bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371112; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Bioactivity assessment of natural compounds using machine learning models based on drug target similarity

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

[1]European Molecular Biology Laboratory, Heidelberg 69117, Germany
[2]Medical Research Council Toxicology Unit, University of Cambridge, United Kingdom
*Corresponding author: kp533@cam.ac.uk

Summary

Natural products constitute a vast yet largely untapped resource of molecules with therapeutic properties. Computational approaches based on structural similarity offer a scalable approach for evaluating their bioactivity potential. However, this remains challenging due to the immense structural diversity of natural compounds and the complexity of structure-activity relationships. We here assess the bioactivity potential of natural compounds using random forest models utilizing structural fingerprints, maximum common substructure, and molecular descriptors. The models are trained with small-molecule drugs for which the corresponding protein targets are known (1,410 drugs, 0.9 million pairs). Using these models, we evaluated circa 11k natural compounds for functional similarity with therapeutic drugs (1.7 million pairs). The resulting natural compound-drug similarity network consists of several links with support from the published literature as well as links suggestive of unexplored bioactivity of natural compounds. As a proof of concept, we experimentally validated the model-predicted Cox-1 inhibitory activity of 5-methoxysalicylic acid, a compound commonly found in tea, herbs and spices. In contrast, a control compound, with the highest similarity score when using the most weighted fingerprint metric, did not inhibit Cox-1. Our results illustrate the importance of complementing structural similarity with the prior data on molecular interactions, and presents a resource for exploring the therapeutic potential of natural compounds.

Introduction

Approximately 65% of the small-molecule drugs in use today have originated from natural compounds or their derivatives (Newman and Cragg, 2020), and their therapeutic potential have been discussed extensively (Atanasov et al., 2015; Harvey et al., 2015; Rodrigues et al., 2016; Shen, 2015). Identification of bioactive natural compounds present in the diet and their effect on health has thus been an active area of research since long (Corbi et al., 2016; Hosseini and Ghorbani, 2015). A number of recent findings have reported that dietary patterns can reduce the risk of many chronic diseases (Alissa and Ferns, 2017; Gu et al., 2019; Hartley et al., 2013), lead to drug and food interactions (Briguglio et al., 2018; Jensen et al., 2015; Rodriguez-Fragoso et al., 2011), and significantly alter or diversify the composition of the human gut microbiome (David et al., 2014; Kolodziejczyk et al., 2019; Sonnenburg and Backhed, 2016; Zmora et al., 2019). Although natural compounds possess rich structural diversity and often have selective biological actions (Clardy and Walsh, 2004; Koehn and Carter, 2005; Pye et al., 2017), they have been generally less accessible, especially in comparison to synthetic compounds, for systematic screening studies due to complex purification methods (Li et al., 2019). Recent technological advances in metabolic engineering

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371112; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and synthetic biology, as well as those in functional assays and phenotypic screens are opening new opportunities for natural compound-based drug discovery (Kang et al., 2016; Li et al., 2019). Yet, systematic experimental exploration of bioactivities of natural compounds is likely to remain unfeasible due to the immense number of compounds involved, their low abundance in natural sources, and technical difficulties in extraction from complex matrices.

An attractive approach to assess the bioactivity potential of a compound is computing its chemical and structural similarity with the molecules with known activity (Cereto-Massague et al., 2015; Muegge and Mukherjee, 2016). Novel methods integrating multiple molecular properties along with the biological activity information are also being deployed to study small-molecule functional similarities (Duran-Frigola et al., 2020). In such virtually screening approaches, machine learning models are being increasingly used to tackle the complex structure-activity relations (Chan et al., 2019; Lavecchia, 2015; Lima et al., 2016; Lo et al., 2018). These approaches have been used to predict, among other, bioactivities (Zhang et al., 2020), shared molecular interactions (Lim et al., 2018; Ryu et al., 2018), toxicity (Yang et al., 2018; Zhang et al., 2018a), and drug-likeness of molecules (Yosipof et al., 2018).

The chemical similarities between drug and natural compounds, esp. dietary compounds, and their association with drug targets have been studied previously (Jensen et al., 2015). Such similarity-based rankings, however, vary considerably depending upon which fingerprint encoding is used (O'Hagan and Kell, 2017b). Indeed, the structural similarity between two molecules is a subjective concept (Maggiora et al., 2014) 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 (Seo et al., 2020) to identify similar properties, establish pairwise relationships (Park et al., 2019), directly compare bioactivities and reduce artefacts (Cereto-Massague et al., 2015; Duran-Frigola et al., 2020).

In this study, we systematically compared FDA-approved drugs with dietary natural compounds based on their structural similarities and physicochemical properties. To make the comparison meaningful in terms of bioactivity, we used a machine learning approach trained on drug molecules with known protein targets. This allowed us to uncover the therapeutic potential of hundreds of natural compounds; and to provide a proof-of-concept validation for 5-methoxysalicylic acid, which showed marked inhibitory activity against Cox-1.

Results

Dataset of drugs with known targets

We utilized 1,410 FDA approved drugs (**Table** S1a) with known curated targets (**Table** S1b) as our machine learning (ML) dataset. The drugs were classified based on the ATC (Anatomical Therapeutic Chemical classification) system and also based on their chemical structures. A large number of these drugs target nervous system (264) followed by cardiovascular (180), anti-infectives (148), multiple ATC (131) and anti-neoplastic (127) (**Fig** S1a). Of all the 18 structural classes, benzenoids and organoheterocyclics constitute the major super-classes of drugs (840) encompassing all therapeutic classes except the nutraceuticals (**Fig** S1a).

For the 1,410 drugs used, there were 1,262 known curated targets (Wishart et al., 2018). The number of drug targets ranged from 1 to 86 (**Table** S1b) highlighting the fact that some drugs are well studied in

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371112; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

terms of their target space. The targets most abundantly found 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 a large number of drugs (264) are targeting the nervous system.

Natural compounds

A catalogue of circa 11,000 compounds was obtained from FooDB (www.foodb.ca) (**Table** S1c). 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 (**Fig** S1b). For the simplicity in representation in **Fig** S1b, the frequency accounts 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 using ClassyFire (Djoumbou Feunang et al., 2016) (**Fig** S1b). Highly represented were lipids and lipid-like molecules (4803), phenylpropanoids and polyketides (2476), organoheterocyclics (1381) and organic oxygen compounds (1120).

Molecular features

For all the drugs and natural compounds considered here, 7 types of molecular fingerprints (distance-based measures; namely Morgan, Featmorgan, Atompair, RDKit, Torsion, Layered and MACCS, maximum common substructure (MCS), and molecular descriptors were computed as described in the methods. We then used drug molecules (all against all pairs) to create a dataset for the machine learning task. All pairwise drug combination fingerprint similarity was scored using Tanimoto Score (TS) measured on a scale of '0-1'; higher the score, more similar the molecules. Consistent with each molecular fingerprint assessing different features of the compound, the Tanimoto score distribution for the ML dataset (i.e. drugs vs drugs) differed across the fingerprints (**Fig** 1B).

When observed on the scale of '0 (no similarity) - 1 (highly similar)' of TS, the Morgan, FeatMorgan and Torsion fingerprints consistently computed a lower score for the majority of the drug pairs (TS_{median} = 0.11, 0.13, 0.08 respectively) as compared to AtomPair, RDKit, Layered, and MACCS (TS_{median} = 0.23, 0.36, 0.45, 0.32 respectively). The drug pairs showed broader distribution across the TS scale in AtomPair, RDKit, Layered and MACCS. A rank-based comparison of drug-pairs (**Fig** S2) called by each fingerprint also showed a low concordance amongst the different fingerprints. As none of the fingerprints alone is universally suited (Baldi and Nasr, 2010), we decided to retain all the 7 computed fingerprints in training the classifier.

