



# SCIENTIFIC REPORTS



OPEN

## DPubChem: a web tool for QSAR modeling and high-throughput virtual screening

Othman Soufan<sup>1</sup>, Wail Ba-alawi<sup>2,3</sup>, Arturo Magana-Mora <sup>4</sup>, Magbubah Essack <sup>5</sup> & Vladimir B. Bajic <sup>5</sup>

Received: 19 February 2018

Accepted: 31 May 2018

Published online: 14 June 2018

High-throughput screening (HTS) performs the experimental testing of a large number of chemical compounds aiming to identify those active in the considered assay. Alternatively, faster and cheaper methods of large-scale virtual screening are performed computationally through quantitative structure-activity relationship (QSAR) models. However, the vast amount of available HTS heterogeneous data and the imbalanced ratio of active to inactive compounds in an assay make this a challenging problem. Although different QSAR models have been proposed, they have certain limitations, e.g., high false positive rates, complicated user interface, and limited utilization options. Therefore, we developed DPubChem, a novel web tool for deriving QSAR models that implement the state-of-the-art machine-learning techniques to enhance the precision of the models and enable efficient analyses of experiments from PubChem BioAssay database. DPubChem also has a simple interface that provides various options to users. DPubChem predicted active compounds for 300 datasets with an average geometric mean and  $F_1$  score of 76.68% and 76.53%, respectively. Furthermore, DPubChem builds interaction networks that highlight novel predicted links between chemical compounds and biological assays. Using such a network, DPubChem successfully suggested a novel drug for the Niemann-Pick type C disease. DPubChem is freely available at [www.cbrc.kaust.edu.sa/dpubchem](http://www.cbrc.kaust.edu.sa/dpubchem).

Comprehensive and expanding public resources, such as the PubChem BioAssay Database (BioAssayDB)<sup>1</sup>, provide access to biological activity information from high-throughput screening (HTS) experiments. The vast amount of available data allows for the development of quantitative structure-activity relationship (QSAR) models to predict biological activities of chemical compounds for individual assays, enabling the so-called virtual (*in silico*) screening. QSAR models for virtual screening are derived by the standard ligand-based computational technique used in drug discovery to examine the compound libraries and to find potential candidates for binding with a specific and known biological target<sup>2–4</sup>. From a computational perspective, virtual screening involves analysis of large amounts of input data, integration of heterogeneous types of data, different statistical measures, and reliable selection of unbiased significant results and predictions<sup>2,5,6</sup>. These challenges, unless addressed in a carefully designed computational setup, cannot be carried out efficiently in later experimental phases in the process of drug discovery. Although several QSAR models implemented as web tools for predicting chemical-protein interactions have been developed<sup>7–15</sup>, they are limited in many aspects, for example, prediction performance is hampered by the imbalanced data in the HTS assays (the number of active compounds is usually significantly smaller than the inactive), the type of models available to the user as well as the flexibility to tune their parameters. Therefore, tools that possess the ability to reduce these limitations are of interest. Here, we introduce Dragon PubChem (DPubChem), a novel web tool to derive QSAR models for virtual screening of biological activity of chemical compounds. DPubChem reduces some of the limitations mentioned above by offering a rich set of options to the user that are easy to choose from (2 input types × 6 types of chemical features × 6 feature selection methods × 7 solutions for addressing class imbalance × 7 types of classifiers). Moreover, by considering the corre-

<sup>1</sup>Institute of Parasitology, McGill University, Montreal, QC, H9X 3V9, Canada. <sup>2</sup>Princess Margaret Cancer Centre, University Health Network, Toronto, ON, M5G 1L7, Canada. <sup>3</sup>Department of Medical Biophysics, University of Toronto, Toronto, ON, M5G 1L7, Canada. <sup>4</sup>Computational Bio Big-Data Open Innovation Laboratory (CBBDO-OIL), National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, 135-0064, Japan. <sup>5</sup>Computational Bioscience Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, Saudi Arabia. Correspondence and requests for materials should be addressed to V.B.B. (email: [vladimir.bajic@kaust.edu.sa](mailto:vladimir.bajic@kaust.edu.sa))

lation of the HTS data, DPubChem allows for multi-label learning, where several HTS assays may be simultaneously used to derive QSAR models for more enriched virtual screening tasks. Since all the different options available in DPubChem tool are easy to use, it is straightforward to run several experiments and compare different models in order to select the optimal model. The results obtained from the 300 selected datasets, composed of 116,751 interactions and characterized by high class imbalance data, show that DPubChem is able to outperform existing QSAR models and that it achieved an average geometric mean (GMean) and  $F_1$  score (referred as  $F_1$ , hereafter) of 76.68% and 76.53%, respectively. Half of the considered 300 datasets represent bioassays with hundreds to thousands of chemical compounds. Nevertheless, other datasets with a fewer number of compounds were also included to show the general applicability of DPubChem and to demonstrate that the implemented recognition models in the tool do not require large datasets for deriving robust models (as opposed to other recognition models that normally require large training data, such as deep learning models<sup>16,17</sup>). To the best of our knowledge, DPubChem is the only tool that provides: 1) an efficient mechanism to retrieve and analyze PubChem BioAssays, 2) an implementation of state-of-the-art machine learning algorithms (i.e., class imbalance and multi-label methods) to build QSAR models, and 3) a tool to rank and visualize unknown activity predictions for hundreds of chemical compounds provided by the user. DPubChem aims to provide an easy to use tool that will help biologists, biochemists, and experimentalists obtain useful insights about the chemicals and drugs of interest.

## Results

The key contribution of our study is the development of DPubChem, a novel and freely available web tool for deriving QSAR models for virtual screening of biologically active compounds from PubChem assays. The DPubChem tool implements the state-of-the-art methods for mining HTS data and provides users with an extensive but easy to use set of options to build robust models without compromising the simplicity of the interface.

In this section, we compare DPubChem to existing tools and provide an overview of its interface. We then show the results obtained when using the state-of-the-art methods for multi-label classification (MLC) and for addressing the data class imbalance. Finally, we provide a case study analysis where we used DPubChem to suggest a drug for the Niemann-Pick type C (NPC) disease.

