LightGBM: An Effective and Scalable Algorithm for Prediction of Chemical Toxicity – Application to the Tox21 and Mutagenicity Datasets

Jin Zhang[1], *Daniel Mucs*[2], *Ulf Norinder*[2,3], *Fredrik Svensson*[4,5]*

1. Department of Chemistry, Umeå University, SE-901 87 Umeå, Sweden

2. Swetox, Unit of Toxicology Sciences, Karolinska Institutet, Forskargatan 20, SE-151 36 Södertälje, Sweden

3. Department of Computer and Systems Sciences, Stockholm University, Box 7003, SE-164 07 Kista, Sweden

4. The Alzheimer's Research UK University College London Drug Discovery Institute, The Cruciform Building, Gower Street, London, WC1E 6BT, UK

5. The Francis Crick Institute, 1 Midland Road, London, NW1 1AT, UK

Corresponding author: Email: f.svensson@ucl.ac.uk, Phone: +44 (0)20 7679 0811

Abstract

Machine learning algorithms have attained widespread use in assessing the potential toxicities of pharmaceuticals and industrial chemicals because of their faster-speed and lower-cost compared to experimental bioassays. Gradient boosting is an effective algorithm that often achieves high predictivity, but historically the relative long computational time limited its applications in predicting large compound libraries or developing *in silico* predictive models that require frequent retraining. LightGBM, a recent improvement of the gradient boosting algorithm inherited its high predictivity but resolved its scalability and long computational time by adopting leaf-wise tree growth strategy and introducing novel techniques. In this study, we compared the predictive performance and the computational time of LightGBM to deep neural networks, random forests, support vector machines, and XGBoost. All algorithms were rigorously evaluated on publicly available Tox21 and mutagenicity datasets using a Bayesian optimization integrated nested 10-fold cross-validation scheme that performs hyperparameter optimization while examining model generalizability and transferability to new data. The evaluation results demonstrated that LightGBM is an effective and highly scalable algorithm offering the best predictive performance while consuming significantly shorter computational time than the other investigated algorithms across all Tox21 and mutagenicity datasets. We recommend LightGBM for applications in *in silico* safety assessment and also in other areas of cheminformatics to fulfill the ever-growing demand for accurate and rapid prediction of various toxicity or activity related endpoints of large compound libraries present in the pharmaceutical and chemical industry.

Introduction

Early prediction of compound toxicity accelerates drug discovery and reduces the safety-related attritions during drug development.[1] One of the predominant approaches for such predictions is machine learning (ML). In this setting, ML uses statistical algorithms to summarize the historic experimental data and predict potential toxicities for the new compounds, as illustrated by examples published in recent literature.[2] Remarkably, in some applications, ML even outperforms animal tests.[3]

For toxicity predictions, as with all ML applications, the operator has to make a decision on which algorithm to deploy. Many different ML techniques have been employed successfully for the prediction of toxicity and there is currently no technique clearly outperforming the others.[4] This is in part due to the high number of different scenarios that can be encountered in toxicity predictions, from small focused datasets to "big data" collected from diverse sources and from specific molecular interactions to organism level toxicities. It is therefore necessary that the computational scientists' toolbox is diverse and emerging methods can be valuable additions to this.

Gradient boosting[5] (GB) is a powerful ML algorithm that has seen multiple uses for toxicity predictions.[6,7] Although capable of generating highly predictive models, the main limitations with GB are the unsatisfactory long training time and scalability.[8] This challenges its application to ever-growing compound datasets with high feature dimensions or its deployment in the drug discovery environment requiring regular retraining of the models.

LightGBM,[8] a recent modified GB algorithm, tackles these limitations by adopting a leaf-wise tree growth strategy and introducing novel techniques, e.g. gradient-based one-side sampling and exclusive feature bundling. This approach results in a faster and less resource intensive implementation of GB suitable for frequent retraining and rapid assessment of larger high-dimensional datasets. LightGBM has been demonstrated to be up to 20 times faster to

train on the same data,[8] compared to the XGBoost[9] implementation of GB. The algorithm has been implemented successfully on issuing peer-to-peer loan in FinTech industry[10] and on forecasting wind power production in smart grid industry[11].

ML model evaluation and selection strategies for cheminformatics applications require the judicial use of both validation and test data, which has been highlighted previously by Tropsha et al.[12] Not only is it important to establish the accuracy of a new algorithm, but also its robustness, transferability, and ease of deployment are important parameters to evaluate. This is a key aspect as methods are often chosen based on their performance on one set and then expected to deliver the same level of performance when applied to new data. Furthermore, factors like the random partitioning of data might influence the results, something that can be counteracted by training multiple models on the same data but using different train and test splits. Nested cross validation strategies have been proposed to provide more robust and generalized evaluation of the model performance.[13] The inner cross validation is used to train the model and tune the model hyperparameter parameters, while the outer cross validation is used to evaluate general performance of the model selected by the inner cross validation. Bayesian optimization is an efficient method for global optimization of the ML algorithm hyperparameters as the method converges faster and requires fewer iterations for hyperparameter tuning than both grid search and random search.[14]

The Tox21 and mutagenicity datasets are two compound datasets commonly used for *in silico* toxicity model development and comparison.[15–18] The Tox21 datasets[19] include the *in vitro* toxicity screening results of approximately 10,000 compounds against a total of 12 Nuclear Receptor (NR) and Stress Response (SR) targets. The mutagenicity dataset published by Hansen et al.[20] contains screening results of approximately 6,500 compounds in the Ames bacteria mutagenicity test that measures if the tested compounds cause mutations in the DNA of the test microbial organism.

In this study, we evaluate the performance of LightGBM algorithm on classification of compound toxicity against a collection of toxicologically relevant endpoints based on the Tox21 and mutagenicity datasets. We compare its predictive performance and the computation time to that of the closely related gradient boosting algorithm XGBoost and three other well-established ML algorithms, deep feedforward neural network (DNN), random forest (RF), and support vector machine classifier (SVC). We also discuss the advantages of Bayesian optimization integrated nested cross validations in proper validation of new ML methods.