We next computed the MCS and molecular descriptors for all drug pairs which would provide more extensive information on substructure similarities and physicochemical properties. The MCS calculation reports a number of statistics, amongst which the MCS size (median = 8), TS (median = 0.19) and OC (overlapping coefficient) (median = 0.43) score are important measures to assess similarity. The TS 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 useful for assessing similarities 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 fingerprints and descriptors (O'Hagan and Kell, 2017a).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371112; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

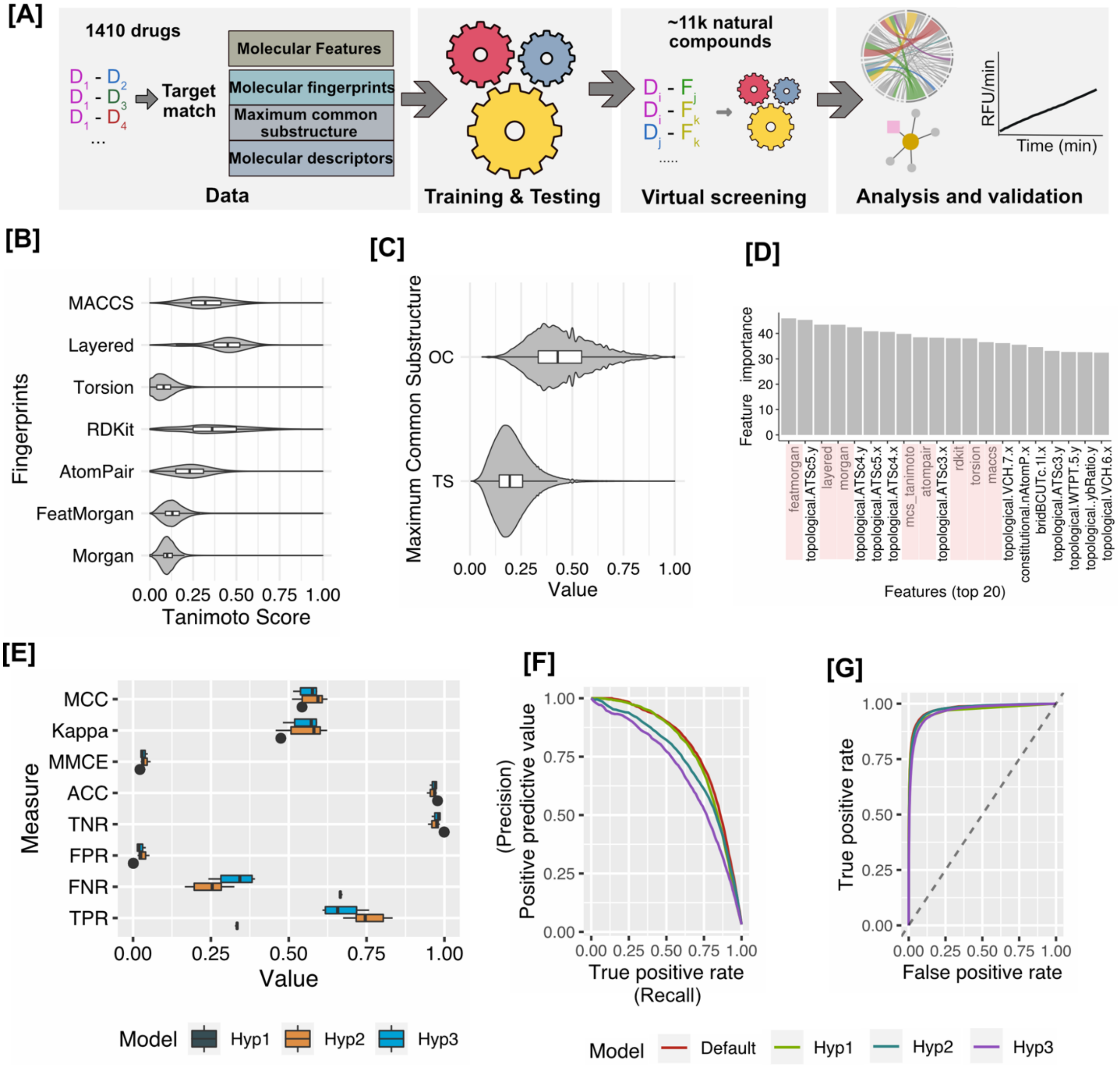

Figure 1 *(A)* ***Overview*** *of workflow deployed. (B-D) -* ***Similarity metrics (ML dataset)****. (B) Molecular fingerprints - the 7 fingerprints generate a different similarity score for the pairs of drug molecules compared. The median value of each is represented in the box plot (in the center) 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 (TS) and the other is the Overlap coefficient (OC). The violin plots were smoothed for density by an adjustment factor of 3. (D) High ranking features - top 20 features are displayed, showing most of the distance-based features provided maximum information gain with 'FeatMorgan' performing best. (E) -* ***Training performance*** *of the three models namely Hyp1, Hyp2, Hyp3 - depicts various performance measures over five random iterations during training. Hyp1 was very consistent over the five iterations during training but appeared less balanced in terms of measurement values whereas Hyp2 and Hyp3 produced more robust measurement values. (F-G) -* ***Performance on the test set.*** *(F) Precision-Recall curve - the default performed better than the hyper-tuned models but closer observation of other statistical measures (****Table*** *1) revealed that the tuned models (Hyp2 and Hyp3) had more robust behavior in terms of balanced measures. (G) ROC curve - all the models had AUC >0.96.*

Additionally, we used 5 different molecular descriptor categories (constitutional, topological, geometrical,

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371112; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

electronic and hybrid) to serve as a means to capture individual physicochemical properties of the drugs. **Table** S1d, 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. Similar to the ML dataset, all molecular features (namely fingerprints, MCS and descriptors) were computed for the natural compounds as well.

Data preprocessing

All pairwise distance-based measures (TS of Morgan, FeatMorgan, AtomPair, RDKit, Torsion, Layered and MACCS), MCS features (MCS size, MCS Tanimoto, MCS overlap coefficient), and the molecular descriptors (constitutional, topological, geometrical, electronic and hybrid) were merged to create a matrix of features for the ML dataset. The unique pairwise combinations for 1,410 drugs resulted in a total of 993,345 drug pairs with 460 molecular features for each pair.

Prior to training the classifier for model building, the data was processed to remove constant feature columns (82 columns with all 0's) and missing data values in rows (NA's). This resulted in a matrix of 961,191 rows and 378 feature columns. The final step in data processing was to randomly split the ML dataset into a training (80%) and an independent validation test set (20%). This split resulted in non-overlapping 768,952 rows in training and 192,239 rows in the test set.

Feature selection

To decrease the overhead of training on noise and to increase the learner's performance, we searched for features most informative for the prediction variable, i.e. outcome (match or no-match). Since feature selection is in itself time-consuming on large datasets, we performed a single run of feature selection on our training dataset (768,952 rows x 378 columns). The importance values of all the features are depicted in **Fig** S3. We used the mean importance value (**Fig** S3) as a threshold for feature selection. This resulted in the retention of 188, or 50% of the features. A sorted list of top 20 features is shown in **Fig** 1D; FeatMorgan was the top performer and interestingly, most of the distance-based features (highlighted) were high ranking performers indicating their high relevance for predicting the target outcome. Amongst the descriptors, topological descriptors were the best performers.

