Post-refereeing version
Final version published in: Environmental Science & Technology, 52, 12953-12955.
doi: 10.1021/acs.est.8b05382

Machine learning for environmental toxicology: a call for integration and innovation

Thomas H. Miller[1,*], Matteo D. Gallidabino[2], James I. MacRae[3], Christer Hogstrand[4], Nicolas R. Bury[5], Leon P. Barron[1], Jason R. Snape[6], Stewart F. Owen[6]

[1] *Department of Analytical, Environmental & Forensic Sciences, School of Population Health & Environmental Sciences, Faculty of Life Sciences and Medicine, King's College London, 150 Stamford Street, London SE1 9NH, UK*

[2] *Department of Applied Sciences, Northumbria University, Newcastle Upon Tyne NE1 8ST, UK*

[3] *Metabolomics Laboratory, The Francis Crick Institute, 1 Midland Road, London, NW1 1AT, UK*

[4] *Division of Diabetes and Nutritional Sciences, Faculty of Life Sciences and Medicine, King's College London, Franklin Wilkins Building, 150 Stamford Street, London SE1 9NH, UK*

[5] *Faculty of Science, Health and Technology, University of Suffolk, James Hehir Building, University Avenue, Ipswich, Suffolk IP3 0FS, UK*

[6] *AstraZeneca, Global Environment, Alderley Park, Macclesfield, Cheshire SK10 4TF, UK*

Abstract

Viewpoint on the current state of art of machine learning in environmental toxicology.

* Corresponding author: thomas.miller@kcl.ac.uk (Tel: +44 20 7848 4978)

Recent advances in computing power have enabled the application of machine learning (ML) across all areas of science. A step change from a data-rich landscape to one where new hypotheses, relationships and knowledge is emerging as a result (Figure 1). Whilst ML is related to artificial intelligence (AI), they are not the same. ML is a branch of AI involving the application of statistical algorithms to enable a system to learn. Learning can involve data interpretation, identification of patterns and decision making. However, application and acceptance of ML within environmental toxicology, and more specifically for our viewpoint, environmental risk assessment (ERA), remains low. ML is an example of a disruptive research technology [1], which is urgently needed to cope with the complexity and scale of work required.

Notable ML achievements in biochemistry and medicine, for example, have aided diagnosis of Alzheimer's disease from magnetic resonance scans [2], survival rates following lymphoma using gene expression profiling [3], chronological age prediction from DNA methylation [4], and more recently in predictive toxicology [5, 6]. ML is rapidly developing and can now solve complex problems in a fraction of the time and cost of laboratory experimentation. In environmental toxicology, complex and highly variable conditions are the norm. ML will be especially valuable here, by disrupting a reliance on hypothesis-driven and systematic approaches exploring simpler linear relationships.

In a recent study, ML outperformed animal testing approaches in chemical safety assessments [6]. In our work, we used ML to predict bioconcentration in aquatic fauna as part of persistent, bioaccumulative and toxicity (PBT) assessments [7]. However, there is a critical lack of literature concerning ML development for environmental exposure and effect assessment. Few reported collaborative initiatives embrace ML approaches in ERA. Given this scale, ML is likely to be the only realistic approach to meet regulatory body requirements for screening, prioritisation and ERA of thousands of chemicals (including mixtures). ML could be used in several ways: (i) incorporation into the ERA process via a weight-of-evidence approach for hazard and exposure; (ii) the eventual substitution of animal testing; (iii) rapid, early decision making on risks posed by a legacy and new chemical; and (iv) the management of risk. This acceptance of ML into an ERA framework is a challenging, but as a research community, we must lead and drive change.

Barriers to the use and acceptance of machine learning

The European Chemicals Agency suggested that toxicology cannot yet be replaced with computers as the underlying science needs improvement [8]. Its concerns relate to (i) feature selection, (ii) model interpretability, (iii) generalisability and (iv) confidence in predictive ability. More exploration of ML

is needed to understand its limitations and value. The demand for it is becoming increasingly apparent. For example, in the UK, substantial research funding is now being directed into ML for benefit of the economy [9]. Importantly, we must improve knowledge and literacy skills in ML to meet such demand. This could be achieved through collaboration, but the disciplinary gap needs to be bridged by cross-sectoral training and learning to improve ML competency for all scientists. This would not only benefit research, but also the peer-review process for research manuscripts and the inter-validation or implementation of models across the field.

While more ML-literate scientists will be essential for driving further funding opportunities and delivering a more predictive approach to environmental protection, ML itself is being driven by 'big data' projects where data/model accessibility and ownership is another progress-limiting challenge. Third-party access has improved, but not in every case. For proprietary data this can become very complex and even taboo for industry-owned data. Researchers must ensure that they are transparent with data, but also their ML models, to further understanding of the science.

Another barrier to ML in ERA is that some regulatory agencies are reluctant to accept and use ML predictions alone for ERA frameworks. The precautionary principle will likely relate to the prediction of false negatives. To understand these (and indeed false positives), algorithms should ideally be unambiguous and interpretable. These principles form part of the Organisation for Economic Co-operation and Development (OECD) 5 Principles for Quantitative Structure Activity Relationship (QSAR) validation, for example, which aims to improve regulatory acceptance of QSAR models. However, these guidelines were established in 2004 [10] and focused most on traditional linear approaches, but were vague concerning ML acceptance criteria. As a priority, we recommend that these guidelines be updated a stronger focus on the spectrum of ML models available now. As a final consideration, we call for tripartite collaborative efforts and initiatives by academia, industry and regulators to enable innovative ways to better protect environmental and public health using ML.

