

Machine learning models for predicting endocrine disruption potential of environmental chemicals

Marco Chierici, Marco Giulini, Nicole Bussola, Giuseppe Jurman & Cesare Furlanello


To cite this article: Marco Chierici, Marco Giulini, Nicole Bussola, Giuseppe Jurman & Cesare Furlanello (2018) Machine learning models for predicting endocrine disruption potential of environmental chemicals, Journal of Environmental Science and Health, Part C, 36:4, 237-251, DOI: [10.1080/10590501.2018.1537155](https://doi.org/10.1080/10590501.2018.1537155)

To link to this article: <https://doi.org/10.1080/10590501.2018.1537155>



Published online: 10 Jan 2019.



Submit your article to this journal 



Article views: 348



View related articles 



View Crossmark data 



Citing articles: 4 View citing articles 



Machine learning models for predicting endocrine disruption potential of environmental chemicals

Marco Chierici^a, Marco Giulini^a, Nicole Bussola^{a,b}, Giuseppe Jurman^a, and Cesare Furlanello^a

^aFondazione Bruno Kessler, Trento, Italy; ^bCentre for Integrative Biology, University of Trento, Trento, Italy

ABSTRACT

We introduce here ML4Tox, a framework offering Deep Learning and Support Vector Machine models to predict agonist, antagonist, and binding activities of chemical compounds, in this case for the estrogen receptor ligand-binding domain. The ML4Tox models have been developed with a 10×5 -fold cross-validation schema on the training portion of the CERAPP ToxCast dataset, formed by 1677 chemicals, each described by 777 molecular features. On the CERAPP “All Literature” evaluation set (agonist: 6319 compounds; antagonist 6539; binding 7283), ML4Tox significantly improved sensitivity over published results on all three tasks, with agonist: 0.78 vs 0.56; antagonist: 0.69 vs 0.11; binding: 0.66 vs 0.26.

KEYWORDS

machine learning; deep learning; toxicology

Introduction

The art of identifying the unexpected, as toxicology was defined by the Royal Chemical Society,¹ has traditionally relied on animal models for the assessment of chemical risk. However, the high costs of animal experiments, ethics and regulatory policies limiting or prohibiting the use of animals, and the need for earlier recognition of toxic molecules are now pushing for re-placing *in vivo* assays with *in silico* toxicological methods developed through different flavors of mathematical models,^{2,3} possibly integrated with animal-free *in vitro* assays.⁴

In particular, algorithmic and technological advances in machine learning have boosted the new paradigm known as predictive toxicology.⁵ This branch of study dates back to late 90's⁶; its relevance is well supported by a steady flow of publications, as well as by initiatives led by public agencies⁷ and societies,¹ with web platforms offering analytical services (e.g. INSPECT⁸).

Despite the key importance – both theoretical and applicative – of predictive toxicology, the scientific community is far from having reached

an optimal shared solution; currently, pitfalls and hurdles range from data quality issues to the low specificity (high number of false-positives) affecting most of the *in silico* methods.⁹ More recently, the rise of neural network-based predictors in this field has provided serious expectation for improvements both in specificity and sensitivity.^{10–12}

The mainstream change proposed by predictive toxicology is the shift from detecting adverse effects at the organism level to the identification of biologically significant disruptions of toxicity pathways at the molecular level.¹ Predictive toxicology is a broad term encompassing four main categories of approaches and resources¹³: data analytics (e.g. toxicoinformatics); chemical and toxicity databases (e.g. toxicogenomics and metabolomics); chemoinformatics (e.g. quantum chemical methods for generating molecular descriptors); and, the quantitative structure-activity relationships (QSAR) modeling framework.

In particular, the QSAR framework is based on the assumption that compounds with similar structures are likely to exhibit similar behavior in terms of biological activity or chemical property, including toxicity. Computational models are trained in order to describe such relationships between chemical structures and toxicological processes, and possibly predict the biological activity of additional chemicals out of the training datasets. Known limits of QSAR are validation challenges, model interpretation issues, and model selection issues.¹⁴ However, QSAR has made available to machine learning frameworks a critical mass of data. Recently, a hazard database of more than 800 thousand chemical properties for about 81,000 chemicals had been used to train supervised statistical machine learning models (logistic regression and random forests) predictive of hazard labels with previously unseen accuracy for purely *in silico* models.¹⁵

With a similar aim to target faster, reproducible, and more cost-effective *in silico* safety assessment, we propose here the machine learning framework ML4Tox, including both deep and shallow learning in a classification pipeline. The framework can employ a deep multilayer network or a Support Vector Machine (SVM, linear or gaussian) as predictive models. The central element in the approach is a Data Analysis Protocol (DAP) that takes care of massive replication of experiments on data partitions (10×5 cross-validation) used to select models separately from the validation phase. Such repeated training/test splitting strategy is endowed with diagnostic indicators to double-test for reproducibility and it has been designed and adopted within the US-FDA led MAQC-II and SEQC initiatives.^{16,17} While in biomarker studies based on high-throughput *omics* features the Matthews Correlation Coefficient^{18–20} (MCC) is adopted as reference error function,²⁰ in this study we also use sensitivity as a target performance metric, in order to give priority to compounds detected as active for potential toxicity.