Random Forest Classification

The observations in the training set were 24,283 matches and 744,669 no-matches. The models were trained on the training set using the target outcome (i.e. match and no-match) as the class label. All learner experiments were done with 3-fold cross-validation and each tree is grown on a different sample of original data with stratification (helpful for imbalanced datasets). Prior experience with ML-based classifiers had suggested that Random Forest works optimally in handling high-throughput chemical compound data classification problems (Periwal et al., 2012; Periwal et al., 2011). Thus, we used random forest classifiers to train the ML models. We trained four different models, one with default parameters (termed 'Default' hereafter) and three with different hyperparameter combinations ('Hyp1', 'Hyp2' and 'Hyp3') (Methods).

The distribution of the performance statistics of the five iterations of all models is shown in **Fig** 1E. Hyp1 produced consistent results over all the iterations but appeared less robust and balanced as compared to

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371112; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Hyp2 and Hyp3 in terms of other performance measures (i.e. MCC (Matthews correlation coefficient), Kappa, TPR and FNR). The result of iterations is aggregated as the mean test value. One of the key considerations for selecting the optimal performance measures for chemical datasets, and in general where predicting the minority class is of more interest (Chen et al., 2004), is to have high prediction accuracy over minority class while maintaining reasonable values for the majority class. Thus, the optimal 'Hyp' combinations were selected based on the best tuned MCC (it gives a balance measure of both classes) and the final model was trained using those parameters.

Table 1 Performance statistics of the four trained models on the test set

Model	MCC	KAPPA	BAC	ACC	TPR	FPR	FNR	TNR
Default (tree, node, weight=default)	0.609	0.558	0.702	0.980	0.406	0.001	0.593	0.999
Hyp1 (tree=70, node=default, weight=8000)	0.607	0.554	0.700	0.980	0.401	0.001	0.599	0.999
Hyp2 (tree=85, node=40, weight=default)	0.666	0.664	0.859	0.977	0.733	0.014	0.266	0.985
Hyp3 (tree=95, node=35, weight=26000)	0.628	0.626	0.834	0.975	0.683	0.014	0.316	0.985

MCC - Matthews correlation coefficient, BAC - Balanced accuracy, ACC - accuracy, TPR – True-positive rate, FPR – False-positive rate, FNR – False-negative rate, TNR – True-negative rate

The behavior and performance of all trained models on unseen data were assessed using the independent test dataset (created by a split of 20%). Interestingly, all four trained models showed good performance in terms of the Precision-Recall curve (**Fig** 1F). Although the Default and Hyp1 models produced higher precision as compared to the Hyp2 and Hyp3, they had lower values for the balanced measures (MCC, Kappa and BAC) (**Table** 1). In the ROC curve (**Fig** 1G) as well, all models behaved equally well. The favorable performance of all the models suggested us to take a consensus approach for predictions in virtual screening of new compound library.

Predicting drug-natural compound similarity

The natural compound library from FooDB was virtually screened against all the four models. Pairwise similarities between drugs and FooDB compounds were computed using the same set of features as for the ML dataset. The MCS computation, which is computationally intensive, was carried out in parallel Drug-food pairs where the food compounds had a perfect match with the drug (i.e. TS of 1 or exact names as of the drug as many drugs are naturally derived) or pairs with missing values for any feature were removed. Thus, the total number of drug-food pairs screened using all the four models were 1,696,695. An overlap of 'match' predictions from the four models is depicted in **Fig** S4. Since all the models performed optimally in their predictions on the test, we considered the union of all positive predictions (i.e. 4,909 drug-food pairs) as our 'matching' compounds (**Fig** S4, **Table** S3a). For illustrative and discussion purposes we chose a subset of these predictions, which is the intersection of all the four models (i.e. 125 unique drug-food pairs) (**Fig** 2 and **Table** S3b). We performed a thorough manual curation of these 125 pairs and categorized them into four different subgroups (Table S3c). Out of the 18 ATC classes, 16 were represented in the prediction results for the 125 pairs highlighting at-least one drug from each therapeutic category sharing similarity with a natural compound. As stated previously, a given food compound can be present in multiple food sources so the drug-food similarity relationship was plotted for more than one food source (extended data in Table S3c (although this list is not exhaustive and sometimes carry only few representative food sources chosen randomly)), for which data was available for these 125 pairs (**Fig** 2). For exhaustive listing of food sources, we recommend querying FooDB using respective compound Ids.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371112; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

