Designing safe and potent herbicides with the cropCSM online resource

Douglas E. V. Pires[1,2,3,4,*], Keith A. Stubbs[5], Joshua S. Mylne[5], David B. Ascher[1,2,3,6,*]

[1] Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne VIC 3004
[2] Department of Biochemistry and Molecular Biology, University of Melbourne, Melbourne, VIC 3010
[3] Computational and Systems Biology, Bio21 Institute, University of Melbourne, 30 Flemington Rd, Parkville VIC 3052
[4] School of Computing and Information Systems, University of Melbourne, Melbourne, VIC 3010
[5] The University of Western Australia, School of Molecular Sciences, 35 Stirling Highway, Crawley, Perth 6009, Australia
[6] Department of Biochemistry, University of Cambridge, 80 Tennis Ct Rd, Cambridge CB2 1GA

*To whom correspondence should be addressed D.B.A. Tel: +61 90354794; Email: david.ascher@unimelb.edu.au. Correspondence may also be addressed to D.E.V.P. douglas.pires@unimelb.edu.au.

Herbicides have revolutionised weed management, increased crop yields and improved profitability allowing for an increase in worldwide food security. Their widespread use, however, has also led to not only a rise in resistance but also concerns about their environmental impact. To help identify new, potent, non-toxic and environmentally safe herbicides we have employed interpretable predictive models to develop the online tool cropCSM (http://biosig.unimelb.edu.au/crop_csm).

Developing herbicides, much like pharmaceuticals, involves a careful balance between efficacy and safety. In the pharmaceutical industry, drug development pipelines have tackled these challenges by modelling and optimising these important parameters early in the development process. This has led, in general, to increased hit rates and decreased attrition due to poor toxicity profiles and, in the process, reduced development time, costs, and animal testing[1-4]. Although many computer-guided approaches have proven invaluable for drug development, by contrast little has been done to aid the development of safe and potent agrochemicals.

Using experimental information on the herbicidal activity of over 4,000 small molecule compounds (22% with herbicidal activity), we investigated what physicochemical properties of the compounds translate to herbicidal activity. Herbicidal molecules were enriched in saturated carbon chains and benzene substructures, compared to the inactive molecules (**Fig. 1a**). The majority (90%) of the active compounds tended to be less than 517 Da, up to 9 acceptors and 4 donors, with fewer than 9 rotatable bonds and a logP between -1.7 and 6.1 (**Supplementary Fig. 1**) (95% less than 700 Da, 11 rotatable bonds, 11 acceptors, 6 donors, and logP -3.0 to 6.1). This is similar, although slightly more lenient, than the widely used Lipinski Rule of Five for orally bioavailable drugs. Interestingly, but consistently, there was no significant distinction in physicochemical properties between herbicides and approved drugs, as illustrated in the t-SNE plot (**Supplementary Fig. 2**). Compared to all FDA approved drugs, however, herbicides were enriched in substructures involving chlorine.

These insights were used as a platform to build a supervised machine learning predictive model, where the small molecule structure was represented as a graph-based signature, termed Cutoff Scanning Matrix (CSM, in which the atoms are represented as nodes, and covalent interactions between them as edges[5,6] (**Supplementary Fig. 3**). Under cross-

validation, we were able to correctly identify 82% of the active molecules with an overall accuracy of 87% and AUC of 0.85 (**Fig. 1b** and **Supplementary Table 1**). When the model was evaluated against a blind test set of 106 active and 345 inactive molecules, we achieved comparable performance (87% accuracy, AUC of 0.87). This provided confidence that the approach can be generalized and used with unknown sets of putative herbicidal molecules active against a target of interest.

Agrochemicals have been linked to a range of unwanted negative effects on both health and the environment. To help identify safe herbicides, complementary models were developed to capture the impact of a small molecule on the honey bee (*Apis mellifera*), mallard (*Anas platyrhynchos*) and flathead minnow (*Pimephales promelas*) toxicity, in addition to measures of human health, including AMES toxicity, rat LD_{50} and oral chronic toxicity. While assessing molecular substructures enriched in toxic compounds, (**Supplementary Fig. 4**), we identified a prevalence of complex ring structures. Of note, structures rich in chlorine, while enriched in herbicides, were also enriched in compounds that were toxic for mallard and minnow, highlighting a potential inherent difficulty in optimising potency and safety when designing herbicides.