Materials and Methods

Computation. The computations were performed in Python v2.7.12 using one 28-thread Intel Xeon E5-2690v4 CPU on a Linux server with 128Gb memory. The following Python packages were installed for the calculations: Keras v1.2.1[21], LightGBM v2.1.0[8], Scikit-learn v0.18.1[22], Scikit-optimize v0.4[23], Tensorflow v0.12.1[24], and XGBoost v0.8.0[9].

Compound Datasets and Features. We downloaded the compounds of the Tox 21[25] and mutagenicity datasets[20] along with their associated activities for toxicological endpoints. The number of active and inactive compounds in each dataset and the descriptions of the assay targets are shown in Table 1. The compound structures were standardized using the IMI eTOX project standardizer[26] in combination with tautomer standardization using the MolVS standardizer[27]. RDKit[28] molecular descriptors and Morgan fingerprints were calculated for all compounds. These two feature sets are referred to in this study as “molecular descriptors” and “fingerprints” respectively. The molecular descriptor set consists of 97 features describing the structural and physicochemical properties of the compounds, e.g. the number of rings, topological polar surface area, and lipophilicity. The Morgan fingerprint is a reimplementation of the extended-connectivity fingerprint (ECFP)[29]. The method generated the fingerprints by parsing each compound atom and obtain all possible paths through this atom with a predefined

radius. Each unique path is hashed to predefined maximum number of bits. Here we set radius=4 when generating the fingerprints and hashed them to 1024 bits. The same structure preparation, descriptors/fingerprints generation, and feature/class preparation protocol have been used for classification problems with good performance in previous studies.[30,31]

Table 1. Number of active and inactive compounds in each the Tox21 and mutagenicity datasets and the target and assay information.

Dataset	Target/Assay	Number of active compounds	Number of inactive compounds
Tox21 Datasets			
nr-ahr	Aryl hydrocarbon receptor	942	7,103
nr-ar	Androgen Receptor	376	8,843
nr-ar-lbd	Androgen receptor (luciferase assay)	302	8,174
nr-aromatase	Aromatase	346	6,759
nr-er	Estrogen receptor	927	6,665
nr-er-lbd	Estrogen receptor (luciferase assay)	441	8,187
nr-ppar-gamma	Peroxisome proliferator-activated receptor gamma	219	7,848
sr-are	Nuclear factor (erythroid-derived 2)-like 2 antioxidant responsive element	1,078	6,003
sr-atad5	Genotoxicity indicated by ATAD5	334	8,628
sr-hse	Heat shock factor response element	419	7,635
sr-mmp	Mitochondrial membrane potential	1,127	6,096
sr-p53	DNA damage p53-pathway	528	7,981

Mutagenicity Data				
Hansen et al	Ames (mutagenicity)	test	3,502	3,007

The molecular descriptors were scaled using the scikit-learn MinMaxScaler to a range between 0 and 1. The datasets were divided into active and inactive and this investigation was accordingly formulated as a binary classification problem. Tox21 datasets are imbalanced with regards to active and inactive compound classes. Class weights of each dataset were calculated under the 'balanced' setting using the scikit-learn package and applied to penalize the ML algorithms for misclassification of the minority class to achieve balanced prediction results.

Machine learning algorithms and modeling scheme. Classification models based on Tox21 and mutagenicity datasets were developed using LightGBM and four regularly used algorithms, namely DNN, SVC, RF, and XGBoost. The models were optimized and extensively evaluated with the following modeling scheme illustrated in Figure 1.

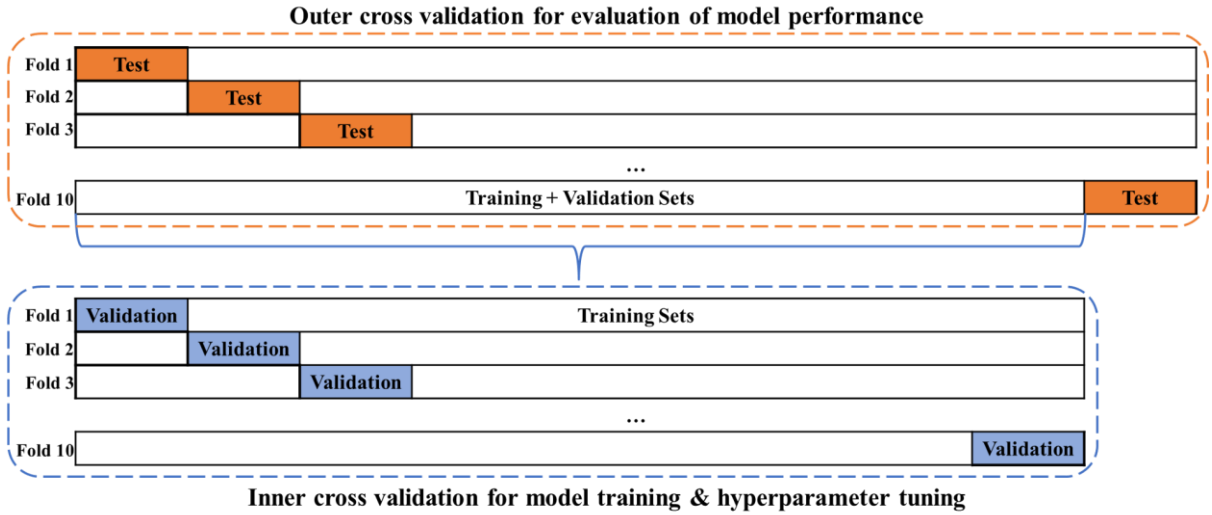

Figure 1. Schematic illustrating how the inner and outer cross validation loops were used to tune the model parameters and perform the predictions on the validation and test data.