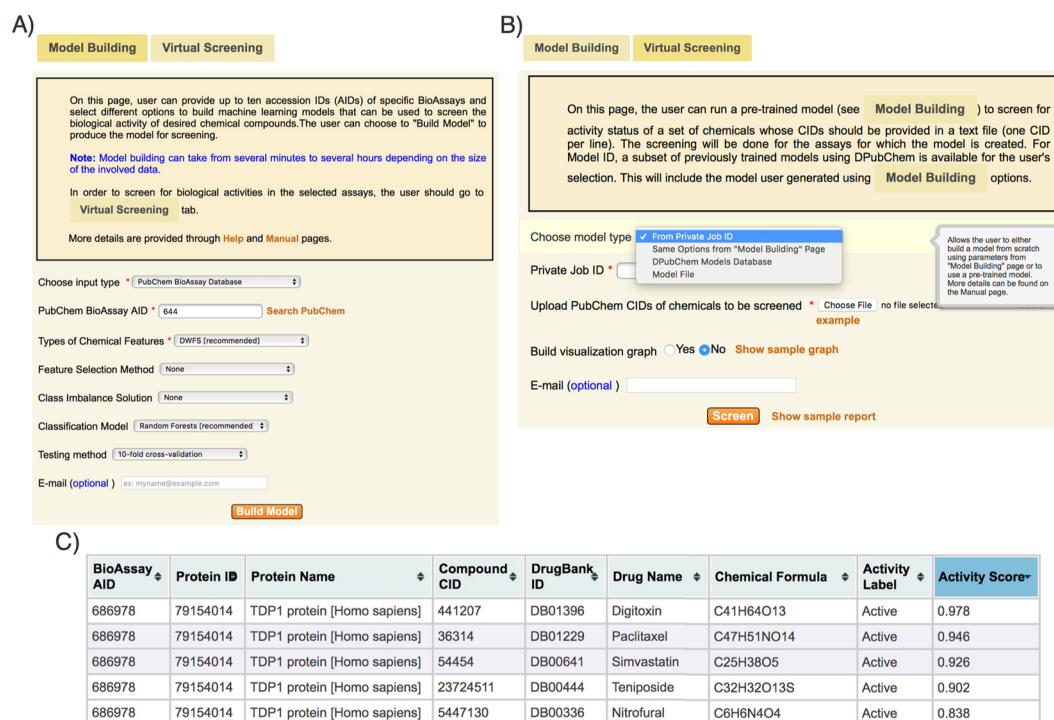
**Comparisons with other web servers and interface overview.** Compared to many existing web servers for 3D docking<sup>18–22</sup>, which rely on ligand-protein docking, a smaller number of data-driven online systems were developed for virtual screening of chemical activities. As opposed to the 3D docking servers, the data-driven approaches do not require any prior knowledge of 3D structures of the target and its ligand. In addition, when data-driven models are trained, they can be used for screening the biological activity status of a set of chemicals faster than ligand-protein docking approaches<sup>23</sup>, which is an issue in screening a large number of compounds.

There are several web tools for predicting chemical-protein interactions<sup>8,10,15</sup>. The OCHEM<sup>14</sup> and ChemBench<sup>12,13</sup> are among the first freely available tools to mine HTS assays and allow users to derive different prediction models based on the user's input. However, these tools require several data processing steps to produce a predictive model, and these may not be straightforward for the user. The HitPick<sup>10</sup> tool has a simpler interface with a fixed model based on 2D molecular fingerprints and a Laplacian-modified naïve Bayes classifier. Later, the STITCH tool<sup>7–9</sup> was developed to facilitate the search for the interactions of chemicals and proteins from a unified database extracted from different databases and literature. However, the main aim of STITCH is to provide a comprehensive database not specifically for the development of QSAR models. The tools mentioned above have certain limitations, for example, they do not address the imbalanced data of HTS assays and thus, are unable to reduce the false positive predictions. Moreover, some of these tools offer the flexibility to select different types of prediction models and parameters but at the expense of the simplicity. Finally, some of these tools also lack integration with a chemical compounds database. With all these shortcomings in mind, DPubChem tool focuses on the usage simplicity, flexibility, and prediction performance. Table 1 summarizes the characteristics of the QSAR tools mentioned above.

The DPubChem tool allows the user to simply provide a bioassay accession number (AID) and the system automatically retrieves all relevant information for processing the HTS data of interest. The user can also provide a set of PubChem compound accession numbers (i.e., CID). Although the primary objective of the DPubChem tool is to derive QSAR models from PubChem data, the user may also input a list of simplified molecular-input line-entry system (SMILES)<sup>24</sup> representing compounds of interest with their corresponding labels for target activity to build a model. Moreover, a list of AIDs may also be submitted for deriving an MLC model, where the correlation given by the common active compounds in different bioassays is exploited. In MLC models, each sample (a chemical compound, in our case) is assigned to multiple labels as opposed to just one label in binary or multi-class classification models<sup>25,26</sup>. This can be thought of as predicting properties of a chemical compound that are not mutually exclusive, such as, a chemical may be an activator in different bioassays (see Methods). MLC models have been applied to solve different problems in the bioscience domain and resulted in improved results compared to single label classification models<sup>27–31</sup>. Additionally, DPubChem implements a set of different feature selection methods to find an optimal subset of features and have demonstrated ability to reduce model complexity while, in some cases, enhances the prediction performance<sup>32–35</sup>. Since chemical compounds may be defined by thousands of different features (e.g., topological fingerprints, MACCS keys, among others), it is likely that not all of these features are relevant for the recognition of compound activities. Therefore, it is possible to derive simpler and possibly more robust QSAR models by removing less relevant features from the initial set of features. Although feature selection methods need not improve the prediction performance for some recognition models, such as random forest (RF), the reduced subset of features still shortens training time and enables better model transparency<sup>36</sup>. Finally, DPubChem implements state-of-the-art solutions for addressing the class imbalance problem, which considerably increased the precision compared to other QSAR models for virtual screening (see Performance Evaluation subsection). Figure 1A shows the screenshot of the DPubChem interface for building a model.

Tool	ChemBench <sup>12,13</sup>	OCHEM <sup>14</sup>	HitPick <sup>10</sup>	MTI-OpenScreen <sup>11</sup>	DPubChem
Approach	HTS assay mining	HTS assay mining	HTS assay mining	Docking based screening	HTS assay mining
Input data type	SDF	SMILES, MOL2, SDF	SMILES	MOL2, SDF	SMILES, PubChem CID, BioAssay ID
Prediction model configuration	Flexible	Flexible	Fixed	Fixed	Flexible
Activity prediction (# of screening chemicals)	Yes (unlimited)	Yes	Yes (100)	Yes (5,000)	Yes (unlimited)
Addressing class imbalance or advanced preprocessing	No	No	Yes	No	Yes
Network visualization	No	No	No	No	Yes

**Table 1.** Characteristics of the virtual screening tools. Note: SDF refers to the structure files which store the structural information of one or more compounds in a dataset. MOL2 is a file containing the information to reconstruct a SYBYL molecule. SMILES stands for simplified molecular input line entry system and is a string of characters that represents a molecule. PubChem CID is a non-zero accession number representing a unique chemical structure.



