Computational Toxicology 15 (2020) 100129

Contents lists available at ScienceDirect

Computational Toxicology

journal homepage: www.elsevier.com/locate/comtox

Active learning efficiently converges on rational limits of toxicity prediction and identifies patterns for molecule design

Ahsan Habib Polash[a,b], Takumi Nakano[b], Christin Rakers[c], Shunichi Takeda[a], J.B. Brown[b,*]

[a] Department of Radiation Genetics, Kyoto University Graduate School of Medicine, Kyoto, Sakyo, Yoshida-konoemachi Building D, 3F, 606-8501, Japan
[b] Laboratory for Molecular Biosciences, Life Science Informatics Research Unit, Kyoto University Graduate School of Medicine, Kyoto, Sakyo, Yoshida-konoemachi Building E, 606-8501, Japan
[c] Graduate School of Pharmaceutical Sciences, Kyoto University, 46-29 Yoshida-shimoadachi-cho, Sakyo-ku, Kyoto 606-8501, Japan

ARTICLE INFO

Keywords:
In vivo toxicity
Active learning
Quantitative structure–property prediction
Toxicity cliff
Regulatory science

ABSTRACT

Legal frameworks to restrict in-vivo animal testing for compound toxicity are now enforced, and regulatory agencies are actively seeking ideas and methods for toxicity prediction. Recent computational drug discovery modeling methods have shown that equivalently performant models can be built from strategically selected subsets of activity data, thus posing the question of method transferability to toxicity prediction. Here, active learning was used to assess predictive convergence using an open toxicity prediction challenge dataset, revealing that a subset of systematically selected data was sufficient for early convergence which approximates predictive performance using all possible data. Exploration of training and external validation data systematically identified varying degrees of "toxicity cliffs" in molecular frameworks and specific compound pairs with structure–-activity discontinuity. Domain of applicability analysis revealed compounds emergently predictable and those which could never be predicted. The removal of compounds positioned on classification borderlines improved the ability to identify very toxic compounds. Fingerprints differentially present in toxic compounds were identified. The combined analyses give clear insights for expectations on predictability and regulatory policy.

1. Introduction

Determining the acute oral systemic toxicity of chemicals is multi-resolution and multi-factorial, and thus challenging as it is caused by multiple mechanisms, including both damage on the epithelial cells of gastrointestinal tracts which leads to systemic symptoms, as well as the toxic effects of blood-circulating chemicals on various tissues. Compared with oral systemic toxicity where mechanisms precluding genotoxicity are much simpler, the vast majority of genotoxic compounds react directly with genomic DNA in cells. Nonetheless, there is no reliable bioassay using human cells; governments in developed countries have widely employed the Ames test [4] which is a bacteria-based bioassay, though it remains unclear whether data from such a test can be extrapolated to human genotoxicity (e.g., the absence of a double strand break repair mechanism in bacteria), and demonstration of oral systemic toxicity requires animal experiments. Still, due to its relatively low cost and ease of implementation, the test remains as a standard in many places, such as its use as part of the Japanese "kashinho" regulatory legislation (meti.go.jp).

Animal-based models are expensive, time-consuming, can yield false positives, and have come under ethical scrutiny [5,13,14,34]. A bit over a decade ago, the EU REACH legislation for registration and minimization of animal testing by read-across data sharing was brought into effect, though it has been reported that animal testing continues to be prevalent [35]. Following in the footsteps of an EU directive that prohibits commercial sale of cosmetics that have undergone any form of animal testing, the US has also announced its strategic road map for establishing new approaches to evaluate the safety of chemicals and pharmaceutical products [15]. Under such a backdrop, research into computational models to predict toxicity is being sponsored at federal levels, with the most common strategy being the creation of quantitative structure–activity/property relationships (QSAR/QSPR), though as Taylor recently reviewed, in vitro and (Q)SAR approaches as standalone replacements still remain low [35].