The modeling scheme was designed to integrate nested cross validation and Bayesian optimization strategies because of the following reasons: 1) nested cross validation reduces bias in model performance evaluation and provides a more robust mean for assessing the model transferability on 'unseen' data, comparing to naïve cross validation;[32] 2) Bayesian

optimization has been suggested as a recommended method for hyperparameter tuning, as it achieves better performance on the test set while requiring fewer iterations than grid search and random search.[33]

Each dataset was split into inner and outer sets using a nested 10-fold cross validation split setting in the scheme. Inner cross validation sets were used to train the models based on the selected ML algorithms, decide the best set of hyperparameters achieving highest balanced accuracy for the models and to perform initial evaluation of model performance (referred as 'validation results'). Hyperparameter selection was performed using 100 iterations of Bayesian optimization with Gaussian processes as surrogate model and expected improvement as acquisition function from the scikit-optimize package. The generalization performance and transferability of the models were further evaluated on the outer cross validation sets (referred as 'test results').

LightGBM

LightGBM is a recent modification of the GB algorithm. It improves the efficiency and scalability of the algorithm without sacrificing its inherited effective performance. Seven hyperparameters governing the performance of the LightGBM classifier were optimized within the following predefined ranges suggested by the package manual: number of leaves ('num_leaves', 30-500), number of feature bins ('max_bin', 250-500), number of iterations ('n_estimators', 50-900), number of samples in one leaf ('min_child_samples', 30-500), model depth ('max_depth', 5-12), learning rate ('learning_rate', 10^{-6}-10^{-1}), and fraction of data used for training ('bagging_fraction', 0.8).

XGBoost

XGBoost is another variant of the GB algorithm. It was designed to be highly scalable by adopting a sparsity-aware algorithm for sparse data and weighted quantile sketch for approximate tree learning.[9] We implemented the XGBoost classifier with the fast histogram

optimized approximate greedy algorithm (tree_method=hist) and optimized the following hyperparameters to achieve a fair comparison between the XGBoost and LightGBM algorithms; number of iterations ('n_estimators', 50-900), model depth ('max_depth', 5-12), and learning rate ('learning_rate', 10^{-6}-10^{-1}).

Deep Neural Networks

Deep feedforward neural network has been identified as the 'most effective' classification algorithm during Merck Kaggle contest[34] and Tox21 challenge[15]. We implemented a three-hidden-layer DNN classification model using Keras with the Tensorflow backend. The model hyperparameters were optimized within the predefined range applied in Korotcov et al.[35]: number of hidden units (200-2,000), epoch (20-100), learning rate (10^{-4}-10^{-2}), size of mini batch size (128), initial weight (random normal), optimization algorithm (Adam), activation function (relu for hidden layers; softmax for output layer), dropout rates for hidden layers (layer 1: 0.25, layer 2: 0.25, layer 3: 0.1).

Support Vector Machine Classifier

SVC performs classification by defining an optimal hyperplane that maximizes the margin between classes.[36,37] Non-linear classification is achieved by transforming the data into a higher dimensional feature space using non-linear kernel function ('kernel trick') and then performing linear separation. LibSVM with radial basis function kernel from scikit-learn was used to develop the classification models with hyperparameters optimized within the predefined ranges suggested by Alvarsson et al.[16]: gamma (10^{-5}-10), and C value:(1-1200).

Random Forest

RF is an ensemble algorithm bagging the results of decision tree classifiers built on subsets of the data.[38] RF models (number of trees: 100-1,000) were developed as baseline models to be compared with the corresponding models derived by the algorithms mentioned before.

The list of optimized hyperparameters for each ML algorithm and their value ranges are provided in Table S1.

Performance evaluation. The performance of the investigated algorithms was evaluated based on their balanced accuracy (BA), Cohen's kappa (Kappa), Matthews correlation coefficient (MCC), positive predictive value (PPV), negative predictive value (NPV), ROC-area under curve (AUC), sensitivity (SE) and specificity (SP) values. The equations of these performance metrics are given in the supporting information. Balanced accuracy was chosen as the scoring function for the Bayesian optimization algorithm since it provides good estimation of the model performance on imbalanced datasets.[39] The computation time of the investigated ML algorithms was also recorded to indicate the speed of each method.

Statistical hypothesis testing using Bonferroni correction was performed in order to compare the differences among algorithms and select the best performing one with the highest balanced accuracy values. The compute times were compared using Wilcoxon Signed-Rank Test corrected for multiple testing using Bonferroni correction (significance level = 0.05).

The percentage of compounds in the test sets outside the model applicability domain was assessed by performing random projection (known for preserving inter-object distances) with a 10-fold reduction in dimensionality for the all datasets.

Results and Discussions

The thirteen Tox21 and mutagenicity datasets were described using both molecular descriptors and fingerprints, resulting in a total of 26 different datasets for each ML algorithm. The model performance was evaluated using subsets from inner 10-fold CV (validation sets) and outer 10-fold CV (test sets). The balanced accuracy values obtained for the 26 different datasets using the five ML algorithms varied between about 0.6 and 0.876 (Table S2). The balanced accuracy values of models developed in this study are comparable to the previous Tox21 Challenge winning model results that reported a span of balanced accuracy between

0.68 to 0.9 for the datasets. Although, the evaluation used a specific leaderboard set not used in this study. The general trend shown in Figures 2, 3 and Table 2 is that LightGBM gave the best prediction for the Tox21 and mutagenicity datasets followed by XGBoost, SVC, DNN, and RF algorithms, when comparing their averaged evaluation metric values on the validation and test sets. We also examined the balance between sensitivity (SE) and specificity (SP) in the prediction of validation and test sets (Figure S1). Table 3 shows the number of instances each algorithm significantly outperformed the others (see Table S5 for p-values). The statistical testing demonstrated that LightGBM had significantly better performance on balanced accuracy than the other four algorithms in the majority of datasets.