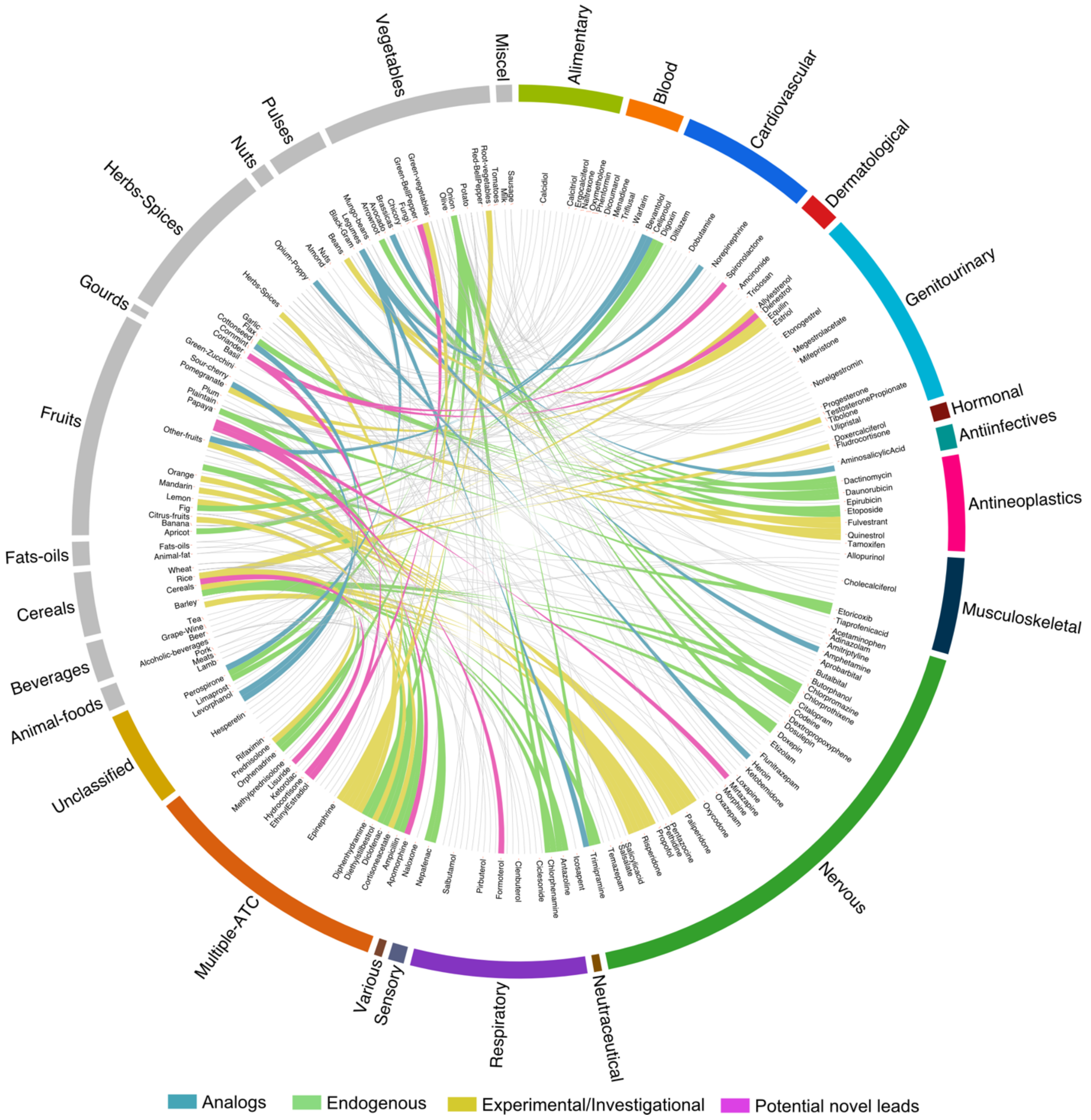

***Figure 2. A consensus subset of drug-food similarity network**. 125 drug-food pairs predicted as 'match' by the consensus result of the four models. A given food compound can be present in more than one food source hence the links here represent all known food source for a food compound (from FooDB). The drugs are arranged according to their therapeutic class and food compounds according to their food source category. The highlighted colored links are the different category examples discussed in detail in the text.*

The 57 unique food compounds constituting the consensus similarity network with 105 unique drugs can be grouped into four categories (Table S3c); drug analogs, host-endogenous metabolites, compounds currently under experimental investigation as drugs, and novel leads for therapeutic applications.

In the first group, 23 food compounds were found to be analogues or sometimes even exactly the same

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371112; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

as other known drugs but annotated with different names or synonyms. This category was informative as a control as well as to highlight their potential presence in various food sources. An interesting example here is the compound 'papain' (named same as the papain enzyme found in papaya and other fruits). Papain is a synonym for ibuprofen, the commonly used as pain, fever and inflammation reliever which is a drug of synthetic origin. It shared similarity with ampicillin, etoricoxib, diclofenac and nepafenac. Podofilox (annotated as lignans in FooDB and is reported to be present in flax and arrow root) is used to treat genital warts and is classified as a dermatological drug with some antineoplastic activity. In our virtual screen, its shared similarity with many anti-neoplastic (limaprost, daunorubicin, etoposide, and dactinomycin) suggesting a potentially much wider role as an anti-neoplastic agent. Indeed, the antineoplastic activity of podofilox and its derivatives has been indicated in various tumor cell lines and is an active area of research for chemotherapy (Zhang et al., 2018b; Zi et al., 2019).

The second group, host-endogenous metabolites, included 7 compounds that are endogenous to human tissues and also reported in food sources. For example, 'desoxycorticosterol' was reported to be present in rice and is endogenously present in amniotic fluid and blood throughout human tissues. 'Estriol', an estrogen produced by the human body, is reported to be present in pomegranate and beans.

The third group comprised of 5 food compounds which are already under experimental investigation category in DrugBank (i.e. under approval to be used as drugs, accessed January 2018). 'diferuloylspermine' (found in ananas) as spermine (DB00127), 'codeine N-oxide' (found in opium) as DB01568, '4-(β-methylaminoethyl) catechol' (found in legumes) as deoxyepinephrine (DB13917) and '3-hydroxyanthranilic acid' (found in brassicas) as DB03644. These 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.

The fourth group comprises of 23 compounds for which, to our knowledge, little or no evidence of their physiological or biological activity have been hitherto reported. Thus, we refer them to as potential lead compounds. Of these, 'Homoeriodictyol' is a trihydroxyflavonone reported in coriander and is found active in inhibiting CYP1B1 and CYP1A1/2 targets (PubChem AID: 311072 and 502473-502475). It was found to be a hit with the three drugs spironolactone (also binds to CYP), dienesterol (not tested for CYP binding) and ethinylestradiol (also binds to CYP1A1 (PubChem AID: 678712)). All the three drugs have different therapeutic target class (cardiovascular, genitourinary and multiple-ATC respectively). Homoeriodictyol has also been reported to possess taste modifying properties i.e. it reduces bitterness (Ley et al., 2005). '(+)-setoclavine' is an alkaloid reported in cereals (ergot found on pearl millet) and shared similarity with the drug lisuride which is a prescription medication for Parkinson's disease. Interestingly, lisuride is also an ergot derivative and binds to dopamine and serotonin receptors. An *in-silico* study also predicted binding of setoclavine to dopamine receptors with a binding affinity comparable to other ergot alkaloids (Paulke et al., 2013). Thus, these four groups highlighted interesting similarity relationships existing between drug and food compounds and their wider therapeutic potential.

Drug - Food compounds - Target network

Since our predictions are based on the hypothesis that molecules binding to the same target share common chemical and structural similarity in a high-dimensional space captured by the machine learning model, we looked into the known targets shared by the predicted 125 drug-food pairs. Some drugs were present as a single node where they have unique targets in the current target space and don't cluster with the others. For instance, the drug digoxin has only one known target 'Na^+/K^+ transporting ATPase subunit-

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371112; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

α1' and the food compound ('Glycosides') paired with it which is actually the drug Ouabain, also shares the same target (**Fig** 3). Additionally, ouabain also binds to two other subunits (α-2 and α-3) of the same protein suggesting that digoxin might also have a binding affinity for these two other subunits. On the other hand, some drugs have a large target range such as spironolactone which is a steroid and has as many as 29 known targets. Many drugs from the nervous, cardiovascular and respiratory category formed very closely clustered target hubs.

An interesting case is that of triflusal, which is an antithrombotic anticoagulant and is considered very important for the secondary prevention of ischemic stroke. As per our predictions, it shares similarity with 5-methoxysalicylic acid (**Fig** 3). 5-methoxysalicylic acid is found in tea, herbs and spices, and is indeed shown to have antiplatelet activity in rats (Yun-Choi et al., 1987). Triflusal has an antagonist effect on prostaglandin G/H synthase 1 (PTGS1) (also called Cox-1) in platelets (Anninos et al., 2009). Another interesting compound also categorized as novel lead is butyl salicylate which shares similarity with the drug salsalate, an anti-inflammatory agent. Salsalate also targets PTGS1 and additionally PTGS2. It would thus be interesting to investigate the inhibitory activity of both 5-methoxysalicylic acid and butyl salicylate on PTGS1 as novel leads.