Interest in prediction of toxicity has garnered increased attention in recent years with the creation of collaborative prediction challenges, such as the Tox21 Data Challenges [37], where participating groups construct estimators (models) of compound toxicity based on chemical structure and biological activity information. Each participating group was free to choose a combination of chemical representation and

* Corresponding author.
E-mail address: jbbrown@kuhp.kyoto-u.ac.jp (J.B. Brown).

https://doi.org/10.1016/j.comtox.2020.100129
Received 25 November 2019; Received in revised form 3 May 2020; Accepted 1 July 2020
Available online 03 July 2020
2468-1113/ © 2020 Elsevier B.V. All rights reserved.

Table 1
Aspects of investigating active learning (AL) for toxicity prediction.

Aspect	Group	Description	Location
1	Datasets	Statistics on rat oral acute toxicity datasets used	Table 2
2		Dataset BM frameworks with mean toxicities in the VT/NT LD_{50} ranges	Fig. 1
3		Full listing of all BM frameworks in datasets used	Supplementary Fig. 1
4	AL performance, descriptors, generalization, and cross-model comparison	Data volume-based prediction performance analysis (MCC, VT/NT) and Active Projection (F1/BA, VT); MACCS keys	Fig. 2
5		Data volume-based prediction performance analysis by F1/BA/TPR/TNR	Supplementary Fig. 2
6		Descriptor comparison: CATS2D/physChem/OE-ECFP/Joint Descriptor	Supplementary Fig. 3
7		Descriptor comparison: OE-ECFP and DRAGON7-ECFP	Supplementary Fig. 4
8		Effect of regularization (by minimum samples per leaf requirement)	Supplementary Fig. 5
9		Active Projection analysis using MCC (VT/NT)	Supplementary Fig. 6
10		Comparison to neural network (ANN) prediction performance	Supplementary Table 1
11		Influence of individual ANN hyperparameters	Supplementary Fig. 7
12	Domain of Applicability	AL-type Domain of Applicability (DoA) analysis, based on external dataset	Fig. 3
13		DoA analysis, based on training data	Supplementary Fig. 8
14		External dataset compounds predictable by only a single descriptor type	Supplementary Fig. 9
15	Structure-activity discontinuities and model impact	Frequency of compound pairwise discontinuities in datasets	Table 3
16		Examples of compound pairs forming discontinuities	Supplementary Fig. 10
17		Consequence of removing borderline non-toxic compounds; discontinuities and exceptional cases	Fig. 4
18	Substructure analyses	Differential fingerprint analysis and example compounds	Fig. 5
19		Bit frequency comparison in training/external datasets	Supplementary Fig. 11
20		Impact of non-BM chemicals as substructures in BM-inclusive compounds	Supplementary Fig. 12

estimator construction algorithm. In the 2016 prediction event, an approach by Mayr et al. to employ a hierarchy of chemical features which were then fed to a deeply-layered neural network produced the highest performance score [20]. Other methods not based on deep learning performed similarly, such as a combination of 10 different chemical descriptors with an Associative Neural Network architecture [2], or a selection of 4071 descriptors which provided good performance [38] when coupled with the Random Forest [7] algorithm. The latter contest organizers concluded that a consensus of model approaches would contribute to optimal performance [12]. Multiple independent reports have commonly reported success from combining multiple representations of chemical data [21,41]. More computational approaches are summarized in a recent review by Tcheremenskaia and colleagues [36].

In a more recent development, the US National Institute for Environmental and Health Sciences has curated a collection of in vivo rat oral acute toxicity data and held a similar community prediction event over the Internet [15,18]. In a preliminary analysis of the data, we identified that individual physicochemical properties, such as lipophilicity and polarizable surface area, trended toward to but were not individually sufficiently predictive of toxicity, nor were combinations of pairs of such properties [25]. An expansion to chemical substructure fingerprints and a repositioning of the predictive question to binary toxicity classification resulted in more satisfactory results, where cross-validated prediction performance was consistent across multiple types of estimator algorithms. One question remaining from that study was the transferability of computational methodology for novel toxicity indications or outcomes, as models were built using many thousands of compounds, which may not be logistically feasible or acceptable in future situations (e.g., the spirit of the REACH legislation). While cross-validation approaches provide insight as to the normality of data, they do not directly indicate if more data leads to better extrapolation on external datasets many times larger than that available for hyperparameter search and estimator construction.