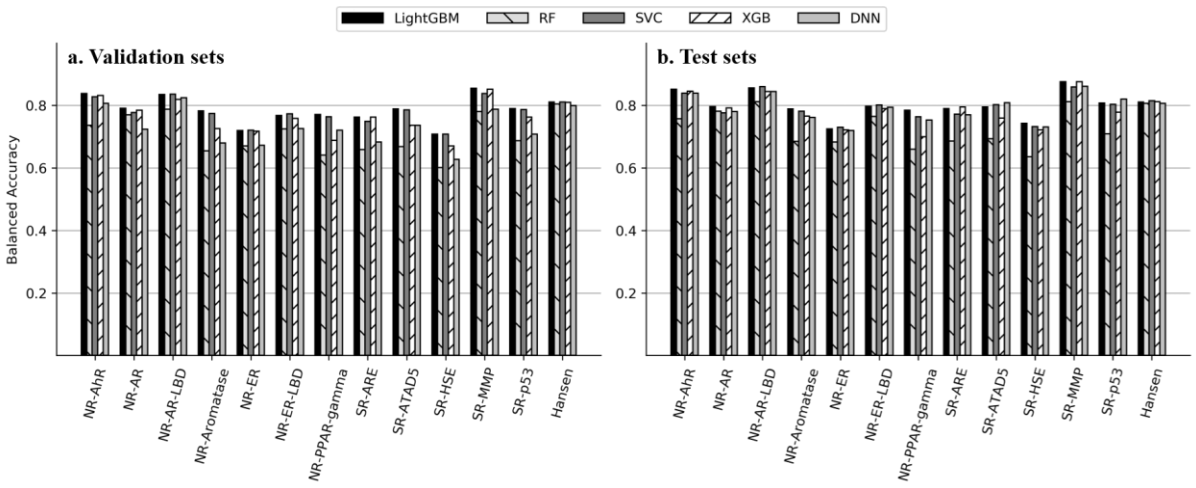

Figure 2. Balanced accuracy values of the five algorithms across validation and test subsets of the Tox21 and mutagenicity datasets described using RDKit molecular descriptors.

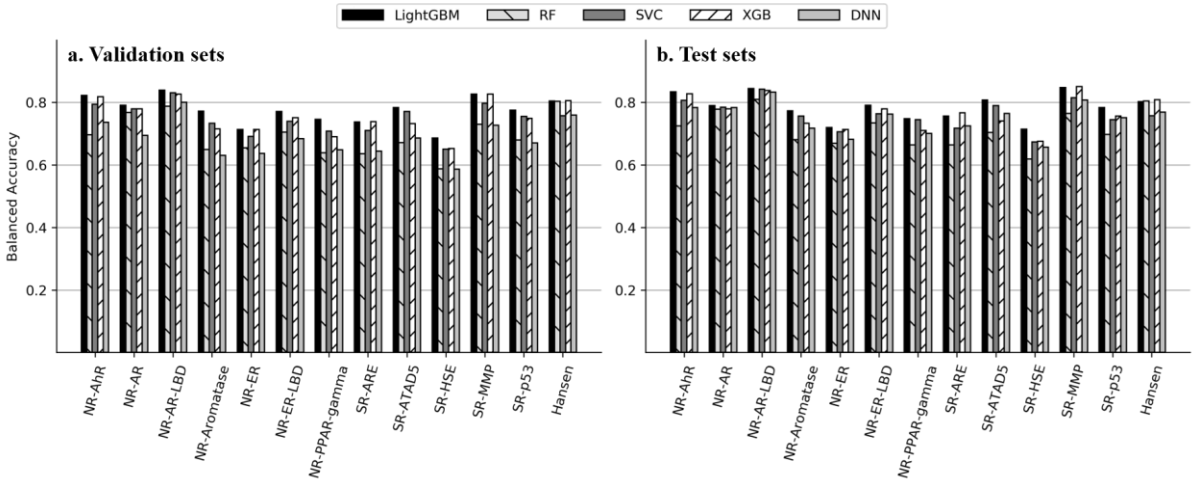

Figure 3. Balanced accuracy values of the five algorithms across validation and test subsets of the Tox21 and mutagenicity datasets described using Morgan fingerprints.

Table 2. Averaged values of area under curve (AUC), balanced accuracy (BA), Cohen's kappa (Kappa), Matthews correlation coefficient (MCC), sensitivity (SE) and specificity (SP) for the five ML algorithms on the validation (V) and test subsets (T) of the Tox21 and mutagenicity datasets described with RDKit molecular descriptors (MD) and Morgan fingerprints (FP).

		LightGBM		RF		SVC		XGB		DNN	
Metrics	Features	V	T	V	T	V	T	V	T	V	T
AUC	FP	0.836	**0.786** [a]	**0.859**	0.723	0.805	0.762	0.835	0.768	0.733	0.736
	MD	0.855	**0.800**	**0.865**	0.728	0.836	0.795	0.857	0.784	0.682	0.694
BA	FP	**0.775**	**0.786**	0.698	0.723	0.747	0.762	0.754	0.768	0.733	0.736
	MD	**0.784**	**0.800**	0.701	0.728	0.781	0.795	0.763	0.784	0.682	0.694
Kappa	FP	0.415	0.424	0.480	0.531	0.367	0.368	**0.504**	**0.534**	0.522	0.532
	MD	0.433	0.440	0.484	0.538	0.393	0.402	**0.494**	**0.524**	0.430	0.463
MCC	FP	0.442	0.448	0.513	**0.556**	0.394	0.396	0.519	0.541	**0.534**	0.544
	MD	0.461	0.467	**0.516**	**0.561**	0.428	0.436	0.511	0.532	0.461	0.491
SE	FP	**0.657**	**0.672**	0.424	0.474	0.618	0.642	0.572	0.595	0.508	0.512
	MD	0.674	0.700	0.430	0.484	**0.689**	**0.708**	0.598	0.636	0.398	0.421
SP	FP	0.892	0.900	**0.972**	**0.972**	0.876	0.881	0.936	0.941	0.958	0.960
	MD	0.893	0.900	**0.973**	**0.973**	0.873	0.882	0.927	0.932	0.966	0.968

[a.] The highest value in each metric for the validation and test sets were highlighted in bold.

Table 3. Statistical comparison results on the balanced accuracy values of the five ML algorithms on the test subsets of the Tox21 and mutagenicity datasets. The performance of each algorithm was compared to that of the other algorithm across the 13 datasets described with RDKit molecular descriptors and Morgan fingerprints.

