frontiers in Environmental Science

ORIGINAL RESEARCH
published: 01 December 2015
doi: 10.3389/fenvs.2015.00077

Prediction of Compounds Activity in Nuclear Receptor Signaling and Stress Pathway Assays Using Machine Learning Algorithms and Low-Dimensional Molecular Descriptors

Filip Stefaniak *

Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, Warsaw, Poland

OPEN ACCESS

Edited by:
Ruili Huang, NIH National Center for Advancing Translational Sciences, USA

Reviewed by:
Keith Shockley, National Institute of Environmental Sciences, USA
Denis Fourches, North Carolina State University, USA
Costica Nitu, University "Politehnica"Bucharest, Romania

***Correspondence:**
Filip Stefaniak
fstefaniak@genesilico.pl

Specialty section:
This article was submitted to Environmental Informatics, a section of the journal Frontiers in Environmental Science

Received: *01 September 2015*
Accepted: *16 November 2015*
Published: *01 December 2015*

Citation:
Stefaniak F (2015) Prediction of Compounds Activity in Nuclear Receptor Signaling and Stress Pathway Assays Using Machine Learning Algorithms and Low-Dimensional Molecular Descriptors. Front. Environ. Sci. 3:77. doi: 10.3389/fenvs.2015.00077

Toxicity evaluation of newly synthesized or used compounds is one of the main challenges during product development in many areas of industry. For example, toxicity is the second reason—after lack of efficacy—for failure in preclinical and clinical studies of drug candidates. To avoid attrition at the late stage of the drug development process, the toxicity analyses are employed at the early stages of a discovery pipeline, along with activity and selectivity enhancing. Although many assays for screening *in vitro* toxicity are available, their massive application is not always time and cost effective. Thus, the need for fast and reliable *in silico* tools, which can be used not only for toxicity prediction of existing compounds, but also for prioritization of compounds planned for synthesis or acquisition. Here I present the benchmark results of the combination of various attribute selection methods and machine learning algorithms and their application to the data sets of the Tox21 Data Challenge. The best performing method: Best First for attribute selection with the Rotation Forest/ADTree classifier offers good accuracy for most tested cases. For 11 out of 12 targets, the AUROC value for the final evaluation set was =0.72, while for three targets the AUROC value was = 0.80, with the average AUROC being 0.784 ± 0.069. The use of two-dimensional descriptors sets enables fast screening and compound prioritization even for a very large database. Open source tools used in this project make the presented approach widely available and encourage the community to further improve the presented scheme.

Keywords: toxicity prediction, machine learning, molecular descriptors, molecular fingerprints, Tox21 Data Challenge 2014

INTRODUCTION

Toxicity evaluation of newly synthesized or used chemicals (pharmaceuticals and its metabolites, cosmetic ingredients, biocides, or anthropogenic pollutants) is one of the main challenges during product development in many areas of industry. For example, it has been estimated that in the pharmaceutical industry the toxicology and clinical safety is accounting for 30% of failures