We were also able to identify toxic molecules as classification and regression tasks with accuracies of up to 92% and Pearson's correlations of up to 0.86, outperforming previous predictive approaches (**Fig. 1b**, **Supplementary Fig. 5** and **Supplementary Tables 2-3**). These results add credence to the tool to rapidly identify potentially hazardous molecules early in the development process, which has the potential to significantly reduce costs and failure rates.

The cropCSM models were then applied to a set of 360 commercial herbicides[7]. Over 97% were correctly identified as herbicidal (**Fig. 2**). Despite being outliers in terms of their physicochemical properties, cropCSM correctly predicted glyphosate and paraquat as herbicides. Of those that weren't, however, they included the natural fatty acid oleic acid, and non-specific fragments like molecules such as dazomet and pentachlorophenol.

Overall, our cropCSM tool provides the first free and easy-to-use *in silico* platform to help develop herbicides that are safe, effective and minimise impact on the environment. We anticipate future iterations of cropCSM that will draw upon larger datasets and as a result will have a higher predictor capability, allowing for a greater increase in accuracy and correlation. The herbicidal and toxicity predictors are freely available via an integrated and easy-to-use web interface (**Supplementary Fig. 6;** http://biosig.unimelb.edu.au/crop_csm).

Acknowledgements

D.B.A. and D.E.V.P. were funded by the Jack Brockhoff Foundation (JBF 4186, 2016) and an Investigator Grant from the National Health and Medical Research Council (NHMRC) of Australia (GNT1174405). Supported in part by the Victorian Government's Operational Infrastructure Support Program. The authors thank Julie Leroux, Joel Haywood, Kirill Sukhoverkov and Kalia Bernath-Levin for acquiring the 4,513 data points used to train cropCSM. To acquire the dataset, we also thank for funding the UWA Office of Industry and Innovation Pathfinder Funding Scheme, the Australian Research Council (ARC) for a Future Fellowship (FT120100013) to J.S.M., Nexgen Plants Pty Ltd and ARC grant DP190101048 to J.S.M. and K.A.S.

Author Contributions

D.E.V.P. and D.B.A. were responsible for method design and website development. K.A.S. and J.S.M. were responsible for development of datasets on herbicidal activity. All authors assisted with manuscript writing.

Competing Interests

The authors declare no competing interests.

Additional Information

Supplementary Information is available.

References

1. Kola, I. & Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery* **3**, 711-716 (2004).
2. Ulrich, R. & Friend, S.H. Toxicogenomics and drug discovery: will new technologies help us produce better drugs? *Nature Reviews Drug Discovery* **1**, 84-88 (2002).
3. Waring, M.J. *et al.* An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nature Reviews Drug Discovery* **14**, 475-486 (2015).
4. Segall, M.D. & Barber, C. Addressing toxicity risk when designing and selecting compounds in early drug discovery. *Drug Discovery Today* **19**, 688-693 (2014).
5. Pires, D.E., Blundell, T.L. & Ascher, D.B. pkCSM: Predicting Small-Molecule Pharmacokinetic and Toxicity Properties Using Graph-Based Signatures. *J Med Chem* **58**, 4066-72 (2015).
6. Pires, D.E.V. & Ascher, D.B. mycoCSM: Using Graph-Based Signatures to Identify Safe Potent Hits against Mycobacteria. *J Chem Inf Model* **60**, 3450-3456 (2020).
7. Sukhoverkov, K.V. *et al.* Refining physico-chemical rules for herbicides using an antimalarial library. *bioRxiv*, 2020.10.27.356576 (2020).
8. Gandy, M.N., Corral, M.G., Mylne, J.S. & Stubbs, K.A. An interactive database to explore herbicide physicochemical properties. *Organic & Biomolecular Chemistry* **13**, 5586-5590 (2015).
9. Wang, F. *et al.* Graph attention convolutional neural network model for chemical poisoning of honey bees' prediction. *Science Bulletin* **65**, 1184-1191 (2020).
10. Zhang, C. *et al.* In silico prediction of chemical toxicity on avian species using chemical category approaches. *Chemosphere* **122**, 280-287 (2015).
11. Weihua, L. & Yun, T. In silico prediction of terrestrial and aquatic toxicities for organic chemicals. *Chinese Journal of Pesticide Science* (2010).
12. Xu, C. *et al.* In silico prediction of chemical Ames mutagenicity. *J Chem Inf Model* **52**, 2840-7 (2012).
13. Zhu, H. *et al.* Quantitative structure-activity relationship modeling of rat acute toxicity by oral exposure. *Chem Res Toxicol* **22**, 1913-21 (2009).
14. Mazzatorta, P., Estevez, M.D., Coulet, M. & Schilter, B. Modeling oral rat chronic toxicity. *J Chem Inf Model* **48**, 1949-54 (2008).
15. Pires, D.E. *et al.* Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. *BMC Genomics* **12 Suppl 4**, S12 (2011).
16. Pires, D.E., Ascher, D.B. & Blundell, T.L. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* **30**, 335-42 (2014).
17. Rodrigues, C.H., Pires, D.E. & Ascher, D.B. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Res* **46**, W350-W355 (2018).