Classification Algorithm	Significant better cases	Total significant cases	Total cases
LightGBM	76	79	104
SVC	60	87	104
XGB	34	61	104
DNN	1	59	104
RF	3	62	104

We also observed that certain Tox21 datasets (e.g. nr-er and sr-hse sets) were more challenging to predict. The LightGBM algorithm outperformed the other algorithms also in prediction of these challenging datasets. For all the ML algorithms in this study, the models built using molecular descriptors resulted in higher balanced accuracies compared to the corresponding models based on fingerprints (Figure 4).

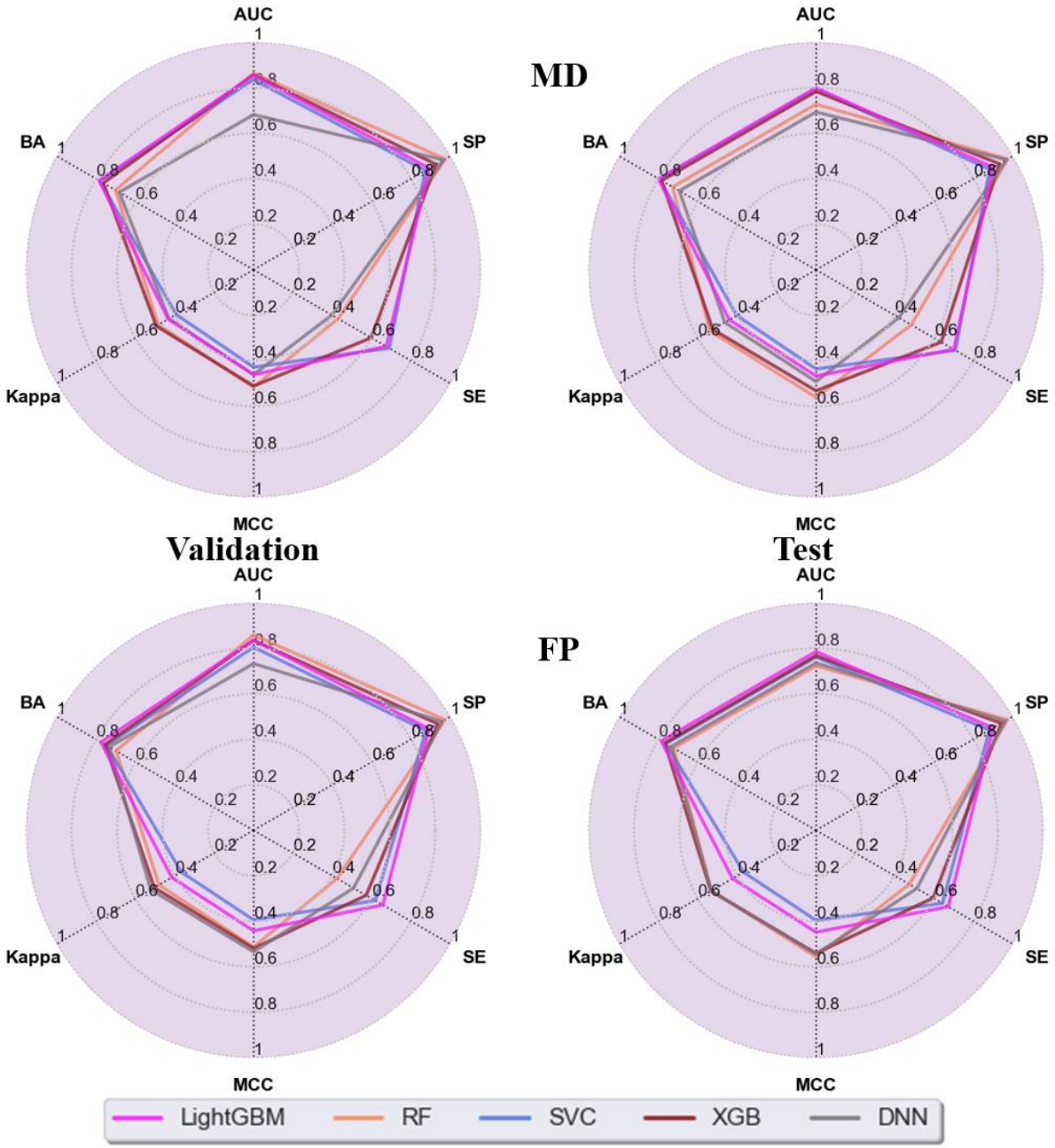

Figure 4. Average values of area under curve (AUC), balanced accuracy (BA), Cohen's kappa (Kappa), Matthews correlation coefficient (MCC), sensitivity (SE) and Specificity (SP) for the five algorithms on the validation and test subsets of the Tox21

and mutagenicity datasets described using RDKit molecular descriptors (MD) and Morgan fingerprints (FP), respectively.

In addition to the predictive performance, we also compared the total computational time consumed by the five algorithms on tuning their hyperparameters using Bayesian optimization and performing model evaluations using nested cross-validation (Figure 5). All differences between the algorithms with respect to compute times are statistically significant (significance level = 0.05) over the 13 endpoints except for RF versus XGBoost (fingerprints) and LightGBM versus SVC (molecular descriptors).

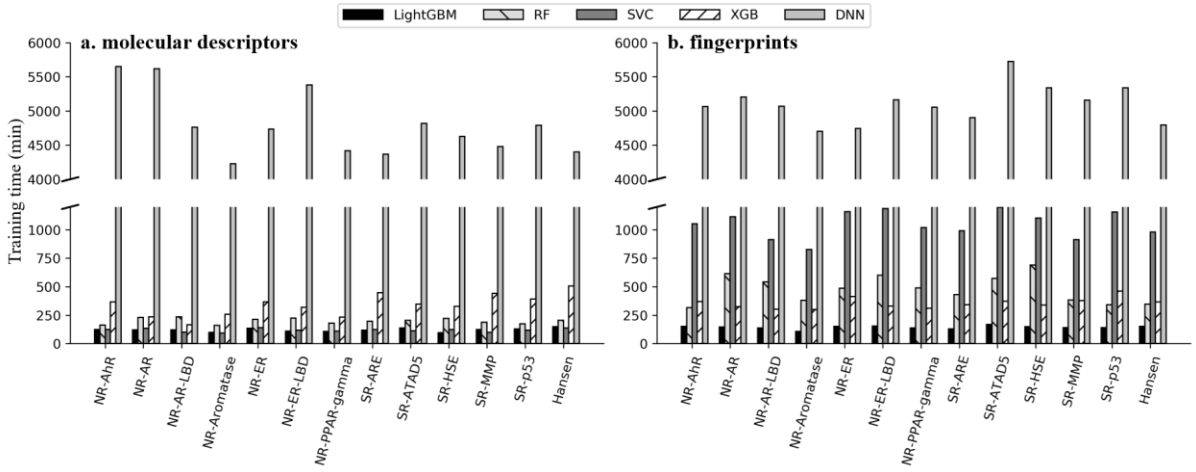

Figure 5. Computational time for the five algorithms across the Tox21 and mutagenicity datasets using molecular descriptors and fingerprints, respectively.