in clinical trials (Kola and Landis, 2004). The risk of attrition can be substantially reduced by the introduction of toxicity testing at the early stages of product development. Such evaluation, especially when performed on a large scale, is neither time/cost effective, nor—in case of tests performed on animals—ethically justified. It is estimated that the introduction of a new pesticide to the market requires testing on 7000 animals and costs tens of millions of dollars (Erickson, 2011). Moreover, animal models are frequently poorly correlated with response on humans (Knight, 2007; Shanks et al., 2009). Although *in vivo* testing seems to be inevitable at the late stage of a product development, many efforts to shift from traditional *in vivo* tests to higher-throughput and less expensive cell-based assays have been made. For example "The Toxicology in the 21st Century" (Tox21) program, is aimed at developing more reliable toxicity assessment methods as well as developing and validating cellular (*in vitro*) toxicity assays. The Tox21 10 K chemical library consists of ∼10,500 plated compound solutions, consisting of 8311 unique chemical substances, including pesticides, industrial chemicals, food-use additives and drugs (Huang et al., 2014). Acquired activity data can serve not only as *in vitro* signatures that could be used to predict *in vivo* toxicity endpoints (Martin et al., 2011; Sipes et al., 2011) and to prioritize chemicals for extensive toxicity testing (Judson et al., 2010), but also to provide the scientific community with training data sets for developing reliable *in silico* toxicity models (Sun et al., 2012). Also, many attempts toward development of new computational methods for high-throughput toxicity prediction have been made and many techniques and algorithms have been proposed (Deeb and Goodarzi, 2012; Bakhtyari et al., 2013; Cheng et al., 2013; Valerio, 2013; Low et al., 2014; Omer et al., 2014; Toropov et al., 2014; Rouquie et al., 2015). In recent years, machine learning methods are gaining more attention as robust and accurate tools for Quantitative structure–activity relationship (QSAR) and Quantitative structure–property relationships (QSPR) modeling (Durrant and Amaro, 2015; Freitas et al., 2015; Liu, 2015). The key to success in building predictive models are: (a) the quality of a training data set, (b) the descriptive power of molecular descriptors, and (c) selecting and tuning machine learning algorithms. Here I present a detailed description of creating activity prediction models using the Tox21 Data Challenge data set (Subchallenges 1–12). It consists of activity data for two panels playing important roles in toxicological pathways. Nuclear Receptor Signaling Panel (nr) included activity data for seven targets: aryl hydrocarbon receptor (ahr), androgen receptor—full length (ar) and Ligand Binding Domain (ar-lbd), aromatase, estrogen receptor alpha—full length (er) and Ligand Binding Domain (er-lbd) and peroxisome proliferator-activated receptor gamma (ppar-gamma). Stress Response Panel (sr) included data for five targets: nuclear factor (erythroid-derived 2)-like 2/antioxidant responsive element (are), ATAD5, heat shock factor response element (hse), the disruption of mitochondrial membrane potential (mmp) and p53. Great emphasis is laid upon the initial performance benchmark of the various combinations of attribute selection methods and classification algorithms. Two-dimensional molecular descriptors set and dictionary-based fingerprints enable fast screening and compound prioritization even for very large databases. All software used during this study is freely available and open source, making the presented approach widely available for the scientific community.

MATERIALS AND METHODS

The training dataset provided by the Challenge organizers (https://tripod.nih.gov/tox21/challenge/data.jsp) consisted of the activity data for ∼10 k compounds (Tox21 10 K compound library, structures provided as SMILES) on 12 targets, with the activity class assigned "Active" or "Not active" (for discussions of activity call procedures, see Shockley, 2012; Tice et al., 2013). The Testing dataset, provided later by the Challenge organizers consisted of activity data for 269 compounds. The final predictions were performed on the evaluation set of 647 compounds with unknown activity.

All calculations were performed on the desktop computer with Intel Core i7-4770 K CPU processor (eight cores) and 16 GB RAM, running Ubuntu 12.04.5 LTS.

Structures Standardization and Preprocessing

The chemical structures in the provided Tox21 Challenge data sets were standardized using the LyChI (Layered Chemical Identifier) program (version 20141028, https://github.com/ncats/lychi). Compounds with ambiguous structure (compound identifier with more than one chemical structure assigned) or activity (compound identifier with activity labels "Active" and "Not active" on a single target) were excluded using KNIME GroupBy node (KNIME 2.10.4, http://www.knime.org/; Berthold et al., 2007). For each compound, only the biggest component was preserved (KNIME component Separator node). For each target, data set was downsized such that the activity values were evenly distributed—all records from the minority class were retained and a random sample from the majority class was added (KNIME Row Sampling node). Standardized and downsized datasets used for modeling are available as Supplementary Materials.

Descriptors Generation

For standardized data sets, two-dimensional molecular descriptors were calculated using KNIME nodes: RDKit (http://rdkit.org/, 117 descriptors), CDK (Beisken et al., 2013; http://sourceforge.net/projects/cdk/, 97 descriptors) and fingerprints [PubChem (881 bits) and MACCS (167 bits)], giving 1262 descriptors for each compound. For the list of used descriptors and literature references see Supplementary Table S5. For each target, Arff weka file was created using KNIME Arff Writer node.

Classification Algorithms Screen

Preprocessing and classification algorithms screen was performed in the Weka Experiment Environment (Weka 3.6.6, Hall et al., 2009), with 10-fold cross validation with 10 repetitions. In each run, data was preprocessed with Remove Useless filter (all constant attributes are deleted, along with

any that exceed the maximum percentage of variance, set to 99%) and Standardize filter (standardizes all numeric attributes to have zero mean and unit variance). Attribute selection was performed with two search methods: Best First and Rank Search, with CfsSubset attribute evaluator. Machine learning algorithms tested were: ADTree (alternating decision tree), FT (functional trees), FURIA (Fuzzy Unordered Rule Induction Algorithm), IBk (*k*-nearest neighbors), J48, Naïve Bayes, REPTree, and SMO (sequential minimal optimization for training a support vector classifier). Ensemble methods tested in the second step of the screen were: Rotation Forest, Decorate, Dagging, Bagging and AdaBoost M1. Unless otherwise stated, all algorithms were used with default settings. The performance of the models was measured using area under the receiver operating characteristic (ROC) curve metrics (AUROC).