An industry perspective on the potential value of machine learning

In medicine, ML models for healthcare are being approved at an increasing rate by the FDA and plays a leading role in Precision Medicine [11]. Regulatory acceptance and knowledge are certainly there, but why does it only appear in certain fields? Confidence in the predictive power and utility of ML is growing within companies. Traditionally, proprietary information and company data lay behind an iron-curtain of confidentiality. The inherent drive to protect data may have been by desire to maintain a competitive advantage; indeed, the cost of generating data has been enormous. Now, through ML and similar technologies, the real value of these closely guarded data may appear on the horizon.

Industry is investing heavily in skilled people, driving competition towards safe and trusted model development in many fields, not least for internal R&D. More environmental regulator engagement is needed before the real value of ML can be realised externally to companies. Would the first approach be via accepting ML tools that identify hazards and exposures? With better understanding of false positives/negatives, confidence should grow regarding predicted risks for new compounds.

Whilst ML accelerates the ability to predict, the limited acceptance and application of the precautionary principle seems to be hindering innovation across all sectors. A paradigm shift is now well underway and given the burgeoning use of this technology in other spheres, we anticipate similar steps in ours. As a community striving to protect the environment, we need to embrace the technology sooner rather than later.

Acknowledgments

TM is funded by the Biotechnology and Biological Sciences Research Council iNVERTOX project (Reference BB/P005187/1) and AstraZeneca Global SHE research programme awarded to LB and NB. JM is supported by the Francis Crick Institute which receives its core funding from Cancer Research UK (FC001999), the UK Medical Research Council (FC001999), and the Wellcome Trust (FC001999). AstraZeneca is a biopharmaceutical company specialising in the discovery, development, manufacturing and marketing of prescription medicines, including some products reported here. SFO and JRS are employees of AstraZeneca and a partner of the Innovative Medicines Initiative Joint Undertaking under iPiE grant agreement no. 115735, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in-kind contribution. The authors declare no financial conflict of interest.

References

1. Sedlak, D.L., *Disruptive Environmental Research.* Environmental Science & Technology, 2018.
2. Klöppel, S., et al., *Automatic classification of MR scans in Alzheimer's disease.* Brain, 2008. **131**(3): p. 681-689.
3. Shipp, M.A., et al., *Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning.* Nature Medicine, 2002. **8**(1): p. 68-74.

4. Vidaki, A., et al., *DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing.* Forensic Science International: Genetics, 2017. **28**: p. 225-236.
5. Liu, R., et al., *Assessing deep and shallow learning methods for quantitative prediction of acute chemical toxicity.* Toxicological Sciences, 2018: p. kfy111.
6. Luechtefeld, T., et al., *Machine learning of toxicological big data enables read-across structure activity relationships (RASAR) outperforming animal test reproducibility.* Toxicological Sciences, 2018: p. kfy152-kfy152.
7. Miller, T.H., et al., *Prediction of bioconcentration factors in fish and invertebrates using machine learning.* Science of The Total Environment, 2019. **648**: p. 80-89.
8. Van Noorden, R., *Software beats animal tests at predicting toxicity of chemicals.* Nature, 2018. **559**(7713): p. 163.
9. Government, H., *Industrial Strategy Building a Britain Fit for the Future*, E.a.I.S. Department for Business, Editor. 2017: London
10. Co-operation, O.f.E. and Development, *The report from the expert group on (Quantitative) Structure-Activity Relationships [(Q) SARs] on the principles for the validation of (Q) SARs*. 2004.
11. FDA. *Press Announcements*. 2018; Available from: https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/default.htm.

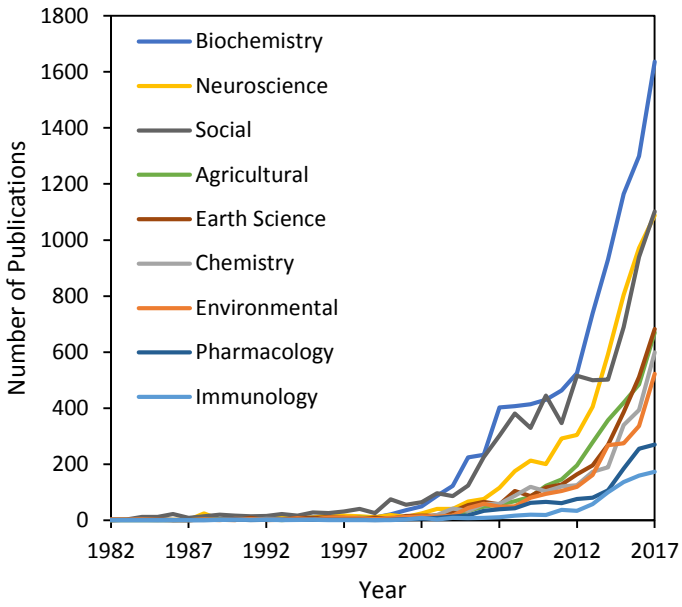

Figure 1 - The number of publications involving ML across different fields. Literature searching was performed using key words "machine learning" through Elsevier's Scopus® and filtering search results through subject categories.