OXFORD

Briefings in Bioinformatics, 2022, 1–11

https://doi.org/10.1093/bib/bbac337

Problem Solving Protocol

toxCSM: comprehensive prediction of small molecule toxicity profiles

Alex G. C. de Sá, Yangyang Long, Stephanie Portelli, Douglas E. V. Pires and David B. Ascher

Corresponding authors: David B. Ascher. Tel.: +61-7-336-53891; E-mail: d.ascher@uq.edu.au; Alex G. C. de Sá. E-mail: alex.desa@baker.edu.au; Douglas E. V. Pires. E-mail: douglas.pires@unimelb.edu.au

Abstract

Drug discovery is a lengthy, costly and high-risk endeavour that is further convoluted by high attrition rates in later development stages. Toxicity has been one of the main causes of failure during clinical trials, increasing drug development time and costs. To facilitate early identification and optimisation of toxicity profiles, several computational tools emerged aiming at improving success rates by timely pre-screening drug candidates. Despite these efforts, there is an increasing demand for platforms capable of assessing both environmental as well as human-based toxicity properties at large scale. Here, we present toxCSM, a comprehensive computational platform for the study and optimisation of toxicity profiles of small molecules. toxCSM leverages on the well-established concepts of graph-based signatures, molecular descriptors and similarity scores to develop 36 models for predicting a range of toxicity properties, which can assist in developing safer drugs and agrochemicals. toxCSM achieved an Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) of up to 0.99 and Pearson's correlation coefficients of up to 0.94 on 10-fold cross-validation, with comparable performance on blind test sets, outperforming all alternative methods. toxCSM is freely available as a user-friendly web server and API at http://biosig.lab.uq.edu.au/toxcsm.

Keywords: graph-based signatures, machine learning, toxCSM, toxicity, toxicity predictions

Introduction

Drug discovery is a costly, time-consuming and uncertain endeavour [1–6], with 80–90% of projects discontinued before even getting tested in humans [6] and almost 95% of drugs entering human trials failing [3, 6, 7].

Although optimising Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) parameters are important during different stages of drug development [8, 9], the assessment of toxicity remains a limiting and crucial step. In fact, the therapeutic utility of a drug is a fine balance between compound efficacy and toxicity [10]. Poor toxicity profiles have been one of the main causes of attrition during pre-clinical and clinical trials, where over 40% of novel drugs fail human clinical trials due to unanticipated human toxicity [11]. Consequently, Van Norman [11] estimated that success rates in clinical trials would increase by approximately 44% if toxicity failures could be mitigated or minimised.

In vivo and *in vitro* screening techniques usually assist in the identification of toxicity during drug development stages. Although helpful, they tend to be expensive, inefficient and time consuming [12–16]. This has driven the emergence of computational approaches as a promising strategy for pre-screening and prioritisation of the large number of potential compounds being investigated via high throughput screening [14, 15, 17–26]. Nevertheless, current methods consider only a limited subset of toxicity endpoints [15, 17, 18, 21–23, 26], missing key potential toxicity properties, with many methods presenting limited accessibility and scalability, hindering their practical utility.

Here, we present toxCSM, a comprehensive and accurate computational platform for the study and optimisation of toxicity profiles of small molecules. toxCSM leverages on the well-established concepts of graph-based signatures, general molecular descriptors and similarity scores to create a web-based machine learning platform, which is composed of 36 models for predicting a wide range of toxicity properties, from nuclear and stress responses to environmental toxicity, which can assist in the development of safer and less toxic drugs as well as herbicides and pesticides.

Alex G. C. de Sá is a postdoctoral researcher at the Computational Biology and Clinical Informatics department at the Baker Institute, with honorary fellowships at the University of Queensland and Systems and Computational Biology at Bio21 Institute. He holds a PhD in computer science, and his research is centred in automated machine learning (AutoML), i.e. on how to select machine learning algorithms for a data set of interest.

Yangyang Long holds an MSc in computer science from the School of Computing and Information Systems at the University of Melbourne. During her MSc, she focused on assessing different approaches for *in silico* small molecule toxicity prediction.

Stephanie Portelli is a postdoctoral researcher at the University Queensland, with honorary fellowships in Computational Biology and Clinical Informatics at the Baker Institute and Systems and Computational Biology at Bio21 Institute. She researches mainly on computational biology and applies machine learning techniques to drug resistance and genetic disease problems.

Douglas E. V. Pires is a senior lecturer in digital health at the School of Computing and Information Systems at the University of Melbourne and group leader at Bio21 Institute. He is a computer scientist and bioinformatician specialising in machine learning and AI and the development of the next generation of tools to analyse omics data, and guide personalised medicine.

David B. Ascher is the deputy director of biotechnology at the University of Queensland and head of Computational Biology and Clinical Informatics at the Baker Institute and Systems and Computational Biology at Bio21 Institute. He is interested in developing and applying computational tools to assist leveraging clinical and omics data for drug discovery and personalised medicine.

Received: April 19, 2022. **Revised:** July 17, 2022. **Accepted:** July 23, 2022

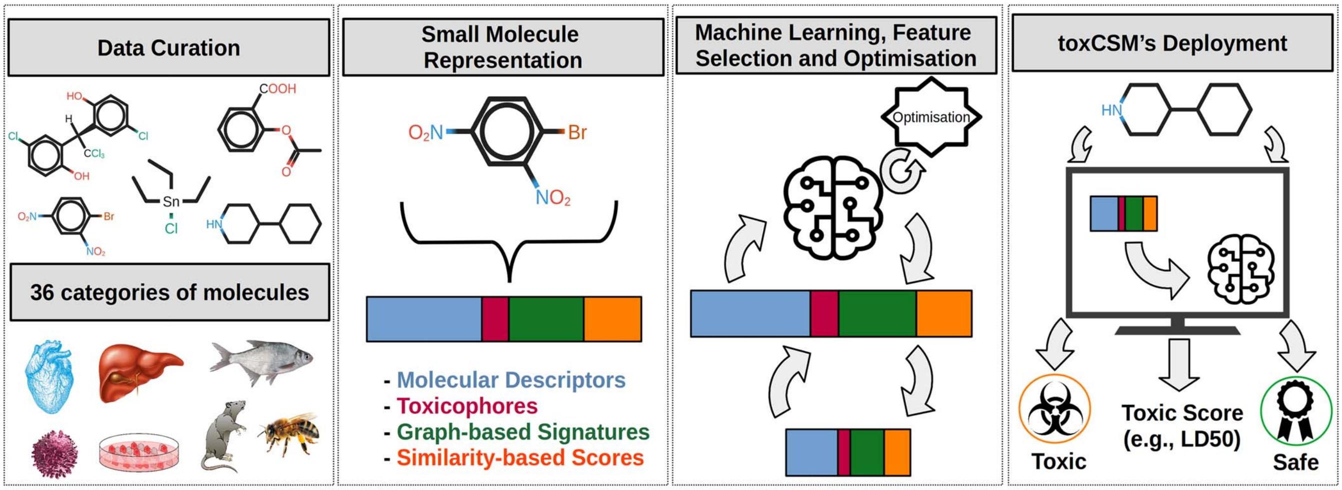


Figure 1. The toxCSM workflow. The toxCSM approach can be subdivided into four main steps. (1) In Data Curation, we have collected experimentally characterised toxicity profiles of small molecules from general toxicity databases and also looked for data on specific toxicity endpoints in the literature, resulting in 36 experimental endpoint data sets. (2) Next, in Small Molecule Representation, feature engineering is performed and collected molecules are represented as feature matrices, which are used as evidence to train machine learning models. Molecules are represented by four types of features, which vary from general molecular properties to graph-based signatures and similarity scores. (3) In Machine Learning, Feature Selection and Optimisation, features are selected in a stepwise way, having a machine learning model as an engine to evaluate the feature space. Machine learning models are also optimised with Bayesian Optimisation and Transfer Learning. Best performing models for each endpoint are then selected based on a trade-off between complexity and predictive performance (i.e. in terms of Matthew's and Pearson's Correlation Coefficient for classification and regression tasks, respectively) and validated using 5-, 10- and 20-fold cross-validation, and blind test sets. (4) Finally, in toxCSM's Deployment, predictive models are made available as a free, scalable and easy-to-use web server and API.