18. Pires, D.E.V., Rodrigues, C.H.M. & Ascher, D.B. mCSM-membrane: predicting the effects of mutations on transmembrane proteins. *Nucleic Acids Res* **48**, W147-W153 (2020).
19. Pires, D.E. & Ascher, D.B. mCSM-AB: a web server for predicting antibody-antigen affinity changes upon mutation with graph-based signatures. *Nucleic Acids Res* **44**, W469-73 (2016).
20. Pires, D.E. & Ascher, D.B. CSM-lig: a web server for assessing and comparing protein-small molecule affinities. *Nucleic Acids Res* **44**, W557-61 (2016).
21. Pires, D.E., Blundell, T.L. & Ascher, D.B. mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. *Sci Rep* **6**, 29575 (2016).
22. Pires, D.E.V. & Ascher, D.B. mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions. *Nucleic Acids Res* **45**, W241-W246 (2017).
23. Rodrigues, C.H.M., Myung, Y., Pires, D.E.V. & Ascher, D.B. mCSM-PPI2: predicting the effects of mutations on protein-protein interactions. *Nucleic Acids Res* **47**, W338-W344 (2019).
24. Myung, Y., Pires, D.E.V. & Ascher, D.B. mmCSM-AB: guiding rational antibody engineering through multiple point mutations. *Nucleic Acids Res* **48**, W125-W131 (2020).
25. Myung, Y., Rodrigues, C.H.M., Ascher, D.B. & Pires, D.E.V. mCSM-AB2: guiding rational antibody design using graph-based signatures. *Bioinformatics* **36**, 1453-1459 (2020).
26. Landrum, G. RDKit: Open-Source Cheminformatics Software.(2016). *URL http://www. rdkit. org/, https://github. com/rdkit/rdkit* (2016).
27. Borgelt, C., Meinl, T. & Berthold, M. MoSS: a program for molecular substructure mining. in *Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations* 6–15 (Association for Computing Machinery, Chicago, Illinois, 2005).
28. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

ONLINE METHODS

Data for Herbicidal Activity

A dataset of 4,513 experimentally characterized, structurally diverse small molecules and their herbicidal activity profiles[7,8]. These were labelled either as active (997 molecules) or inactive (3,516 molecules). They had an average molecular weight of 380 Da and logP of 2.4 (**Supplementary Fig. 1**). A database of 360 commercial herbicides was also used to evaluate cropCSM[7,8].

Environmental and Human Toxicity

We have developed new predictors based on six environmental and human toxicity data sets with experimentally characterised molecules. Environmental toxicity data sets included (i) honey bee (*A. mellifera*) toxicity, which was composed of 247 toxic and 353 atoxic molecules[9]; (ii) avian toxicity, composed of 461 small molecules and their effects on mallard duck (66 toxic and 395 atoxic)[10] and (iii) flathead minnow toxicity, with lethal concentration values (LC50) for a diverse set of 554 molecules[11]. Human toxicity data sets included (i) AMES toxicity, with compounds labelled based on their carcinogenic potential (4,632 carcinogenic and 3,470 not-carcinogenic)[12]; (ii) oral acute toxicity in rats, denoted as lethal dose (LD50) values for 10,145 compounds[13] and (iii) oral chronic toxicity in rats values for 567 compounds[14].