LightGBM was the fastest algorithm among the investigated algorithms as it consumed the shortest computation time for all datasets in this study (Table S3), whereas DNN was the most time-consuming algorithm, requiring, on average, approximately 37 times longer computational time than LightGBM due to large number of algorithm parameters (Table 4). We also observed that the investigated algorithms required more computational time to model higher dimensional fingerprints compared to molecular descriptors. Interestingly, gradient boosting algorithms, i.e. LightGBM and XGBoost, showed highly scalable characteristics, as they required similar amount of computation time in developing models using molecular descriptors and fingerprints. This is contrasted by the other three investigated algorithms, as

exemplified by SVC that required up to approximately nine times longer computational time to model fingerprint compared to molecular descriptor (Table 4 and Figure 5).

Table 4. The average computation time for the five ML algorithms on the Bayesian optimization integrated nested cross validation scheme of the Tox21 and mutagenicity datasets described with RDKit molecular descriptors and Morgan fingerprints.

	Average computation time (mins)				
Feature type	LightGBM	RF	SVC	XGB	DNN
Molecular descriptors	121	199	118	339	4,790
Fingerprints	144	476	1,047	354	5,096

Although recent trends in cheminformatics have seen a rise in more computationally intensive approaches by using different varieties of deep neural networks that have demonstrated their values in the area of image recognition, SMILES based sequence modeling, multi-tasking, and generative models,[40] there is still a need for algorithms that are fast to train and easy to deploy while delivering high and robust predictive performance in the classical CPU computation environment. In this study we have shown through a thorough evaluation procedure that LightGBM fulfills these criteria and reliably outperformed the DNN algorithm in terms of both balanced accuracy as well as computation time, which emphasizes LightGBM as a suitable option for rapid *in silico* assessment of the toxicity of ever-growing number of industrial or pharmaceutical compounds.[41,42] While we have applied LightGBM to predictions of toxicity data we expect this to be representative of the performance across many areas of cheminformatics as well.

The percentage of test compounds outside the model applicability domain was assessed by performing random projection with a 10-fold reduction in dimensionality for the datasets. According to the assessment results (Table S6), the maximum percentage of test compounds outside the applicability domain is 3.7 % using the mean minimum distance to the closest

neighbor for the test folds of the inner validation loop + 2*std (within 95 % confidence interval) of these minimum distances as cut-off when comparing to the minimum distance of the outer test compounds.

The performance and speed of the LightGBM algorithm are mainly determined by the following six hyperparameters (given in Table S1) and the interquartile ranges and medians of suggested hyperparameter settings are provided in Table 5 and the exact values are shown in Table S4.

1; the number of leaves in the classification tree controls the model complexity. Larger values could improve model accuracy but may lead to overfitting. The LightGBM models developed using either molecular descriptors or fingerprints adopted similar median values for the number of leaves ranging between 140 and 250 (except for nr-er-lbd set) that could improve model accuracy and also avoid overfitting.

2; learning rate regulates the step size on the weights with respect to the loss gradient. The LightGBM models developed using the two feature types adopted comparable learning rates in the range of 0.004 to 0.042 to that could avoid local minima but reduce convergence time.

3; model depth limits the complexity and prevents overfitting. A majority (18 out of 26) of the models were developed with depth smaller than 10. The models (11 out of 13) developed with molecular descriptors are more likely to choose smaller values (median = 7) of the 'max_depth' compared to models developed with fingerprints (median = 10), which might be related to their higher dimensionality.

4; The LightGBM models in this study tend to require a significant number of iterations (approximate median = 570) to improve balanced accuracy. However, the computation times still remained short due to the leaf-wise growth structure and good parallel computational performance of this algorithm.

5; Larger number of samples in one leaf prevent model overfitting. LightGBM models developed based on descriptors resulted in more samples (median = 298) in each leaf than the corresponding fingerprint based models (median = 55), which might be due to molecular descriptors having continuous values whereas fingerprints are represented as binary.

6; LightGBM models used comparable number of feature bin values for both descriptor and fingerprint datasets, which could explain its similar computational time between the two feature types.

Table 5. The interquartile ranges and medians of suggested hyperparameter settings for LightGBM models developed based on Tox21 and mutagenicity datasets described with RDKit molecular descriptors and Morgan fingerprints by the Bayesian optimization integrated nested cross validation scheme.

Hyperparameter	Definition	Molecular descriptors interquartile range (median)	Fingerprints interquartile range (median)
'num_leaves'	number of leaves	159-231 (190)	151-248 (212)
'learning_rate'	learning rate	0.009-0.02 (0.016)	0.01-0.02 (0.013)
'max_depth'	max model depth	7-8 (7)	8-11 (10)
'n_estimators'	number of iterations	448-625 (574)	421-627 (572)
'min_child_samples'	minimum number of samples in one leaf	153-318 (298)	36-111 (55)
'max_bin'	number of feature bins	355-373 (365)	345-398 (372)

We integrated a robust evaluation approach that validates the model performance and ensures the model transferability to new data with a hyperparameter tuning component to identify suitable setting for hyperparameters governing the model performance.