Material and methods

The toxCSM methodological workflow depicted in Figure 1 is composed of four main steps: (1) Data Curation; (2) Small Molecule Representation; (3) Machine Learning, Feature Selection and Optimisation; (4) toxCSM Web Server Deployment. These steps will be detailed in the following subsections. To better understand toxCSM, Figure S1 presents a flow chart, incorporating toxCSM's bio(chem)informatics workflow.

Data curation

Collecting small molecule endpoints

Experimental data for 36 distinct toxicity endpoints were collected from the literature (Tables S1 and S2), which fall into six broad mechanistic categories: (i) nuclear response, (ii) stress response, (iii) genomic, (iv) environmental, (v) human dose response and (vi) organic (directly related to organs). These endpoints are typically used in small molecule research and have been benchmarks of computational methods [14, 15, 17–23] or involved in specific *in silico* toxicity evaluations of independent approaches [27–31]. Table S1 demonstrates that toxCSM includes the largest set of endpoints in its evaluation to date.

In Table S2, each toxicity endpoint is further subdivided based on the nature of the experimental data available, into categorical or continuous target variables, which are associated with classification or regression problems, respectively. In total, these endpoints consist of 43,236 unique compounds represented as SMILES strings, after filtering valid molecules using the RDkit cheminformatics toolkit [32].

Before splitting each toxicity endpoint into training and blind test sets, we performed an analysis of the molecular (sub)structures of each classification endpoint to verify the reliability of using structural clustering information on splitting. Nevertheless, clustering molecules using the Butina clustering algorithm [33] by Tanimoto similarity [34] on Morgan fingerprints [35] revealed that even structurally very similar molecules could have opposing toxicity effects (see Table S3 and Supplementary Material and Methods), showing that clustering is not always an appropriate alternative to be incorporated into splitting schemes across machine learning pipelines on molecular data. If the molecules have diverging opposing toxicity effects, they should not be grouped together. This means the outputs from the clustering model are not entirely correct, and their inclusion to split the data can induce a training and evaluation bias on toxCSM models. Therefore, we do not rely on clustering information to develop, evaluate and validate the toxCSM models. Given the high level of imbalance found in the endpoint data sets, clustering molecules based on structural similarity to derive training and blind test sets could also further skew data sets and introduce more biases. By not considering clustering on data splitting (e.g. cross-validation, training and blind testing) schemes, we intend to provide the toxCSM machine learning models with the ability of better distinguishing toxic molecules (commonly rare in data sets) from those considered safe (commonly more frequent), consequently yielding to better generalisation performance.

Thereafter, data within each toxicity endpoint was divided into training (90%/95%—utilised for internal validation via stratified 10-fold cross-validation) and independent blind test (10%/5%) sets for assessing predictive performance and generalisation of toxCSM models. The sizes of both training and blind test sets were chosen according to data availability, but also to guarantee fair comparison to alternative methods. For categorical endpoints, splitting occurred in a stratified manner to ensure a similar distribution between training and test sets. For continuous endpoints, splitting into training and blind test was performed randomly, with a subsequent analysis to ensure comparable distributions. Data sets are available at http://biosig.lab.uq.edu.au/toxcsm/data.

Molecular substructure mining

To guide characterisation of what structural aspects contribute to toxicity, the molecular substructure miner (MoSS) algorithm [36] was used to identify substructures that were enriched in toxic molecules. MoSS aims to find substructures that can distinguish a target from a complement group. In the case of toxicity, the target groups would be those containing the toxic molecules, whereas the complement groups would be linked to safe molecules. Different minimum and maximum percentage thresholds were tested for the toxic and safe molecules, respectively. Relatively, we aimed to find substructures appearing in the toxic class at least five or 10 times more than in the safe class. This analysis can be found in the Supplementary Materials. It is worth noting that the substructures discovered by MoSS (Tables S4 and S5) were used to guide the development and interpretation of toxCSM's models.

Small molecule representation

Representing small molecules with the objective of using them as inputs to machine learning models is considered a challenging task and critical decision step. Several approaches have been proposed in the literature to properly characterise the complexity across distinct compounds [22, 37–39], including fingerprints, graph-based representations and deep learning-based features. To build toxCSM models, small molecules were represented in this work by four classes of interpretable properties (Tables S6 and S7), which have been successfully applied in other types of small molecule prediction tasks [40–50]. We presented an overview of these features in Figure S2 and summarised all their four classes next.

General molecular descriptors

Over 150 molecular descriptors quantitatively describing molecular structures, including physicochemical properties, molecular surface and functional groups, were calculated using the RDkit cheminformatics toolkit.

Toxicophores

Toxicophores are molecular substructures that are commonly linked to toxicity [51]. Their bit-vector representations were used to identify the presence or absence of 36 experimentally validated and statistically tested toxicophores [52] within each molecule. Other toxicophores, such as the ones derived in FAF-Drugs4 [51], have not been included in the molecular representation as they have not been validated and statistically tested as well as the toxicophores employed in our representation.

Graph-based signatures

toxCSM's graph-based signatures were adapted from pkCSM [22]. These signatures are generated from molecule graphs, where atoms are regarded as vertices, while covalent bonds are set as edges. Within the toxCSM graph-based signatures, atoms were labelled by their respective molecular properties (pharmacophores), including as aromatic, hydrophobic, acceptor, donor, positive/negative ionisable. Distance patterns identified between pharmacophoric atom pairs are summarised as cumulative distributions of distances and represented as a feature vector. These signatures have been demonstrated to be a robust approach to represent geometry and physicochemical properties of macromolecules and small molecules and, as a consequence, they have been successfully employed in a range of predictive tasks [40–50].

Similarity-based scores

toxCSM also includes a set of similarity-based features. These features are based on Tanimoto similarity scores [34] across Morgan/Circular fingerprints [35] between a given input molecule and known toxic molecules, which are found in the training set of each endpoint. For continuous endpoints, similarity was only considered to the 20th percentile of most toxic compounds. The main purpose of these features is to define reference molecules for toxicity. In toxCSM's endpoints, the number of similarity scores, which were defined as features, varies from 51 (for the Carcinogenesis endpoint) to 4,167 (for the AMES Mutagenesis endpoint).

Machine learning, feature selection and optimisation

In this study, we evaluated ten distinct machine learning algorithms using a 10-fold cross-validation procedure on all endpoints [53], including Random Forest, Extremely Randomised Trees, Gradient Boosting, Extreme Gradient Boosting, Adaptive Boosting, Support Vector Machines, Gaussian Processes, Multilayer Perceptron for Artificial Neural Networks, K-Nearest Neighbours and Decision Trees, using the implementations available on the Scikit-Learn library [54]. Tables S8 and S9 define the initial attempts on default and non-default hyper-parameters used for training toxCSM's models.

To better explore and optimise the hyper-parameters on toxCSM's models for each endpoint data, we also used Hyperopt-Sklearn [55] based on Matthew's and Pearson's correlation coefficients for classification and regression problems, respectively. Hyperopt-Sklearn relies on a Bayesian optimisation method named as Tree of Parzen Estimator (TPE) [55, 56]. Briefly, in its first step, TPE works by randomly sampling from the search space of hyper-parameter configurations. This initialisation step is performed to initialise two density distributions (one fitting good hyper-parameters and the other maintaining bad hyper-parameter configurations), which are used to guide a Parzen estimator to hierarchically search for the next algorithm hyper-parameter configuration. TPE updates the distributions given this configuration performance and continues with this iterative process until the defined stopping criterion is satisfied. For each endpoint, we optimised the hyper-parameters for 2,500 iterations, restricting the runtime of each machine learning model to 10 minutes. The hyper-parameter space for each employed machine learning algorithm on Scikit-Learn can be found at https://github.com/hyperopt/hyperopt-sklearn.