We applied ML4Tox in three toxicology tasks defined in the Collaborative Estrogen Receptor Activity Prediction Project (CERAPP).²¹ Environmental exposure to endocrine-disrupting chemical compounds (EDC) poses high risks for human health, with potential impact on the endocrine system causing adverse immune, neurological and developmental effects. The interest of the toxicological community in datasets describing effects of estrogen receptor (ER) related compounds has steadily grown in the last few years, due to the key role of the ER molecular complex in the reproductive function.

Supported by accuracy improvements reported for autoencoder architectures trained on the ToxCast *invitrodb* dataset,²² we aim in particular at extending the application of the QSAR data-driven approach to deep learning architectures to improve sensitivity of CERAPP tasks. Specifically, the CERAPP assessed the application of predictive modeling to evaluate the binding interactions of environmental chemicals to the ligand-binding domain of human ER from *in vitro* high-throughput screening (HTS) assay data. These interactions are differentiated into three classes: agonist, antagonist and binding. The CERAPP has defined a training set of 5031 compounds (1677 per class) and the “Literature Evaluation set” (6319, 6539, and 7283, respectively, for agonist, antagonist, and binding) labeled as positive or negative. The data defines three distinct learning tasks, which are tackled by ML4Tox with a deep multilayer network, a linear SVM and a Gaussian SVM, respectively, yielding superior performances over those published. We present here the general architecture of the ML4Tox framework, its main methods and experimental application to the CERAPP tasks, finally discussing the potential for the future development of deep learning architectures in predictive toxicology.

Materials and methods

Data sets

In this study, for model development we used the CERAPP training set (TR in the following), derived from ToxCast and Tox21 programs and consisting of 1677 chemicals^{21,23,24} (Table 1). Each chemical was assigned a binary label representing their agonist, antagonist, and binding activities for the ligand-binding domain of ER. Models were tested on the CERAPP “Literature” evaluation set (EV), consisting of 6319 chemicals for agonist,

Table 1. CERAPP Training set for binary classification tasks.

Class	Active	Inactive	Total
Agonist	219	1458	1677
Antagonist	41	1636	1677
Binding	237	1440	1677

Table 2. CERAPP “Literature” evaluation set for binary classification tasks.

Class	Active	Inactive	Total
Agonist	350	5969	6319
Antagonist	284	6255	6539
Binding	1982	5301	7283

6539 for antagonist, and 7283 for binding (see Table 2). EV was derived from the 7547 compounds in the full CERAPP evaluation set after exclusion of chemicals with relatively high (>20%) disagreement amongst literature sources.²¹

Feature extraction and filtering

All chemicals were described by 777 molecular descriptors extracted from their bi-dimensional chemical structure by the software Mold2,^{25,26} following previous analyses.²¹ Additionally, we explored the use of Extended-Connectivity Fingerprints²⁷ (ECF), a class of circular topological fingerprints, as an alternative set of descriptors. The ECF features were generated from the canonical SMILES of each compound using the Morgan algorithm²⁸ with four iterations, as implemented in the DeepChem²⁹ and RDKit³⁰ Python 2.7 modules. As a filtering step, features with constant values across the training samples were removed from the TR and EV sets before further analysis.

Machine learning

We applied Deep Learning (DL) and Support Vector Machine (SVM) models to predict compound agonist, antagonist and binding activities. The general ML4Tox architecture is sketched in Figure 1. We first developed ML4Tox-Agonist, a deep multilayer neural network to predict Agonist activity with five fully-connected hidden layers of 644, 644, 420, 30, and 5 nodes. As activation functions, we considered the Rectified Linear Unit³¹ for the inner layers and a SoftMax one for the output layer. The optimizer was Adam³² with a learning rate of $LR = 5 \times 10^{-6}$ and Cross Entropy as the loss function, with weights proportional to the class sizes. The batch size was 300 and the number of epochs 1500. To avoid overfitting, dropout layers were added with rate 0.5 after each of the first three inner layers and rate 0.25 after the fourth layer, with no dropout applied to the last hidden layer. Optimizer type, LR, number of epochs, and the dropout strategy were selected amongst alternatives by training over 4000 epochs on 70% of TR (randomly chosen) and evaluating the performance on the left-out 30%. Using a grid of LR values, we compared the training and validation performance and losses of the net using stochastic gradient descent (SGD)