Graph-based Signatures and Feature Engineering

Graph modelling has an invaluable tool to model biological entities, including small molecules. Over the years we have proposed and developed the concept of graph-based signatures (based on Cutoff Scanning Matrix concept[15]) to represent physicochemical and geometrical properties of a range of macromolecules[5,16-18] and their interactions[19-25]. These have also been successfully adapted to represent small molecules pharmacokinetics, toxicity and bioactivity[5,6]. **Supplementary Fig. 3** depicts the main steps involved in feature engineering with graph-based signatures. Small molecules are modelled as unweighted, undirected graphs where nodes represent atoms and edges represent covalent bonds. Via pharmacophore modelling[5], atoms are labelled based on their properties and all-pairs shortest paths distances are calculated. Molecules are then represented as cumulative distribution functions of atom distances labelled based on their respective physicochemical properties (pharmacophores) and converted as a feature vector used as evidence to train and test predictive methods. Complementary physicochemical properties are calculated and included using the RDKit cheminformatics library[26] and included in the feature vector. Frequent substructure mining was performed using MoSS[27].

Model Selection and Validation

Different supervised learning algorithms available on the scikit-learn Python library[28] were assessed with best performing models selected based on Matthew's Correlation Coefficient (MCC) and the Area under the ROC curve (AUC) for classification tasks and Pearson's correlation and Root Mean Squared Error (RMSE) for regression tasks. Performance was assessed under 10-fold cross validation as well as using non-redundant blind tests. A feature selection step was used to reduce dimensionality and improve performance via a Forward Greedy Selection approach.

Web server

The backend of the cropCSM web server was developed using the Python Flask framework version 0.12.3 and the front end using Bootstrap framework version 3.3.7. The system is hosted by a Linux server running Apache.