Thus, where our previous analysis had an unmet need which is also important to regulatory agencies in practice – the need to reliably quantify a minimal amount of toxicity data needed for establishing an estimator with defined predictive performance goals and to evaluate such an estimator with best practices in metric interpretation, leading toward method and decision intelligence for regulatory practices. Toward this end, adaptive (or active) machine learning and its application to chemical property prediction has witnessed a revival in the past few years after a preliminary proposal of the concept in 2003 [39]. In short, active learning starts with a minimal set of examples for computing an estimator, makes predictions on a set of examples, and applies a selection function to pick new examples for adding to the training data and re-computation of the estimator. The goal is to find as few examples as possible that provide the same performance as having built an estimator using all possible data. By design, the process is highly dynamic and can be coupled with experimental environments for adaptive exploration of QSPRs. Multiple groups have independently reported the ability of active learning to effectively navigate ligand-receptor interaction data and display early convergence on performance equivalent to models computed from full bioactivity datasets, where the key difference is that models built using active learning only use 5–50% of the data available [16,22,26,30]. Active learning distinctly satisfies the recommendations provided by Rusyn and Daston, which is that models should be capable of being computed fast and cheap (with respect to data volume) [31].

Whereas active learning in life sciences has been investigated primarily in the context of drug discovery, here we study the question of how effectively active learning can be applied to building extrapolative estimators of rat acute oral toxicity with a minimum of data. In addition to examining performance per data volume, we also employ a state of the art visualization technique known as Active Projection [8] for enhanced interpretation of active learning dynamics, and consider the predictability of individual compounds as well as compounds grouped by scaffold frameworks. Results demonstrate effective convergence on limits of predictability regardless of estimator algorithm, quantitative reasoning for such limits, and the revealing of “toxicity cliff” chemical scaffolds, all of which can contribute to future toxicity study size considerations, regulatory agency method advancements, and pharmaceutical or other molecule design. As the number of aspects investigated in this report is considerable, we encourage the reader to refer to Table 1 for a complete list of all hypotheses tested and where to find the corresponding results.

2. Materials

Modeling experiments were based on the 2018 Rat Oral Acute

Toxicity community prediction event [15]. Organizers of the prediction challenge provided approximately 9000 compounds available for training data and approximately 50,000 compounds as an external challenge. Later, a subset of approximately 3000 of the external prediction compounds were released for method development and validation purposes, and used here. The data contains both continuously-valued LD_{50} values as well as explicit binary classification of "Very Toxic" (VT) and "Non-Toxic" (NT) for each compound (respective conditions of $LD_{50} \leq 50$ mg/kg and ≥ 2000 mg/kg). Compounds in the intermediate range were thus both "not very toxic" and "not non-toxic". The validation dataset was designed by the organizers to have the same binary classification data ratios as the training data.

Compounds and activities were provided in tabular format, where SMILES representations of compounds were included. For each compound, its formulation inclusive of salt or solvent and corresponding toxicity value was given. If a compound was tested with different formulations, each compound-formulation-response trio was provided as a separate entity in the table. Rather commonly in chemoinformatics, salts or solvents are stripped from computer representation of compounds, in an aim to reduce the occurrence of spurious structure–-activity correlations. Here as well, we stripped the salt or solvent formulations from entries in the raw data. In the case that this led to a contradiction (i.e., a pair of identical primary structures but with LD50 values on both sides of a toxicity threshold), the data was removed. An example of this is the pair of compounds (CASRN ID) 2610-86-8 (VT = false, K+) and 81–81-2 (VT = true, salt-solvent unlisted) The size of the post-cleaned dataset is as given in Table 2. With respect to molecular size, distributions of the number of atoms per compound were strongly overlapped between the training and validation data (data not shown).