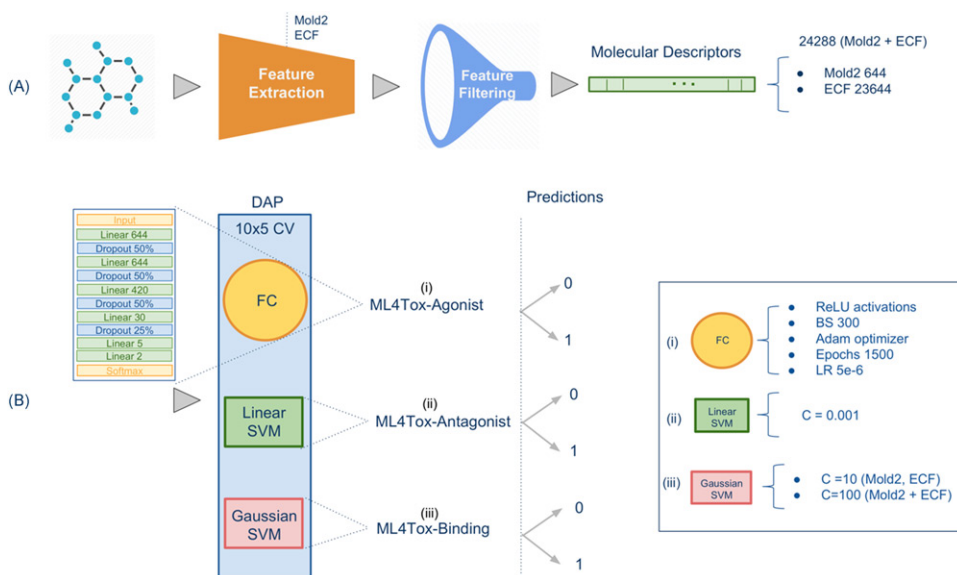


Figure 1. Overview of ML4Tox machine learning framework. A: Preprocessing of molecular descriptors from chemical compounds; features are extracted by Mold2 or the Extended-Connectivity Fingerprint (ECF) algorithms. Descriptors constant across all training data are filtered out. B: Model training; descriptors and binary labels (agonist, antagonist, binding) for each compound are used to train three *in-silico* models: ML4Tox-Agonist, ML4Tox-Antagonist, and ML4Tox-Binding. DAP: Data analysis protocol; FC: fully-connected layers; 10x5 CV: 10 rounds of 5-fold cross-validation.

(Figure 2) and Adam (Figure 3) as optimizers, with no dropout at first. For SGD, the net was trained with $\text{LR} \in [10^{-4}, 5 \times 10^{-4}, 10^{-3}, 2 \times 10^{-3}]$; with the two lowest LR values (Figure 2, panels A-B) we observed a slow decay of training/validation losses combined with lower MCC (defined below) on training; with the two highest LR values (Figure 2, panels C-D), the net is slowly learning from data ($\text{MCC} \sim 0.8$), while also quickly overfitting. For Adam, we trained the net with $\text{LR} \in [10^{-6}, 5 \times 10^{-6}, 7.5 \times 10^{-6}, 10^{-5}]$, as Adam requires smaller LR with respect to SGD.³² We chose $\text{LR} = 5 \times 10^{-6}$ as the best compromise for the LR; further, to delay overfitting and to narrow the gap between training and validation MCC, we added the dropout layers, with different dropout rates. The overall training performance with Adam and the optimal dropout rates are displayed in Figure 4. We fixed the number of epochs to 1500, e.g. a value large enough to ensure proper learning and small enough to prevent overfitting.

For the Antagonist task, which is highly unbalanced with less than 3% of positive labels in the training set (Table 1), we developed ML4Tox-Antagonist, a linear SVM with a regularization parameter $C = 0.001$ and balanced class weights. Linear, polynomial and Gaussian kernels were tested as possible kernels with a cross-validation strategy only on the training set.

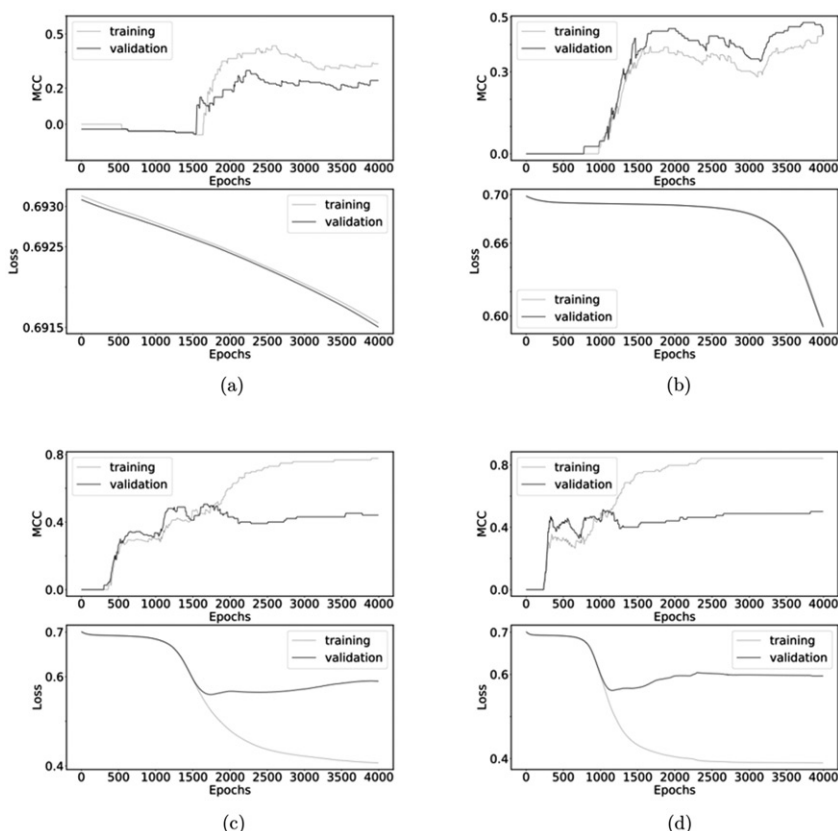


Figure 2. ML4Tox-Agonist: training with the SGD optimizer at different learning rates (LRs). (a) LR = 0.0001; (b) LR = 0.0005; (c) LR = 0.001; (d) LR = 0.002. Upper panels: training (grey) and validation (black) MCC; lower panels: training and validation losses (cross-entropy).