Nested cross validation has been proposed to give the lower generalization error and less degradation in model performance by overcoming the bias and overfitting compared to single loop cross validation.[32] Here we observed comparable model performance of the investigated

algorithms on the internal validation and the external test sets for the individual datasets (Figure 2 and 3) and also on average (Figure 4). This was true for models built using molecular descriptors and fingerprints as well. The comparable model performance on both validation and test sets indicates that the applied nested cross validation strategy offers generalizability and transferability of the developed model to previously unseen new data. In order to obtain the best performance from each investigated algorithm, we performed hyperparameter optimization to select the best hyperparameter settings for the modeling tasks at hand. This was performed using 100 iterations of Bayesian optimization with Gaussian process as the surrogate model and expected improvement acquisition functions[23], identifying the settings which yielded the best predictive balanced accuracies on the inner validation data. The model performance under the same settings were then assessed using outer test data as well. This evaluation strategy has the following advantages:

1) Bayesian optimization method allows faster and more robust hyperparameters optimization, comparing to grid search and random search, as the method keeps track of past evaluations that are used to form a probabilistic model mapping hyperparameters to a probability of the balanced accuracy scoring function.[43]

2) In the 10-fold nested cross validation, the input dataset was split into training sets that were used to select the 'best' hyperparameter settings giving the 'highest' balanced accuracies on the inner validation data. Model performance under the same hyperparameter settings was assessed using outer loop test data. The Bayesian optimization integrated nested cross validation scheme incorporating the rapid hyperparameter tuning and robust evaluation features allowing us to select the hyperparameter set that resulted in the lowest generalization error and prevented the common risk of over-optimizing the algorithm hyperparameters for the internal test data, resulting in overestimation of performance on new independent data.

When evaluating the results from this study, it is important to be aware that cytotoxicity has been detected in the Tox21 screening assays. Approximately 6-8% of the testing compounds were affected by the cytotoxicity and potentially incorrectly labeled as positives (false positives).[44,45] The presence of the mislabeled false positives in the activity results might mislead the modeling algorithms causing misprediction of negative compounds to positives and increase false positive rate as well as reduce sensitivity in the prediction results.

Conclusions

In this study, we evaluated the emerging LightGBM classification algorithm using a model evaluation and selection scheme incorporating Bayesian optimization and nested 10-fold cross validations. The scheme allows rapid hyperparameter tuning and robust assessment of model evaluation and transferability. When applying this scheme to compare the predictive power of ML algorithms using Tox21 and mutagenicity datasets, LightGBM offered the best performance in terms of balanced accuracy for both internal validation and external test sets, compared to four widely used algorithms, DNN, RF, SVC, and XGBoost. In addition to the excellent predictive performance, LightGBM is also faster and more scalable in model development. This is especially apparent when using high dimensional data matrices such as fingerprint features. In conclusion, LightGBM is a more effective and scalable ML algorithm that is able to fulfill the ever-growing demand for rapid *in silico* toxicological assessment of emerging industrial or pharmaceutical compounds.

Supporting Information

Equations for calculating performance metrics, tabulated metrics results including balanced accuracy, Cohen's kappa, Matthews correlation coefficient, positive predictive values, negative predictive value, ROC- area under curve (AUC), sensitivity and specificity values, computation

time for each of the Tox21 and mutagenicity dataset using the five ML algorithms, and p-values for algorithm comparisons.

Acknowledgements

The ARUK UCL Drug Discovery Institute is core funded by Alzheimer's Research UK (registered charity No. 1077089 and SC042474). The Francis Crick Institute receives its core funding from Cancer Research UK (FC001002), the UK Medical Research Council (FC001002), and the Wellcome Trust (FC001002).

Notes

The Authors declare no competing financial interest.

References

(1) Cumming, J. G.; Davis, A. M.; Muresan, S.; Haeberlein, M.; Chen, H. Chemical Predictive Modelling to Improve Compound Quality. *Nat. Rev. Drug Discov.* **2013**, *12*, 948–962.

(2) Raies, A. B.; Bajic, V. B. In Silico Toxicology: Computational Methods for the Prediction of Chemical Toxicity. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2016**, *6*, 147–172.

(3) 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*, 198–212.

(4) Maltarollo, V. G.; Gertrudes, J. C.; Oliveira, P. R.; Honorio, K. M. Applying Machine Learning Techniques for ADME-Tox Prediction: A Review. *Expert Opin. Drug Metab. Toxicol.* **2015**, *11*, 259–271.

(5) Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232.

(6) Sheridan, R. P.; Wang, W. M.; Liaw, A.; Ma, J.; Gifford, E. M. Extreme Gradient Boosting as a Method for Quantitative Structure–Activity Relationships. *J. Chem. Inf.*

Model. **2016**, *56*, 2353–2360.

(7) Bowles, M.; Shigeta, R. Statistical Models for Predicting Liver Toxicity from Genomic Data. *Syst. Biomed.* **2013**, *1*, 144–149.

(8) Ke, G.; Meng, Q.; Wang, T.; Chen, W.; Ma, W.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Adv. Neural Inf. Process. Syst. 30* **2017**.

(9) Chen, T.; Guestrin, C. XGBoost : Reliable Large-Scale Tree Boosting System. *arXiv* **2016**, arXiv:1603.02754.

(10) Ma, X.; Sha, J.; Wang, D.; Yu, Y.; Yang, Q.; Niu, X. Study on a Prediction of P2P Network Loan Default Based on the Machine Learning LightGBM and XGboost Algorithms According to Different High Dimensional Data Cleaning. *Electron. Commer. Res. Appl.* **2018**, *31*, 25–39.

(11) Ju, Y.; Sun, G.; Chen, Q.; Zhang, M.; Zhu, H.; Rehman, M. U. A Model Combining Convolutional Neural Network and Lightgbm Algorithm for Ultra-Short-Term Wind Power Forecasting. *IEEE Access* **2019**.

(12) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inform.* **2010**, *29*, 476–488.

(13) Krstajic, D.; Buturovic, L. J.; Leahy, D. E.; Thomas, S. Cross-Validation Pitfalls When Selecting and Assessing Regression and Classification Models. *J. Cheminform.* **2014**, *6*, 10.1186/1758-2946-6–10.

