds was obtained from FooDB (www.foodb.ca) (S1C **Table**). These correspond to 261 unique food sources and are categorized into 15 main food types such as vegetables, fruits, herbs and spices, and milk products (S1B **Fig**). For the simplicity in representation in S1B **Fig**, the frequency accounts for only one source per compound; however, a particular compound can be present in multiple food sources. The food compounds were structurally classified into 21 classes (see Material and methods) (S1B **Fig**). Highly represented were lipids and lipid-like molecules (4803), phenylpropanoids and polyketides (2476), organoheterocyclics (1381) and organic oxygen compounds (1120). All these natural compounds were used to create an assessment library, where each chemical pair comprised of a drug and a natural compound (**Fig** 1A) (now referred to as drug-food pair). Pairwise similarities between each natural compound (S1C **Table**) and all the drugs were computed using the same set of predictors (i.e., using same predictors molecular fingerprints, MCS and molecular descriptors) as was for training dataset.

Similarity predictions by RF model. The similarity of each of the 11k natural compounds paired to each of the 1410 drugs was evaluated using the trained RF models. Since all 5 split sets performed optimally, we chose to accommodate the drug-food similarity predictions from all 5 RF models. As many drugs have originated from natural compounds, drug-food pairs with a very high similarity fingerprint score (i.e., tanimoto score > 0.9) were removed (n = 1,850 pairs) considering them to be the same compound. Overall, the number of drug-food pairs compared on each RF model were 1,941,762. Pairs which passed the threshold of 0.5 probability are considered as a match. The number of hits with the 5 RF models are shown in Fig 2A. We picked drug-food pairs which were predicted as match by at-least 3 models (686 pairs) in further analysis. The full list of these 686 pairs is provided in S2B **Table**.

These 686 pairs comprise of 329 unique food compounds and 289 unique drugs (full annotated list in S4B **Table**). Note that a drug can share similarity with more than one food compound and vice-versa. Also, a food compound can be present in multiple food sources, for exhaustive listing of known sources, we recommend querying the FooDB using respective compound Ids or names.

We performed manual curation of 30% of these 686 drug-food pairs (200 pairs) and categorized the food compounds into five custom defined groups based on the meta information available in the public domain (S4A **Table** and **Fig** 2B). Group 1, Analogs: food compounds which themselves represent a drug. Group 2, Endogenous: compounds reported as a metabolite in humans. Group 3, Experimental: food compounds currently under investigational or clinical trial as a therapeutic. Group 4, Probable lead: compounds with potentially novel bioactivity. Group 5, Others: compounds currently used in industrial application or used as additive or flavor enhancer. Each group is discussed below with case examples.

Group 1-analogs (56 pairs, 20 natural compounds)—food compounds in this group were found to be structural analogues to other known drugs i.e., the food compound paired with the drug is apparently another drug itself. This observation is consistent with the fact that the origin of many drugs is from natural sources. Yet, presence of some of these compounds in food is intriguing. A compound referred to as 'Satiomem' in FooDB (reported in barley and onions) resembles the drug Carbinoxamine, an antihistamine. It shared similarity with other antihistamines (such as Chlorprothixene, Doxepin, Antazoline and Chlorphenamine, **Fig** 2B, highlighted in blue) belonging to the nervous ATC category which are used as antipsychotics. These drugs share 'Histamine H1 receptor' as a target but there was no evidence found for Satiomem/Carbinoxamine having antipsychotic activity so it could potentially serve as a good candidate for further testing as an antipsychotic or resulting in drug interactions when used in combination with these drugs. This group also provides an opportunity to explore drug-repurposing.

Group 2-endogenous (49 pairs, 24 natural compounds)—compounds that are endogenous to human tissues but also reportedly present in various food sources. For example, 'desoxycorticosterol' a.k.a. 21-Hydroxyprogesterone was reported to be present in rice and is endogenously present in amniotic fluid and blood throughout human tissues. It's predicted to be similar to other Hormonal and Genitourinary drugs (**Fig** 2B, highlighted in pink). 'Estriol', an estrogen produced by the human body, is reported to be present in pomegranate and beans.

Group 3-experimental (10 pairs, 5 natural compounds)—these food compounds are already under experimental investigation category (i.e., under approval to be used as drugs, reported in DrugBank accessed January 2018). 'Higenamine' is reported to be present in opium and coffee. This compound is in clinical trial (DrugBank id: DB12779) and has been patented for various therapeutic applications (**Fig** 2B, highlighted in purple). This group serve as a proof of principle that we could recall natural compounds with similar activity as currently used human-targeted drugs, which are being actively investigated pre-clinically.

Group 4-probable lead (76 pairs, 58 natural compounds)–to our knowledge, the compounds in this group have little or no hitherto reported evidence of their physiological or biological activity. The drug Papaverine is an alkaloid which is a vasodilator. 'Annocherine B' reportedly