Details on the SVM metaparameter selection are not reported here for brevity. For the Binding Task, we designed ML4Tox-Binding, a Gaussian kernel SVM trained on Mold2 and ECF features used alone or in combination (“Mold2 + ECF”). The SVM regularization parameter was $C = 10$ for Mold2 and Mold2 + ECF, and $C = 100$ for ECF features; the kernel coefficient $\gamma = \frac{1}{n_f}$, for n_f the number of features. The optimal SVM regularization parameter was chosen from the grid [0.001, 0.01, 0.1, 1, 10, 100, 1000].

Predictive modeling strategy

All machine learning models were trained and evaluated within a Data Analysis Protocol (DAP) workflow. The protocol had been previously developed by FBK within the MAQC-II and SEQC challenges,^{16,17} the U.S. FDA initiatives for reproducibility of biomarkers from microarray and sequencing expression studies. Briefly, given a dataset split in TR and EV portions, the former undergoes 10 rounds of a fivefold stratified

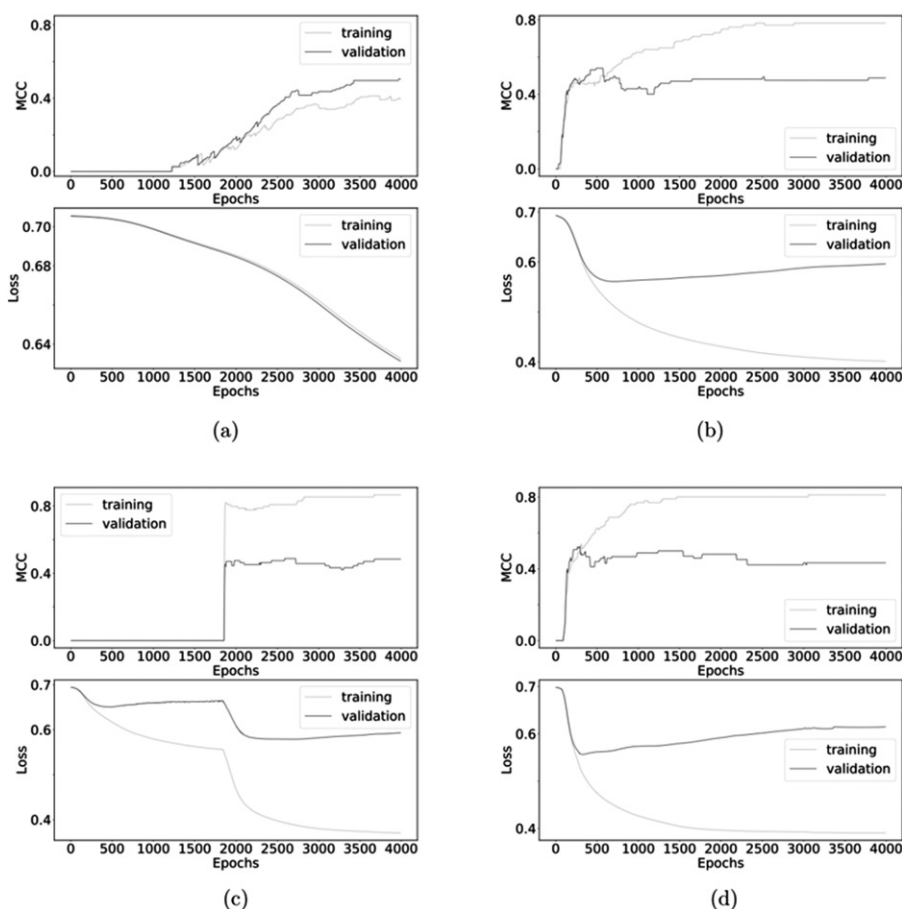


Figure 3. ML4Tox-Agonist: training with the Adam optimizer at different learning rates (LRs). (a) $LR = 10^{-6}$; (b) $LR = 5 \times 10^{-6}$; (c) $LR = 7.5 \times 10^{-6}$; (d) $LR = 10^{-5}$. Upper panels: training (grey) and validation (black) MCC; lower panels: training and validation losses (cross-entropy).

Cross-Validation (CV); models are trained and applied to the data splits, resulting in a set of metrics, including balanced accuracy (BA),³³ sensitivity (SN), specificity (SP), and the Matthews Correlation Coefficient (MCC). In the binary case, BA and MCC are defined as $BA = \frac{1}{2} \left(\frac{TP}{AP} + \frac{TN}{AN} \right)$ and $MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$, respectively, for TN, TP, FN, FP the entries of the binary confusion matrix and $AP = TP + FN$, $AN = TN + FP$ (TN: true negatives; TP: true positives; FN: false negatives; FP: false positives). Features were rescaled in the interval $[0, 1]$ before undergoing classification: to avoid information leakage, rescaling parameters from TR were used for rescaling both TR and EV sets. Random label experiments were also run to test against selection bias. After obtaining CV performance estimates, models were retrained on the whole TR set and evaluated on the EV set.