**Figure 1.** (A) Screenshot of the Model Building page in DPubChem. Several options are available to build different types of machine learning models to predict biological activities of provided chemical compounds. (B) Screenshot of the DPubChem virtual Screening page. The user can simply provide Job ID of a previously trained model and submit a list of compounds for activity screening. (C) List of the screening outputs.

For building QSAR models and predicting the biological activities of new compounds, DPubChem provides the following options: 1) a model is trained by using the parameters specified in the Model Building tab, 2) if a model has already been trained, the model is stored on our server and the jobID obtained from the Model Building page (Fig. 1A) may be directly used, 3) the user can select from a list of pre-trained models, and 4) the user can upload the files of a previously trained model using our system. The implemented options to upload model files or to input a jobID of an already trained model facilitate both the collaboration across different teams and the reproducibility of the results. Finally, DPubChem can build a visualization graph that measures chemical-chemical similarity, protein-protein similarity, bioassay-bioassay similarity, and assign activity screening scores for chemical-bioassay interactions<sup>37</sup>. Figure 1B shows a screenshot of the virtual screening page. Finally, DPubChem generates several statistical measures, an interactive network and a list of the screening outcome (Fig. 1C). Supplementary Material 1 shows the steps for building and testing a screening model in the tool.

**Performance evaluation.** To evaluate the utility of the DPubChem tool, we first describe the results obtained by the state-of-the-art methods for addressing class imbalance and multi-label classification, followed by the performance obtained from 300 selected HTS assays.

Statistical measure	Equation
$F_1$	$\frac{2 \times TP}{2 \times TP + FP + FN}$
GMean	$\sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}}$

**Table 2.** Selected statistical measures. TP, TN, FN, and FP refer to true positives, true negatives, false negatives, and false positives, respectively.

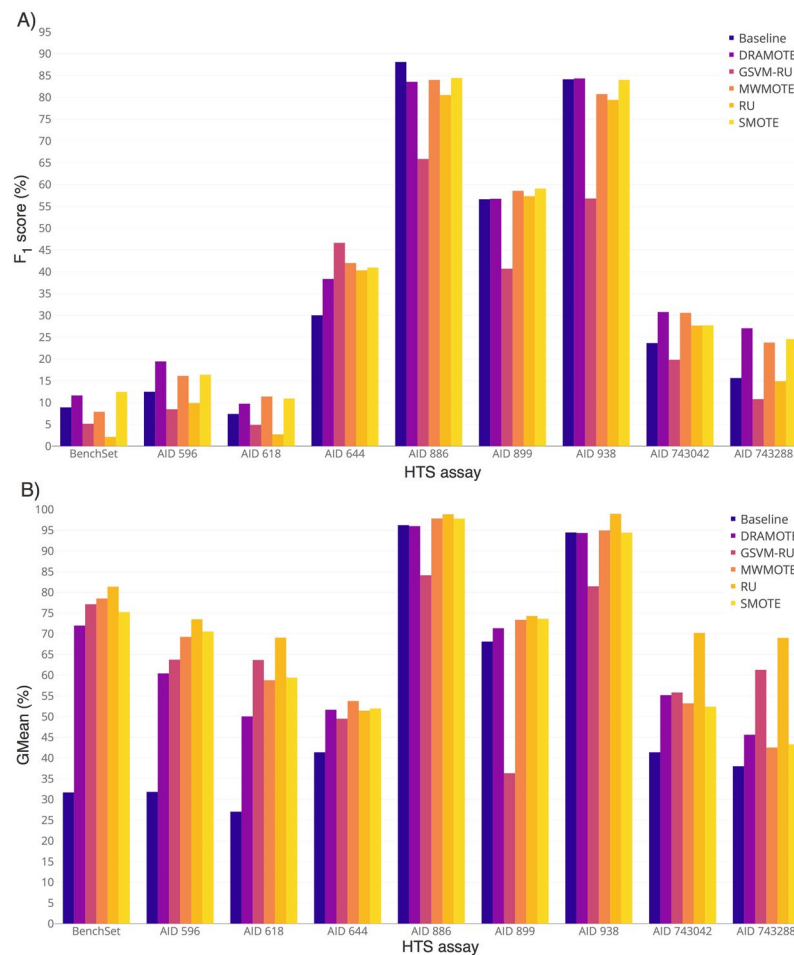
In the context of virtual screening, a novel predicted interaction by the QSAR model may require further experimental validation. Therefore, it is crucial for the QSAR models to reduce the number of falsely predicted active compounds (false positives). In this case, precision is a meaningful statistical measure since a higher precision score indicates a lower number of false positives, or in other words, represents the proportion of all predictions denoted as active that are actually active. However, a very stringent QSAR model that only predicts few active compounds may achieve a high precision score while failing to identify most of the active compounds. For this, sensitivity is also important to consider and it represents the proportion of the active compounds correctly identified by the model. Therefore, we report the results in terms of the  $F_1$ , which accounts for both precision and sensitivity. We also show the results in terms of the GMean of sensitivity and specificity to summarize prediction accuracy over both the true positive as well as the true negative rates. Both  $F_1$  and GMean incorporate false positives with different weight of importance. Specifically,  $F_1$  gives more preference to a lower number of false positives, while GMean reflects more the ability to identify the inactive class (i.e., true negatives). These two metrics are defined in Table 2.

As mentioned above, reducing the number of false positive and false negative predictions is of critical importance for the virtual screening, but also a major challenge for the machine learning methods. As HTS assays usually contain a much higher number of inactive than active compounds (imbalanced labels), machine learning models tend to bias the majority class<sup>38</sup>. For this, we integrated the state-of-the-art methods for solving the class imbalance problem (see Methods) into the DPubChem tool, namely, the Dragon Oversampling Technique (DRAMOTE)<sup>39</sup>, granular support vector machine for under-sampling (GSVM-RU)<sup>40,41</sup>, majority weighted minority over-sampling technique (MWMOTE)<sup>42</sup>, synthetic minority over-sampling technique (SMOTE)<sup>43</sup>, and the simple random under-sampling technique (RU). The methods were tested on 11 bioassays representing 487,557 active and inactive compounds. These datasets are characterized by different class imbalance ratios. Figure 2A,B show the average  $F_1$  and GMean from six classifiers, namely, support vector machines with linear kernel (SVM-L), support vector machine with radial basis function (SVM-RBF), K-nearest neighbor (KNN), linear discriminant analysis (LDA), naïve Bayes classifier (NBC), and RF using a 5-fold cross-validation, respectively. Supplementary Table S1 shows the number of compounds and the imbalance ratios of the HTS assays. From Fig. 2A, one observes that addressing the class imbalance enhanced the results showing an improvement of up to 55% compared to the baseline models (where the class imbalance is not addressed). The DRAMOTE and SMOTE, on average, achieved the best results except in assay AID 886. Figure 2B clearly demonstrates the effects of the unbalanced class labels when deriving machine learning models. In all the considered HTS assays, addressing the class imbalance considerably increased both the specificity and sensitivity and therefore, the GMean of the models. Notably, the BenchSet (AIDs 773, 1006, and 1379) showed the largest increase of the GMean by 247% compared to the baseline model. Although classification models are affected differently by the imbalance problem, an improvement of sensitivity (how well the model is able to recognize active compounds) was always observed when applying a class imbalance solution for the BenchSet dataset. Supplementary Table S2 shows the results obtained for each classification model on the BenchSet dataset. For instance, without addressing the class imbalance, we observe that RF was able to recognize all inactive compounds (100% specificity) but failed to identify the active compounds (only 4.06% sensitivity). Conversely, the sensitivity of the RF model when the class imbalance issue is addressed, increased to ~35–85% while conserving a high specificity (~85–100%).