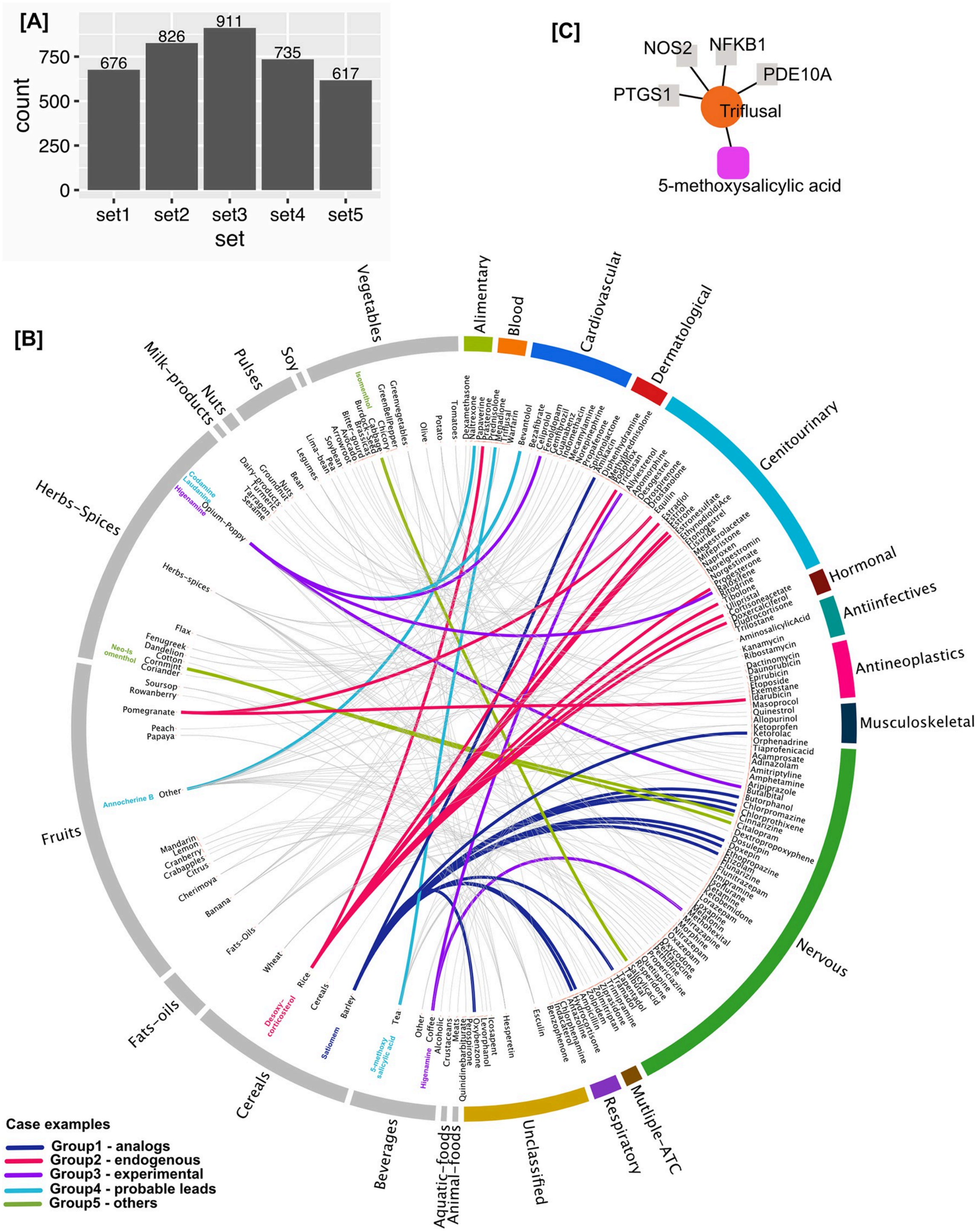


Fig 2. Drug-food compound similarity. (A) Number of hits retrieved from each split-sets model. (B) 200 drug-food pairs predicted as 'match' at the probability threshold of >0.5. The drugs are arranged according to their therapeutic class and food compounds according to their food source. The highlighted colored links represent the case examples in the five author defined groups (details in the text). (C) Group4-probable lead example taken up for experimental validation. The food compound 5-methoxysalicylic acid was a hit with the drug triflusal which has 4 known targets. We validated the inhibitory activity of triflusal and 5-methoxysalicylic acid against the target PTGS1 (also known as Cox-1).

https://doi.org/10.1371/journal.pcbi.1010029.g002

present in many fruits showed high similarity with Papaverine, however no evidence or reports of this compound about its action or use was found. 'Bevantolol' is a cardiovascular drug which shared similarity with two novel compounds 'Codamine' and 'Laudanine' (both reported in opium).

Group 5-others (9 pairs, 5 natural compounds)–food compounds found in this group are reported to be used as food additives such as flavor enhancers or have other industrial applications such as emulsifiers. 'Neoisomenthol' and 'Isomenthol' are used as a flavoring agent are similar with nervous category drugs 'Codeine', 'Dezocine', and 'Tapentadol'. Thus, these five groups highlighted interesting similarity relationships existing between drug and food compounds and their wider therapeutic potential.

Experimental validation of Cyclooxygenase-1 (Cox-1) inhibition by 5-methoxysalicylic acid

With the growing number of new and improved learners [66,67] there's always a possibility to benchmark and compare the performances with different learners. There are multiple scenarios under which one learner can be chosen over another for a given problem. For instance, the performance of nonlinear RF model over the linear learner logistic regression models has been benchmarked previously on a large number of experimental datasets [68]. While this complements our observation, our major goal was to present a proof-of-concept of the learner's predictions in real and not merely restrict to their performance comparisons. As has been proposed, complementary experiments that can validate any model's predictions can help build trust in the method or it's outputs [26].

An interesting case for experimental follow-up in our 200 curated high-confidence similarity pairs is that of triflusal and 5-methoxysalicylic acid (**Fig** 2C). Triflusal is an antithrombotic anticoagulant and is considered very important for the secondary prevention of ischemic stroke. Triflusal has an antagonist effect on prostaglandin G/H synthase 1 (PTGS1) (also called Cox-1) in platelets [69]. Consistent with our prediction, 5-methoxysalicylic acid, which is found in tea, herbs and spices, has been shown to have antiplatelet activity in rats [70]. Yet, the molecular target of 5-methoxysalicylic acid is not known. We therefore chose to assess Cox-1 binding activity of 5-methoxysalicylic acid. To attest the utility of ML approach in comparison to using a single similarity measure (such as a selected fingerprint), we also tested a 'negative control' molecule, 4-isopropylbenzoic acid, which is found in cumin, herbs and spices. This molecule had the highest similarity with triflusal based on Featmorgan, which was the most important predictor (**Fig** 1G). While 4-isopropylbenzoic acid was predicted as 'Nomatch' by the RF model, our test compound (5-methoxysalicylic acid) was predicted as a 'Match' but had a lower Tanimoto Score for Featmorgan (S3A **Fig**). Structural representation of all the tested compounds and their shared MCS (maximum common substructure) is depicted in **Fig** 3A.

We tested the Cox-1 inhibitory activity of the three compounds, triflusal (positive control), 5-methoxysalicylic acid (test compound), and 4-isopropylbenzoic acid (negative control) using an enzymatic assay based on fluorometric detection of prostaglandin G2, which is an intermediate product generated by the Cox enzyme (S3B **Fig**).

Triflusal and 5-methoxysalicylic acid exhibited highly similar inhibition profiles over different concentrations (**Fig** 3B and S5 **Table**). The IC50 (is a measure of potency of a substance in inhibiting a biological or biochemical function) of triflusal was estimated to be ~100 μM, while 5-methoxysalicylic acid showed ~40% inhibition at 100μM. Maximum inhibition achieved with SC560 –a positive control included in the assay kit–was 42%, comparable to that of 5-methoxysalicylic acid. In stark contrast, 4-isopropylbenzoic acid did not show any inhibition at 100 μM. At higher concentrations, it caused a color formation (bright pink) and hence was