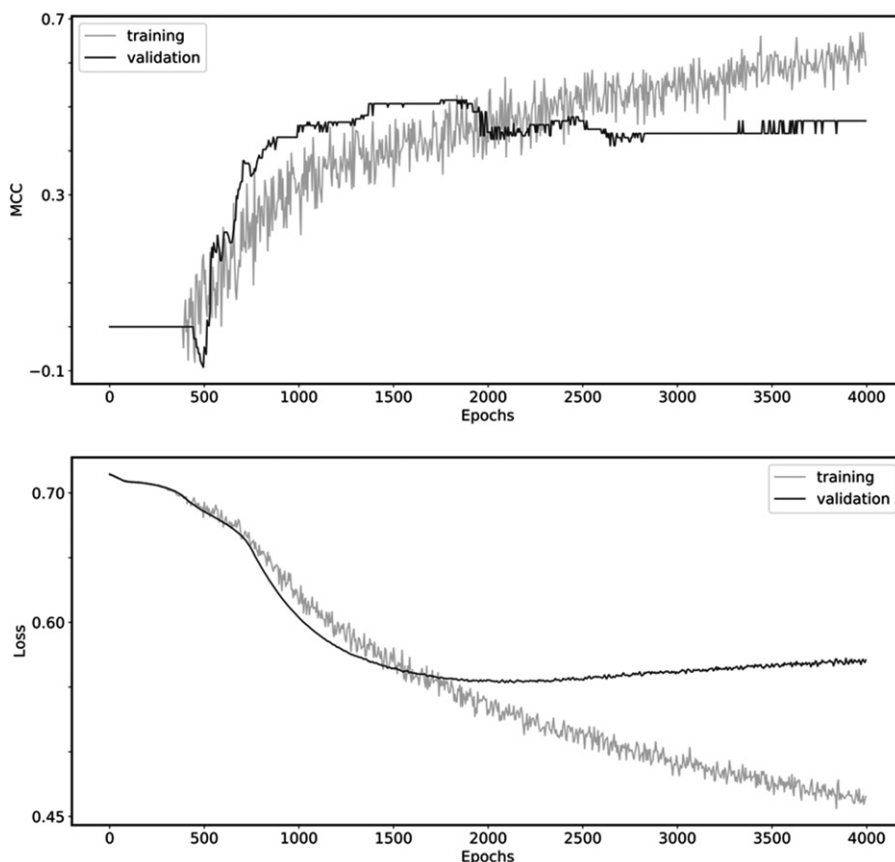


Figure 4. ML4Tox-Agonist model: training convergence curves. Optimizer = Adam; $LR = 5 \times 10^{-6}$ and dropout rate = 0.5 (inner layers 1–4) and 0.25 (layer 5).

Computational details

The ML4Tox-Agonist models were implemented in PyTorch v0.4.1,³⁴ with code available at <https://gitlab.fbk.eu/toxpred/DL4Tox>; ML4Tox-Antagonist and ML4Tox-Binding were built on top of the scikit-learn v0.19.1³⁵ Python library. The whole DAP was written in Python based on scikit-learn functions. Computations were run on the FBK KORE high-performance computing cluster for shallow learning models, and on a Linux workstation with $2 \times$ NVIDIA GeForce GTX 1080 cards and on a cloud instance with $4 \times$ NVIDIA Tesla K80 GPU cards, funded by the Azure Research grant “Deep Learning for Precision Medicine”.

Results and discussion

After filtering out descriptors with constant values across training samples, a total of 24288 features were kept (644 Mold2 features, 23644 ECF features) for model selection, training and evaluation as described above.

Table 3. Classification performance of ML4Tox models in cross-validation on the training set, and in evaluation.

Task	Feature set	MCC (CI)	BA (CI)	SN (CI)	SP (CI)	MCC _{EV}	BA _{EV}	SN _{EV}	SP _{EV}
Agonist	M2	0.32 (0.29, 0.35)	0.71 (0.69, 0.73)	0.60 (0.54, 0.65)	0.82 (0.80, 0.83)	0.22	0.73	0.78	0.68
Antagonist	M2	0.07 (0.05, 0.08)	0.60 (0.58, 0.62)	0.49 (0.44, 0.54)	0.70 (0.69, 0.72)	0.19	0.71	0.69	0.73
Binding	M2	0.37 (0.36, 0.39)	0.74 (0.73, 0.75)	0.70 (0.68, 0.72)	0.79 (0.78, 0.80)	0.22	0.60	0.37	0.84
Binding	ECF	0.42 (0.40, 0.44)	0.67 (0.66, 0.68)	0.38 (0.36, 0.40)	0.96 (0.96, 0.96)	0.25	0.59	0.23	0.94
Binding	M2 + ECF	0.24 (0.23, 0.26)	0.67 (0.66, 0.68)	0.76 (0.74, 0.79)	0.59 (0.57, 0.60)	0.24	0.64	0.66	0.61

CI: 95% studentized bootstrap confidence interval; MCC: Matthews correlation coefficient; BA: balanced accuracy; SN: sensitivity; SP: specificity. EV subscript indicates performance on the independent evaluation set. M2: Mold2 feature set.