Besides using hyper-parameter optimisation to avoid overfitting and improve predictive performance, a bottom-up greedy feature selection method was employed with the same two aims and also to ensure low complexity of the produced machine learning models for toxCSM. This feature selection method works as follows. Greedy feature selection starts with zero features in the feature set and adds one feature at time across its iterative process. To include one feature in the feature set, this method evaluates all features (except those already selected) using a 10-fold cross-validation procedure on a machine learning algorithm. The evaluation of each feature relies on Matthew's and Pearson's correlation coefficients for classification and regression problems, respectively. These evaluation coefficients are detailed in Supplementary Materials. After this step, the best performing feature is combined with the current set. At the end, best performing models in terms of greedy feature selection were also chosen based on Matthew's and Pearson's correlation coefficients for classification and regression problems, respectively. It is worth noting that Matthew's correlation coefficient was chosen as it helps to select models which are more resilient to class imbalance.

Feature selection was guided by transfer learning [57], where the knowledge acquired on the selected features by one machine learning algorithm is transferred to the others. To do this, all sets of features selected with one machine learning algorithm were tested (in order) by all other algorithms. Next, an analysis is made to decide the algorithm and the set of features to be used. If the performance on 10-fold cross-validation of a particular configuration (i.e. set of features and learning algorithm) is better than the previous configuration, the current configuration is kept. This process is done to ensure the best model for each endpoint in terms of predictive performance and complexity is chosen.

The best models, which were selected for toxCSM for each endpoint, are shown in Table S10. Supplementary Material provides a section analysing the selected features in terms of the number of features, percentual of each type of features and top 10 most important features, which are described in Tables S11–S14. This analysis can be found in the Supplementary Results. Given hyper-parameter optimisation and feature selection approaches, the predictive results of each one of the 10 machine learning algorithms across the 10-fold cross-validation procedure and blind test assessment (in terms of Matthew's and Pearson's correlation coefficient for classification and regression endpoints, respectively) are shown in Tables S15–S18.

toxCSM web server deployment

To provide a robust, scalable, reliable and easy-to-use toxicity prediction platform, toxCSM's web server front end was developed via Bootstrap framework version 3.3.7, and the back end was built in Python 2.7 with the use of the Flask framework (version 0.12.3). The web server is hosted on a Linux server running Apache. The most important components of toxCSM are depicted in Figure 2: landing page (Figure 2**A**), prediction page (Figure 2**B**) and tabular results page (Figure 2**C**). Furthermore, the complementary components of toxCSM are depicted in Figures S3–S8.

Input

toxCSM can be used to evaluate small molecule toxicity profiles with four types of inputs (Figure 2**B**): (1) an SDF (Structural Data File) covering a list of molecular structures (up to 1000 molecules); (2) a SMILES file containing a list of compounds (up to 1000 molecules); (3) a single SMILES (Simplified Molecular Input Line Entry System) string; (4) a single molecular drawing, where its respective SMILES is retrieved from the drawn molecule structure. We utilised these formats (SDF and SMILES) as they are standard ways to represent chemical compounds. Examples of formatted files and a help page to assist users can be found at http://biosig.lab.uq.edu.au/toxcsm/prediction and http://biosig.lab.uq.edu.au/toxcsm/help, respectively. In addition, users can choose to receive prediction results via email (Figure 2**B**/5). Finally, users can select which toxicity endpoints they would like to get predictions for their molecules (Figure 2**B**/6). These have been categorised into nuclear response, stress response, human dose response, organic, environmental or genomic toxicities. It is also possible for users to run all the prediction modes at once or an example.

Output

Toxicity predictions for the input molecules are presented in tabular format (Figure 2**C**), which can be downloaded as a comma-separated-values (csv) file (Figure 2**C**/8). These involve either numerical or categorical values accompanied by an interpretation of the outcome and by the confidence/probability scores from the predictions (if available). Users can further interpret the predictions of each molecule by clicking on a 'View Details' button (Figure 2**C**/7). This leads to an informative analysis page that presents besides the molecule depiction and SMILES, a comprehensive list of physicochemical properties, drug-likeness properties and adherence to druggability rules (within tables and radar plots) and presence of toxicophores. Figures S1–S6 illustrate the analysis page for toxCSM.

Application programming interface (API)

An Application Programming Interface (API) to assist users to seamlessly integrate our predictive tool into their cheminformatics analytical pipelines is also available. Input fields follow the same format previously described for our web server implementation (i.e. toxCSM's API supports as inputs an SDF, a SMILE file or a SMILE string). All jobs submitted are labelled with a unique identifier, which is used to query the status of the job. Results are outputted with these identifiers as a JavaScript object notation (JSON) standard tabular file. A complete description of toxCSM's API, which includes tutorial examples in both curl and python scripting languages, can be found at http://biosig.lab.uq.edu.au/toxcsm/api_docs.

Results

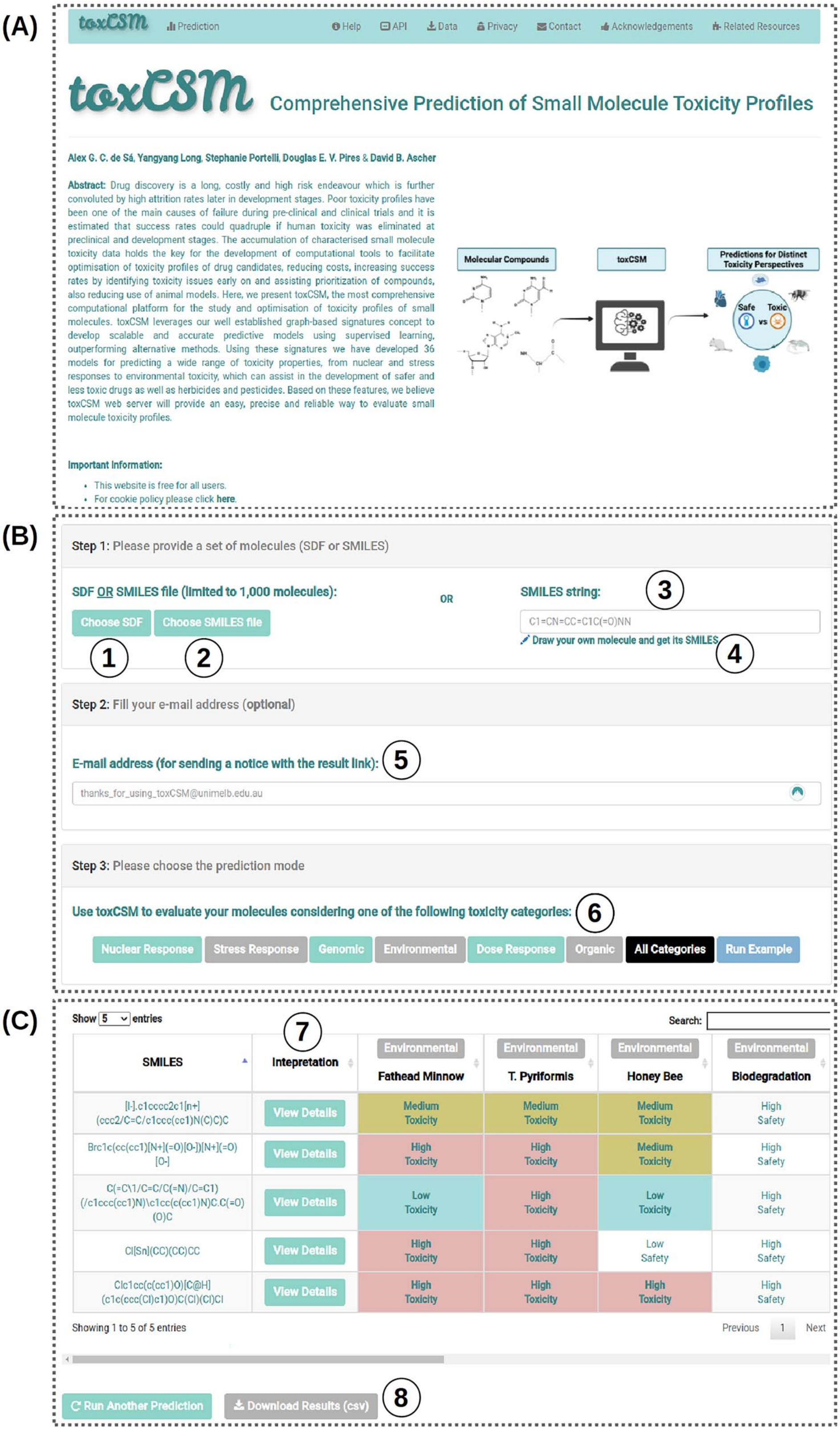
The comprehensive collection of predictive models for assessing human and environmental toxicity of small molecules in toxCSM was assessed using internal and external validation procedures. In this section, we have also contrasted performance of toxCSM models with alternative methods in respect of the Area Under the ROC Curve (AUC) and Coefficient of Determination (R^2) for classification and regression endpoints, respectively. Both metrics are commonly used for toxicity prediction analysis and are properly defined in the Supplementary Materials. Finally, we have investigated molecular determinants of toxicity.