(14) Snoek, J.; Larochelle, H.; Adams, R. P. Practical Bayesian Optimization of Machine Learning Algorithms. *Arxiv* **2012**, arXiv:1206.2944.

(15) Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity Prediction Using Deep Learning. *Frontiers in Environmental Science*. 2016, p 80.

(16) Alvarsson, J.; Eklund, M.; Andersson, C.; Carlsson, L.; Spjuth, O.; Wikberg, J. E. S. Benchmarking Study of Parameter Variation When Using Signature Fingerprints Together with Support Vector Machines. *J. Chem. Inf. Model.* **2014**, *54*, 3211–3217.

(17) Banerjee, P.; Eckert, A. O.; Schrey, A. K.; Preissner, R. ProTox-II: A Webserver for the Prediction of Toxicity of Chemicals. *Nucleic Acids Res.* **2018**, *46*, W257–W263.

(18) Yang, H.; Lou, C.; Sun, L.; Li, J.; Cai, Y.; Wang, Z.; Li, W.; Liu, G.; Tang, Y. AdmetSAR 2.0: Web-Service for Prediction and Optimization of Chemical ADMET Properties. *Bioinformatics* **2018**, *35*, 1067–1069.

(19) Huang, R.; Xia, M.; Nguyen, D.-T.; Zhao, T.; Sakamuru, S.; Zhao, J.; Shahane, S. A.; Rossoshek, A.; Simeonov, A. Tox21Challenge to Build Predictive Models of Nuclear Receptor and Stress Response Pathways as Mediated by Exposure to Environmental Chemicals and Drugs. *Frontiers in Environmental Science*. 2016, p 85.

(20) Hansen, K.; Mika, S.; Schroeter, T.; Sutter, A.; ter Laak, A.; Steger-Hartmann, T.; Heinrich, N.; Müller, K.-R. Benchmark Data Set for in Silico Prediction of Ames Mutagenicity. *J. Chem. Inf. Model.* **2009**, *49*, 2077–2081.

(21) Chollet, F. Keras. *https://keras.io* **2015**.

(22) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(23) Scikit-Optimize. *https://github.com/scikit-optimize* **2018**.

(24) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I. J.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Józefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mané , D.; Monga, R.; Moore, S.; Murray, D. G.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P. A.; Vanhoucke, V.; Vasudevan, V.; Viégas, F. B.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; Zheng, X. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *CoRR* **2016**, *abs/1603.0*.

(25) NIH. Tox21 Data Challenge. *https//tripod.nih.gov/tox21/challenge/data.jsp, accessed April 30, 2018*.

(26) IMI ETOX Project Standardizer. *version 0.1.7. https://pypi.python.org/pypi/standardiser*.

(27) MolVS Standardizer. *version 0.0.9. https://pypi.python.org/pypi/MolVS*.

(28) RDKit: Open-Source Cheminformatics (http://www.rdkit.org).

(29) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(30) Svensson, F.; Norinder, U.; Bender, A. Improving Screening Efficiency through Iterative Screening Using Docking and Conformal Prediction. *J. Chem. Inf. Model.* **2017**, *57*, 439–444.

(31) Svensson, F.; Norinder, U.; Bender, A. Modelling Compound Cytotoxicity Using Conformal Prediction and PubChem HTS Data. *Toxicol. Res. (Camb).* **2017**, *6*, 73–80.

(32) Cawley, G. C.; Talbot, N. L. On Over-Fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *J. Mach. Learn. Res.* **2010**, *11*, 2079–2107.

(33) Bergstra, J.; Yamins, D.; Learning, D. C. B. T. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. PMLR February 13, 2013, pp 115–123.

(34) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships. *J. Chem. Inf. Model.* **2015**, *55*, 263–274.

(35) Korotcov, A.; Tkachenko, V.; Russo, D. P.; Ekins, S. Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. *Mol. Pharm.* **2017**, *14*, 4462–4475.

(36) Boser, B. E.; Guyon, I. M.; Vapnik, V. N. A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory - COLT '92*; 1992.

(37) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297.

(38) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.

(39) García, V.; Mollineda, R. A.; Sánchez, J. S. Index of Balanced Accuracy: A Performance Measure for Skewed Class Distributions BT - Pattern Recognition and Image Analysis; Araujo, H., Mendonça, A. M., Pinho, A. J., Torres, M. I., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2009; pp 441–448.

(40) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discov. Today* **2018**, *23*, 1241–1250.

(41) Blomme, E. A. G.; Will, Y. Toxicology Strategies for Drug Discovery: Present and Future. *Chem. Res. Toxicol.* **2016**, *29*, 473–504.

(42) Richard, A. M. Future of Toxicology-Predictive Toxicology: An Expanded View of "Chemical Toxicity." *Chem. Res. Toxicol.* **2006**, *19*, 1257–1262.

(43) Bergstra, J. S.; Bardenet, R.; Bengio, Y.; Kégl, B. Algorithms for Hyper-Parameter Optimization BT - Advances in Neural Information Processing Systems 24. In *Advances in Neural Information Processing Systems 24*; 2011; pp 2546–2554.

(44) Judson, R. S.; Magpantay, F. M.; Chickarmane, V.; Haskell, C.; Tania, N.; Taylor, J.; Xia, M.; Huang, R.; Rotroff, D. M.; Filer, D. L.; Houck, K. A.; Martin, M. T.; Sipes, N.; Richard, A. M.; Mansouri, K.; Setzer, R. W.; Knudsen, T. B.; Crofton, K. M.; Thomas, R. S. Integrated Model of Chemical Perturbations of a Biological Pathway Using 18 In Vitro High-Throughput Screening Assays for the Estrogen Receptor. *Toxicol. Sci.* **2015**, *148*, 137–154.

(45) Hsieh, J.-H.; Sedykh, A.; Huang, R.; Xia, M.; Tice, R. R. A Data Analysis Pipeline Accounting for Artifacts in Tox21 Quantitative High-Throughput Screening Assays. *J. Biomol. Screen.* **2015**, *20*, 887–897.

Table of Contents (TOC) graphics