Predictions

The final models were built in KNIME with Weka 3.6 nodes, using the Best First attribute selection method with Rotation Forest/ADTree classifier (for parameters of the classifier see Supplementary Table S6). For each target, 10 models were built using randomly selected subset of 95% of training set. Each model was evaluated on the remaining 5% of the training set and on the testing set. The model with the best AUROC value was selected for the final predictions. The estimation of probability of a chemical being active was rounded to three decimal places.

RESULTS AND DISCUSSION

The data processing workflow is shown in **Figure 1**. It involved six main steps: data preprocessing, descriptors calculation, feature selection and classification algorithms screen, training, testing, and predictions.

Data Preprocessing

The first stage of data preprocessing included data sanitization. First, SMILES were standardized with the LyChi program. For the training dataset, out of 11,764 unique input compounds, 9231 (78%) had fixed structure. Among the most frequent modifications were: unifying aromaticity model, neutralization and small counterions removal. Next, structures containing more than one component were separated and only the biggest component was preserved. This was the most vague reduction of the initial data, but this step was necessary for proper descriptors calculations. Also, an analysis of the most frequently removed components showed that these were mainly inorganic acids, metal ions and water molecules (see **Table 1**), which are frequent components of pharmaceutical mixtures and should not be treated as a factors determining activity on investigated targets. Finally, each subset of the training data set was downsized such that the activity values are equally distributed. The selection of the majority class members (inactives) was random (see Sections Structures Standardization and Preprocessing: Materials and Methods), which means that the output from this step could influence the results of further predictions. Here, the downsizing was a single-time procedure and the influence of various sets of majority class on models' performance was not investigated. For the initial and final compositions of the training data set (see **Table 2**).

Molecular Descriptors Calculation

Generation of higher-dimensional molecular descriptors (3D, 4D, 5D) is time consuming and may be prone to conformer generation errors. To avoid these shortcomings, low-dimensional (0D, 1D, 2D) descriptors and dictionary-based fingerprints were

TABLE 1 | Top 10 most frequently removed minor components from an initial training data set.

Removed component	Count	% of all removed components
HCl	955	32.6
Na^+	533	18.2
H_2O	254	8.7
Cl^-	157	5.4
Br^-	110	3.8
Sulphuric acid	83	2.8
Methylsulfonic acid	54	1.8
K^+	50	1.7
Maleic acid	47	1.6
I^-	41	1.4

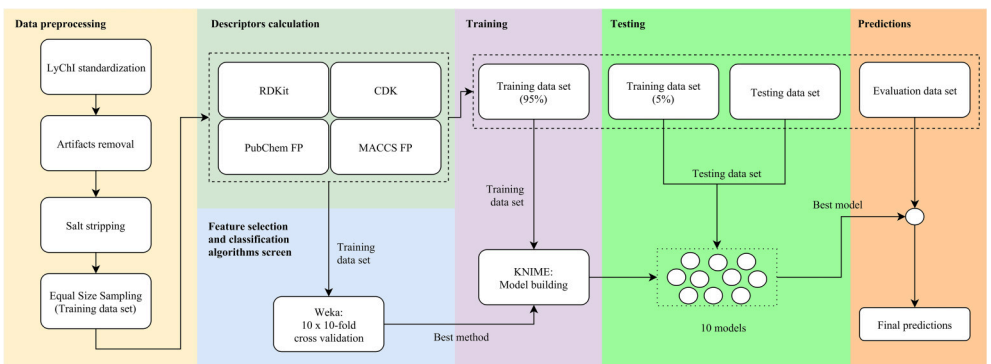

FIGURE 1 | Activity prediction workflow.

TABLE 2 | Initial and final training data sets composition.