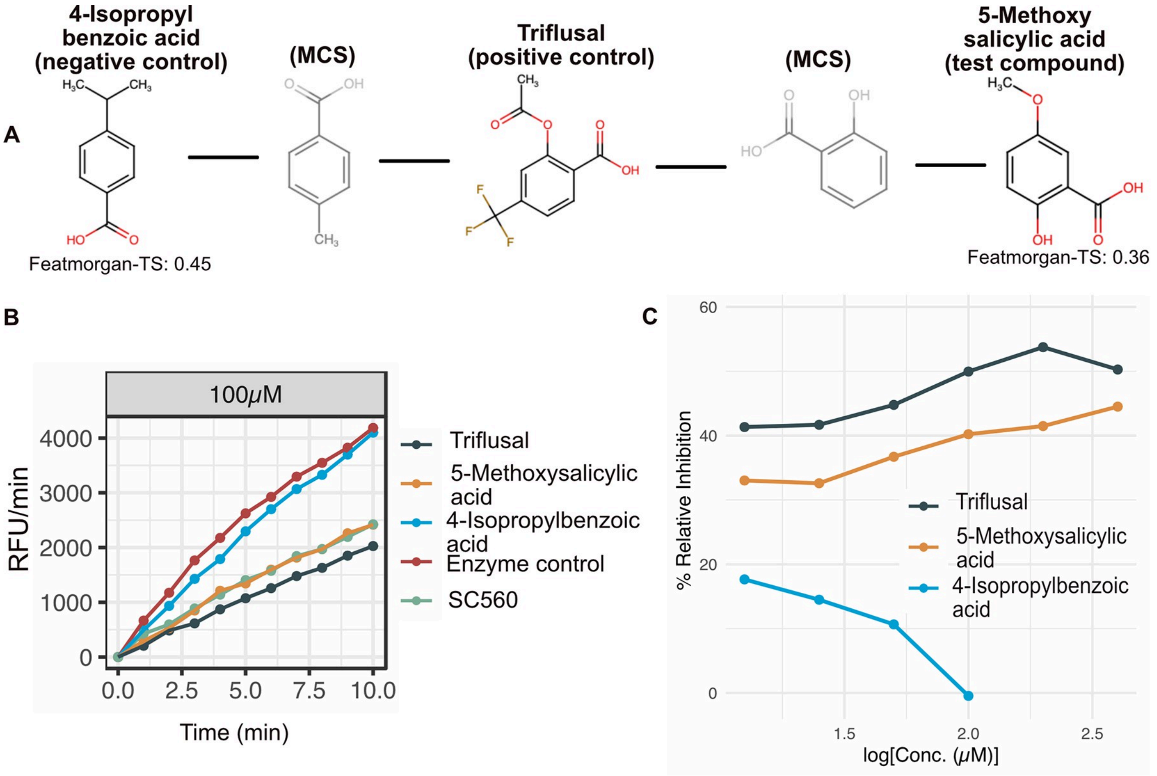


Fig 3. Cox-1 inhibitor assay. (A) Chemical structures of all the tested compounds. MCS structures are also depicted which helped to intuitively assess the structural similarity between the tested compounds (B) An example relative fluorescent units (RFU) plot of the tested compounds at 100μM (other tested conc.: 12.5μM to 400μM serial dilutions). SC560 is a positive control provided by the assay kit supplier (Materials and methods). (C) Relative inhibition of the positive control (drug triflusal), test compound (5-methoxy salicylic acid) and negative control (4-isopropyl benzoic acid) at different tested concentrations. 5-methoxy salicylic acid showed similar inhibition of Cox-1 as the drug triflusal whereas no such inhibition was observed for 4-isopropyl benzoic acid. 4-isopropyl benzoic acid showed strong color change (bright pink) reaction beyond 100μM and thus was found unsuitable for being tested at higher concentration with this assay.

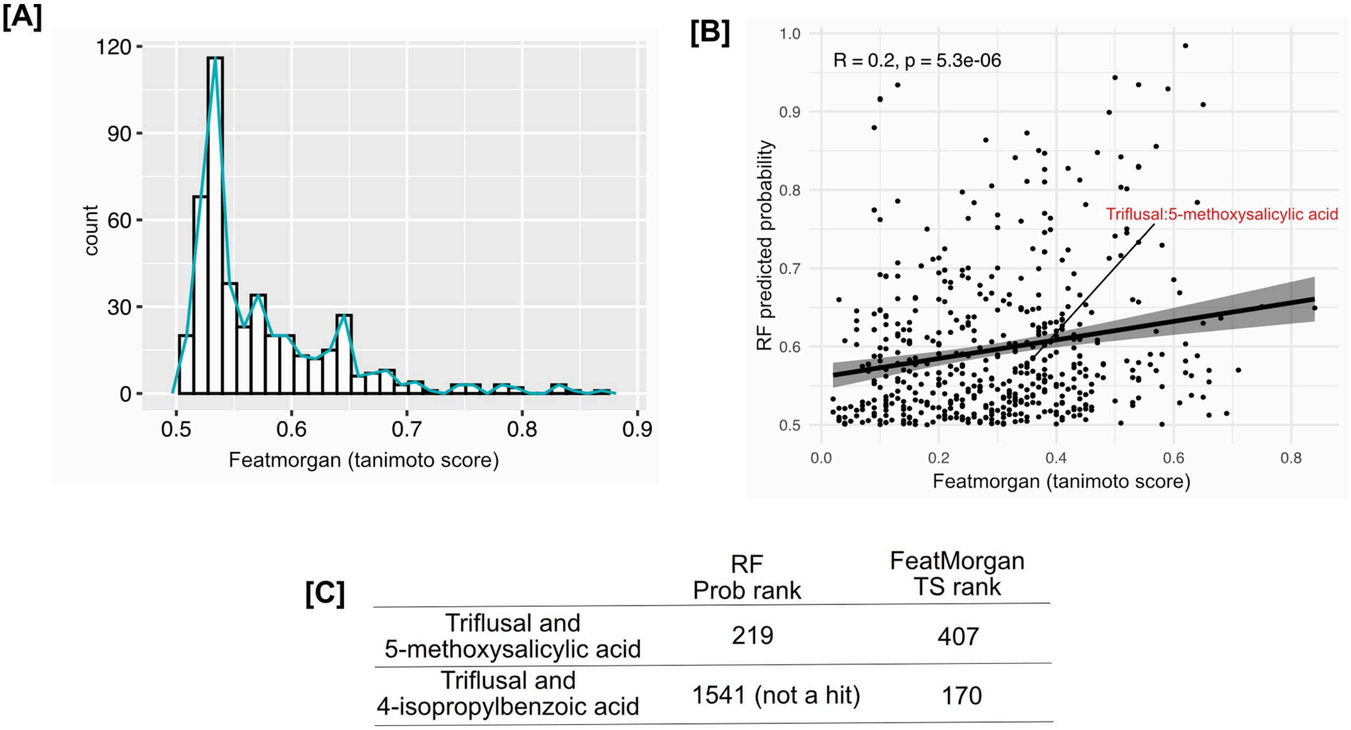
https://doi.org/10.1371/journal.pcbi.1010029.g003

not tested beyond 100 μM. It only showed a small (10–17%) inhibitory effect at a lower concentration, possibly be due to non-specific binding. Taken together, the ML predicted drug-natural compound pair experimentally showed the same target binding with quantitatively similar activity, supporting the underlying model.

ML vs single fingerprint

We compared the performance of a single fingerprint alone ('featmorgan' as it was the best predictor in our analysis and is also popularly used) versus our RF models on all split-sets trained in this study.

Featmorgan based hits grew significantly in number as the tanimoto score threshold was lowered (> 0.7–21 compounds, > 0.6 tanimoto score– 112 compounds, > 0.5–451 compounds) (Fig 4A). This behavior is not helpful in a drug discovery setting where computational shortlisting is meant to reduce the number of hits that can be thereafter tested experimentally. The validation pair triflusal-5-methoxysalicylic acid was predicted as a match with an averaged probability (across all five split-sets) of 0.58 but had a much lower tanimoto score for Featmorgan (0.36) (Fig 4B). This pair could have been easily missed (would be a false negative) if Featmorgan was used alone to rank molecule similarity and thereby estimate their activity profile.



	RF Prob rank	FeatMorgan TS rank
Triflusal and 5-methoxysalicylic acid	219	407
Triflusal and 4-isopropylbenzoic acid	1541 (not a hit)	170

Fig 4. RF vs featmorgan. (A) Number of hits retrieved by using tanimoto score with featmorgan as similarity measure, which grows markedly as threshold is reduced (lower threshold means less similarity). (B) correlation between RF models' average probability predictions >0.5 with corresponding tanimoto score of featmorgan of drug-food pairs. Our hit pair triflusal and 5-methoxysalicylic acid (highlighted in red) was predicted a hit by RF models (as top 219th pair) would be missed by featmorgan if used alone. (C) Rank comparison between hit pair (Triflusal:5-methoxysalicylic acid) and the negative control (Triflusal:4-isopropylbenzoic acid). The negative control was not a hit using RF models although had a higher rank with featmorgan than the hit pair and vice versa.

https://doi.org/10.1371/journal.pcbi.1010029.g004

In contrast, the negative control, despite having a higher rank with featmorgan (Fig 4C), wasn't even picked by the RF model.

Discussion and conclusion

This study has addressed three goals: identification of potential molecular targets of ingested natural compounds and exploring their therapeutic potential; evaluating the utility of a comprehensive ML based approach to deconvolute the complex SAR between molecules as opposed to restricting to a single similarity measure; and, lastly, complementing in-silico model predictions with experimental validation to build trust in model's predictions.