However, it is not always necessary to address the class imbalance for certain HTS assays as they do not contain significant class imbalances. In these cases, it is possible to further improve the accuracy of the QSAR models by considering the correlation between assays given by the active compounds that are common in different HTS assays. As such, DPubChem implements the state-of-the-art technique Dragon Bayesian Active Learning (DRABAL)<sup>44</sup>, which consists of an MLC for modeling the correlations between several BioAssayDB assays (see Methods). The performance of DRABAL was tested on five assays (AIDs 1458, 485297, 485313, 588342, and 686978) representing 1,448,403 interactions with 7.7% hit rate indicating positive interactions. Using the 5-fold cross-validation, DRABAL achieved an  $F_1$  and GMean of 51.11% and 61.05%, respectively. These results represent a relative improvement of 14.27% and 9.91% for  $F_1$  and GMean, respectively, compared to the multi-label state-of-the-art methods.

Finally, to illustrate the applicability of DPubChem, we derived models for 300 datasets with different imbalance ratios and number of reported activities. The 300 selected datasets, reporting 116,751 activities, include bioassays with few compounds up to 11,000 (see Methods for the bioassay selection criteria). The imbalance ratio for these datasets ranged from 1:204 to 1:1. For each HTS assay, we used 80% of the data for training the model and the remaining 20% for testing with the default DPubChem options. The average  $F_1$  and GMean over the 300 assays are 76.68% and 76.53%, respectively. These results indicate a reasonable performance over the comprehensive set of HTS. Supplementary Table S3 shows the performance of the individual HTS assays. It is worth noting that the DPubChem tool provides a set of options for testing the QSAR model performance including cross-validation technique and holdout settings. The tool generates a report highlighting several performance measures, such as Cohen's kappa coefficient, sensitivity, specificity,  $F_1$ , and GMean, to help user judge the validity of the model and impact of potential noise.

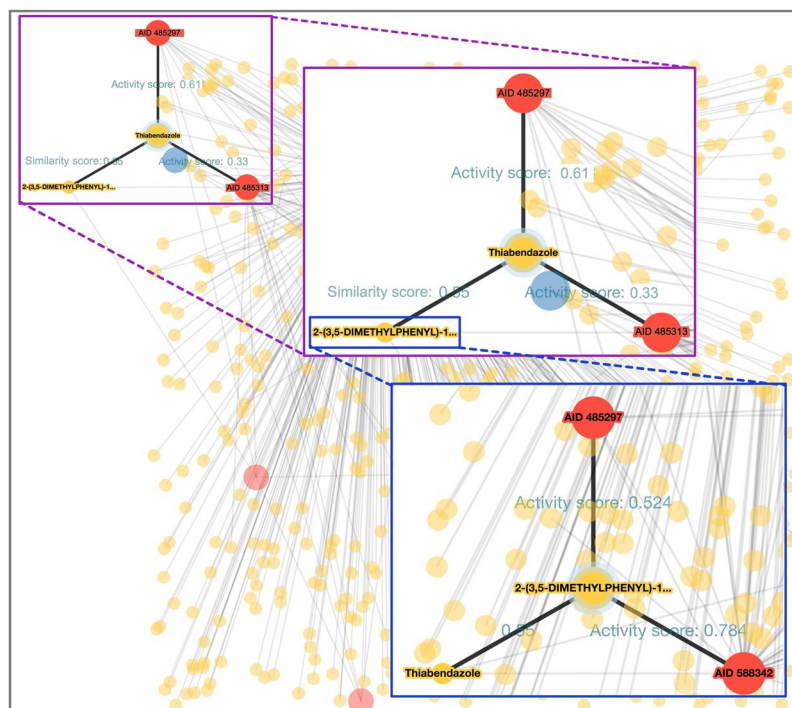




**Figure 2.** Average performance of the SVM-L, SVM-RBF, KNN, LDA, NBC, and RF from a 5-fold cross-validation of the implemented methods for reducing the effects of the class imbalance. **(A)** The results in terms of  $F_1$  for the 11 considered HTS assays. BenchSet refers to the pooled assays AIDs 773, 1006, and 1379 as described by Li *et al.*<sup>70</sup>. **(B)** The GMean results for the 11 HTS assays.

**Case study analysis.** In order to show the utility of DPubChem and the interaction networks, we screened the FDA approved drugs in five HTS assays, namely, AID 1458, AID 485313, AID 485297, AID 588342, and AID 686978. Based on the interaction network produced by DPubChem, we generated Fig. 3 aiming to highlight the interactions of relevance for the case study. Figure 3 shows the interactions between HTS assays and compounds with their predicted activity scores (see Methods) that enable us to suggest potential drug-target interactions of interest. Specifically, we focused on the results for bioassays AID 485313 (target protein being: Ras-related protein Rab-9A) and AID 485297 (target protein being: Niemann-Pick C1 protein precursor) as both of these target proteins that could potentially serve as targets of drugs for Niemann-Pick type C (NPC) disease. The prediction of interactions by DRABAL<sup>44</sup> component of DPubChem and resultant interaction network showed that Thiabendazole (DrugBank Database ID: DB00730) is the strongest common activator of HTS assays AID 485313 and AID 485297. Since overexpression of both Rab-A9 and NPC1 proteins have been shown to reduce the symptoms of the NPC disease<sup>45,46</sup>, we hypothesized that the common predicted activator (Thiabendazole) may inhibit the progression of the NPC disease. Additionally, Thiabendazole is an aryl hydrocarbon receptor ligand that has been shown to reduce levels of cathepsin D<sup>47</sup>, a protein which overexpression has been implicated in some of the symptoms of the NPC disease, apoptosis<sup>48</sup>, and liver fibrosis<sup>49</sup>. Finally, we note that both Thiabendazole (predicted activator) and Benzoic Acid (approved drug DB03793 to target Rab-9A protein) belong to the same Benzenoid superclass. Although Thiabendazole is deemed slightly toxic and is actively used as a pesticide, pharmacokinetics studies of Thiabendazole report that ~87% of the oral dose in humans excretes within 24 hours and similarly in animals<sup>50</sup>. Therefore, Thiabendazole may be beneficial if administered conservatively<sup>44</sup>.