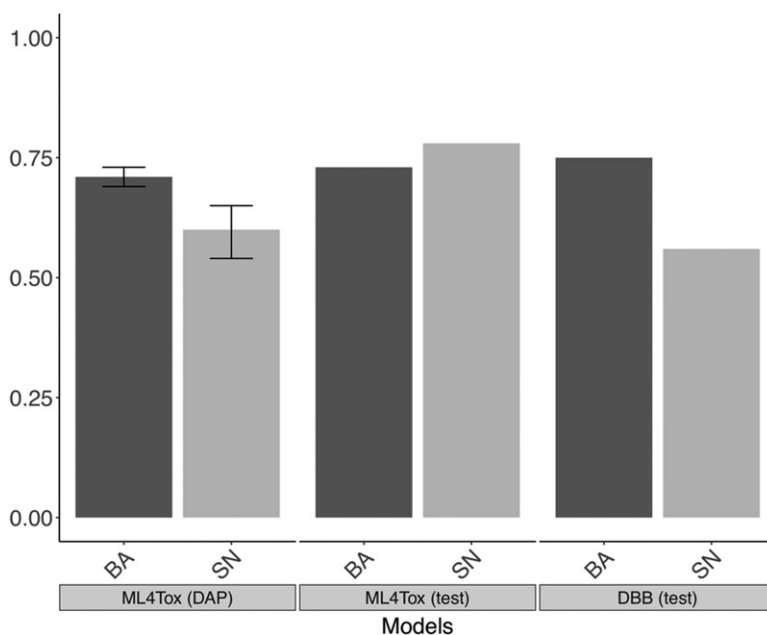


Figure 5. Agonist prediction task. ML4Tox-Agonist classification metrics in cross-validation on training ("DAP", leftmost panel) and on independent evaluation set ("test", middle panel), compared with evaluation metrics of the FDA-NCTR DBB model (rightmost panel). Vertical bars represent 95% studentized confidence intervals. BA: balanced accuracy; SN: sensitivity.

Classification results by ML4Tox on the CERAPP tasks are reported in Table 3, for Training (CV estimates) and Literature data. We compared ML4Tox with the FDA_NCTR_DBB model (DBB), which was among the best performing models on the same CERAPP tasks and datasets²¹ (see Figures 5–7).

On the Agonist task, the deep learning ML4Tox model scored a balanced accuracy $BA = 0.71$ (CI: 0.69, 0.73; 95% studentized bootstrap confidence interval) on Training and $BA_{EV} = 0.73$ on Evaluation. Notably, the model had fair sensitivity both in training $SN = 0.60$ (CI: 0.54, 0.65) and evaluation $SN_{EV} = 0.78$, improving on the original DBB model ($SN_{EV}^{DBB} = 0.56$), as shown in Figure 5.

On the Antagonist task, with $BA_{EV} = 0.71$ ($SN_{EV} = 0.69$, $SP_{EV} = 0.73$) in evaluation, the ML4Tox significantly improved over DBB both for balanced accuracy and sensitivity: $BA_{EV}^{DBB} = 0.55$ ($SN_{EV}^{DBB} = 0.11$, $SP_{EV}^{DBB} = 0.98$), as shown in Figure 6.

On the binding task, the most accurate ML4Tox model was Gaussian SVM over combined Mold2 and ECF features (M2 + ECF), with $BA = 0.67$ (CI: 0.66, 0.68) and $SN = 0.76$ (CI: 0.74, 0.79) in cross-validation, and $BA_{EV} = 0.64$ ($SN_{EV} = 0.66$, $SP_{EV} = 0.61$), also significantly improved the tradeoff between sensitivity and specificity of the DBB model ($BA_{EV}^{DBB} = 0.60$, $SN_{EV}^{DBB} = 0.26$,

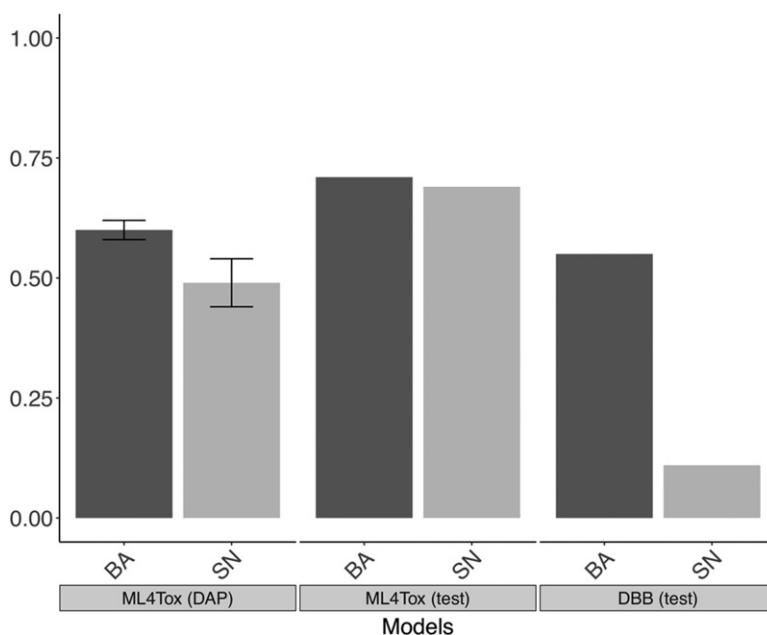


Figure 6. Antagonist prediction task. ML4Tox-Antagonist classification metrics in cross-validation (“DAP”, left- most panel) and on independent evaluation set (“test”, middle panel), compared with evaluation set metrics by NCTR DBB model (rightmost panel). Vertical bars represent confidence intervals. BA: balanced accuracy; SN: sensitivity.