Systematically integrating computational chemistry approaches can help deconvolute the intricate structure-activity relationship between small molecules and their biological targets. The data fusion approach–i.e., integrating multiple similarity metrics based on fingerprints, maximum common substructures, and physicochemical descriptors–used in this study proved effective in identifying natural compounds functionally similar to known therapeutic drugs. The RF models achieved a good performance with an average AUC of 0.9. Analysis of the curated 200 drug-food pairs predicted from the models helped to capture drug analogs, host endogenous metabolites, some investigational drugs, as well as novel molecular leads present in various food sources which are deemed to share the same target as the drugs.

Data fusion approaches can aid in reducing the amounts of artefacts [30]. In line with this, various methods are being proposed to accelerate the performance of in-silico similarity searches such as inclusion of bioactivity profiles [27,71,72], multiple fingerprint algorithms [73], and similarity ensemble approach [74]. Our approach serves as another promising addition to these strategies. A recent study has also demonstrated similar strategy to identify molecules using pairwise similarity concept and target engagement [35] which supports our strategy and showcases the utility and success rate of such methods in early-stage molecule discovery.

In the context of natural compounds search space, two closely related issues remain open before the proposed approach could be more broadly applied: the estimation of true-positive rate, and the availability of curated molecular target information for an additional, structurally more diverse, set of small molecules. The former could be addressed through high-throughput testing of bioactivity profile of natural compounds against a set of protein targets using cell-based or enzymatic assays [75,76]. The generated data could then be used to address the latter issue. Further, new structured and curated datasets such as those recently reported by Duran et al [27] would be valuable to this end. Our current study is limited to single target sharing between drugs which has been the classical and predominant starting point of many drugs discovery programs [77]. However, to deconvolute the complex interplay of multiple targets such as in phenotypic drug discovery [77] would require incorporation of additional level of information which is beyond the scope of current study. ML models trained using knowledge-based graphs and networks would be pivotal at this end [26].

The ML approach used here was notable in capturing complex high-dimensional similarity that would not be accessible based on any structural similarity metric used in isolation. Indeed, we could show that a molecule that is highly similar based on a single fingerprint's tanimoto score did not show any appreciable activity. In contrast, the natural compound identified using RF model trained on multiple features showed the predicted enzyme inhibitory activity. In further support, the model could also identify several drug-food compound relations including compounds that are currently under investigation or have been ascribed with related bioactivity in the literature. Taken together, this study has implications for efficient exploration of drug-like properties of natural compounds.

Materials and methods

Data source and processing

All the FDA approved drugs which had target information (S1A and S1B **Table**) associated with them were taken from DrugBank [51] (accessed January 2018). The natural compound library used for virtual screening was obtained from FooDB (www.foodb.ca; freely available and accessed June 2017, (~11k compounds). It was curated and formatted to be smoothly integrated into our analysis. The full list of compounds with their annotations is provided in S1C **Table**. It included compounds from both raw and processed foods. We used drug classification codes from ATC (https://www.whocc.no/) to therapeutically classify all the drugs and ClassyFire [50] to structurally classify the drugs and the natural compounds.

Computing predictor variables

In order to predict the molecular targets for the natural compounds based on their pairwise structural similarities with the drugs, we needed to create a ML dataset that contains molecular features (i.e., predictors) computed from the chemical structures of the drugs and gives out a binary response for the target. This model can turn then be applied to natural compounds and drug pairs to predict the potential targets for natural compounds. For this, a number of pairwise distance measures and molecule specific physicochemical descriptors were generated for drugs and all the natural compounds. These included distance-based fingerprint similarities, maximum common substructure similarities and physicochemical descriptors. These molecular features formed the basis of our ML dataset (explained in later section) created from drugs and their known respective targets.

Fingerprint calculation, tanimoto score estimation and molecular descriptor generation.

The INCHIs from drugs and natural compound library were used to generate 2D structural information in **S**tructural **D**ata **F**ormat (SDF). These SDF files were used to calculate 7 different molecular fingerprints (Morgan (circular fingerprints, ECFP-like), Featmorgan (circular fingerprints, FCFP-like), Atompair, RDKit (daylight-like topological fingerprints), Torsion (topological-torsion), Layered (substructure-matching fingerprint) and MACCS (RDKit implementation of 166 public MACCS keys)) to gather theoretical 2D structural information from the molecules. The entire workflow was designed using the KNIME [78] analytics platform utilizing RDKit [79] plugin (for fingerprints) with default parameters. The pairwise structural similarity from the fingerprints was scored using the widely used Tanimoto similarity metric (computed as $TS_{AB} = (A \cap B)/(A \cup B)$). We also computed maximum common substructure (MCS) shared between each chemical pair. The MCS is a graph-based similarity search wherein the largest substructure shared between query and target is identified and gives out various parameters such as number of MCS generated, size of each molecule, size of MCS, tanimoto score, overlapping coefficient (computed as $OC_{AB} = (A \cap B)/min(A, B)$). It was computed using the 'ChemmineR' [80] and 'fmcsR' [53] packages available for R. Although MCS calculation is computationally intensive and time-consuming, but they are more sensitive, accurate and intuitive, thus we implemented this computation in batch-mode on high-performance computing cluster for faster processing. In addition to the distance-based features, five different types of molecular descriptors (constitutional, topological, geometrical, electronic and hybrid) were computed for all compounds using R package: RCDK [81].

Data preprocessing and machine learning

All predictors (computed above) for the paired drug data were combined into a single matrix with each observation associated with a response variable ('Match' or 'Nomatch'). This data was then split into an 80% train-cum-validation set and a 20% test set by performing random selection of drugs from the highly represented Superclasses such that both sets were mutually exclusive and had a balanced distribution of the drug classes. Following this the training set was pre-processed to remove constant predictors and missing value observations to avoid training on noise.

We explored both linear and nonlinear learners to build our binary classification models. The linear learners used were the two types of regularized logistic regression: called as L1R and L2R ('LiblineaR' package in R) and the nonlinear learners were naïve bayes (NB) and random forest ('mlr' package in R). The binary classifiers were trained for two classes which are referred to as 'Match' and 'Nomatch' indicating whether they share a target or not. In order not to increase the system complexity in terms of protein target similarity (as drugs can share multiple targets), each drug pair which shared even at-least one target were considered as 'Match' and the rest were considered as 'Nomatch'. Modern machine learning algorithms require tuning various parameters in order to achieve their best performance [82]. Thus, the classifiers used were optimized accordingly by tuning their parameters.

L1R and L2R logistic regression are extremely fast learners and benefit when input data is centered and scaled in a *nxp* numerical matrix form and a response variable *(1xn)* containing class labels. The two types of regularization used were L1 (type 6) and L2 (type 0) of LiblineaR package which can give out probability estimates of prediction. As our paired data was highly imbalanced, we used class weights where the positive class received a higher weight ratio

(Match—0.97) then the negative class (Nomatch—0.03). The weights were derived from ratio of positive/negative classes. To handle misclassification errors, costs were determined by using 'heuristicC' function on balanced sub-sample of the dataset [60]. Naïve bayes classifier is also has fast processing times and works well with large datasets. It assumes feature independence and is based on the Bayes theorem of conditional probabilities [57]. The rationale for testing multiple classifiers was to cover a range of inductive biases and to pick a simpler model if it shows good performance (e.g., in case L1R and RF show similar performance then L1R would be selected).