Molecular substructural determinants of toxicity

In this section, we investigate the overall properties of toxic molecules in terms of enriched substructural patterns, which were identified by MoSS (Tables S4 and S5). When comparing the substructures that were majorly found in toxic molecules for environmental endpoints to those found in human-based endpoints (Table S2), we observed that the substructures were strikingly more enriched in the former for phosphate, phosphorothioate or other, sulphur-, phosphorus- and oxygen atom combinations. On the other hand, those substructures found in toxic classes in human-based endpoints were consistently more enriched in ether, ketone, hydroxyl and nitrogen-enriched substructures. Similarities between the environmental and human-based endpoints include the presence of chloride and nitroso groups. However, within the environmental endpoints, these represented more notable contributions to toxicity. Finally, when comparing substructures enriched across the different endpoints to well-known, experimentally validated and statistically tested toxicophores [52], we have observed that most toxicophores have been identified by MoSS in the curated endpoints, and our results are congruent with known toxic compounds, highlighting their importance for predictive purposes. This was particularly the case for genomic endpoints, where specific examples include the sulfonic acid moiety that was enriched in the carcinogenesis endpoint, and nitrosobenzene, which was mainly observed across the AMES and micronucleus genomic endpoints. This is one of the main reasons why these toxicophores were included into the feature space representing the endpoint molecules.

Performance assessment on internal and external validation

The predictive performances of the final toxCSM's classification and regression models for the 36 distinct toxicity endpoints across



toxCSM Comprehensive Prediction of Small Molecule Toxicity Profiles

Alex G. C. de Sá, Yangyang Long, Stephanie Portelli, Douglas E. V. Pires & David B. Ascher

Abstract: Drug discovery is a long, costly and high risk endeavour which is further convoluted by high attrition rates later in development stages. Poor toxicity profiles have been one of the main causes of failure during pre-clinical and clinical trials and it is estimated that success rates could quadruple if human toxicity was eliminated at preclinical and development stages. The accumulation of characterised small molecule toxicity data holds the key for the development of computational tools to facilitate optimisation of toxicity profiles of drug candidates, reducing costs, increasing success rates by identifying toxicity issues early on and assisting prioritization of compounds, also reducing use of animal models. Here, we present toxCSM, the most comprehensive computational platform for the study and optimisation of toxicity profiles of small molecules. toxCSM leverages our well established graph-based signatures concept to develop scalable and accurate predictive models using supervised learning, outperforming alternative methods. Using these signatures we have developed 36 models for predicting a wide range of toxicity properties, from nuclear and stress responses to environmental toxicity, which can assist in the development of safer and less toxic drugs as well as herbicides and pesticides. Based on these features, we believe toxCSM web server will provide an easy, precise and reliable way to evaluate small molecule toxicity profiles.

Important Information:

- This website is free for all users.
- For cookie policy please click here.

Step 1: Please provide a set of molecules (SDF or SMILES)

SDF OR SMILES file (limited to 1,000 molecules):

OR

SMILES string: C1=CN=CC=C1C(=O)NN

Draw your own molecule and get its SMILES

Step 2: Fill your e-mail address (optional)

E-mail address (for sending a notice with the result link): thanks_for_using_toxCSM@unimelb.edu.au

Step 3: Please choose the prediction mode

Use toxCSM to evaluate your molecules considering one of the following toxicity categories:

Show 5 entries

Search:

SMILES	Intepretation	Environmental Fathead Minnow	Environmental T. Pyriformis	Environmental Honey Bee	Environmental Biodegradation
[I-].c1cccc2c1[n+](ccc2/C=C/c1ccc(cc1)N(C)C)C	View Details	Medium Toxicity	Medium Toxicity	Medium Toxicity	High Safety
Brc1c(cc(cc1)[N+](=O)[O-])[N+](=O)[O-]	View Details	High Toxicity	High Toxicity	Medium Toxicity	High Safety
C(=C\1/C=C/C(=N)/C=C1)(/c1ccc(cc1)N)\c1cc(c(cc1)N)C.C(=O)(O)C	View Details	Low Toxicity	High Toxicity	Low Toxicity	High Safety
Cl[Sn](CC)(CC)CC	View Details	High Toxicity	High Toxicity	Low Safety	High Safety
Clc1cc(c(cc1)O)[C@H](c1c(ccc(Cl)c1)O)C(Cl)(Cl)Cl	View Details	High Toxicity	High Toxicity	High Toxicity	High Safety

Showing 1 to 5 of 5 entries

Figure 2. toxCSM web server. (**A**) presents the landing page for toxCSM. By clicking on ‘Prediction’ at the top menu, users are directed to the job submission page (**B**). Users have four input options to provide their molecules: (1) SDF; (2) SMILES file; (3) a single SMILES string; (4) molecular drawing. Users can opt in (5) to include their respective email addresses so toxCSM’s web server can send a link with the prediction results. Given the input molecules, users can choose among six prediction models, run all of them at once or run an example (6). As a result, the predictions in terms of toxicity properties are shown in (**C**) by using different levels of toxicity and safety (i.e. low, medium and high) for each input molecule (represented by its respective SMILES). Furthermore, by clicking on ‘View Details’ in the column ‘Interpretation’ (7), toxCSM will show an extra level of details for this molecule. Finally, these results can be downloaded by clicking on the button ‘Download Results (csv)’ (8).