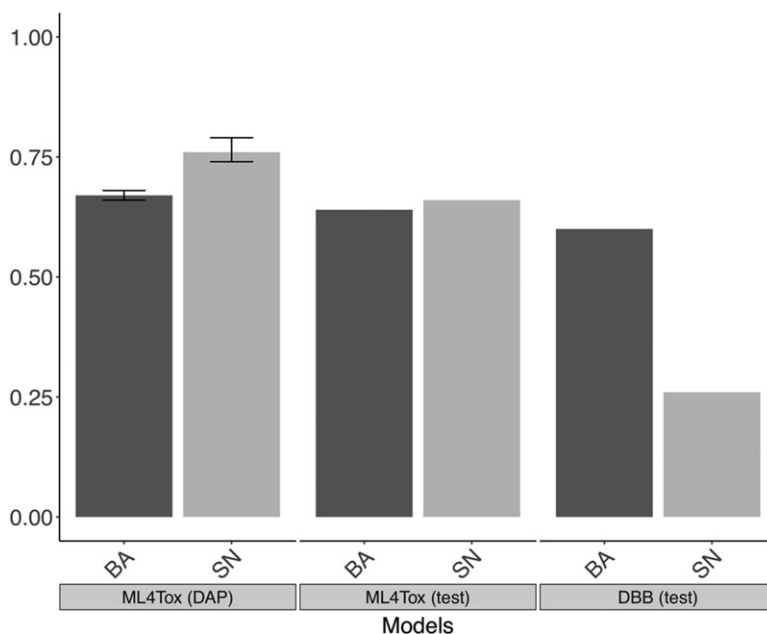


Figure 7. Binding prediction task. ML4Tox-Binding (Mold2 + ECF features) classification metrics in cross-validation (“DAP”, leftmost panel) and on independent evaluation set (“test”, middle panel), compared with evaluation set metrics by NCTR DBB model (rightmost panel). Vertical bars represent confidence intervals. BA: balanced accuracy; SN: sensitivity.

$SP_{EV}^{DBB} = 0.94$), see Figure 7. Notably, the improvement in sensitivity of ML4Tox-Binding model was obtained with the M2 + ECF feature combination.

The consistent gain in sensitivity with ML4Tox in the antagonist and binding tasks is important in the context of predictive toxicology, where the prioritization of chemicals for *in vivo* risk assessment is one of the main goals.

Conclusions

In this study, we introduced ML4Tox, a predictive toxicology computational framework for modeling the potential endocrine disruption of environmental chemicals, based on machine learning. Our approach is motivated by the recent availability of QSAR-ready datasets from the literature.

We demonstrated ML4Tox by developing Deep Learning and Support Vector Machine models to predict chemical compound agonist, antagonist and binding activity for the human estrogen receptor ligand-binding domain, using data sourced from the CERAPP initiative.

We have explored the use of Deep Learning models for the first task (Agonist), and of shallow methods on the other two (Antagonist and Binding). We also test the potential of a different class of molecular descriptors (circular topological fingerprints) in the Binding task. While the sensitivity of these methods remains fair at best, all three ML4Tox models improve over published results (Agonist: 0.78 vs 0.56; Antagonist: 0.69 vs 0.11; Binding: 0.66 vs 0.26).

We did not yet test the use of the richer set of features for predictive toxicology with deep architectures: combining the two improvements is expected to provide an advantage. In general, the recent availability of curated datasets has been used until now only with shallow statistical machine learning models¹⁴ and thus it opens new potential applications for deep learning. In particular, as already proposed in the combination of diagnostic–prognostic tasks,³⁶ we aim to exploit multi-task learning architectures to simultaneously solve the agonist, antagonist and binding tasks. The hypothesis that supports multi-task architectures in predictive toxicology is that when a shared core structure is trained to target several tasks simultaneously, each covering different aspects of toxicity, the model may be driven to better describe pathway disruption, thus improving its potential for prioritizing chemicals.

Acknowledgments

The authors would like to thank Luca Coviello (FBK and Université Nice Sophia Antipolis) for useful technical comments on deep learning models. They particularly thank Drs.

Huixiao Hong and Bohu Pan (FDA/NCTR) for kindly providing the Mold2-preprocessed CERAPP training and evaluation data sets. The Microsoft Azure resources used in training models was funded by the Azure Research grant “Deep Learning for Precision Medicine”, endowed to CF.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The authors gratefully acknowledge financial support of the FBK institutional program for Data Science (Big Data Analytics 2018).

References

1. Royal Chemical Society. Royal Chemical Society view on predictive toxicology, 2016. [cited 2018 Sept 29]. Available from: http://www.rsc.org/images/predictive-toxicology-position-consultation_tcm18-223078.pdf.
2. Raies AB, Bajic VB. *In silico* toxicology: computational methods for the prediction of chemical toxicity. *WIREs Comput Mol Sci*. 2016;6(2):147–172.
3. Reisfeld B, Mayeno AN. What is Computational Toxicology? In: Reisfeld B., Mayeno A. eds., *Computational Toxicology. Methods in Molecular Biology (Methods and Protocols)*, vol 929. Humana Press, Totowa, NJ; 2012: 3–7
4. Nelms MD, Mellor CL, Enoch SJ, et al. A mechanistic framework for integrating chemical structure and high-throughput screening results to improve toxicity predictions. *Comput Toxicol*. 2018;8:1–12.
5. Malloy T, Beryt E. Leveraging the new predictive toxicology paradigm: alternative testing strategies in regulatory decision-making. *Environ Sci Nano*. 2016;3(6):1380–1395.
6. Yang RS, Thomas RS, Gustafson DL, et al. Approaches to developing alternative and predictive toxicology based on PBPK/PD and QSAR modeling. *EnvironHealth Persp*. 1998;106(Suppl 6):1385–1393.
7. U.S. Food and Drug Administration. 2017. FDA’s predictive toxicology roadmap. [cited 2018 Sept 29]. Available from: <https://www.fda.gov/downloads/scienceresearch/specialtopics/regulatoryscience/ucm587831.pdf>.
8. Fera Science Limited, London, UK. 2018. In Silico Predictive Toxicology (INSPECT). Available from: <https://www.fera.co.uk/chemical-regulation/consultancy/in-silico-predictive-toxicology>.
9. Yang H, Sun L, Li W, Liu G, Tang Y. *In Silico* prediction of chemical toxicity for drug design using machine learning methods and structural alerts. *Front Chem*. 2018;6:30.
10. Ekins S, Clark AM, Perryman AL, Freundlich JS, Korotcov A, Tkachenko V. Accessible machine learning approaches for toxicology. 1st ed. In: Ekins S, ed., *Computational toxicology: risk assessment for chemicals*. John Wiley & Sons, Inc: Wiley, NJ; 2018: 1–29.
11. Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: toxicity prediction using deep learning. *Front Environ Sci*. 2016;3:80.