We defined a search space to tune the hyperparameters for random forest (RF) to obtain settings suitable for the data at hand. Four broadly used hyperparameters were tuned: the number of trees (*ntree*), the number of observations at terminal nodes (*nodesize*), number of variables to split at each node (*mtry*) and class weights (*weight*). We limited the *ntree* to 300 as setting higher *ntree* values would result in increased computational overhead (i.e., training time and memory usage). The default value for the parameter *nodesize* is 1, but with low values of tree depth, the tree can fail to recognize useful signals from the data. We searched for *nodesize* value in the range 20–50. Lower *nodesize* can result in lower detection signals of the true positives and a high false-negative rate. The default $mtry = \sqrt{p}$ where p is the number of features in the input data. In our ML dataset, the default *mtry* was ($\sqrt{378}$) 19 features. We searched *mtry* in the range of 15–30. Lastly, as our training data set had high class imbalance (Match-Nomatch ratio of 0.03), we also tuned the class weight parameter starting with a minimum weight of 300 for the positive class ('Match') and searching up to 10 times i.e., 3000. The hyperparameters were additionally optimized for 10 random iterations with 5-fold cross-validation each using stratification. Owing to the size of the data and to speed up the iterations and parameter search the tuning was performed in a cluster environment with parallel backend.

We used several standard performance measures (mean test values for MCC (Matthews correlation coefficient), Balanced accuracy (BAC), Kappa, MMCE (mean classification error), ACC (accuracy), TPR (true positive rate/Recall/Sensitivity), FPR (false positive rate), TNR (true negative rate), (false negative rate), PPV (Precision/Positive predictive value)) to evaluate the learner's performance.

Compound preparation and assay protocol

Briefly, all tested compounds were dissolved in their respective solvents. Triflusal and 5-methoxysalicylic acid were dissolved in DMSO and 4-isoproplybenzoic acid was dissolved in ethanol. Compounds supplied with the assay kit were prepared as per the manufacturer's protocol and the assay was also performed according to the instructions present in the kit from Abcam (CAT#ab204698). The kit included the Cox-1 enzyme (source: ovine) and had a positive control Cox-1 inhibitor (SC560).

Literature evidence showed that triflusal binds to purified Cox-2 at 240–320 μM [83]. Thus, we assayed all compounds at different concentrations starting at 400μM and going down to 12.5μM with serial dilutions and in triplicates. Relative Fluorescence Units (RFU) were measured immediately after starting the reaction by using microplate reader (Tecan infinite M1000Pro) at Ex/Em = 535/587 nm in a kinetic mode for 40 minutes at 25° C. All fluorescence readings for triplicates under a given concentration were averaged and initial time point RFU reading was used to shift the measurements to start from 0. We took the first 10 time-points of RFU readings to assess the inhibitory effect of the tested compounds. Slopes for all samples (triflusal (positive control), 5-methoxysalicylic acid (test compound) and 4-isopropylbenzoic acid (negative control)), enzyme control, and kit supplied positive control (SC560) were calculated by fitting linear equations, respectively. Percent relative inhibition for samples was calculated as

$$\% \; Relative \; Inhibition = \frac{slope \; of \; enzyme \; control - slope \; of \; sample}{slope \; of \; enzyme \; control} * 100$$

Supporting information

S1 Table. List of drugs (S1A Table), drug targets (S1B Table), food compounds (S1C Table) and count of molecular descriptors (S1D Table).
(XLSX)

S2 Table. Training and test sets description (S2A Table), superclass classification of different input sets (S2B Table) and full performance metrics of random forest models on 5 split-sets (S2C Table).
(XLSX)

S3 Table. Model performance metrics and their description.
(DOCX)

S4 Table. Author curated list of 200 drug-food pairs (S4A Table) and full list of 686 drug-food pairs predicted as hits (S4B Table).
(XLSX)

S5 Table. Cox-assay results. Relative Inhibition (%)
(DOCX)

S1 Fig. Structural classification of drugs (a) and foods (b).
(TIFF)

S2 Fig. Concordance at top (CAT plots) of fingerprints. 'Concordance at top' (CAT plots)[55] of fingerprints. We performed a rank-based assessment of drug-pairs called by each fingerprint as top scoring pairs and found a low concordance between them. The CAT plots here depict the concordance of each fingerprint (when taken as reference) against all other fingerprints. Each fingerprint was taken as a reference and the top 100 high scoring drug-pairs were ranked in decreasing order to estimate the overlapping proportions between each fingerprint when compared with the reference. Mathematically, for *ith* top-ranked molecules, concordance is defined as $length(intersect(list1[1:i], list2[1:i]))/i$.
(TIFF)

S3 Fig. Cox-assay description. (A) Table describes the compounds tested in the cox-1 inhibitor assay, their source and usage. Triflusal is the drug which is known to bind to Cox-1 and it was a positive control in our experiment. The food compound which came out as hit with our prediction models is 5-methoxysalicylic acid (referred to as test compound) for which target engagement is being studied. Additional inclusion was 4-isopropylbenzoic acid (selected based on high FM score as compared to test compound but deemed no match by prediction models) as a negative control. (B) The reaction mechanism involves fluorometric detection of intermediate product (prostaglandin G2) generated by cox-enzyme.
(TIFF)

Author Contributions

Conceptualization: Vinita Periwal, Sergej Andrejev, Athanasios Typas, Kiran Raosaheb Patil.

Data curation: Vinita Periwal.

Formal analysis: Vinita Periwal, Kaustubh Raosaheb Patil.

Funding acquisition: Vinita Periwal, Stefan Bassler, Natalia Gabrielli, Kiran Raosaheb Patil.

Investigation: Vinita Periwal, Stefan Bassler, Natalia Gabrielli.

Methodology: Vinita Periwal, Stefan Bassler, Sergej Andrejev, Natalia Gabrielli, Kaustubh Raosaheb Patil.

Project administration: Kiran Raosaheb Patil.

Resources: Athanasios Typas, Kiran Raosaheb Patil.

Supervision: Kaustubh Raosaheb Patil, Athanasios Typas, Kiran Raosaheb Patil.

Validation: Vinita Periwal, Stefan Bassler, Natalia Gabrielli.

Visualization: Vinita Periwal.

Writing – original draft: Vinita Periwal, Kiran Raosaheb Patil.

Writing – review & editing: Vinita Periwal, Kaustubh Raosaheb Patil, Kiran Raosaheb Patil.