Another interesting hub in the network is that of the 5 drugs (calcitriol, ergocalciferol, calcidiol, doxercalciferol, and cholecalciferol) binding to the vitamin D3 receptor (**Fig** 3). All the drugs in this hub except calcitriol have only one known target. Only one food compound in this hub was categorized as potential lead and others were either analogues or endogenously present in humans. The compound is annotated as '22E, 24x-ergosta-4,6,8,22-tetraen-3-one' and is reported in mushrooms and fungi. It appears to be an analogue of ergone which is also found in medicinal fungi and lichens, and has been shown to prevent chronic kidney disease (Zhao et al., 2014).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371112; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

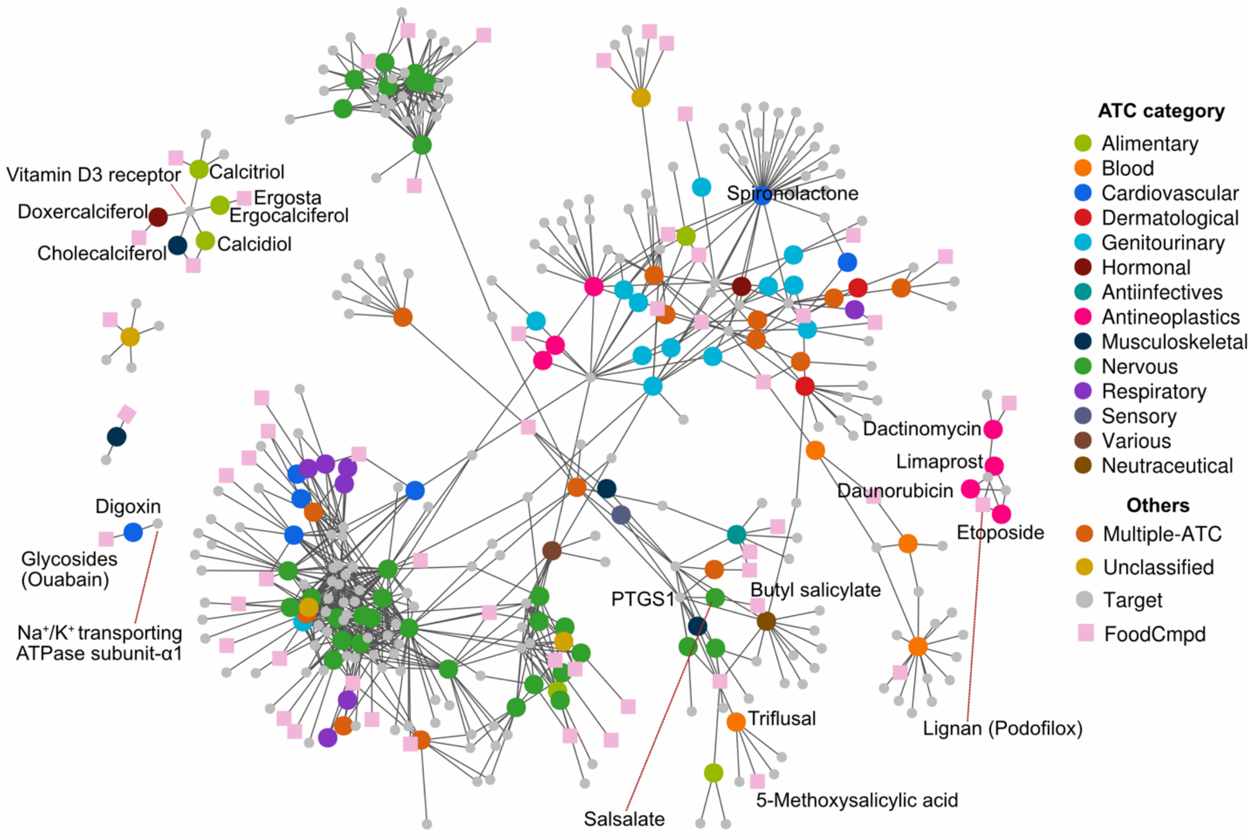

Figure 3 Drug-Target-Food network. *Network of 125 drug-food pairs shown along with the drug targets. The drugs are central nodes colored by their therapeutic class. The known targets of the drugs (grey circles) are proposed to be the potential target space of the respective food compounds as well (light pink squares).*

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

As a proof-of-concept experimental validation, we chose the prediction of similarity between food compound 5-methoxysalicylic acid and triflusal, and the Cox-1 inhibition by 5-methoxysalicylic acid implied therein. Triflusal is prescribed as an antithrombotic anticoagulant and new natural compound/s with similar activity could be highly relevant. To additionally test whether the machine learning approach used here is more useful in identifying true positives in comparison to using a single similarity measure (such as fingerprint alone), we also tested a 'negative control' molecule, 4-isopropylbenzoic acid, which is found in cumin, herbs and spices. This molecule had higher TS for FeatMorgan, which was the most important feature during feature selection (**Fig** 2D), but was predicted as no-match by all four machine learning models. In contrast, our test compound (5-methoxysalicylic acid) was predicted as a match by all machine-learning models and had a lower TS for FeatMorgan (**Fig** S4A). Structural representation of all the tested compounds and their shared MCS (maximum common substructure) is depicted in **Fig** 4A.

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 (**Fig** S4B).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371112; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

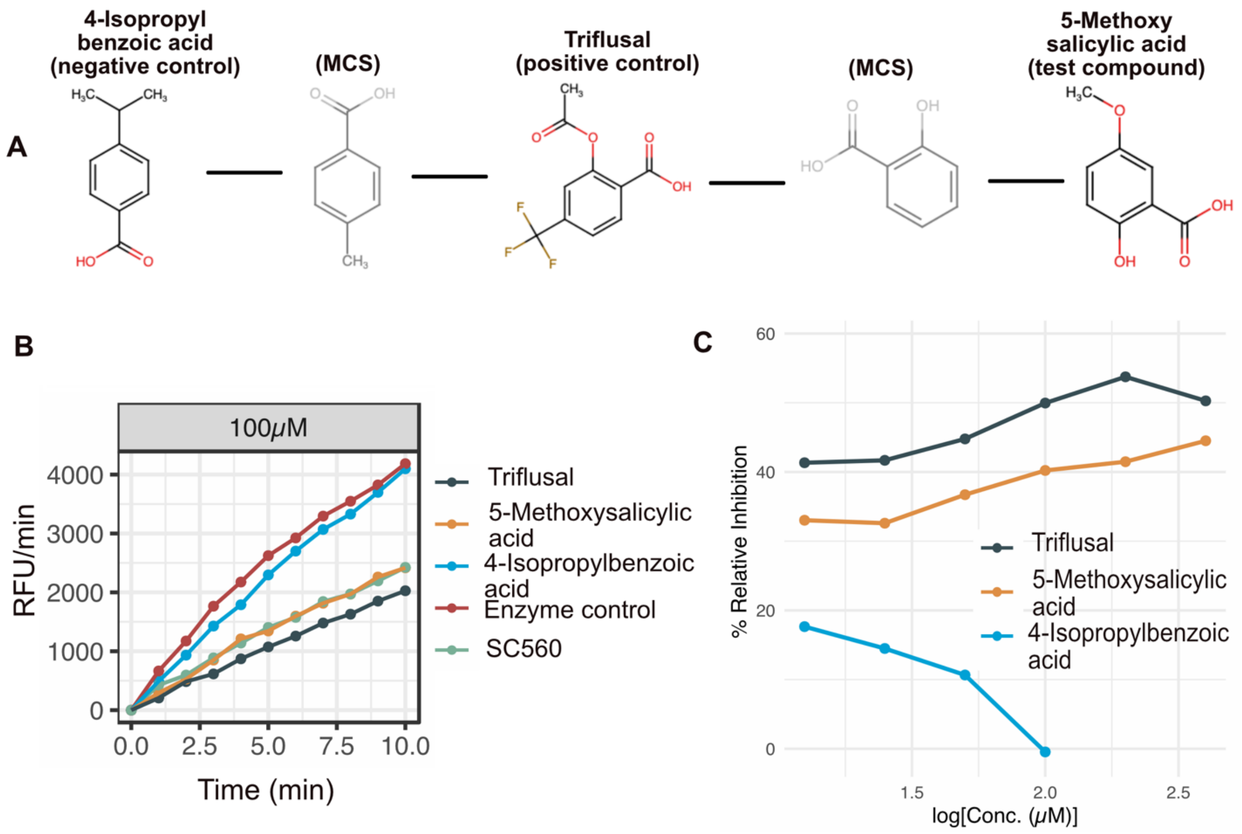

Figure 4 Cox-1 inhibitor assay. *(A) Chemical structures of all the tested compounds. MCS structures are also depicted which helped to intuitively assess the structural similarity between the tested compounds (B) An example relative fluorescent units (RFU) plot of the tested compounds at 100μM (other tested conc.: 12.5μM to 400μM serial dilutions). SC560 is a positive control provided by the assay kit supplier (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.*