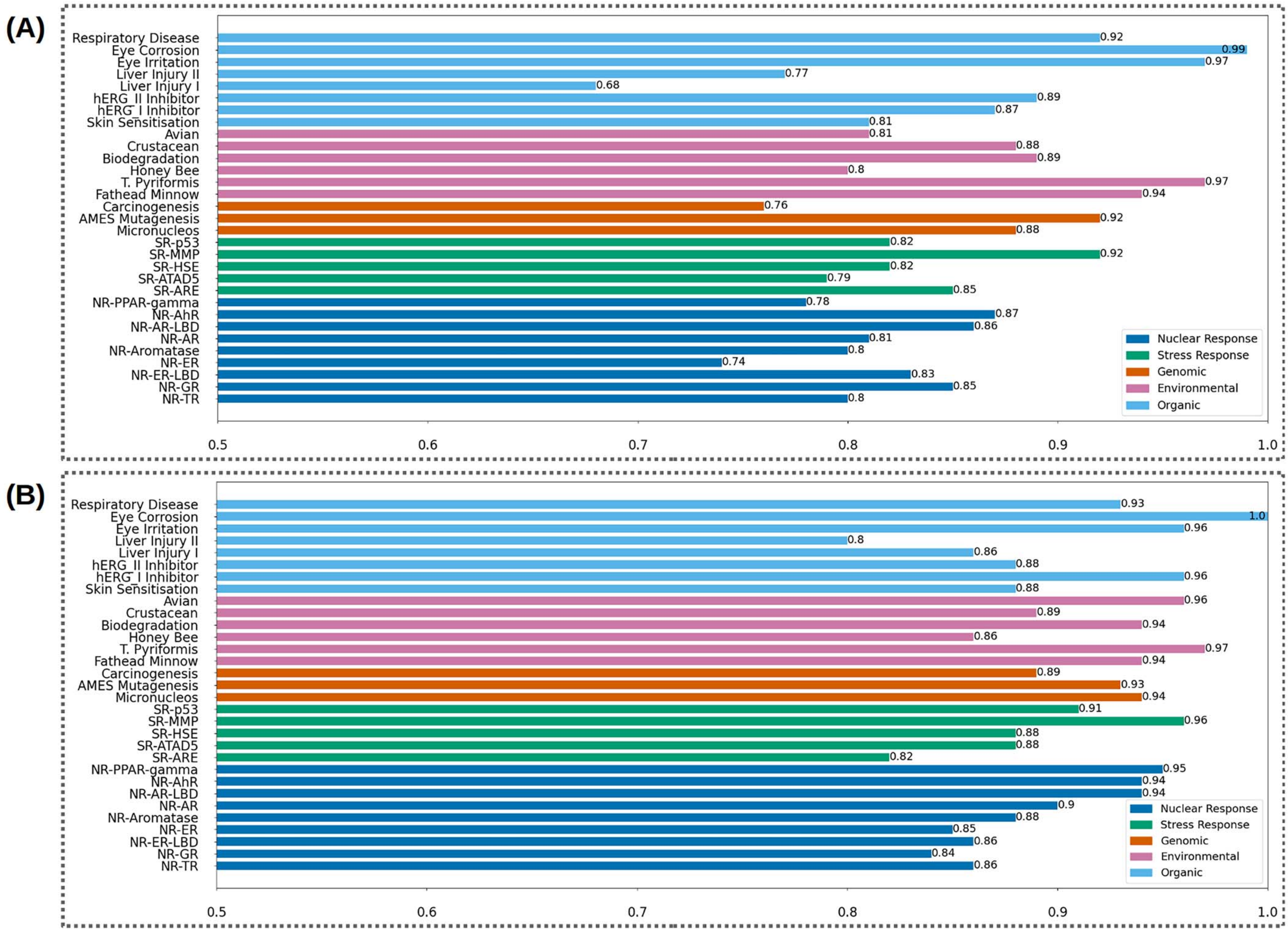


Figure 3. toxCSM's predictive classification performances. The summary of AUC results achieved by toxCSM on 10-fold cross-validation (**A**) and blind test sets (**B**) across the 31 classification endpoints.

internal (using different cross-validation schemes—5-fold, 10-fold and 20-fold) and external (blind test sets) evaluation procedures are shown in Figures 3 and 4, Figures S9–S16 and S18–S22 and Tables S19–S28. These results have been calculated and generated after feature selection, hyper-parameter optimisation and transfer learning on all toxicity endpoint models present in toxCSM.

We built classification models for 31 distinct categorical toxicity endpoints (Table S2, including six environmental-based and 25 human-related toxicity endpoints). The toxCSM classification models were able to accurately identify compounds that were likely to be toxic, achieving AUCs from 0.68 to 0.99 under 10-fold cross-validation (Figure 3**A**, Figures S9–S16 and Table S20), with comparable performance on repeated 10-fold, and on 5- and 20-fold cross-validation procedures, providing confidence in the trained models (Tables S19, S21 and S22). When the models were evaluated against independent blind tests (i.e. unseen data), the AUC performances ranged from 0.80 to 1.00 (Figure 3**B**, Figures S9–S16 and Table S23), consistent with performance on cross-validation, demonstrating toxCSM's robustness and generalisation capabilities. A similar trend on generalisation of the models was also observed in other classification measures, such as Matthew's Correlation Coefficient (MCC) and Balanced Accuracy (BACC).

Figure 3 summarises toxCSM's classification results, providing its strong AUC performances. These results are supported by Figure S17, which can make us conclude that a correlation between data set size and predicted performance on the blind test set was not observed, even when the endpoint data sets had similar sizes. In other words, the predictive performances achieved by toxCSM tend to be independent of the endpoint data, assuring those different data distributions do not affect the predictive performances of the machine learning models on toxCSM. Finally, when looking at incorrectly predicted compounds across the blind test sets, there was a major enrichment for large values of molecular logP, number of rings and molecular weight properties, which are underrepresented in most data sets, on average.

We have also built five regression models on continuous toxicity endpoints (Table S2, including four environmental and one human toxicity endpoints), capable of quantitatively assessing toxicity levels and ranking compounds accordingly. Under 10-fold cross-validation, models achieved Pearson's, Spearman's and Kendall's correlations of up to 0.94, 0.93 and 0.80, respectively (Figure 4**A**, Figures S18–S22 and Table S25). Pearson's, Spearman's and Kendall's correlation performances increased to up to 0.97, 0.97 and 0.86, respectively, after 10% of outliers were removed. The outliers are defined as the 10% worst predicted data points, which are the most distant molecular samples to the regression line defining the best fitting between actual and predicted toxicity effect values. It is important to emphasise that these outliers were removed only for analysis purposes.

Taking a closer look at outliers revealed that they usually had larger values of molecular logP, number of rotatable bonds and