3. Methods

3.1. Compound structure processing

Compound structures were washed to remove salts and metal atoms (OEChem, OpenEye Scientific Software). Contradiction removal was done as described above. Bemis-Murcko fragmentation of compounds was also executed (OEMedChem module).

3.2. Compound representation

A collection of 97 physicochemical properties of compounds, including but not limited to topological polarizable surface area, Moriguchi octanol–water partition coefficient, and the packing density index, were computed (DRAGON 7, Talete S.R.L., descriptor blocks 1/2/20 corresponding to descriptors 1–79 and 4839–4855). Three toxicity predictors (BLTF96, BLTD48, BLTA96) included in the descriptor blocks were removed; analysis demonstrated that they did not show a correlation with the LD_{50} values in the dataset (data not shown). Separately, the structural MACCS keys and the extended circular fingerprint representations (ECFP) were also computed (independent ECFP implementations in OEChem and DRAGON7). ECFP representations were tested using atom radii 0-2 and 512/1024/4096 bits. DRAGON ECFPs were also tested for the use of 1 or 2 bits per pattern. Finally, the CATS-2D pharmacophoric representation of compounds [32] was computed as well (DRAGON7).

Table 2
Rat oral acute toxicity datasets used in this study. While the focus of this work was on binary prediction, we include counts of compounds where LD_{50} is also available.

Training data statistics				
Endpoint	Very Toxic		Non-Toxic	
	Compounds	LD_{50} Given	Compounds	LD_{50} Given
TRUE	741 (8%)	721	3787 (43%)	2137
FALSE	8133 (92%)	6013	5076 (57%)	4597
Validation data statistics				
Endpoint	Very Toxic		Non-Toxic	
	Compounds	LD_{50} Given	Compounds	LD_{50} Given
TRUE	243 (8%)	235	1235 (43%)	687
FALSE	2651 (92%)	1939	1655 (57%)	1487

3.3. Estimator construction and active learning

Active learning was implemented and executed as reported in previous literature [30,28]. Iterative random selection of compounds to update toxicity models and their predictive performance was used as a control experiment, after which selection by uncertainty/explorative and greedy/exploitative picking methods were evaluated. The uncertainty picking method is also referred to as the curiosity-based picking method [30,29], as compounds with the maximum disagreement (ensemble estimator, including decision trees) or uncertainty in a model can be interpreted as curious examples to learn from. Greedy picking selects compounds with the highest probability or number of votes. In each dataset and picking strategy setting, an active learning modeling experiment is repeated 10 times, where each execution uses a different random seed to start the process with one positive and one negative compound (e.g., one VT = true and one VT = false). The underlying classification estimator used was the Random Forest algorithm. Experiments were performed to observe the effect of regularization via a minimum number of samples per decision leaf . Both retrospective active learning of the available training data and evaluation of predictive ability on the external validation data were performed. In the retrospective context, one would expect perfect prediction performance at the final iteration for a random forest with minimum of one sample per decision leaf, where all training data is used for estimator computation and subsequent recall prediction evaluation.

In follow-up experiments to check the impact of different underlying estimators on predictive ability, neural networks were constructed by computing models using the scikit-learn [23] and TensorFlow [1] packages independently. A grid of regularization parameters, network topologies, learning rates, epochs, and learning batch sizes were systematically tested.

For the scikit-learn implementation, the "relu" activation function and "adam" solver were applied to a grid search using L2 regularization values of 0.001, 0.01, 0.1, 0.5, 1 and 5. In using a joint descriptor representation of compounds, models were computed using the following topologies: 1 × 200, 500–200, 500–500–200, 3 × 100, 4 × 50, and 5 × 50 (where YxZ means Y layers fully forward connected, Z units per layer). For ANN models using MACCS keys, the topologies were: 6 × 50, 6 × 100, 8 × 50, 8 × 100, 10 × 50, 10 × 100. The models were built using the full training dataset available, where here a protocol decision was made to discard the final 5% of compounds picked by an execution of active learning, so as to investigate any variance in ANN predictions resulting from minor changes in the dataset. As above with active learning, 10 executions of modeling and prediction were executed.