Triflusal and 5-methoxysalicylic acid exhibited highly similar inhibition profiles at different concentrations (Fig. 4B, Table S4). From this assay, the IC50 of triflusal could be estimated to be around 100μM. 5-methoxysalicylic acid was able to achieve 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 inhibition at 100μM or higher concentration. It rather showed strong colored reaction (bright pink) which resulted in exceptionally high RFU values. Thus, we found it unsuitable for testing at higher concentrations with this assay and only considered its inhibition values up to 100μM. It only showed a small (10-17%) inhibitory effect at a lower concentration, possibly due to non-specific binding. Taken together, the biological activity predicted by the machine learning could be confirmed experimentally, while that based on a high-scoring single similarity metric (i.e. fingerprints) was a false-positive.

Discussion

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371112; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The data fusion approach (i.e. integrating similarity metrics based on fingerprints, maximum common substructures and physico-chemical descriptors) used in this study appeared to be an effective way of identifying closely related molecules which have the potential to bind to the same protein target. In particular, the use of drug-target similarity matrix for training the machine learning classifier appeared useful in capturing complex high-dimensional similarity that would not be accessible based on any single metric alone. Indeed, we could show that a molecule that deemed similar based solely on a single fingerprint's Tanimoto score showed no significant activity while the one predicted by machine learning models trained on multiple features showed the predicted enzyme inhibitory activity. In further support of our approach, we could identify several drug-food compound relations including compounds that are currently under investigation or have been ascribed with related bioactivity in the literature.

Data fusion approaches can aid in reducing the amounts of artefacts (Cereto-Massague et al., 2015). In line with this, various methods are being proposed to accelerate the performance of in-silico similarity searches such as inclusion of bioactivity profiles (Duran-Frigola et al., 2020; Petrone et al., 2012; Yu et al., 2015), multiple fingerprint algorithms (Montaruli et al., 2019), and similarity ensemble approach (Wang et al., 2016). 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 (Park et al., 2019) which supports our strategy and showcases the utility and success rate of such methods in early-stage molecule discovery.

In the context of natural compound 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 (An and Tolliday, 2010; Wang et al., 2012). 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 (Duran-Frigola et al., 2020) would be immensely valuable to this end.

In summary, we report a computational approach utilizing the structural, physicochemical and protein target information of approved drugs to uncover the therapeutic potential of natural compounds. Our methodology combines various distance- and similarity-scores from molecular fingerprints and common substructures, as well as descriptor-based features to train a random forest classifier on a drug dataset. The predictions not only captured compounds with literature support for their therapeutic potential, but also provided novel leads, one of which we experimentally validated for target binding. Taken together, our study supports the using target-similarity for uncovering drug-like properties of natural compounds.

Materials and methods

Data Source and processing

All the FDA approved drugs which had target information associated with them were taken from DrugBank (Law et al., 2014) (accessed January 2018). The drugs and their target information were later used to create ML models. The natural compound library used for virtual screening was FooDB (www.foodb.ca; freely available and accessed June 2017, (~11k compounds). It required additional annotation, curation and formatting in order to be smoothly integrated into our analysis. It included compounds from both raw and processed foods. We used drug classification codes from ATC (https://www.whocc.no/) to

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371112; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

therapeutically classify all the drugs and ClassyFire (Djoumbou Feunang et al., 2016) to structurally classify the drugs and the natural compounds.

Molecular features generation

A number of pairwise distance measures and molecule specific 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 dataset created from drugs and their known respective targets. This matrix was later used to train the drug structure-target relationship where each drug pair was labelled 'match' if they share at least one molecular target and a 'no-match' if no target was reported common between them.

Fingerprint calculation, Tanimoto Score (TS) estimation and Molecular descriptor generation

The INCHIs from natural compound libraries and drugs 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, FeatMorgan, AtomPair, RDKit, Torsion, Layered and MACCS) to gather theoretical 2D structural information from the molecules. The pairwise structural similarity from the fingerprints was scored using the widely used Tanimoto similarity metric (computed as $TS_{AB} = (A \cap B)/(A \cup B)$). The entire workflow was designed using the KNIME (Berthold MR, 2007) analytics platform utilizing RDKit (Greg, 2006) plugin (for fingerprints) with default parameters. 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 (Guha, 2007).

Maximum Common Substructure (MCS) based TS estimation

The data features thus generated for each molecule pair were further complimented by computing maximum common substructure (MCS) shared between them. 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, TS, overlapping coefficient (computed as $OC_{AB} = (A \cap B)/\min(A, B)$). It was computed using the 'ChemmineR' (Cao et al., 2008) and 'fmcsR' (Wang et al., 2013) packages available for R. The MCS calculation is computationally intensive and time-consuming but they are more sensitive, accurate and intuitive, thus we implemented parallel scripts in batch-mode on high-performance computing cluster for faster processing.

Data preprocessing and Machine Learning

All computed molecular features of drug pairs were combined into a single matrix and preprocessed to build the classification model using the Random Forest classifier in R (using 'mlr' package). The binary tree-based classifier was trained for two classes which are referred to as target 'match' and 'no-match'. Each drug pair which shared at-least one target were labelled as 'match' and the rest were labelled as 'no-match'. We used feature selection and hyperparameter tuning to specify the search space of random forest in order to find the optimal set of parameters for our dataset. The importance of each feature was calculated using 'randomForest importance' method which is based on calculating OOB-accuracy (Out-Of-Bag).

Hyperparameters tuned to identify the optimal set of parameters for building the random forest classifier were the number of trees (*ntree*), the number of observations at terminal nodes (*nodesize*) and class

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371112; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

weights (*weight*). In our initial runs, we fine-tuned different values for the number of trees to grow ($ntree = 50\ to\ 1001$). It was observed from our preliminary runs (**Table** S2) that setting higher *ntree* values didn't result in any significant difference in performance rather only increased the computational overhead. The number of variables to split at each node (*mtry*) was kept at default (i.e. $mtry = \sqrt{p}$ where p is the number of features in the input data). In our ML dataset, the default *mtry* was ($\sqrt{188}$) 13 features. 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 10-300. Lower *nodesize* can result in lower detection signals of the true positives and a high false-negative rate. In our tuning experiments, we found setting *nodesize* to a higher number resulted in increased learner robustness without compromising on overall prediction performance. The systematic hyperparameter search was performed in a cluster environment with parallel backend.

All the 'Hyp' models were optimized for 5 random iterations with 3-fold cross-validation each using stratification. The search space of the 'Hyp' models was: Hyp1 ($ntree = 50 - 151, nodesize = default\ (1)\ and\ weight = 1000 - 20000\ (step\ of\ 1000)$, Hyp2 ($ntree = 50 - 151, nodesize = 30 - 150\ and\ weight = default\ (1{,}1)$) and Hyp3 ($ntree = 50 - 101, nodesize = 30 - 150, weight = 1000 - 30000\ (step\ of\ 5000)$.

We used the 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 in each iteration and model assessment.