Although this case study focuses on Thiabendazole for the potential inhibition of NPC disease, interaction networks are a powerful tool for the identification of other chemical compounds common to multiple bioassays. Moreover, while QSAR models predict the activity of a candidate chemical compound for a specific bioassay, the similarities between these compounds provide another layer of information. Inspired by the underlying idea that similar compounds are likely to interact with similar proteins, these networks provide a graphical representation that may facilitate the identification of similar compounds to those known to be active within a given bioassay. Additionally, based on Fig. 3 (bottom right), we may hypothesize that



**Figure 3.** DPubChem interaction network obtained from applying DRABAL on the five selected HTS assays (AIDs 1458, 485313, 485297, 588342, and 686978). Thiabendazole (DB00730) is the top common prediction for assays AIDs 485297 and 485313. In the graph, red, yellow, and blue colored nodes indicate HTS assays from PubChem database, chemicals, and the biological target entities like protein targets, respectively. The interaction network is accessible at [www.cbrc.kaust.edu.sa/dpubchem/DraPubChemGraph/drapubchemgraph.html](http://www.cbrc.kaust.edu.sa/dpubchem/DraPubChemGraph/drapubchemgraph.html) and is obtained by running DRABAL MLC method on the selected HTS assays.

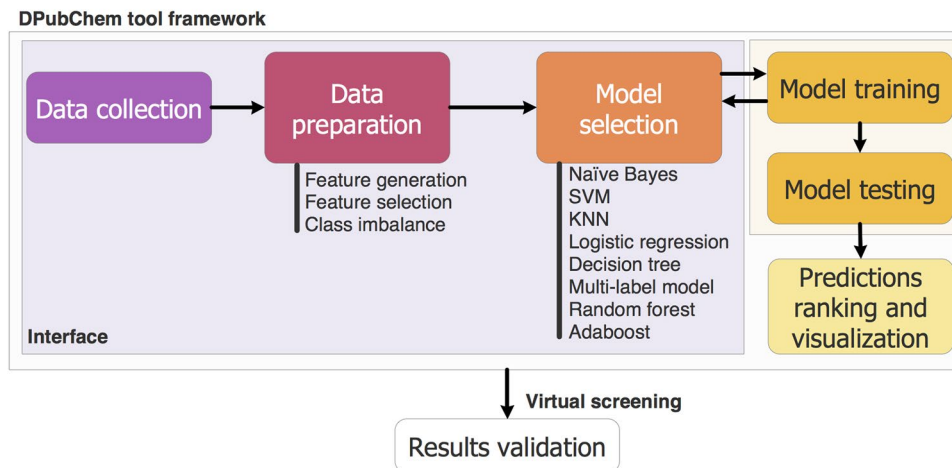
compound 2-(3,5-Dimethylphenyl)-1,3-Benzoxazole may also have an effect in the NPC disease as it is similar to Thiabendazole (similarity score of 0.55) and has a predicted activity in AID 485297 (activity score of 0.524). Clearly, a compound with a higher similarity score to Thiabendazole would indicate a better candidate.

## Discussion

With the vast amount of data from public repositories that provide access to biological activity information from HTS experiments (e.g., BioAssayDB), there is an opportunity to develop categorical models to predict the biological activities of millions of chemical compounds that remain untested. Although several QSAR models have been proposed, they remain limited in many aspects. Some of these tools lack flexibility or impose a set of different steps for data processing that are complicated and time-consuming. Moreover, these tools have a low precision of predictions, i.e., high false positive rates. Consequently, we developed DPubChem, a tool that enables sophisticated virtual screening strategies based on state-of-the-art machine learning methods for feature selection, class imbalance, and MLC. The DPubChem tool focuses on the simplicity of the interface without compromising the flexibility and in the precision of the QSAR models to reduce false positive predictions. Because DPubChem uses a single easy-to-use workflow that supports a different set of models and options, it is straightforward to generate different models for the same screening task. Each of these models provides the performance statistics, predictions ranking, and the graph visualization, which allows the user to easily select the best performing model. The notable ease-of-use of this tool is essential for users. The prediction results from DPubChem may be further examined by published results in the literature and other computational techniques like 3D docking simulations. Although docking simulations are prone to false positives, they can indirectly support the top predictions from our data-driven approach<sup>39</sup>, especially, if an experimental validation (i.e., *in vitro* and *in vivo* based) is prohibitive to run. Currently, DPubChem does not explicitly allow for such docking simulation types of verifications. Nevertheless, it provides visualization graphs with reference links that can be used to get more background information. We believe that DPubChem will contribute to the progress of bioinformatics and biomedical research.

## Methods

The framework in Fig. 4 depicts the core architectural design of DPubChem. In particular, we propose four modules to enable virtual screening of HTS assays using machine learning methods. The first module is the data collection, in where the user may input 1) a BioAssay accession number (AID), 2) a list of AIDs for an MLC model, 3) a list of PubChem compound accession numbers (CID), and 4) a list of SMILES records with their corresponding labels. The next is the data preparation module, where the chemical compounds are described by a set of features (feature generation), which can be further reduced by a feature selection method. The third module is for the model selection, where different classifiers may be selected. Finally, the QSAR models for virtual screening



**Figure 4.** DPubChem framework for virtual screening. Different colors indicate the four modules that represent the core architectural design of DPubChem.

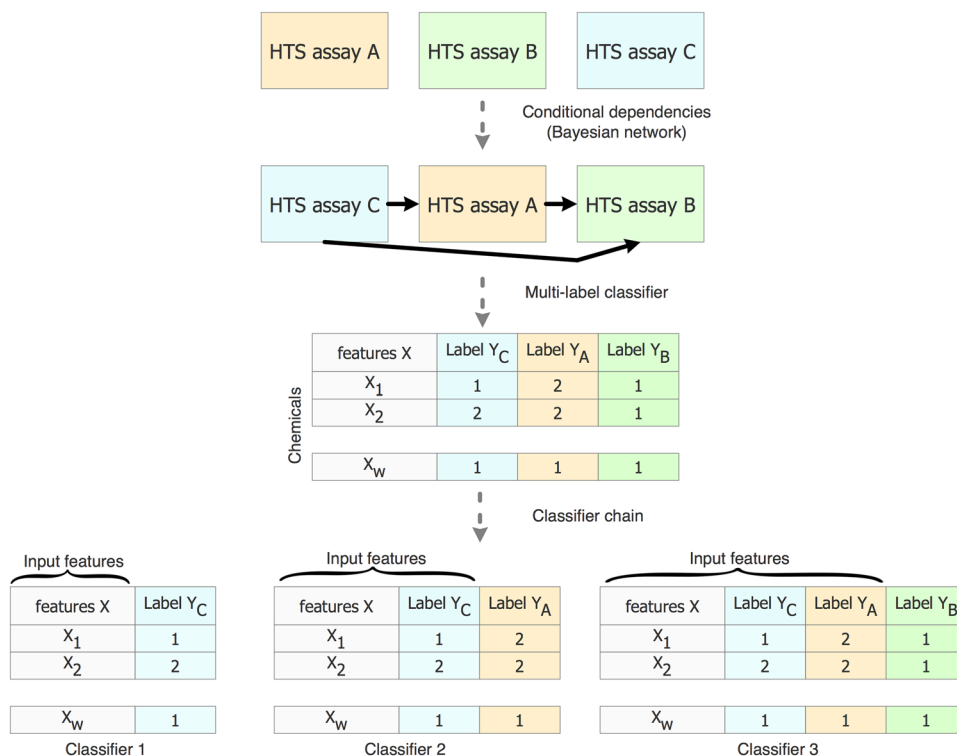
are derived, and the predictions are ranked and visualized in an interaction network. The following subsections describe these modules in more details.