References

1. Newman DJ, Cragg GM. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. J Nat Prod. 2020; 83(3):770–803. Epub 2020/03/13. https://doi.org/10.1021/acs.jnatprod.9b01285 PMID: 32162523.
2. Atanasov AG, Zotchev SB, Dirsch VM, International Natural Product Sciences T, Supuran CT. Natural products in drug discovery: advances and opportunities. Nat Rev Drug Discov. 2021; 20(3):200–16. Epub 2021/01/30. https://doi.org/10.1038/s41573-020-00114-z PMID: 33510482; PubMed Central PMCID: PMC7841765.
3. Harvey AL, Edrada-Ebel R, Quinn RJ. The re-emergence of natural products for drug discovery in the genomics era. Nat Rev Drug Discov. 2015; 14(2):111–29. Epub 2015/01/24. https://doi.org/10.1038/nrd4510 PMID: 25614221.
4. Rodrigues T, Reker D, Schneider P, Schneider G. Counting on natural products for drug design. Nat Chem. 2016; 8(6):531–41. Epub 2016/05/25. https://doi.org/10.1038/nchem.2479 PMID: 27219696.
5. Shen B. A New Golden Age of Natural Products Drug Discovery. Cell. 2015; 163(6):1297–300. Epub 2015/12/08. https://doi.org/10.1016/j.cell.2015.11.031 PMID: 26638061; PubMed Central PMCID: PMC5070666.
6. Corbi G, Conti V, Davinelli S, Scapagnini G, Filippelli A, Ferrara N. Dietary Phytochemicals in Neuroimmunoaging: A New Therapeutic Possibility for Humans? Front Pharmacol. 2016; 7:364. Epub 2016/10/30. https://doi.org/10.3389/fphar.2016.00364 PMID: 27790141; PubMed Central PMCID: PMC5062465.
7. Hosseini A, Ghorbani A. Cancer therapy with phytochemicals: evidence from clinical studies. Avicenna J Phytomed. 2015; 5(2):84–97. Epub 2015/05/08. PMID: 25949949; PubMed Central PMCID: PMC4418057.
8. Alissa EM, Ferns GA. Dietary fruits and vegetables and cardiovascular diseases risk. Crit Rev Food Sci Nutr. 2017; 57(9):1950–62. Epub 2015/07/21. https://doi.org/10.1080/10408398.2015.1040487 PMID: 26192884.
9. Gu HF, Mao XY, Du M. Prevention of breast cancer by dietary polyphenols-role of cancer stem cells. Crit Rev Food Sci Nutr. 2019:1–16. Epub 2019/01/12. https://doi.org/10.1080/10408398.2017.1355775 PMID: 28799777.
10. Hartley L, Flowers N, Holmes J, Clarke A, Stranges S, Hooper L, et al. Green and black tea for the primary prevention of cardiovascular disease. Cochrane Database Syst Rev. 2013;(6):CD009934. https://doi.org/10.1002/14651858.CD009934.pub2 PMID: 23780706.
11. Briguglio M, Hrelia S, Malaguti M, Serpe L, Canaparo R, Dell'Osso B, et al. Food Bioactive Compounds and Their Interference in Drug Pharmacokinetic/Pharmacodynamic Profiles. Pharmaceutics. 2018; 10(4). Epub 2018/12/19. https://doi.org/10.3390/pharmaceutics10040277 PMID: 30558213; PubMed Central PMCID: PMC6321138.

12. Jensen K, Ni Y, Panagiotou G, Kouskoumvekaki I. Developing a molecular roadmap of drug-food interactions. PLoS Comput Biol. 2015; 11(2):e1004048. Epub 2015/02/11. https://doi.org/10.1371/journal.pcbi.1004048 PMID: 25668218; PubMed Central PMCID: PMC4323218.

13. Rodriguez-Fragoso L, Martinez-Arismendi JL, Orozco-Bustos D, Reyes-Esparza J, Torres E, Burchiel SW. Potential risks resulting from fruit/vegetable-drug interactions: effects on drug-metabolizing enzymes and drug transporters. J Food Sci. 2011; 76(4):R112–24. Epub 2012/03/16. https://doi.org/10.1111/j.1750-3841.2011.02155.x PMID: 22417366.

14. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, et al. Diet rapidly and reproducibly alters the human gut microbiome. Nature. 2014; 505(7484):559–63. Epub 2013/12/18. https://doi.org/10.1038/nature12820 PMID: 24336217; PubMed Central PMCID: PMC3957428.

15. Kolodziejczyk AA, Zheng D, Elinav E. Diet-microbiota interactions and personalized nutrition. Nat Rev Microbiol. 2019; 17(12):742–53. Epub 2019/09/22. https://doi.org/10.1038/s41579-019-0256-8 PMID: 31541197.

16. Sonnenburg JL, Backhed F. Diet-microbiota interactions as moderators of human metabolism. Nature. 2016; 535(7610):56–64. https://doi.org/10.1038/nature18846 PMID: 27383980.

17. Zmora N, Suez J, Elinav E. You are what you eat: diet, health and the gut microbiota. Nat Rev Gastroenterol Hepatol. 2019; 16(1):35–56. Epub 2018/09/29. https://doi.org/10.1038/s41575-018-0061-2 PMID: 30262901.

18. Clardy J, Walsh C. Lessons from natural molecules. Nature. 2004; 432(7019):829–37. https://doi.org/10.1038/nature03194 PMID: 15602548.

19. Koehn FE, Carter GT. The evolving role of natural products in drug discovery. Nat Rev Drug Discov. 2005; 4(3):206–20. https://doi.org/10.1038/nrd1657 PMID: 15729362.

20. Pye CR, Bertin MJ, Lokey RS, Gerwick WH, Linington RG. Retrospective analysis of natural products provides insights for future discovery trends. Proc Natl Acad Sci U S A. 2017; 114(22):5601–6. Epub 2017/05/04. https://doi.org/10.1073/pnas.1614680114 PMID: 28461474; PubMed Central PMCID: PMC5465889.

21. Wang S, Dong G, Sheng C. Structural Simplification of Natural Products. Chem Rev. 2019; 119 (6):4180–220. Epub 2019/02/08. https://doi.org/10.1021/acs.chemrev.8b00504 PMID: 30730700.

22. Li F, Wang Y, Li D, Chen Y, Dou QP. Are we seeing a resurgence in the use of natural products for new drug discovery? Expert Opin Drug Discov. 2019; 14(5):417–20. Epub 2019/02/28. https://doi.org/10.1080/17460441.2019.1582639 PMID: 30810395.

23. Yao H, Liu J, Xu S, Zhu Z, Xu J. The structural modification of natural products for novel drug discovery. Expert Opin Drug Discov. 2017; 12(2):121–40. Epub 2016/12/23. https://doi.org/10.1080/17460441.2016.1272757 PMID: 28006993.

24. Kang J, Hsu CH, Wu Q, Liu S, Coster AD, Posner BA, et al. Improving drug discovery with high-content phenotypic screens by systematic selection of reporter cell lines. Nat Biotechnol. 2016; 34(1):70–7. Epub 2015/12/15. https://doi.org/10.1038/nbt.3419 PMID: 26655497; PubMed Central PMCID: PMC4844861.

25. Lo YC, Rensi SE, Torng W, Altman RB. Machine learning in chemoinformatics and drug discovery. Drug Discov Today. 2018; 23(8):1538–46. Epub 2018/05/12. https://doi.org/10.1016/j.drudis.2018.05.010 PMID: 29750902; PubMed Central PMCID: PMC6078794.

26. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019; 18(6):463–77. Epub 2019/04/13. https://doi.org/10.1038/s41573-019-0024-5 PMID: 30976107; PubMed Central PMCID: PMC6552674.

27. Duran-Frigola M, Pauls E, Guitart-Pla O, Bertoni M, Alcalde V, Amat D, et al. Extending the small-molecule similarity principle to all levels of biology with the Chemical Checker. Nat Biotechnol. 2020. Epub 2020/05/23. https://doi.org/10.1038/s41587-020-0502-7 PMID: 32440005.

28. Moret N, Clark NA, Hafner M, Wang Y, Lounkine E, Medvedovic M, et al. Cheminformatics Tools for Analyzing and Designing Optimized Small-Molecule Collections and Libraries. Cell Chem Biol. 2019; 26 (5):765–77 e3. Epub 2019/04/09. https://doi.org/10.1016/j.chembiol.2019.02.018 PMID: 30956147; PubMed Central PMCID: PMC6526536.

29. Sorokina M, Steinbeck C. Review on natural products databases: where to find data in 2020. J Cheminform. 2020; 12(1):20. Epub 2021/01/13. https://doi.org/10.1186/s13321-020-00424-9 PMID: 33431011; PubMed Central PMCID: PMC7118820.

30. Cereto-Massague A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallve S, Pujadas G. Molecular fingerprint similarity search in virtual screening. Methods. 2015; 71:58–63. Epub 2014/08/19. https://doi.org/10.1016/j.ymeth.2014.08.005 PMID: 25132639.