Target	Initial training data set			Preprocessed training data set		
	Data set size	Actives count	% actives	Data set size	Actives count	% actives
nr-ahr	8169	950	11.6	1900	950	50.0
nr-ar	9362	380	4.1	756	378	50.0
nr-ar-lbd	8599	303	3.5	604	302	50.0
nr-aromatase	7226	360	5.0	712	356	50.0
nr-er	7697	937	12.2	1866	933	50.0
nr-er-lbd	8753	446	5.1	882	441	50.0
nr-ppar-gamma	8184	222	2.7	442	221	50.0
sr-are	7167	1098	15.3	2188	1094	50.0
sr-atad5	9091	338	3.7	674	337	50.0
sr-hse	8150	428	5.3	850	425	50.0
sr-mmp	7320	1142	15.6	2246	1123	50.0
sr-p53	8634	537	6.2	1064	532	50.0

used here. It was shown earlier that such descriptors may carry the similar information-level to higher dimensional ones (Estrada et al., 2001; Oprea, 2002; Roy and Das, 2014) and can be successfully used in building predictive QSAR models (Roy and Roy, 2009; Garcia et al., 2011; Chavan et al., 2014; Su et al., 2015).

Feature Selection and Classification Algorithms Screen

Various attribute selection, data preprocessing and classification algorithms are available (Witten et al., 2011). It is not known *a priori* which combination of the above is optimal for the problem under consideration, as for different data sets the accuracy of algorithms varies (Smusz et al., 2013). This is why an initial methods assessment was conducted, evaluating the performance (expressed as the AUROC value) of the combination of:

- Attribute selection methods: two search methods were evaluated: Best First and Rank Search
- Classifiers: 14 classifiers setups were evaluated

Most classifiers were used with default settings. For IBk, four values of k were probed (1, 3, 5, and 10), as this parameter may significantly influence the performance of this classifier. SMO algorithm was probed with three kernels (RBF kernel, polynomial kernel, and normalized polynomial kernel). To validate various modeling approaches, a 10-fold cross validation with 10 repetitions was used. In each run, training data were preprocessed independently (removal of a constant attribute, data standardization, attribute selection). This allowed an estimation of how the procedures under the investigation will generalize to an independent data set. Results of the initial evaluation are summarized in **Figure 2** (for values obtained in the initial methods evaluation see Supplementary Table S1).

As expected, the performance of evaluated classifiers varied. For the tested set of the descriptors, among the best performing ones were ADTree, IBk, and Naïve Bayes. Performance of IBk classifier varied slightly for various values of k, with better AUROC values for the higher k (5 and 10). The worst performance was observed for SMO (Sequential Minimal Optimization). However, the parameters for these methods (C, gamma) were not optimized and certainly such optimization would increase their performance. As for the attribute selection methods, in most cases there were no significant differences in performance between algorithms. The exception is the Naïve Bayes classifier, where the differences are substantial. Generally, the Best First method was slightly better than Rank Search (mean AUROC for all experiments: 0.778 ± 0.056 and 0.768 ± 0.055, respectively). In the studied descriptors space, the overall "target predictability" also varied. The sr-mmp and nr-ar-lbd are "the most predictable" targets while sr-hse and nr-er are "the least predictable" ones. The latter observation may be caused by the insufficient descriptive power of calculated molecular features to describe the nature of binding small molecule ligands to these targets.

After initial algorithms screen, the four best performing methods (Naïve Bayes, ADTree, and IBk) were evaluated in combination with ensemble methods: Rotation Forest, Decorate, Dagging, Bagging, and AdaBoost. The SMO classifier was treated as the "negative control." The Best First attribute selection method was used. Results are summarized in **Figure 3** (for AUROC values obtained in this experiment see Supplementary Table S2).

The application of the ensemble methods in most cases caused increase of the obtained AUROC values. The average AUROC for all targets for Naïve Bayes classifier increased from 0.79 to 0.80 (when combined with Bagging, Dagging, Decorate and Rotation Forest) but decreased to 0.78 in case of AdaBoostM1. For ADTree, the AUROC values increased from the initial 0.79–0.82 (in combination with Decorate) and 0.83 (for Rotation Forest). For comparison of the performance of the selected ensemble classifiers see Supplementary Table S4. The best and most stable performance for all targets was observed for Rotation Forest ensemble method with two classifiers: ADTree and IBk ($k = 10$) (Mean AUROC for all experiments: 0.831 ± 0.038 and 0.820 ± 0.038 respectively). Based on these results, the Best First attribute selection method with Rotation Forest/ADTree classifier was used for the final activity predictions for all targets.