FIGURES

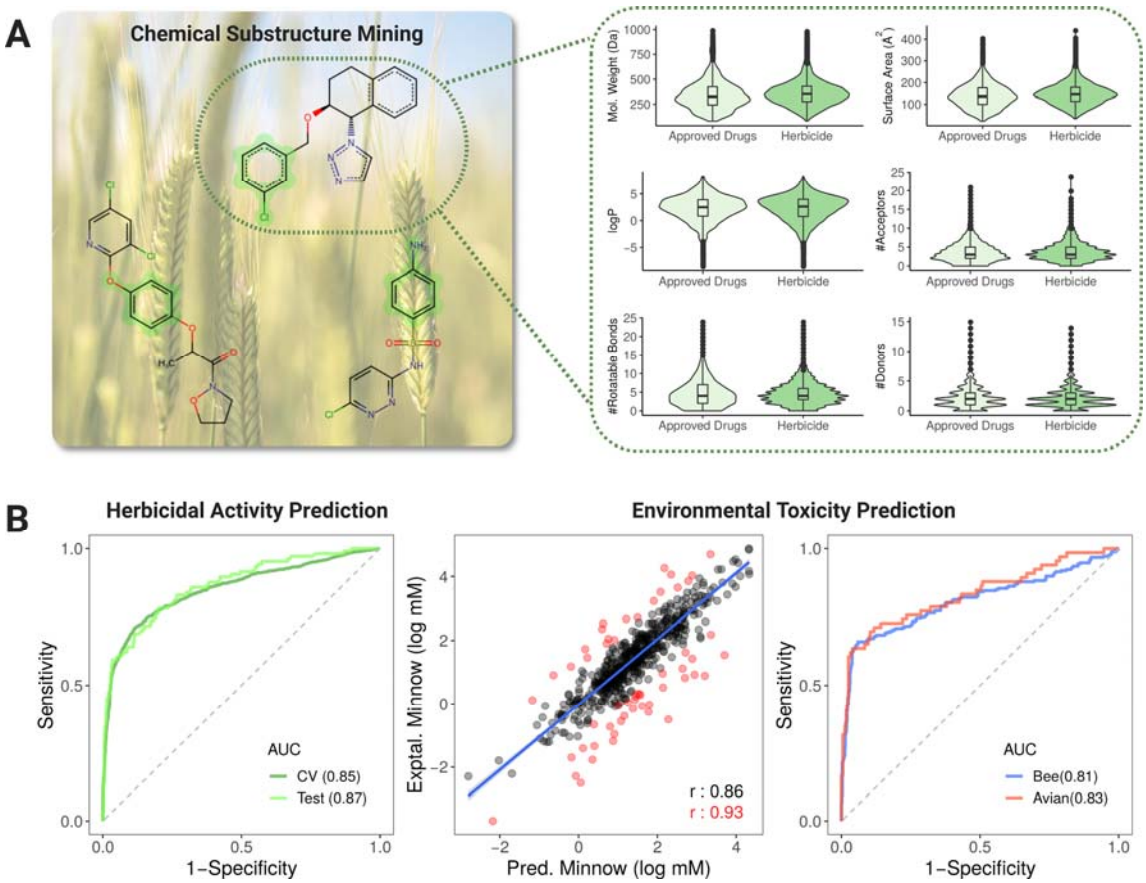

Figure 1. **cropCSM: predicting safe and potent herbicides.** Using chemical substructure mining we identified common enriched substructures in compounds with herbicidal activity (A-left). Active compounds presented similar molecular properties of approved drugs (A-right). Performance of herbicide and environmental-toxicity predictors is shown in (B). Our herbicide predictor was able to accurately identify active compounds with AUC>0.85 on cross-validation and blind test. Three environmental toxicity models have been developed and were capable of successfully measuring minnow toxicity (as a regression task, center graph) as well as identifying potentially harmful compounds for Bees and Mallard (right-hand side graph).

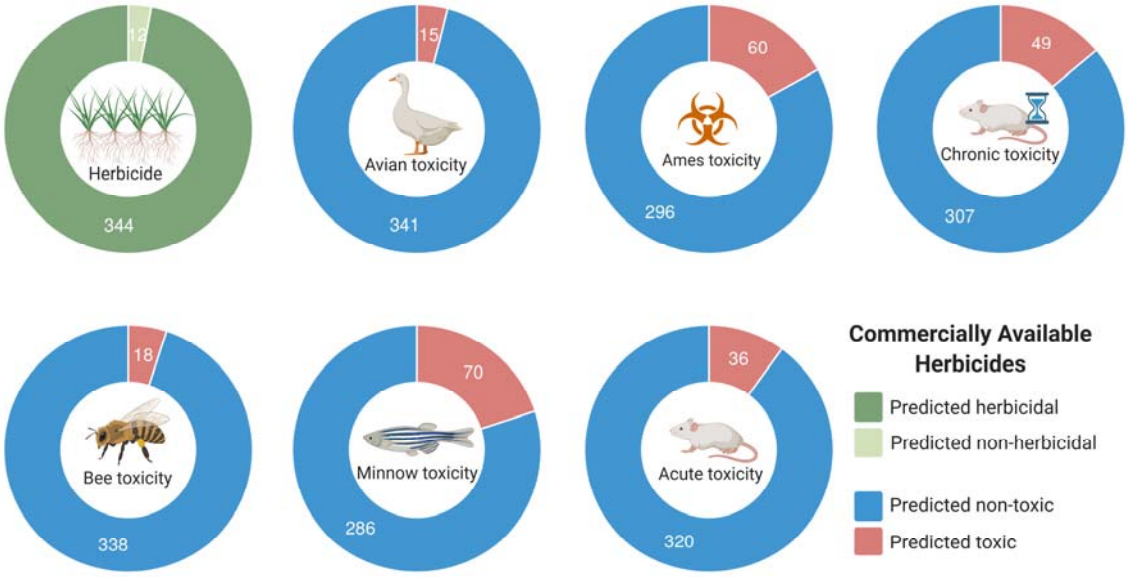

Figure 2. Performance of cropCSM on commercially available herbicides. Our method was able to correctly classify 97% of commercial herbicides (344 out of 356, top-left graph). The figure also shows the proportion of compounds predicted to be environmental or human toxic. Molecules were more frequently predicted as AMES toxic (17%, 60 out of 356) and Minnow toxic (20%, 70 out of 356).