31. Muegge I, Mukherjee P. An overview of molecular fingerprint similarity search in virtual screening. Expert Opin Drug Discov. 2016; 11(2):137–48. Epub 2015/11/13. https://doi.org/10.1517/17460441.2016.1117070 PMID: 26558489.

32. O'Hagan S, Kell DB. Analysis of drug-endogenous human metabolite similarities in terms of their maximum common substructures. J Cheminform. 2017; 9:18. Epub 2017/03/21. https://doi.org/10.1186/s13321-017-0198-y PMID: 28316656; PubMed Central PMCID: PMC5344883.

33. S OH, Swainston N, Handl J, Kell DB. A 'rule of 0.5' for the metabolite-likeness of approved pharmaceutical drugs. Metabolomics. 2015; 11(2):323–39. Epub 2015/03/10. https://doi.org/10.1007/s11306-014-0733-z PMID: 25750602; PubMed Central PMCID: PMC4342520.

34. Maggiora G, Vogt M, Stumpfe D, Bajorath J. Molecular similarity in medicinal chemistry. J Med Chem. 2014; 57(8):3186–204. Epub 2013/10/25. https://doi.org/10.1021/jm401411z PMID: 24151987.

35. Park K, Ko YJ, Durai P, Pan CH. Machine learning-based chemical binding similarity using evolutionary relationships of target genes. Nucleic Acids Res. 2019; 47(20):e128. Epub 2019/09/11. https://doi.org/10.1093/nar/gkz743 PMID: 31504818; PubMed Central PMCID: PMC6846180.

36. Seo M, Shin HK, Myung Y, Hwang S, No KT. Development of Natural Compound Molecular Fingerprint (NC-MFP) with the Dictionary of Natural Products (DNP) for natural product-based drug development. Journal of Cheminformatics. 2020; 12(1). ARTN 6. https://doi.org/10.1186/s13321-020-0410-3 WOS:000513586100001. PMID: 33431009

37. Chan HCS, Shan H, Dahoun T, Vogel H, Yuan S. Advancing Drug Discovery via Artificial Intelligence. Trends Pharmacol Sci. 2019; 40(8):592–604. Epub 2019/07/20. https://doi.org/10.1016/j.tips.2019.06.004 PMID: 31320117.

38. Lavecchia A. Machine-learning approaches in drug discovery: methods and applications. Drug Discov Today. 2015; 20(3):318–31. Epub 2014/12/03. https://doi.org/10.1016/j.drudis.2014.10.012 PMID: 25448759.

39. Lima AN, Philot EA, Trossini GH, Scott LP, Maltarollo VG, Honorio KM. Use of machine learning approaches for novel drug discovery. Expert Opin Drug Discov. 2016; 11(3):225–39. Epub 2016/01/28. https://doi.org/10.1517/17460441.2016.1146250 PMID: 26814169.

40. Rodrigues T, Bernardes GJL. Machine learning for target discovery in drug development. Curr Opin Chem Biol. 2020; 56:16–22. Epub 2019/11/18. https://doi.org/10.1016/j.cbpa.2019.10.003 PMID: 31734566.

41. Piazza I, Beaton N, Bruderer R, Knobloch T, Barbisan C, Chandat L, et al. A machine learning-based chemoproteomic approach to identify drug targets and binding sites in complex proteomes. Nat Commun. 2020; 11(1):4200. Epub 2020/08/23. https://doi.org/10.1038/s41467-020-18071-x PMID: 32826910; PubMed Central PMCID: PMC7442650.

42. Zhang R, Li X, Zhang X, Qin H, Xiao W. Machine learning approaches for elucidating the biological effects of natural products. Nat Prod Rep. 2020. Epub 2020/09/02. https://doi.org/10.1039/d0np00043d PMID: 32869826.

43. Lim S, Lee K, Kang J. Drug drug interaction extraction from the literature using a recursive neural network. PLoS One. 2018; 13(1):e0190926. Epub 2018/01/27. https://doi.org/10.1371/journal.pone.0190926 PMID: 29373599; PubMed Central PMCID: PMC5786304.

44. Ryu JY, Kim HU, Lee SY. Deep learning improves prediction of drug-drug and drug-food interactions. Proc Natl Acad Sci U S A. 2018; 115(18):E4304–E11. Epub 2018/04/19. https://doi.org/10.1073/pnas.1803294115 PMID: 29666228; PubMed Central PMCID: PMC5939113.

45. Yang H, Sun L, Li W, Liu G, Tang Y. In Silico Prediction of Chemical Toxicity for Drug Design Using Machine Learning Methods and Structural Alerts. Front Chem. 2018; 6:30. Epub 2018/03/09. https://doi.org/10.3389/fchem.2018.00030 PMID: 29515993; PubMed Central PMCID: PMC5826228.

46. Zhang L, Zhang H, Ai H, Hu H, Li S, Zhao J, et al. Applications of Machine Learning Methods in Drug Toxicity Prediction. Curr Top Med Chem. 2018; 18(12):987–97. Epub 2018/07/28. https://doi.org/10.2174/1568026618666180727152557 PMID: 30051792.

47. Yosipof A, Guedes RC, Garcia-Sosa AT. Data Mining and Machine Learning Models for Predicting Drug Likeness and Their Disease or Organ Category. Front Chem. 2018; 6:162. Epub 2018/06/06. https://doi.org/10.3389/fchem.2018.00162 PMID: 29868564; PubMed Central PMCID: PMC5954128.

48. Wright MH, Sieber SA. Chemical proteomics approaches for identifying the cellular targets of natural products. Nat Prod Rep. 2016; 33(5):681–708. Epub 2016/04/22. https://doi.org/10.1039/c6np00001k PMID: 27098809; PubMed Central PMCID: PMC5063044.

49. Chen X, Wang Y, Ma N, Tian J, Shao Y, Zhu B, et al. Target identification of natural medicine with chemical proteomics approach: probe synthesis, target fishing and protein identification. Signal Transduct Target Ther. 2020; 5(1):72. Epub 2020/05/22. https://doi.org/10.1038/s41392-020-0186-y PMID: 32435053; PubMed Central PMCID: PMC7239890.

50. Djoumbou Feunang Y, Eisner R, Knox C, Chepelev L, Hastings J, Owen G, et al. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. J Cheminform. 2016; 8:61. Epub 2016/11/22. https://doi.org/10.1186/s13321-016-0174-y PMID: 27867422; PubMed Central PMCID: PMC5096306.

51. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018; 46(D1):D1074–D82. Epub 2017/11/11. https://doi.org/10.1093/nar/gkx1037 PMID: 29126136; PubMed Central PMCID: PMC5753335.

52. Baldi P, Nasr R. When is chemical similarity significant? The statistical distribution of chemical similarity scores and its extreme values. J Chem Inf Model. 2010; 50(7):1205–22. Epub 2010/06/15. https://doi.org/10.1021/ci100010v PMID: 20540577; PubMed Central PMCID: PMC2914517.

53. Wang Y, Backman TW, Horan K, Girke T. fmcsR: mismatch tolerant maximum common substructure searching in R. Bioinformatics. 2013; 29(21):2792–4. Epub 2013/08/22. https://doi.org/10.1093/bioinformatics/btt475 PMID: 23962615.

54. Yuan Y, Zheng F, Zhan CG. Improved Prediction of Blood-Brain Barrier Permeability Through Machine Learning with Combined Use of Molecular Property-Based Descriptors and Fingerprints. AAPS J. 2018; 20(3):54. Epub 2018/03/23. https://doi.org/10.1208/s12248-018-0215-8 PMID: 29564576; PubMed Central PMCID: PMC7737623.