For the TensorFlow neural networks, layer depth, units per layer, number of epochs, batch size, and learning rate were all varied and assessed for impact on performance. ECFP descriptors were used. Parameter ranges include: layer count = 2/3/4/5/10; units per layer = 10/25/50/100/200/300; number of epochs = 10/40/80/100/120; batch size = 10/20/40/80/200/500; learning rate = 0.001/0.005/0.01/0.1. A direct combinatorial grid search was substituted with the systematic investigation of the impact of each parameter when

holding other parameters constant, after which a candidate set of parameters was obtained and further optimization in the ranges above was attempted.

Further, models using the mathematically optimal SVM algorithm [33], as implemented in scikit-learn, were computed. The same protocol as the scikit-learn ANN model computations (using 95% of training data per experiment) was repeated. A grid search using the following parameters was tested: tolerance/loss/C = 0.001/0.01/0.1/1/10/100; kernels = linear/RBF; RBF kernel parameter gamma = 0.001/0.01/0.1/1.

3.4. Evaluation metrics

Per recommendation from previous studies [9,26] active learning was principally evaluated using a three-metric approach. First, the Positive Predictive Value (PPV) was used to assess the correct prediction rate when an estimator predicted a compound to have a "positive label". Especially for the Very Toxic dataset, this directly addresses key regulatory policy decision making when computational models predict compounds to be very toxic.

Also important is to quantify when an estimator fails to detect a positive label (Type-II error, False Negative). While measures such as the Power Metric [17] or G-Mean (as used for toxicity classification evaluation by Moorthy et al. [21]) can address this, application of the Metric Surface method [9] shows that these metrics have the potential for over-estimation of performance, and we instead use the more strict F1 criteria which, similar to the Power Metric and G-mean values, incorporates the True Positive Rate (TPR) value, but with more penalty for poor negative class prediction performance.

Finally, we employ Matthews' Correlation Coefficient [19] which incorporates all four results of the confusion matrix with multiplicative penalty for misclassification.

In a comparison of the DRAGON and OEChem implementations of ECFPs, we also adapt the Enrichment Factor (EF) metric used in drug discovery. In short, this metric quantifies how many positives (e.g., very toxic compounds) one can detect against the background of the number of negatives in the dataset. If a dataset is highly imbalanced towards negatives, a common scenario, then a predictor with high EF values signals the ability to correctly recognize positives despite the imbalance.

We also considered the Balanced Accuracy metric. This is a data ratio-invariant metric, and was used as a way to assess models in the 2018 AcuteTox prediction event.

The formulas used to compute the metrics are as follows.

$$PPV = TP / (TP + FP)$$

$$F1 = (2 * PPV * TPR) / (PPV + TPR),$$
$$where\ TPR = TP/(TP + FN)$$

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

$$EF = PPV / [(TP + FN)/(TP + TN + FP + FN)]$$

$$BA = (TPR + TNR)/2,$$
$$where\ TNR = TN/(TN + FP)$$

3.5. Active projection

Any binary classification metric used can be visually assessed under the context of positive–negative data ratio for the relationship between all true possible positive rates, all possible True Negative Rates (TNR), and the resulting metric value, a technique called the Metric Surface method [9]. As an extension of this, Brown proposed to project the results of active learning onto metric surfaces [8], which provides a dynamic visual delineation of an estimator's ability to distinguish between the two classes as the volume of training data grows. The active projection used here is an in-house implementation.

3.6. Pareto front calculation

Models are selected for analysis at intermediate stages of active learning by applying Pareto Front analysis [10]. In pareto analysis, the set of conditions by which optimal tradeoff amongst multiple objectives is selected. Here, we employ the technique to identify the amount of (non-)toxicity data required to yield maximum performance with an optimal balance between TPR and TNR in an active learning modeling experiment (in-house implementation).