**Data collection.** The datasets retrieved by DPubChem are based on the PubChem BioAssay protocol, where datasets represent HTS assays that can be referenced by a unique AID identifier. We considered bioassays that report experimental activity results for a set of chemical compounds over a specific biological target (e.g., a protein). Therefore, a bioassay dataset contains a list of chemical compounds to which we assign labels, where label '1' indicates that the compound appears active in the assay, while '2' relates to inactive compounds. The probe designation was considered as active (label of 2) as it indicates that the activity of the test result has been tested and confirmed through multiple rounds of experimental inquiry<sup>1</sup>. Inconclusive or unspecified activity types are ignored. The criteria for selecting the 300 considered datasets are: 1) since sufficient information about both classes is needed to build meaningful recognition models only confirmatory assays with more than five samples for both active and inactive classes were considered, 2) to have a reasonable processing time, we selected assays containing at most 11,000 reported active compounds, and 3) the 300 datasets were randomly selected. Although some of these datasets contain a relatively small number of compounds, models derived from these datasets allow for HTS.

**Data preparation and feature selection.** In order to build QSAR models, chemicals are encoded into a set of features. The generation and selection of a representative subset of features are critical for developing accurate QSAR models<sup>33</sup>. DPubChem implements different types of chemical features, namely, 1) the PubChem fingerprints<sup>51</sup> (881 features from [ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem\\_fingerprints.txt](ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt)), 2) MACCS keys fingerprints from the toolkit OpenBabel<sup>52</sup> (166 features), 3) the topological fingerprints from the toolkit RDKit<sup>53</sup> (1,024 features), 4) chemical descriptors, i.e., the number of H-acceptors and donors, molecular weight, and Log-P, among others (166 features), 5) a standard set of features, representing a combination of all the previous types of features, and 6) a recommended set of chemical features, DWFS. The DWFS set of features is the result of an extensive analysis and feature selection process, which resulted in an optimized subset of 1,064 chemical features that enabled superior predictor performance<sup>39,44,54</sup>. However, there is no guarantee that this optimized set of features will be optimal for all cases.

Feature selection is crucial for removing irrelevant or redundant features that do not contribute to the performance of the QSAR models. This results in simpler and possibly more accurate models. For this, DPubChem provides efficient solutions to select the most relevant features. The current implementation incorporates the following state-of-the-art methods for feature selection from the MATLAB Feature Selection Tool (FEAST)<sup>55</sup>: minimum redundancy maximum relevance (mRMR)<sup>34</sup>, joint mutual information (JMI)<sup>34</sup>, conditional mutual information maximization (CMIM)<sup>56</sup>, and RELIEF<sup>57</sup>. Moreover, DPubChem also includes the simpler algorithms for feature selection based on the correlation of features and the standard deviation.

**Classification model selection.** Seven widely used classifiers are available in DPubChem as a basis for building prediction models for PubChem assays. These include SVM-L and SVM-RBF<sup>58</sup>, K-NN<sup>59</sup>, decision trees<sup>60</sup>, NBC<sup>61</sup>, LDA, and ensemble classifiers RF<sup>62</sup> and Adaboost<sup>63</sup> from the Scikit learn machine learning package<sup>64</sup>. There are several justifications for the selection of the implemented classifications models. Although deep neural networks have achieved superior performance over shallow models in some applications<sup>38,65</sup>, the training and tuning of such complex models are computationally-demanding, especially for large datasets as are many in the PubChem assays. Therefore, we have included less computationally-demanding models. Additionally, an extensive empirical study<sup>66</sup> tested the performance of different classification models over the UCI machine-learning repository database<sup>67</sup> and showed that SVM and ensemble models (random forest and Adaboost) were among the top-ranked models.



**Figure 5.** Multi-label classification model using classification chain transformation.

**Multi-label classification model.** Since various HTS assays in BioAssayDB are correlated by sharing some portion of the same set of active compounds, we have included DRABAL<sup>44</sup>, an MLC technique for modeling correlations between several BioAssayDB assays to enhance prediction performance. DRABAL uses a problem transformation method to derive MLC models. This method transforms the MLC problem into a chain of  $n$  single label classifiers, where  $n$  denotes the number of labels assigned to a chemical compound. For this, the first classifier is derived by using the input data to fit one label, and then each of the next  $n$  classifiers is trained on the original input data and the labels of the previous classifiers (i.e., target labels are concatenated to the original set of features). For example, the last classifier in the chain, classifier  $n$ , would be derived by using the original input data and the  $n-1$  target labels. This implies that the order of the labels has to be specified. For this, DRABAL uses a Bayesian network to learn the correlation of the target labels of the assays<sup>44</sup>. Figure 5 shows an example of an MLC model for three HTS assays.

**Class imbalance problem in HTS assays.** Given the nature of HTS assays, which are often characterized by a small number of active chemical compounds obtained after screening a big compound set library, we implemented several state-of-the-art solutions in DPubChem to address the class imbalance problem. In several cases, addressing the class imbalance has shown to overcome possible bias to the majority class and achieved considerably better results than without any data preprocessing. Depending on the imbalance ratio and size of active compounds, different solutions can lead to different QSAR model performance. DPubChem offers seven combinations that can result in a different effect on both sensitivity and precision of virtual screening. The current options include the following approaches: RU<sup>39</sup>, SMOTE<sup>43</sup>, MWMOTE<sup>42</sup>, and the precision-aware method called DRAMOTE<sup>39</sup>.