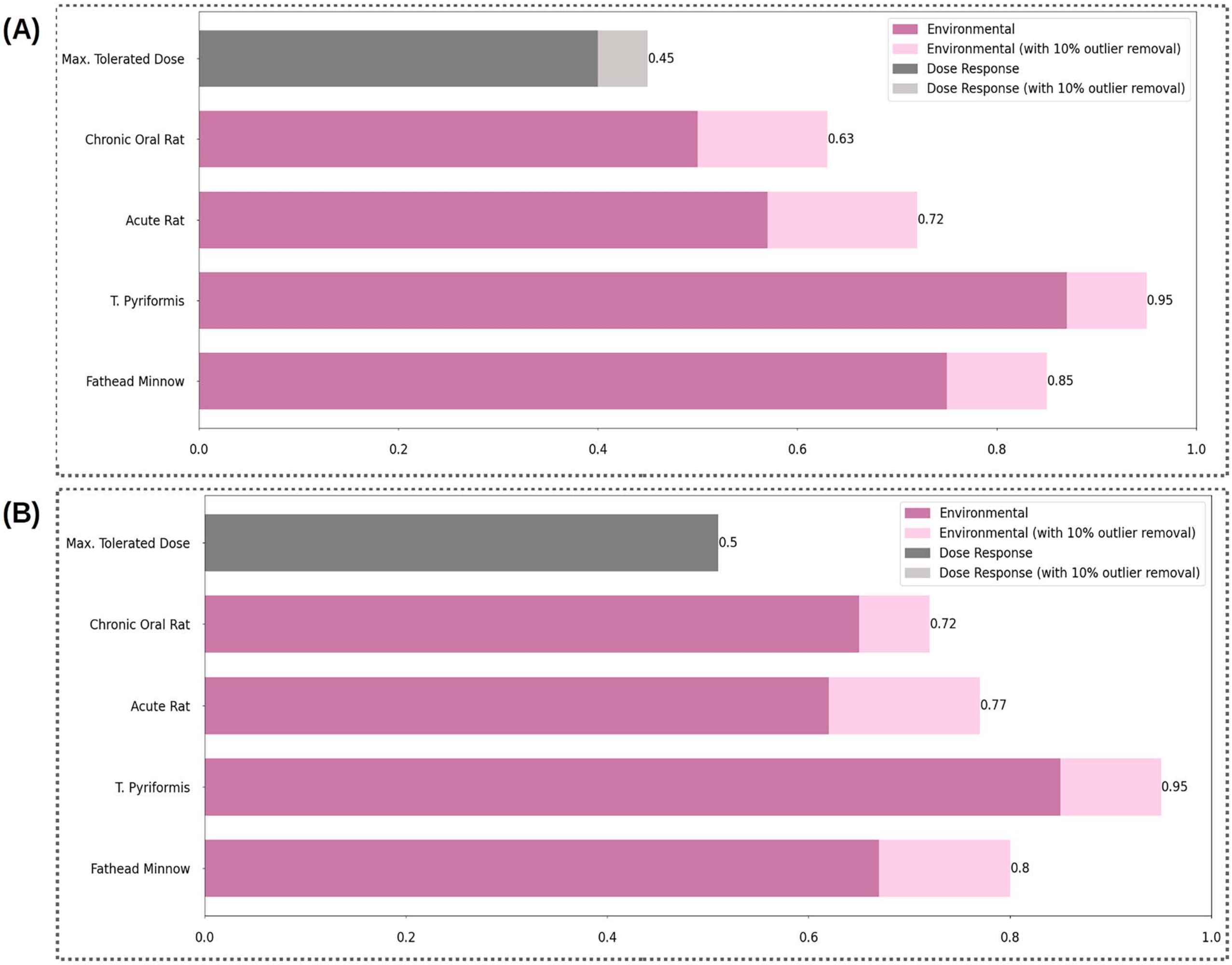


Figure 4. toxCSM's regression performances before outlier removal. The summary of coefficient of determination (R^2) results achieved by toxCSM on 10-fold cross-validation (**A**) and blind test sets (**B**) across the five regression endpoints before and after 10% of outlier removal.

number of rings, consistent with outliers identified for classification tasks, denoting underrepresented molecules. We believe these behaviours are part of the reasons toxCSM's regression models tended to slightly underestimate the prediction of numerical toxicity properties (Figures S18–S22). Given this analysis, we conclude that such features can be used to better understand/interpret toxCSM models in a way to have a closer look at their predictions while the molecules have larger or normal values of molecular logP, number of rotatable bonds and number of rings, for example. Nonetheless, toxCSM demonstrated comparable performance across alternate cross-validation strategies (Tables S24, S26 and S27) and, also, on independent blind tests (Figure 4**B** and Table S28). For instance, before the procedure of removing 10% of the outliers, toxCSM presented Pearson's, Spearman's and Kendall's correlation coefficients up to 0.92, 0.91 and 0.78 on the blind test sets, respectively. After outlier removal, the Pearson's, Spearman's and Kendall's correlation coefficients could reach values up to 0.97, 0.97 and 0.86 on the blind test evaluation, respectively. This analysis shows that, on average, toxCSM regression models will perform exceptionally well while predicting continuous targets. The regression results of toxCSM are also summarised in terms of coefficient of determination in Figure 4, providing clear evidence of the robustness, consistency and generalizability of the models.

Accordingly, given the overall predictive performance achieved by toxCSM models on the different cross-validation schemes and blind test sets, we believe toxCSM reached a suitable the applicability domain for each toxicity endpoint, when applying the models to an unseen set of compounds. Furthermore, the applicability domain of the models can be defined by the wide range of toxicity categories involved in toxCSM (e.g. stress response, nuclear response, environmental, organic, dose response and genomic), making it reliable for distinct types of compounds.

Comparison with alternative methods

In order to put the performance of toxCSM into context, we compared it with six alternative methods available in the literature, including DeepTox [17], ProTox II [20], admetSAR 2.0 [23], ADMETlab 2.0 [15], pkCSM [22] and Interpretable-ADMET [26]. We show toxCSM models perform better than alternative methods for all comparisons across classification and regression endpoints (Tables 1–6 and Tables S29–S32).

Across the classification models, toxCSM was compared to pkCSM, ADMETLab 2.0, DeepTox, Protox II, admetSAR 2.0 and Interpretable-ADMET. As these methods follow different validation procedures and use a different number of endpoints, we limit comparisons of the Wilcoxon signed-rank test [58] to the blind

Table 1. Comparison between toxCSM and ADMETLab 2.0 AUCs across the 21 equivalent classification endpoints; the best results in this comparison are indicated in bold

Endpoint name	**Methods**	
	toxCSM	**ADMETLab 2.0**
NR-AhR	0.939	**0.943**
NR-AR	**0.902**	0.886
NR-AR-LBD	**0.943**	0.915
NR-Aromatase	**0.877**	0.852
NR-ER	**0.851**	0.771
NR-ER-LBD	**0.858**	0.850
NR-PPAR-gamma	**0.946**	0.893
SR-ARE	0.823	**0.863**
SR-ATAD5	**0.882**	0.874
SR-HSE	0.876	**0.907**
SR-MMP	**0.957**	0.927
SR-p53	**0.906**	0.881
AMES Mutagenesis	**0.929**	0.902
Carcinogenesis	**0.886**	0.788
Skin Sensitisation	**0.883**	0.707
hERG_I Inhibitor	**0.957**	0.943
Liver Injury I	**0.864**	0.814
Liver Injury II	0.797	**0.924**
Eye Irritation	0.959	**0.982**
Eye Corrosion	**0.999**	0.983
Respiratory Disease	**0.930**	0.828
Average	**0.903**	0.878

Table 2. Comparison between toxCSM and ProTox II AUCs across 15 equivalent classification endpoints; the best results in this comparison are indicated in bold

Endpoint name	**Methods**	
	toxCSM	**ProTox II**
NR-AhR	**0.939**	0.900
NR-AR	**0.902**	0.730
NR-AR-LBD	**0.943**	0.750
NR-Aromatase	**0.877**	0.750
NR-ER	**0.851**	0.790
NR-ER-LBD	**0.858**	0.800
NR-PPAR-gamma	**0.946**	0.840
SR-ARE	**0.823**	0.790
SR-ATAD5	**0.882**	0.800
SR-HSE	**0.876**	0.870
SR-MMP	**0.957**	0.920
SR-p53	**0.906**	0.870
AMES Mutagenesis	**0.929**	0.900
Carcinogenesis	**0.886**	0.850
Liver Injury I	**0.864**	0.860
Average	**0.896**	0.828

Table 3. Comparison between toxCSM and DeepTox AUCs across the 12 equivalent classification endpoints; the best results in this comparison are indicated in bold

Endpoint name	**Methods**	
	toxCSM	**DeepTox**
NR-AhR	**0.939**	0.928
NR-AR	**0.9026**	0.807
NR-AR-LBD	**0.943**	0.879
NR-Aromatase	**0.877**	0.834
NR-ER	**0.851**	0.810
NR-ER-LBD	**0.858**	0.814
NR-PPAR-gamma	**0.946**	0.861
SR-ARE	0.823	**0.840**
SR-ATAD5	**0.882**	0.793
SR-HSE	**0.876**	0.865
SR-MMP	**0.957**	0.942
SR-p53	**0.906**	0.862
Average	**0.897**	0.853

Table 4. Comparison between toxCSM and Interpretable-ADMET AUCs across the 24 equivalent classification endpoints; the best results in this comparison are indicated in bold