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 (Fernandez de Arriba et al., 1999). 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 (**Fig** 4B). 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

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371112; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Data and Code availability

The data files and codes used for generating all main and supplementary figures are available at https://github.com/periwal45/periwaletal2020.

Acknowledgements

VP and NG were supported by the EMBL Interdisciplinary Postdoc (EIPOD) program under Marie Skłodowska-Curie actions COFUND (grant number 664726).

References

Alissa, E.M., and Ferns, G.A. (2017). Dietary fruits and vegetables and cardiovascular diseases risk. Crit Rev Food Sci Nutr *57*, 1950-1962.

An, W.F., and Tolliday, N. (2010). Cell-based assays for high-throughput screening. Mol Biotechnol *45*, 180-186.

Anninos, H., Andrikopoulos, G., Pastromas, S., Sakellariou, D., Theodorakis, G., and Vardas, P. (2009). Triflusal: an old drug in modern antiplatelet therapy. Review of its action, use, safety and effectiveness. Hellenic J Cardiol *50*, 199-207.

Atanasov, A.G., Waltenberger, B., Pferschy-Wenzig, E.M., Linder, T., Wawrosch, C., Uhrin, P., Temml, V., Wang, L., Schwaiger, S., Heiss, E.H.*, et al.* (2015). Discovery and resupply of pharmacologically active plant-derived natural products: A review. Biotechnol Adv *33*, 1582-1614.

Baldi, P., and Nasr, R. (2010). When is chemical similarity significant? The statistical distribution of chemical similarity scores and its extreme values. J Chem Inf Model *50*, 1205-1222.

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

Briguglio, M., Hrelia, S., Malaguti, M., Serpe, L., Canaparo, R., Dell'Osso, B., Galentino, R., De Michele, S., Dina, C.Z., Porta, M.*, et al.* (2018). Food Bioactive Compounds and Their Interference in Drug Pharmacokinetic/Pharmacodynamic Profiles. Pharmaceutics *10*.

Cao, Y., Charisi, A., Cheng, L.C., Jiang, T., and Girke, T. (2008). ChemmineR: a compound mining framework for R. Bioinformatics *24*, 1733-1734.

Cereto-Massague, A., Ojeda, M.J., Valls, C., Mulero, M., Garcia-Vallve, S., and Pujadas, G. (2015). Molecular fingerprint similarity search in virtual screening. Methods *71*, 58-63.

Chan, H.C.S., Shan, H., Dahoun, T., Vogel, H., and Yuan, S. (2019). Advancing Drug Discovery via Artificial Intelligence. Trends Pharmacol Sci *40*, 592-604.

Chen, C., Liaw, A., and Breiman, L. (2004). Using Random Forest to Learn Imbalanced Data (Conference Proceedings: UC Berkeley).

Clardy, J., and Walsh, C. (2004). Lessons from natural molecules. Nature *432*, 829-837.

Corbi, G., Conti, V., Davinelli, S., Scapagnini, G., Filippelli, A., and Ferrara, N. (2016). Dietary Phytochemicals in Neuroimmunoaging: A New Therapeutic Possibility for Humans? Front Pharmacol *7*, 364.

David, L.A., Maurice, C.F., Carmody, R.N., Gootenberg, D.B., Button, J.E., Wolfe, B.E., Ling, A.V.,

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371112; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Devlin, A.S., Varma, Y., Fischbach, M.A., *et al.* (2014). Diet rapidly and reproducibly alters the human gut microbiome. Nature *505*, 559-563.

Djoumbou Feunang, Y., Eisner, R., Knox, C., Chepelev, L., Hastings, J., Owen, G., Fahy, E., Steinbeck, C., Subramanian, S., Bolton, E., *et al.* (2016). ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. J Cheminform *8*, 61.

Duran-Frigola, M., Pauls, E., Guitart-Pla, O., Bertoni, M., Alcalde, V., Amat, D., Juan-Blanco, T., and Aloy, P. (2020). Extending the small-molecule similarity principle to all levels of biology with the Chemical Checker. Nat Biotechnol.

Fernandez de Arriba, A., Cavalcanti, F., Miralles, A., Bayon, Y., Alonso, A., Merlos, M., Garcia-Rafanell, J., and Forn, J. (1999). Inhibition of cyclooxygenase-2 expression by 4-trifluoromethyl derivatives of salicylate, triflusal, and its deacetylated metabolite, 2-hydroxy-4-trifluoromethylbenzoic acid. Mol Pharmacol *55*, 753-760.

Greg, L. (2006). RDKit: Open-source cheminformatics (Online ToolKit).

Gu, H.F., Mao, X.Y., and Du, M. (2019). Prevention of breast cancer by dietary polyphenols-role of cancer stem cells. Crit Rev Food Sci Nutr, 1-16.

Guha, R. (2007). Chemical Informatics Functionality in R. Journal of Statistical Software *18*.

Hartley, L., Flowers, N., Holmes, J., Clarke, A., Stranges, S., Hooper, L., and Rees, K. (2013). Green and black tea for the primary prevention of cardiovascular disease. Cochrane Database Syst Rev, CD009934.

Harvey, A.L., Edrada-Ebel, R., and Quinn, R.J. (2015). The re-emergence of natural products for drug discovery in the genomics era. Nat Rev Drug Discov *14*, 111-129.

Hosseini, A., and Ghorbani, A. (2015). Cancer therapy with phytochemicals: evidence from clinical studies. Avicenna J Phytomed *5*, 84-97.

Jensen, K., Ni, Y., Panagiotou, G., and Kouskoumvekaki, I. (2015). Developing a molecular roadmap of drug-food interactions. PLoS Comput Biol *11*, e1004048.

Kang, J., Hsu, C.H., Wu, Q., Liu, S., Coster, A.D., Posner, B.A., Altschuler, S.J., and Wu, L.F. (2016). Improving drug discovery with high-content phenotypic screens by systematic selection of reporter cell lines. Nat Biotechnol *34*, 70-77.

Koehn, F.E., and Carter, G.T. (2005). The evolving role of natural products in drug discovery. Nat Rev Drug Discov *4*, 206-220.

Kolodziejczyk, A.A., Zheng, D., and Elinav, E. (2019). Diet-microbiota interactions and personalized nutrition. Nat Rev Microbiol *17*, 742-753.

Lavecchia, A. (2015). Machine-learning approaches in drug discovery: methods and applications. Drug Discov Today *20*, 318-331.

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A.C., Liu, Y., Maciejewski, A., Arndt, D., Wilson, M., Neveu, V., *et al.* (2014). DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res *42*, D1091-1097.

Ley, J.P., Krammer, G., Reinders, G., Gatfield, I.L., and Bertram, H.J. (2005). Evaluation of bitter masking flavanones from Herba Santa (Eriodictyon californicum (H. and A.) Torr., Hydrophyllaceae). J Agric Food Chem *53*, 6061-6066.

Li, F., Wang, Y., Li, D., Chen, Y., and Dou, Q.P. (2019). Are we seeing a resurgence in the use of natural products for new drug discovery? Expert Opin Drug Discov *14*, 417-420.

Lim, S., Lee, K., and Kang, J. (2018). Drug drug interaction extraction from the literature using a recursive neural network. PLoS One *13*, e0190926.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371112; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Lima, A.N., Philot, E.A., Trossini, G.H., Scott, L.P., Maltarollo, V.G., and Honorio, K.M. (2016). Use of machine learning approaches for novel drug discovery. Expert Opin Drug Discov *11*, 225-239.