12. Baskin II. Machine Learning Methods in Computational Toxicology. In: Nicolotti O. ed., *Computational Toxicology. Methods in Molecular Biology*, vol 1800. Humana Press: New York, NY; **2018**: 119–139.
13. Parthasarathi A, Dhawan A. *In silico* approaches for predictive toxicology. In: Dhawan A, Kwon S, editors. *In vitro toxicology*. Academic Press, **2018**. Ch. 5, p. 91–109.
14. Luechtefeld T, Rowlands C, Hartung T. Big-data and machine learning to revamp computational toxicology and its use in risk assessment. *Toxicol Res.* **2018**;7(5): 732–744.
15. Luechtefeld T, Marsh D, Rowlands C, Hartung T. Machine learning of toxicological big data enables read- across structure activity relationships (RASAR) outperforming animal test reproducibility. *Toxicol Sci.* **2018**;165(1):198–212.
16. The MicroArray Quality Control (MAQC) Consortium. The MAQC-II Project: A comprehensive study of common practices for the development and validation of microarray-based predictive models. *Nat Biotechnol.* **2010**;28(8):827–838.
17. The SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequence Quality Control consortium. *Nat Biotechnol.* **2014**;32:903–914.
18. Matthews B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta.* **1975**;405(2):442–451.
19. Baldi P, Brunak S, Chauvin Y, et al. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics.* **2000**;16(5):412–424.
20. Jurman G, Riccadonna S, Furlanello C. A comparison of MCC and CEN error measures in multi-class prediction. *PLoS One.* **2012**;7(8):e41882
21. Mansouri K, Abdelaziz A, Rybacka A, et al. CERAPP: collaborative estrogen receptor activity prediction project. *Environ Health Persp.* **2016**;124(7):1023.
22. Burgoon LD. Autoencoder Predicting Estrogenic Chemical Substances (APECS): an improved approach for screening potentially estrogenic chemicals using in vitro assays and deep learning. *Comput Toxicol.* **2017**;2:45–49.
23. Dix DJ, Houck KA, Martin MT, Richard AM, Setzer RW, Kavlock RJ. The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicol Sci.* **2007**;95(1):5–12.
24. Huang R, Sakamuru S, Martin MT, et al. Profiling of the Tox21 10K compound library for agonists and antagonists of the estrogen receptor alpha signaling pathway. *Sci Rep.* **2014**;4:5664.
25. Hong H, Slavov S, Ge W, et al. Mold2 molecular descriptors for QSAR. In: *Statistical modelling of molecular descriptors in QSAR/QSPR*. Wiley-Blackwell; **2012**. Ch. 3, p. 65–109.
26. Hong H, Xie Q, Ge W, et al. Mold(2), molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *J Chem Inf Model.* **2008**;48(7):1337–1344.
27. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inform Model.* **2010**; 50(5):742–754.
28. Morgan H. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *J Chem Doc.* **1965**;5(2):107–113.
29. DeepChem Team. DeepChem: Democratizing deep-learning for drug discovery, quantum chemistry, materials science and biology, **2016**. [cited 2018 Sept 29]. Available from: <https://github.com/deepchem/deepchem>, accessed: 2018-08-29.
30. RDKit Team. RDKit: Open-source cheminformatics, 2018. [cited 2018 Sept 29]. Retrieved from: <http://www.rdkit.org>.

31. Nair V, Hinton GE. Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10), 2010. p. 807–814.
32. Kinga D, Adam JB. A method for stochastic optimization. In: Proceedings of the 3rd international conference on learning representations (ICLR-15). 2015, arXiv:1412.6980.
33. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The balanced accuracy and its posterior distribution. In: Proceedings of the 20th International Conference on Pattern Recognition (ICPR-10), IEEE; 2010. p. 3121–3124.
34. Paszke A, Gross S, Chintala S, et al. Automatic differentiation in PyTorch; 2017. Available at: <https://openreview.net/pdf?id=BJJsrmfCZ>, NIPS 2017 Autodiff Workshop.
35. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12:2825–2830.
36. Maggio V, Chierici M, Jurman G, Furlanello C. A multiobjective deep learning approach for predictive classification in neuroblastoma, arXiv preprint arXiv:1711.08198.