**Interaction networks.** Given a set of chemical compounds  $C = \{c_1, c_2, \dots, c_n\}$ , a set of proteins  $P = \{p_1, p_2, \dots, p_n\}$ , and a set of assays  $A = \{a_1, a_2, \dots, a_n\}$ , the interaction network can be generated to represent nodes from  $C$ ,  $P$ , and  $A$  and their links. Chemical-chemical links (weighted edges between  $c_1, c_2 \in C$ ) and protein-protein links (weighted edges between  $p_1, p_2 \in P$ ) represent the similarity scores between the two nodes and are computed by using the SIMComp<sup>68</sup> and Smith-Waterman<sup>69</sup> methods, respectively. The chemical compound-bioassay interaction links (weighted edges between  $c \in C, a \in A$ ) denote the predicted probability of a compound to be active by the QSAR model. The similarity scores and the probability of interactions are values within the range of  $[0, 1]$ .

## References

1. Wang, Y. *et al.* PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic acids research*, gkp456 (2009).
2. McInnes, C. Virtual screening strategies in drug discovery. *Current opinion in chemical biology* **11**, 494–502 (2007).
3. Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews Drug discovery* **3**, 935–949 (2004).



4. Roy, A. & Skolnick, J. LIGSIFT: an open-source tool for ligand structural alignment and virtual screening. *Bioinformatics* **31**, 539–544 (2015).
5. Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research* **40**, D1100–D1107 (2012).
6. Loging, W., Harland, L. & Williams-Jones, B. High-throughput electronic biology: mining information for drug discovery. *Nature Reviews Drug discovery* **6**, 220–230 (2007).
7. Kuhn, M. *et al.* STITCH 4: integration of protein-chemical interactions with user data. *Nucleic Acids Res* **42**, D401–407, <https://doi.org/10.1093/nar/gkt1207> (2014).
8. Kuhn, M., von Mering, C., Campillos, M., Jensen, L. J. & Bork, P. STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res* **36**, D684–688, <https://doi.org/10.1093/nar/gkm795> (2008).
9. Szklarczyk, D. *et al.* STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Research* **44**, <https://doi.org/10.1093/nar/gkv1277> (2015).
10. Liu, X., Vogt, I., Haque, T. & Campillos, M. HitPick: a web server for hit identification and target prediction of chemical screenings. *Bioinformatics* **29**, 1910–1912, <https://doi.org/10.1093/bioinformatics/btt303> (2013).
11. Labbé, C. M. *et al.* MTiOpenScreen: a web server for structure-based virtual screening. *Nucleic acids research* **43**, W448–W454 (2015).
12. Capuzzi, S. J. *et al.* Chembench: A Publicly Accessible, Integrated Cheminformatics Portal. *J. Chem. Inf. Model* **57**, 105–108, <https://doi.org/10.1021/acs.jcim.6b00462> (2017).
13. Walker, T., Grulke, C. M., Pozefsky, D. & Tropsha, A. Chembench: a cheminformatics workbench. *Bioinformatics* **26**, 3000–3001, <https://doi.org/10.1093/bioinformatics/btq556> (2010).
14. Sushko, I. *et al.* Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J Comput Aided Mol Des* **25**, 533–554, <https://doi.org/10.1007/s10822-011-9440-2> (2011).
15. Sakakibara, Y. *et al.* COPICAT: a software system for predicting interactions between proteins and chemical compounds. *Bioinformatics* **28**, 745–746, <https://doi.org/10.1093/bioinformatics/bts031> (2012).
16. Liu, B., Wei, Y., Zhang, Y., & Yang, Q. Deep neural networks for high dimension, low sample size data. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17 (pp. 2287–2293), (2017).
17. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *The Journal of Machine Learning Research* **15**, 1929–1958 (2014).
18. Grosdidier, A., Zoete, V. & Michielin, O. SwissDock, a protein-small molecule docking web service based on EADock DSS. *Nucleic Acids Res* **39**, W270–277, <https://doi.org/10.1093/nar/gkr366> (2011).
19. Li, H. *et al.* TarFisDock: a web server for identifying drug targets with docking approach. *Nucleic Acids Res* **34**, W219–224, <https://doi.org/10.1093/nar/gkl114> (2006).
20. Wang, J. C., Chu, P. Y., Chen, C. M. & Lin, J. H. idTarget: a web server for identifying protein targets of small chemical molecules with robust scoring functions and a divide-and-conquer docking approach. *Nucleic Acids Res* **40**, W393–399, <https://doi.org/10.1093/nar/gks496> (2012).
21. Luo, H. *et al.* DPDR-CPI, a server that predicts Drug Positioning and Drug Repositioning via Chemical-Protein Interactome. *Sci Rep* **6**, 35996, <https://doi.org/10.1038/srep35996> (2016).
22. Labbe, C. M. *et al.* AMMOS2: a web server for protein-ligand-water complexes refinement via molecular mechanics. *Nucleic Acids Res*, <https://doi.org/10.1093/nar/gkx397> (2017).
23. Xie, X. Q. & Chen, J. Z. Data mining a small molecule drug screening representative subset from NIH PubChem. *J Chem Inf Model* **48**, 465–475, <https://doi.org/10.1021/ci700193u> (2008).
24. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling* **28**, 31–36, <https://doi.org/10.1021/ci00057a005> (1988).
25. Tsoumakas, G. & Katakis, I. Multi-label classification: An overview. *Dept. of Informatics, Aristotle University of Thessaloniki, Greece* (2006).
26. Zhang, M.-L. & Zhou, Z.-H. A review on multi-label learning algorithms. *Knowledge and Data Engineering, IEEE Transactions on* **26**, 1819–1837 (2014).
27. Afzal, A. M., Mussa, H. Y., Turner, R. E., Bender, A. & Glen, R. C. A multi-label approach to target prediction taking ligand promiscuity into account. *Journal of Cheminformatics* **7**, 24, <https://doi.org/10.1186/s13321-015-0071-9> (2015).
28. Gonen, M. & Margolin, A. A. Drug susceptibility prediction against a panel of drugs using kernelized Bayesian multitask learning. *Bioinformatics* **30**, i556–563, <https://doi.org/10.1093/bioinformatics/btu464> (2014).
29. Heider, D., Senge, R., Cheng, W. & Hullermeier, E. Multilabel classification for exploiting cross-resistance information in HIV-1 drug resistance prediction. *Bioinformatics* **29**, 1946–1952, <https://doi.org/10.1093/bioinformatics/btt331> (2013).
30. Michielan, L., Terloth, L., Gasteiger, J. & Moro, S. Comparison of multilabel and single-label classification applied to the prediction of the isoform specificity of cytochrome p450 substrates. *J Chem Inf Model* **49**, 2588–2605, <https://doi.org/10.1021/ci900299a> (2009).
31. Wang, X., Zhang, W., Zhang, Q. & Li, G. Z. MultiP-SChlo: multi-label protein subchloroplast localization prediction with Chou's pseudo amino acid composition and a novel multi-label classifier. *Bioinformatics* **31**, 2639–2645, <https://doi.org/10.1093/bioinformatics/btv212> (2015).
32. Soufan, O., Klefogiannis, D., Kalnis, P. & Bajic, V. B. DWFS: a wrapper feature selection tool based on a parallel genetic algorithm. *PLoS One* **10**, e0117988, <https://doi.org/10.1371/journal.pone.0117988> (2015).
33. Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *The Journal of Machine Learning Research* **3**, 1157–1182 (2003).
34. Peng, H., Long, F. & Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence* **27**, 1226–1238 (2005).
35. Alshahrani, M., Soufan, O., Magana-Mora, A. & Bajic, V. B. DANNP: an efficient artificial neural network pruning tool. *PeerJ Computer Science* **3**, <https://doi.org/10.7717/peerj-cs.137> (2017).
36. Eklund, M., Norinder, U., Boyer, S. & Carlsson, L. Choosing Feature Selection and Learning Algorithms in QSAR. *Journal of Chemical Information and Modeling* **54**, 837–843, <https://doi.org/10.1021/ci400573c> (2014).
37. Ba-Alawi, W., Soufan, O., Essack, M., Kalnis, P. & Bajic, V. B. DASPfind: new efficient method to predict drug-target interactions. *Journal of Cheminformatics* **8**, 15 (2016).
38. Magana-Mora, A. & Bajic, V. B. OmniGA: Optimized Omnivariate Decision Trees for Generalizable Classification Models. *Scientific Reports* **7**, <https://doi.org/10.1038/s41598-017-04281-9> (2017).
39. Soufan, O. *et al.* Mining Chemical Activity Status from High-Throughput Screening Assays. *PLoS One* **10**, e0144426, <https://doi.org/10.1371/journal.pone.0144426> (2015).
40. Tang, Y. & Zhang, Y. Q. Granular SVM with repetitive undersampling for highly imbalanced protein homology prediction. *IEEE International Conference on Granular Computing* (2006).
41. Tang, Y., Zhang, Y. Q., Chawla, N. V. & Krasser, S. SVMs modeling for highly imbalanced classification. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*. **39**, 281–288 (2009).
42. Barua, S., Islam, M. M., Yao, X. & Murase, K. MWMOTE—majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Transactions on Knowledge and Data Engineering* **26**, 405–425 (2014).
43. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* **16**, 321–357 (2002).