4. Results and discussion

We again refer the reader to Table 1 to follow the questions addressed below with convenient reference to the relevant figure or table demonstrating results.

4.1. Molecular scaffolds and forecasts of predictability

Compounds were organized according to their Bemis-Murcko (BM) frameworks, where a framework is the collection of ring systems and the minimum number of linker atoms required to connect the ring systems. The spread of toxicity values per framework was evaluated for frameworks with multiple compounds, as shown in Fig. 1. This resulted in subdividing frameworks into groups that were very toxic, non-toxic on average yet containing modifications that could result in a very toxic compound and therefore leading to the group label of "volatile", non-toxic on average but still with cautionary levels of toxicity, and finally non-toxic frameworks with all non-toxic members (deemed "safe"). We provide examples of each type of framework classification group in Fig. 1.

In particular, the volatile group indicates the presence of "toxicity cliffs", which are analogous to the concepts of activity cliffs for medicinal chemistry [40]. Since the framework of two compounds in a volatile group is identical and would lead to identical descriptor representation for at least the common framework, we forecasted that these types of compounds would present a challenge for estimator calculation and external prediction. Since we principally employed the random forest algorithm as the estimator method, we expect that a pair of compounds in a toxicity cliff group would result in a pair of decision leaves that are within a proximal distance to each other. In Supplementary Fig. 1, we have provided an expanded version of Fig. 1 which demonstrates per-scaffold average toxicity for all 271 scaffolds in the training dataset.

4.2. Data volume, external predictability, and picking strategy evaluation

Control experiments using retrospective active learning (MACCS key representation) demonstrated that random selection of compounds required nearly the full dataset in order to achieve high predictive MCC (Fig. 2, top). This was true for either the very toxic or non-toxic datasets. In contrast, selection by either the curiosity or greedy strategies was substantially more effective at picking compounds in the very toxic dataset, achieving MCC values of more than 0.8 within 30% of the available data (Fig. 2, top-left). The curiosity picker is also the most effective in the non-toxic dataset, achieving an MCC of 0.8 within 40% of the available data (Fig. 2, top-right). Analysis of the same time-series performances using each of the BA, F1, TPR, and TNR metrics is shown in Supplementary Fig. 2.

Evaluation on the external dataset, which is never made available to active learning for picking and inclusion in training, demonstrates limits of predictive ability in the range of 20–40% of the available training data. No amount of additional data beyond such yields an improvement in external classification (Fig. 2, top). For the imbalanced