Endpoint name	**Methods**	
	toxCSM	**Interpretable-ADMET**
NR-AhR	**0.939**	0.727
NR-AR	**0.902**	0.730
NR-AR-LBD	**0.943**	0.802
NR-Aromatase	**0.877**	0.679
NR-ER	**0.851**	0.644
NR-ER-LBD	**0.858**	0.669
NR-PPAR-gamma	**0.946**	0.590
SR-ARE	**0.823**	0.689
SR-ATAD5	**0.882**	0.619
SR-HSE	**0.876**	0.685
SR-MMP	**0.957**	0.769
SR-p53	**0.906**	0.655
AMES Mutagenesis	**0.929**	0.815
Carcinogenesis	**0.886**	0.637
Fathead Minnow	**0.937**	0.847
Honey Bee	**0.860**	0.640
Biodegradation	**0.935**	0.821
Skin Sensitisation	**0.883**	0.814
hERG_I Inhibitor	**0.957**	0.750
Liver Injury I	**0.864**	0.622
Liver Injury II	**0.797**	0.653
Eye Irritation	**0.959**	0.913
Eye Corrosion	**0.999**	0.956
Respiratory Disease	**0.930**	0.765
Average	**0.904**	0.729

test sets, when they were available, to avoid any biases (Tables 1–4 and 6 and Tables S29–S32). Therefore, pkCSM and admetSAR 2.0 are not part of these statistical comparisons. In general, toxCSM presented statistically better results than all alternative methods (i.e. DeepTox, ADMETLab 2.0, ProTox and Interpretable-ADMET) on the assessed blind test sets. It is worth noting that although admetSAR 2.0 and pkCSM did not provide results on the blind test, toxCSM's results were comparable to or better than them across cross-validation procedures, indicating good classification standards for toxCSM. Furthermore, as Interpretable-ADMET encompasses two models (i.e. graph attention and graph convolutional neural networks), we compared the best of each endpoint model in terms of ROC AUC.

toxCSM regression models were compared with ADMETLab 2.0, Interpretable-ADMET, pkCSM and admetSAR2.0 (Tables 5 and 6), as the other methods did not provide results for any regression endpoint. However, we only compared toxCSM on the blind test sets with ADMETLab 2.0 and Interpretable-ADMET because of result unavailability from the two other methods under this

Table 5. Comparison of toxCSM to ADMETLab 2.0 and Interpretable-ADMET across three toxicity endpoints; performance values are reported as coefficients of determination (R^2); two values are shown per column for toxCSM, denoting the performance on the entire data set and the performance after 10% outlier removal; the best results of toxCSM in this comparison are indicated in bold

Endpoint name	**Methods**		
	toxCSM	**ADMETLab 2.0**	**Interprebble-ADMET**
Fathead Minnow (LC50)	**0.636/0.795**	0.745	0.546
T. Pyriformis (pIGC50)	**0.849/0.950**	0.723	0.832
Acute Rat (LD50)	**0.617/0.774**	–	0.575
Average	**0.701/0.840**	0.734	0.651

Table 6. Comparative performance of toxCSM to alternative methods; pairwise comparisons between toxCSM and the alternative methods for toxicity property predictions across the blind test sets. toxCSM results are shown considering an average across all endpoints; comparisons are shown in terms of the percentage of toxCSM's improvement of toxCSM to alternative methods, employing also a Wilcoxon signed rank test to verify the statistical differences among them; two values are shown per column for the regression, representing the performance on the entire blind test set and the performance after 10% outlier removal

Endpoint type (Evaluation measure)		**Alternative methods**			
	toxCSM	**ADMETLab 2.0**	**ProTox II**	**DeepTox**	**Interpretable-ADMET**
Classification (ROC AUC)	0.905	2.5%*	6.8%*	4.4%*	17.5%*
Regression (R^2)	0.651/ 0.746	+0.1%/ +13.9%	N.A.	N.A.	+5.0%/ +18.9%

N.A. denotes cases where the authors of the methods did not provide results on those particular endpoints or blind test sets; * significantly different with *P*-value < 0.05.

evaluation procedure. When compared to ADMETLab 2.0 and Interpretable-ADMET, toxCSM was able to achieve reasonable better predictive performances in terms of R^2, although no statistical comparison was made due to the low number of endpoint samples utilised by the alternative method. Furthermore, by looking at pkCSM and admetSAR 2.0 (cross-)validation performances, toxCSM was able to achieve as good as or better predictive coefficients when compared to them.

Conclusion

In this work, we develop toxCSM, a comprehensive and scalable web-based platform to assess toxicity profiles of small molecules to date, accounting for 36 different endpoints. toxCSM relies on our well-established graph-based signatures, molecular descriptors and similarity scores to provide accurate and robust predictors that have been thoroughly evaluated and validated, presenting statistically better predictive performances than alternative methods across the assessed toxicity endpoints. We believe toxCSM will be an invaluable tool for the study and optimisation of toxicity profiles of small molecules at early stages of development. We made toxCSM available as an easy-to-use and reliable web-based platform at http://biosig.lab.uq.edu.au/toxcsm as well as API, which allows its complete integration with analytical chemoinformatics pipelines.

Key Points

- toxCSM is a comprehensive and accurate platform to assess small molecule toxicity profiles.
- toxCSM uses graph-based signatures, molecular descriptors and similarity calculations to predict 36 toxicity endpoints, outperforming alternative methods.
- toxCSM is freely available via a user-friendly web server and API to provide a seamless integration with cheminformatics and bioinformatics pipelines.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Availability

toxCSM's web server prediction interface and API were made available at http://biosig.lab.uq.edu.au/toxcsm. The interface is free for all users, with no requirements of login or licence. In addition, all the experimental data used to train, (cross-)validate and test toxCSM's models can be downloaded at https://biosig.lab.uq.edu.au/toxcsm/data.

Acknowledgements

We would like to thank Carlos Rodrigues for his assistance on web server deployment.

Funding

Joe White Bequest Fellowship (to A.G.C.S.); Investigator Grant from the National Health and Medical Research Council of Australia (GNT1174405); Victorian Government's Operational Infrastructure Support Program (in part).

References

1. Khanna I. Drug discovery in pharmaceutical industry: productivity challenges and trends. *Drug Discov Today* 2012;**17**:1088–102.
2. Moffat JG, Vincent F, Lee JA, *et al.* Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nat Rev Drug Discov* 2017;**16**:531–43.
3. Hutchinson L, Kirk R. High drug attrition rates – where are we going wrong? *Nat Rev Clin Oncol* 2011;**8**:189–90.
4. Waring MJ, Arrowsmith J, Leach AR, *et al.* An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nat Rev Drug Discov* 2015;**14**:475–86.