Lo, Y.C., Rensi, S.E., Torng, W., and Altman, R.B. (2018). Machine learning in chemoinformatics and drug discovery. Drug Discov Today *23*, 1538-1546.

Maggiora, G., Vogt, M., Stumpfe, D., and Bajorath, J. (2014). Molecular similarity in medicinal chemistry. J Med Chem *57*, 3186-3204.

Montaruli, M., Alberga, D., Ciriaco, F., Trisciuzzi, D., Tondo, A.R., Mangiatordi, G.F., and Nicolotti, O. (2019). Accelerating Drug Discovery by Early Protein Drug Target Prediction Based on a Multi-Fingerprint Similarity Search. Molecules *24*.

Muegge, I., and Mukherjee, P. (2016). An overview of molecular fingerprint similarity search in virtual screening. Expert Opin Drug Discov *11*, 137-148.

Newman, D.J., and Cragg, G.M. (2020). Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. J Nat Prod *83*, 770-803.

O'Hagan, S., and Kell, D.B. (2017a). Analysis of drug-endogenous human metabolite similarities in terms of their maximum common substructures. J Cheminform *9*, 18.

O'Hagan, S., and Kell, D.B. (2017b). Consensus rank orderings of molecular fingerprints illustrate the ‘most genuine’ similarities between marketed drugs and small endogenous human metabolites, but highlight exogenous natural products as the most important ‘natural’ drug transporter substrates. ADMET & DMPK *5*, 85-125.

Park, K., Ko, Y.J., Durai, P., and Pan, C.H. (2019). Machine learning-based chemical binding similarity using evolutionary relationships of target genes. Nucleic Acids Res *47*, e128.

Paulke, A., Kremer, C., Wunder, C., Achenbach, J., Djahanschiri, B., Elias, A., Schwed, J.S., Hubner, H., Gmeiner, P., Proschak, E.*, et al.* (2013). Argyreia nervosa (Burm. f.): receptor profiling of lysergic acid amide and other potential psychedelic LSD-like compounds by computational and binding assay approaches. J Ethnopharmacol *148*, 492-497.

Periwal, V., Kishtapuram, S., Open Source Drug Discovery, C., and Scaria, V. (2012). Computational models for in-vitro anti-tubercular activity of molecules based on high-throughput chemical biology screening datasets. BMC Pharmacol *12*, 1.

Periwal, V., Rajappan, J.K., Open Source Drug Discovery, C., Jaleel, A.U., and Scaria, V. (2011). Predictive models for anti-tubercular molecules using machine learning on high-throughput biological screening datasets. BMC Res Notes *4*, 504.

Petrone, P.M., Simms, B., Nigsch, F., Lounkine, E., Kutchukian, P., Cornett, A., Deng, Z., Davies, J.W., Jenkins, J.L., and Glick, M. (2012). Rethinking molecular similarity: comparing compounds on the basis of biological activity. ACS Chem Biol *7*, 1399-1409.

Pye, C.R., Bertin, M.J., Lokey, R.S., Gerwick, W.H., and Linington, R.G. (2017). Retrospective analysis of natural products provides insights for future discovery trends. Proc Natl Acad Sci U S A *114*, 5601-5606.

Rodrigues, T., Reker, D., Schneider, P., and Schneider, G. (2016). Counting on natural products for drug design. Nat Chem *8*, 531-541.

Rodriguez-Fragoso, L., Martinez-Arismendi, J.L., Orozco-Bustos, D., Reyes-Esparza, J., Torres, E., and Burchiel, S.W. (2011). Potential risks resulting from fruit/vegetable-drug interactions: effects on drug-metabolizing enzymes and drug transporters. J Food Sci *76*, R112-124.

Ryu, J.Y., Kim, H.U., and Lee, S.Y. (2018). Deep learning improves prediction of drug-drug and drug-food interactions. Proc Natl Acad Sci U S A *115*, E4304-E4311.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371112; this version posted November 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Seo, M., Shin, H.K., Myung, Y., Hwang, S., and No, K.T. (2020). Development of Natural Compound Molecular Fingerprint (NC-MFP) with the Dictionary of Natural Products (DNP) for natural product-based drug development. Journal of Cheminformatics *12*.

Shen, B. (2015). A New Golden Age of Natural Products Drug Discovery. Cell *163*, 1297-1300.

Sonnenburg, J.L., and Backhed, F. (2016). Diet-microbiota interactions as moderators of human metabolism. Nature *535*, 56-64.

Wang, L., Li, X., Zhang, S., Lu, W., Liao, S., Liu, X., Shan, L., Shen, X., Jiang, H., Zhang, W.*, et al.* (2012). Natural products as a gold mine for selective matrix metalloproteinases inhibitors. Bioorg Med Chem *20*, 4164-4171.

Wang, Y., Backman, T.W., Horan, K., and Girke, T. (2013). fmcsR: mismatch tolerant maximum common substructure searching in R. Bioinformatics *29*, 2792-2794.

Wang, Z., Liang, L., Yin, Z., and Lin, J. (2016). Improving chemical similarity ensemble approach in target prediction. J Cheminform *8*, 20.

Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z.*, et al.* (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res *46*, D1074-D1082.

Yang, H., Sun, L., Li, W., Liu, G., and Tang, Y. (2018). In Silico Prediction of Chemical Toxicity for Drug Design Using Machine Learning Methods and Structural Alerts. Front Chem *6*, 30.

Yosipof, A., Guedes, R.C., and Garcia-Sosa, A.T. (2018). Data Mining and Machine Learning Models for Predicting Drug Likeness and Their Disease or Organ Category. Front Chem *6*, 162.

Yu, X., Geer, L.Y., Han, L., and Bryant, S.H. (2015). Target enhanced 2D similarity search by using explicit biological activity annotations and profiles. J Cheminform *7*, 55.

Yun-Choi, H.S., Kim, J.H., and Lee, J.R. (1987). Potential inhibitors of platelet aggregation from plant sources, III. J Nat Prod *50*, 1059-1064.

Zhang, L., Zhang, H., Ai, H., Hu, H., Li, S., Zhao, J., and Liu, H. (2018a). Applications of Machine Learning Methods in Drug Toxicity Prediction. Curr Top Med Chem *18*, 987-997.

Zhang, R., Li, X., Zhang, X., Qin, H., and Xiao, W. (2020). Machine learning approaches for elucidating the biological effects of natural products. Nat Prod Rep.

Zhang, X., Rakesh, K.P., Shantharam, C.S., Manukumar, H.M., Asiri, A.M., Marwani, H.M., and Qin, H.L. (2018b). Podophyllotoxin derivatives as an excellent anticancer aspirant for future chemotherapy: A key current imminent needs. Bioorg Med Chem *26*, 340-355.

Zhao, Y.Y., Chen, H., Tian, T., Chen, D.Q., Bai, X., and Wei, F. (2014). A pharmaco-metabonomic study on chronic kidney disease and therapeutic effect of ergone by UPLC-QTOF/HDMS. PLoS One *9*, e115467.

Zi, C.T., Gao, Y.S., Yang, L., Feng, S.Y., Huang, Y., Sun, L., Jin, Y., Xu, F.Q., Dong, F.W., Li, Y.*, et al.* (2019). Design, Synthesis, and Biological Evaluation of Novel Biotinylated Podophyllotoxin Derivatives as Potential Antitumor Agents. Front Chem *7*, 434.

Zmora, N., Suez, J., and Elinav, E. (2019). You are what you eat: diet, health and the gut microbiota. Nat Rev Gastroenterol Hepatol *16*, 35-56.