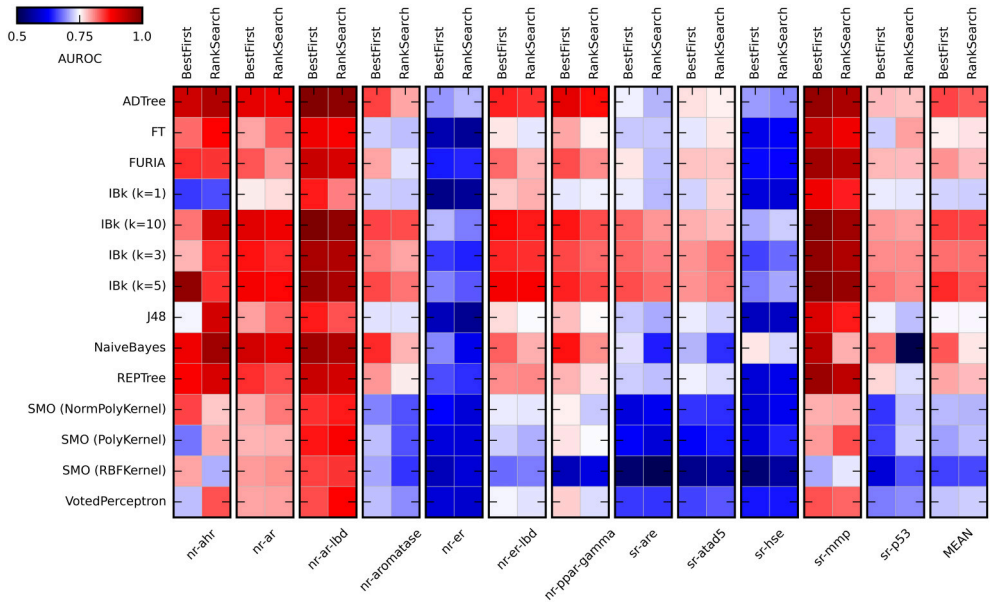

FIGURE 2 | Heat maps presenting results of the initial methods evaluation. Color coded AUROC values are presented for 14 classifiers (Y axis) in combination with two attribute selection methods (top X axis), grouped by the target (down X axis). Additional group presenting mean AUROC values is added for classifiers comparison.

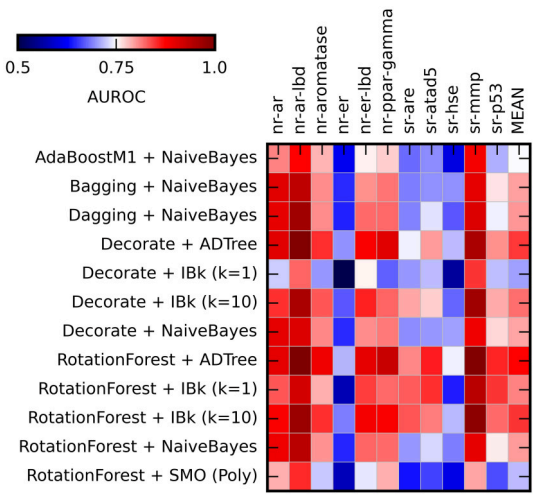

FIGURE 3 | Heat maps presenting results of the ensemble methods assessment. Color coded AUROC values for 12 classifiers (Y axis) for each challenge target (X axis) are shown. Aditional column presenting mean AUROC values is added for classifiers comparison.

TABLE 3 | AUROC values obtained for the best models selected for final predictions.

Target	AUROC testing		AUROC evaluation set
	Training set 5%	Testing set	
nr-ahr	0.92	0.84	0.89
nr-ar	0.76	0.50	0.73
nr-ar-lbd	0.91	0.82	0.79
nr-aromatase	0.92	0.79	0.78
nr-er	0.85	0.67	0.77
nr-er-lbd[a]	0.95	0.70	0.78
nr-ppar-gamma[a]	0.97	0.71	0.67
sr-are	0.87	0.80	0.72
sr-atad5[a]	0.91	0.65	0.76
sr-hse	0.90	0.74	0.80
sr-mmp	0.92	0.86	0.93
sr-p53	0.88	0.72	0.79

[a] *These models were not submitted to the final evaluation of the Tox21 Challenge.*

Training, Testing, and Final Predictions

For each target, 10 models were built using randomly selected subsets of 95% of the training set. Each model was tested on two sets: the remaining 5% of the training set and the provided testing set. The use of the 5%-random subset, apart from the constant testing set, helped to assure that the performance of the selected model is obtained not due to chance, but by merit inherent to the method. The model with the highest AUROC value was selected for the final predictions on the evaluation set. The performance on the testing and evaluation data sets of selected best models is summarized in **Table 3**. For AUROC statistics of all generated models see Supplementary Table S3.