44. Soufan, O. *et al.* DRABAL: novel method to mine large high-throughput screening assays using Bayesian active learning. *Journal of Cheminformatics* **8**, 64 (2016).
45. NCBI, A. *PubChem BioAssay Database AID 485313*.
46. NCBI, A. *PubChem BioAssay Database AID 485297*.
47. Ramadoss, P., Marcus, C. & Perdew, G. H. Role of the aryl hydrocarbon receptor in drug metabolism. *Expert Opin Drug Metab Toxicol* **1**, 9–21, <https://doi.org/10.1517/17425255.1.1.9> (2005).
48. Heinrich, M. *et al.* Cathepsin D links TNF-induced acid sphingomyelinase to Bid-mediated caspase-9 and -3 activation. *Cell Death Differ* **11**, 550–563 (2004).
49. Moles, A. *et al.* Acidic sphingomyelinase controls hepatic stellate cell activation and *in vivo* liver fibrogenesis. *Am. J. Pathol* **177**, 1214–1224 (2010).
50. Cochran, R. Thiabendazole: Risk Characterization Document. (2001).
51. PubChem. *PubChem Substructure Fingerprint* (2009).
52. O'Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **3**, 33, <https://doi.org/10.1186/1758-2946-3-33> (2011).
53. Landrum, G. RDKit: Open-source Cheminformatics. **3**, 2012 (2006).
54. Soufan, O. M. *Novel Data Mining Methods for Virtual Screening of Biological Active Chemical Compounds* PhD thesis, King Abdullah University of Science and Technology, (2016).
55. Brown, G., Pocock, A., Zhao, M.-J. & Luján, M. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *Journal of Machine Learning Research* **13**, 27–66 (2012).
56. Fleuret, F. Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research* **5**, 1531–1555 (2004).
57. Kononenko, I., Šimec, E. & Robnik-Šikonja, M. Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF. *Applied Intelligence* **7**, 39–55, <https://doi.org/10.1023/A:1008280620621> (1997).
58. Boser, B. E., Guyon, I. M., & Vapnik, V. N. A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory (pp. 144–152). ACM, (1992).
59. Cover, T. M. & Hart, P. E. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on* **13**, 21–27 (1967).
60. Quinlan, J. R. Induction of decision trees. *Machine learning* **1**, 81–106 (1986).
61. Mitchell, T. M. *Machine learning*. 1997. *Burr Ridge, IL: McGraw Hill* **45**, 870–877 (1997).
62. Breiman, L. Random forests. *Machine learning* **45**, 5–32 (2001).
63. Freund, Y. & Schapire, R. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences* **55**, 119–139 (1997).
64. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* **12**, 2825–2830 (2011).
65. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521** (2015).
66. Fernandez-Delgado, M., Cernadas, E. & Barro, S. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research* **15**, 3133–3781 (2014).
67. Bache, K. & Lichman, M. UCI Machine Learning Repository. *Irvine, CA: University of California, School of Information and Computer Science*. (2013).
68. Hattori, M., Okuno, Y., Goto, S. & Kanehisa, M. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J Am Chem Soc* **125**, 11853–11865 (2003).
69. Smith, T. F. & Waterman, M. Identification of common molecular subsequences. *J Mol Biol* **147**, 195–197 (1981).
70. Li, Q., Wang, Y. & Bryant, S. H. A novel method for mining highly imbalanced high-throughput screening data in PubChem. *Bioinformatics* **25**, 3310–3316, <https://doi.org/10.1093/bioinformatics/btp589> (2009).

## Acknowledgements

This work has been supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. FCC/1/1976-02-01, and KAUST Base Research Fund (BAS/1/1606-01-01) to VBB. This research made use of the resources of CBRC at KAUST, Thuwal, Saudi Arabia.

## Author Contributions

O.S., M.E. and V.B.B. conceptualized and designed the study. O.S. performed the experiments. O.S., W.B.A. and A.M.M. analyzed the data. W.B.A., A.M.M., M.E. and V.B.B. contributed to the discussion. O.S. and A.M.M. wrote the manuscript. W.B.A., M.E. and V.B.B. edited the manuscript.

## Additional Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s41598-018-27495-x>.

**Competing Interests:** Vladimir Bajic, corresponding author of this article is an Editorial Board member of *Scientific Reports*.

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



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018