55. Kumari C, Abulaish M, Subbarao N. Exploring Molecular Descriptors and Fingerprints to Predict mTOR Kinase Inhibitors using Machine Learning Techniques. IEEE/ACM Trans Comput Biol Bioinform. 2020; PP. Epub 2020/01/07. https://doi.org/10.1109/TCBB.2020.2964203 PMID: 31905145.

56. Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression. 3rd ed: John Wiley & Sons; 2013.

57. Friedman N, Geiger D, Goldszmidt M. Bayesian Network Classifiers. Machine Learning. 1997; 29:131–63.

58. Leo B. Random Forests. Machine Learning. 2001; 45:5–32. https://doi.org/10.1023/A:1010933404324.

59. Salehi F, Abbasi E, Hassibi B. The Impact of Regularization on High-dimensional Logistic Regression. Proceedings of NeurIPS 2019 [Internet]. 2019.

60. Joachims T, editor SVM Light: Support Vector Machine2002.

61. Meyer JG, Liu S, Miller IJ, Coon JJ, Gitter A. Learning Drug Functions from Chemical Structures with Convolutional Neural Networks and Random Forests. J Chem Inf Model. 2019; 59(10):4438–49. Epub 2019/09/14. https://doi.org/10.1021/acs.jcim.9b00236 PMID: 31518132; PubMed Central PMCID: PMC6819987.

62. Banerjee P, Siramshetty VB, Drwal MN, Preissner R. Computational methods for prediction of in vitro effects of new chemical structures. J Cheminform. 2016; 8:51. Epub 2017/03/21. https://doi.org/10.1186/s13321-016-0162-2 PMID: 28316649; PubMed Central PMCID: PMC5043617.

63. Shi H, Liu S, Chen J, Li X, Ma Q, Yu B. Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure. Genomics. 2019; 111(6):1839–52. Epub 2018/12/15. https://doi.org/10.1016/j.ygeno.2018.12.007 PMID: 30550813.

64. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020; 21(1):6. Epub 2020/01/04. https://doi.org/10.1186/s12864-019-6413-7 PMID: 31898477; PubMed Central PMCID: PMC6941312.

65. Jeni LA, Cohn JF, De La Torre F. Facing Imbalanced Data Recommendations for the Use of Performance Metrics. Int Conf Affect Comput Intell Interact Workshops. 2013; 2013:245–51. Epub 2013/01/01. https://doi.org/10.1109/ACII.2013.47 PMID: 25574450; PubMed Central PMCID: PMC4285355.

66. Zhang L, Tan J, Han D, Zhu H. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. Drug Discov Today. 2017; 22(11):1680–5. Epub 2017/09/08. https://doi.org/10.1016/j.drudis.2017.08.010 PMID: 28881183.

67. Rifaioglu AS, Atas H, Martin MJ, Cetin-Atalay R, Atalay V, Dogan T. Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. Brief Bioinform. 2019; 20(5):1878–912. Epub 2018/08/08. https://doi.org/10.1093/bib/bby061 PMID: 30084866; PubMed Central PMCID: PMC6917215.

68. Couronne R, Probst P, Boulesteix AL. Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics. 2018; 19(1):270. Epub 2018/07/19. https://doi.org/10.1186/s12859-018-2264-5 PMID: 30016950; PubMed Central PMCID: PMC6050737.

69. Anninos H, Andrikopoulos G, Pastromas S, Sakellariou D, Theodorakis G, Vardas P. Triflusal: an old drug in modern antiplatelet therapy. Review of its action, use, safety and effectiveness. Hellenic J Cardiol. 2009; 50(3):199–207. Epub 2009/05/26. PMID: 19465361.

70. Yun-Choi HS, Kim JH, Lee JR. Potential inhibitors of platelet aggregation from plant sources, III. J Nat Prod. 1987; 50(6):1059–64. Epub 1987/11/01. https://doi.org/10.1021/np50054a008 PMID: 3127544.

71. Petrone PM, Simms B, Nigsch F, Lounkine E, Kutchukian P, Cornett A, et al. Rethinking molecular similarity: comparing compounds on the basis of biological activity. ACS Chem Biol. 2012; 7(8):1399–409. Epub 2012/05/19. https://doi.org/10.1021/cb3001028 PMID: 22594495.

72. Yu X, Geer LY, Han L, Bryant SH. Target enhanced 2D similarity search by using explicit biological activity annotations and profiles. J Cheminform. 2015; 7:55. Epub 2015/11/20. https://doi.org/10.1186/s13321-015-0103-5 PMID: 26583046; PubMed Central PMCID: PMC4648974.

73. Montaruli M, Alberga D, Ciriaco F, Trisciuzzi D, Tondo AR, Mangiatordi GF, et al. Accelerating Drug Discovery by Early Protein Drug Target Prediction Based on a Multi-Fingerprint Similarity Search. Molecules. 2019; 24(12). Epub 2019/06/19. https://doi.org/10.3390/molecules24122233 PMID: 31207991; PubMed Central PMCID: PMC6631269.

74. Wang Z, Liang L, Yin Z, Lin J. Improving chemical similarity ensemble approach in target prediction. J Cheminform. 2016; 8:20. Epub 2016/04/26. https://doi.org/10.1186/s13321-016-0130-x PMID: 27110288; PubMed Central PMCID: PMC4842302.

75. An WF, Tolliday N. Cell-based assays for high-throughput screening. Mol Biotechnol. 2010; 45(2):180–6. Epub 2010/02/13. https://doi.org/10.1007/s12033-010-9251-z PMID: 20151227.

76. Wang L, Li X, Zhang S, Lu W, Liao S, Liu X, et al. Natural products as a gold mine for selective matrix metalloproteinases inhibitors. Bioorg Med Chem. 2012; 20(13):4164–71. Epub 2012/06/05. https://doi.org/10.1016/j.bmc.2012.04.063 PMID: 22658537.

77. Moffat JG, Vincent F, Lee JA, Eder J, Prunotto M. Opportunities and challenges in phenotypic drug discovery: an industry perspective. Nat Rev Drug Discov. 2017; 16(8):531–43. Epub 2017/07/08. https://doi.org/10.1038/nrd.2017.111 PMID: 28685762.

78. Berthold MR CN, Dill F, Gabriel TR, Kötter T, Meinl T, Ohl P, Sieb C, Thiel K, Wiswedel B. KNIME: The Konstanz Information Miner.: Berlin: Springer-Verlag; 2007.

79. Greg L. RDKit: Open-source cheminformatics: Online ToolKit; 2006. Available from: http://www.rdkit.org/.

80. Cao Y, Charisi A, Cheng LC, Jiang T, Girke T. ChemmineR: a compound mining framework for R. Bioinformatics. 2008; 24(15):1733–4. https://doi.org/10.1093/bioinformatics/btn307 PMID: 18596077; PubMed Central PMCID: PMC2638865.

81. Guha R. Chemical Informatics Functionality in R. Journal of Statistical Software. 2007; 18.

82. Probst P, Boulesteix A-L, Bischl B. Tunability: Importance of Hyperparameters of Machine Learning Algorithms. Journal of Machine Learning Research. 2019; 20:1–32.

83. Fernandez de Arriba A, Cavalcanti F, Miralles A, Bayon Y, Alonso A, Merlos M, et al. Inhibition of cyclooxygenase-2 expression by 4-trifluoromethyl derivatives of salicylate, triflusal, and its deacetylated metabolite, 2-hydroxy-4-trifluoromethylbenzoic acid. Mol Pharmacol. 1999; 55(4):753–60. Epub 1999/04/01. PMID: 10101034.