The average AUROC value for the final predictions for all 12 targets was 0.784 ± 0.069. The best results were obtained for nr-ahr and sr-mmp (AUROC values: 0.89 and 0.93, respectively). The lowest AUROC value was obtained for nr-ppar-gamma (0.67), despite good performance of the model on the testing

sets. As stated earlier, lower performance for some targets may be caused by the insufficient descriptive power of calculated molecular features to describe the complex nature of binding small molecule ligands to these targets.

In general, one can observe the correlation between AUROC values for testing and evaluation data sets. Most prominent examples include sr-mmp and nr-ahr (good performance in both testing and final evaluation) and nr-ar (moderate performance in both cases). On the other hand, for nr-ppar-gamma, the results obtained on the testing data sets are very good, while the final performance is moderate. In this case, one of the reasons could be that the chemical space of the evaluation set is out of the applicability domain of the selected model.

Computational Performance

Descriptors Calculation

The choice of low-dimensional descriptors guaranteed a high speed of calculations. A test run, carried for randomly selected 50 k clean drug like compounds fetched from ZINC database, showed a calculation rate at 12.65 s/1000 compounds ($\pm$1.33 s). The workflow for the descriptors calculation may be further optimized by applying a better parallelization scheme and by using all available CPUs on all stages of calculations.

Classification Performance

The biggest influence on the training time has the attribute selection step. Results from initial algorithms assessment (10-fold cross validation with 10 repetitions) shows that, for Best First, the average time of a single run was 10.197 $\pm$ 6.359 s, while for Rank Search it was 80.983 $\pm$ 66.302 s. Although the differences between these algorithms are high, in many cases training is a one-time procedure and training time is not a main factor for consideration. The average testing time for Best First method was 0.034 $\pm$ 0.063 s, while for Rank Search it was 0.119 $\pm$ 0.242 s. For the setup used for final evaluation (Best First attribute selection method with Rotation Forest/ADTree classifier) the average training time for all targets was 13.084 $\pm$ 8.627 s, while the testing time was 0.042 $\pm$ 0.033 s. For training and testing time values see Supplementary Tables S1, S2).

Related Works

Recently, a few papers describing various classification methods applied to the Tox21 dataset have been published. Drwal et al. described a successful approach of applying similarity comparison and machine learning for activity prediction (Drwal et al., 2015). These authors also used two dimensional descriptors sets in the form of 2929 bit-long bitvector, encoding molecular features, properties and connectivity information. The training dataset was enriched by adding activity data fetched from the literature (when available). Various parameters of similarity searching (Tanimoto fingerprint similarity to active or inactive compounds), of machine learning (Naïve Bayes) and of the combination of these methods were evaluated. The established methodology applied to the Tox21 dataset gave comparable results to the ones shown in this work (for four targets, the methods presented here gave better AUROC values, for two, the values were equal).

Deep learning methods were also applied to the Tox21 classification challenge. Unterthiner et al. used deep neural network with 40,000 input features describing molecules (Unterthiner et al., 2015). The presented scheme allowed the team to get the highest AUROC values in most of the Tox21 sub-challenges. The drawback of this methodology is the high demand for computational resources. Ramsundar et al. used simple two-dimensional descriptors and fingerprints in connection with Massively Multitask Networks (Ramsundar et al., 2015). Comparison to other classification algorithms (logistic regression, random forest) showed better performance for the deep learning method. Again, this methodology is computationally very expensive.

CONCLUSIONS

The presented method uses fast to calculate, two-dimensional descriptors and, in most cases, shows good predictive performance. Moreover, the use of free and open source tools makes the presented approach widely available for the community. To further improve the described workflow, a wider set of descriptors may be used, including fingerprints basing on connectivity information (like ECFP4 or Morgan fingerprints) or recently presented ToxPrint fingerprint, which cover substructures associated with toxicity (Yang et al., 2015). Also, other classification methods, including ensemble methods and deep learning techniques, should be investigated.

The Tox21 Data Challenge 2014 has offered the opportunity to compare and benchmark various approaches for toxicity prediction. The results clearly show that the very accurate *in silico* methods are now, or soon will be, at our fingertips. However, there is still a lot of work to be done to improve the quality of models to fully supersede traditional, *in vitro* assays.

ACKNOWLEDGMENTS

FS was supported by the European Research Council (ERC, StG grant RNA+P=123D project number 261351, to Janusz M. Bujnicki.). I would like to thank Krzysztof Szczepaniak, Grzegorz Lach, Catarina Almeida and Janusz M. Bujnicki for carefully reading the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fenvs.2015.00077