5. Pankevich DE, Altevogt BM, Dunlop J, *et al.* Improving and accelerating drug development for nervous system disorders. *Neuron* 2014;**84**:546–53.
6. Seyhan AA. Lost in translation: the valley of death across preclinical and clinical divide – identification of problems and overcoming obstacles. *Transl Med Commun* 2019;**4**:18.
7. Moreno L, Pearson AD. How can attrition rates be reduced in cancer drug discovery? *Expert Opin Drug Discov* 2013;**8**:363–8.
8. Li AP. Screening for human ADME/Tox drug properties in drug discovery. *Drug Discov Today* 2001;**6**:357–66.
9. Alqahtani S. In silico ADME-Tox modeling: progress and prospects. *Expert Opin Drug Metab Toxicol* 2017;**13**:1147–58.
10. Muller PY, Milton MN. The determination and interpretation of the therapeutic index in drug development. *Nat Rev Drug Discov* 2012;**11**:751–61.
11. Van Norman GA. Limitations of animal studies for predicting toxicity in clinical trials: is it time to rethink our current approach? *JACC Basic Transl Sci* 2019;**4**:845–54.
12. Van de Waterbeemd H. From in vivo to in vitro/in silico ADME: progress and challenges. *Expert Opin Drug Metab Toxicol* 2005;**1**: 1–4.
13. Lave T, Parrott N, Grimm HP, *et al.* Challenges and opportunities with modelling and simulation in drug discovery and drug development. *Xenobiotica* 2007;**37**:1295–310.
14. Dong J, Wang NN, Yao ZJ, *et al.* ADMETlab: a platform for systematic ADMET evaluation based on a comprehensively collected ADMET database. *J Chem* 2018;**10**:29.
15. Xiong G, Wu Z, Yi J, *et al.* ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties. *Nucleic Acids Res* 2021;**49**:W5–14.
16. Pires DEV, Veloso WNP, Myung Y, *et al.* EasyVS: a user-friendly web-based tool for molecule library selection and structure-based virtual screening. *Bioinformatics* 2020;**36**:4200–2.
17. Mayr A, Klambauer G, Unterthiner T, *et al.* DeepTox: toxicity prediction using deep learning. *Front Environ Sci* 2016;**80**:1–15.
18. Banerjee P, Eckert AO, Schrey AK, *et al.* ProTox-II: a webserver for the prediction of toxicity of chemicals. *Nucleic Acids Res* 2018;**46**:W257–63.
19. Cheng F, Li W, Zhou Y, *et al.* admetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties. *J Chem Inf Model* 2012;**52**:3099–105.
20. Drwal MN, Banerjee P, Dunkel M, *et al.* ProTox: a web server for the in silico prediction of rodent oral toxicity. *Nucleic Acids Res* 2014;**42**:W53–8.
21. Hao Y, Moore JH. TargetTox: a feature selection pipeline for identifying predictive targets associated with drug toxicity. *J Chem Inf Model* 2021;**61**:5386–94.
22. Pires DE, Blundell TL, Ascher DB. pkCSM: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. *J Med Chem* 2015;**58**:4066–72.
23. Yang H, Lou C, Sun L, *et al.* admetSAR 2.0: web-service for prediction and optimization of chemical ADMET properties. *Bioinformatics* 2019;**35**:1067–9.
24. Pires DEV, Kaminskas LM, Ascher DB. Prediction and optimization of pharmacokinetic and toxicity properties of the ligand. *Methods Mol Biol* 2018;**1762**:271–84.
25. Pires DEV, Portelli S, Rezende PM, *et al.* A comprehensive computational platform to guide drug development using graph-based signature methods. *Methods Mol Biol* 2020;**2112**: 91–106.
26. Wei Y, Li S, Li Z, *et al.* Interpretable-ADMET: a web service for ADMET prediction and optimization based on deep neural representation. *Bioinformatics* 2022;**38**:2863–71.
27. Mulliner D, Schmidt F, Stolte M, *et al.* Computational models for human and animal hepatotoxicity with a global application scope. *Chem Res Toxicol* 2016;**29**:757–67.
28. Fan D, Yang H, Li F, *et al.* In silico prediction of chemical genotoxicity using machine learning methods and structural alerts. *Toxicol Res (Camb)* 2018;**7**:211–20.
29. Cao Q, Liu L, Yang H, *et al.* In silico estimation of chemical aquatic toxicity on crustaceans using chemical category methods. *Environ Sci Process Impacts* 2018;**20**:1234–43.
30. Zhang C, Cheng F, Sun L, *et al.* In silico prediction of chemical toxicity on avian species using chemical category approaches. *Chemosphere* 2015;**122**:280–7.
31. Wang Z, Zhao P, Zhang X, *et al.* In silico prediction of chemical respiratory toxicity via machine learning. *Comput Toxicol* 2021;**18**:100155.
32. Landrum G. RDKit: Open-source cheminformatics. 2006.
33. Butina D. Unsupervised data base clustering based on daylight's fingerprint and tanimoto similarity: a fast and automated way to cluster small and large data sets. *J Chem Inf Comput Sci* 1999;**39**: 747–50.
34. Tanimoto TT. Elementary mathematical theory of classification and prediction. New York: International Business Machines Corporation, 1958, 1–10.
35. Glem RC, Bender A, Arnby CH, *et al.* Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs* 2006;**9**:199–204.
36. Borgelt C, Meinl T, Berthold M. *Proceedings of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations*. Chicago, Illinois: Association for Computing Machinery, 2005, 6–15.
37. Raghunathan S, Priyakumar UD. Molecular representations for machine learning applications in chemistry. *Int J Quantum Chem* 2022;**122**:e26870.
38. David L, Thakkar A, Mercado R, *et al.* Molecular representations in AI-driven drug discovery: a review and practical guide. *J Chem* 2020;**12**:56.
39. Wigh DS, Goodman JM, Lapkin AA. A review of molecular representation in the age of machine learning. *WIREs Comput Mol Sci* n/a, e1603, 1–19.
40. Al-Jarf R, de Sa AGC, Pires DEV, *et al.* pdCSM-cancer: using graph-based signatures to identify small molecules with anticancer properties. *J Chem Inf Model* 2021;**61**:3314–22.
41. Pires DEV, Ascher DB. mycoCSM: using graph-based signatures to identify safe potent hits against Mycobacteria. *J Chem Inf Model* 2020;**60**:3450–6.
42. Rodrigues CHM, Pires DEV, Ascher DB. pdCSM-PPI: using graph-based signatures to identify protein-protein interaction inhibitors. *J Chem Inf Model* 2021;**61**:5438–45.
43. Velloso JPL, Ascher DB, Pires DEV. pdCSM-GPCR: predicting potent GPCR ligands with graph-based signatures. *Bioinform Adv* 2021;**1**:vbab031.
44. Kaminskas LM, Pires DEV, Ascher DB. dendPoint: a web resource for dendrimer pharmacokinetics investigation and prediction. *Sci Rep* 2019;**9**:15465.
45. Myung Y, Pires DEV, Ascher DB. CSM-AB: graph-based antibody-antigen binding affinity prediction and docking scoring function. *Bioinformatics* 2021;**38**:1141–3.
46. Nguyen TB, Pires DEV, Ascher DB. CSM-carbohydrate: protein-carbohydrate binding affinity prediction and docking scoring function. *Brief Bioinform* 2022;**23**:1–8.
47. Pires DEV, Rodrigues CHM, Ascher DB. mCSM-membrane: predicting the effects of mutations on transmembrane proteins. *Nucleic Acids Res* 2020;**48**:W147–53.

48. Pires DEV, Stubbs KA, Mylne JS, Ascher DB. cropCSM: designing safe and potent herbicides with graph-based signatures. *Briefings in Bioinformatics* 2022;**23**(2):1–9.
49. Zhou Y, Al-Jarf R, Alavi A, *et al.* kinCSM: using graph-based signatures to predict small molecule CDK2 kinase inhibitors. *Research Square (Preprint)* 2021;1–19.
50. Pires DE, Ascher DB. CSM-lig: a web server for assessing and comparing protein-small molecule affinities. *Nucleic Acids Res* 2016;**44**:W557–61.
51. Lagorce D, Bouslama L, Becot J, *et al.* FAF-Drugs4: free ADME-tox filtering computations for chemical biology and early stages drug discovery. *Bioinformatics* 2017;**33**:3658–60.
52. Kazius J, McGuire R, Bursi R. Derivation and validation of toxicophores for mutagenicity prediction. *J Med Chem* 2005;**48**: 312–20.
53. Raschka S. *Python machine learning*. Birmingham, UK: Packt publishing ltd., 2015.
54. Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011;**12**:2825–30.
55. Komer B, Bergstra J, Eliasmith C. In: Hutter F, Kotthoff L, Vanschoren J (eds). *Automated Machine Learning: Methods, Systems, Challenges*. Cham: Springer International Publishing, 2019, 97–111.
56. Bergstra J, Bardenet R, Bengio Y, *et al. Proceedings of the 24th International Conference on Neural Information Processing Systems*. Granada, Spain: Curran Associates Inc., 2011, 2546–54.
57. Pan SJ, Yang Q. A Survey on Transfer Learning. *IEEE Trans Knowl Data Eng* 2010;**22**:1345–59.
58. Demšar J. Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res* 2006;**7**:1–30.