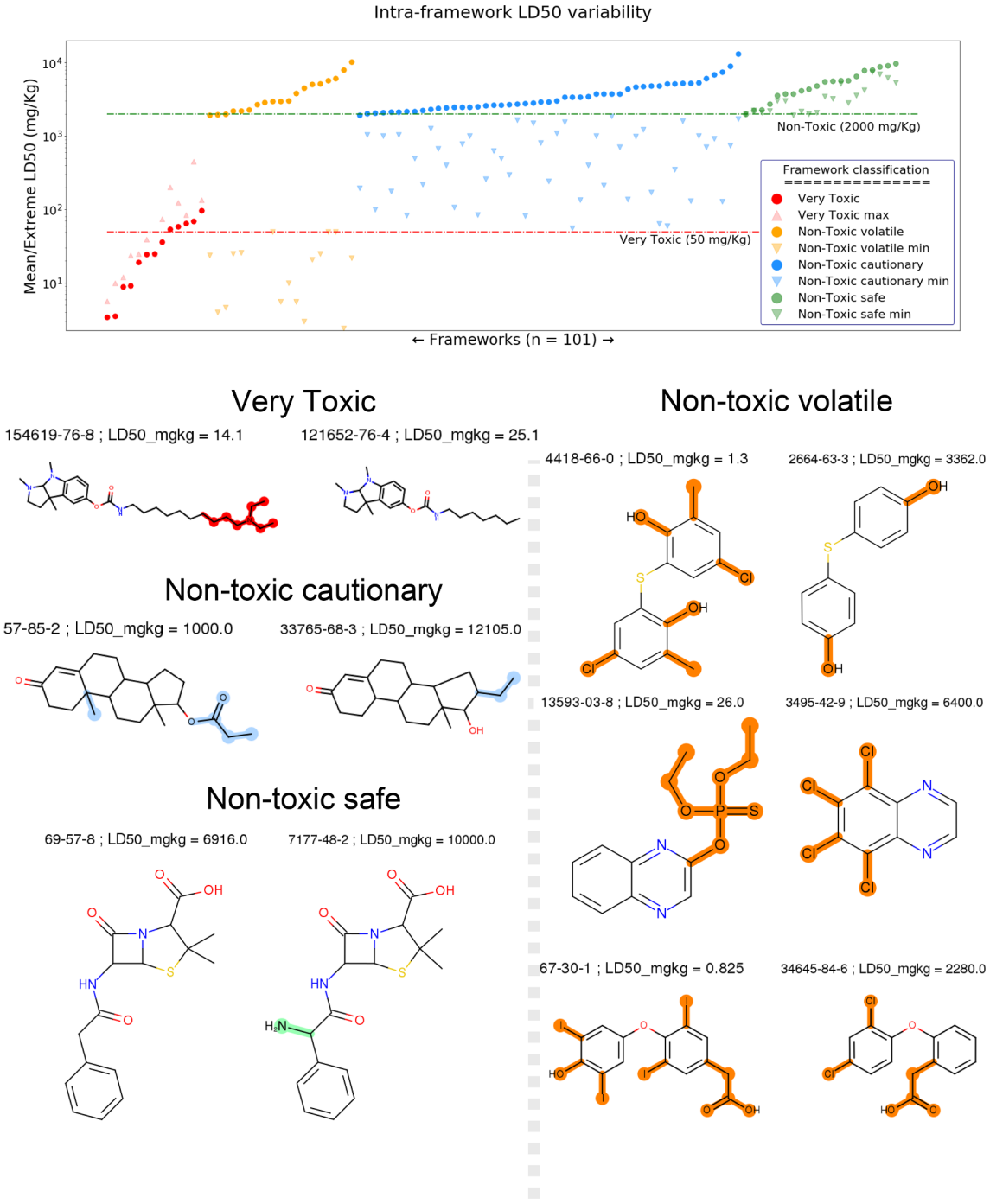

Fig. 1. Chemical frameworks and their toxicity classes. BM frameworks may be very toxic altogether, non-toxic on average but containing "toxicity cliffs", non-toxic but requiring caution, or generally safe. For illustration purposes, a framework was placed in the very toxic class if its mean LD_{50} was less than 100 mg/kg. Differences in structure pairs are highlighted using the color of the toxicity class.

very toxic dataset, either selection function demonstrably outperforms the control experiment using random picking. This was consistent across all of the descriptors tested (Supplementary Fig. 3); similar to Fig. 1 for the NT classification problem, picking via the greedy strategy was less performant than random picking in early stages but was on par or superior in middle to late stages of active learning when testing different descriptors.

Another observation from the descriptor comparison experiments was that the structural fingerprint descriptions of compounds (ECFP and MACCS) yielded higher performance on the external data than the physicochemical or pharmacophore representations (Supplementary Fig. 3). However, it was interesting to note that a combined descriptor of all four individual types was more predictive on the external dataset than any individual descriptor. As the combined descriptor represents both continuously-valued properties of compounds as well as both experience- and data-based substructure fingerprints, the extra resolution provided alternative decision tree boundary formulation, and the improved external performance signals that the higher-resolution rulesets transferred better to the external dataset.

We queried if different implementations in the ECFP algorithm would yield different results. A head-to-head comparison of the ECFP implementations in the DRAGON and OEChem packages, using an

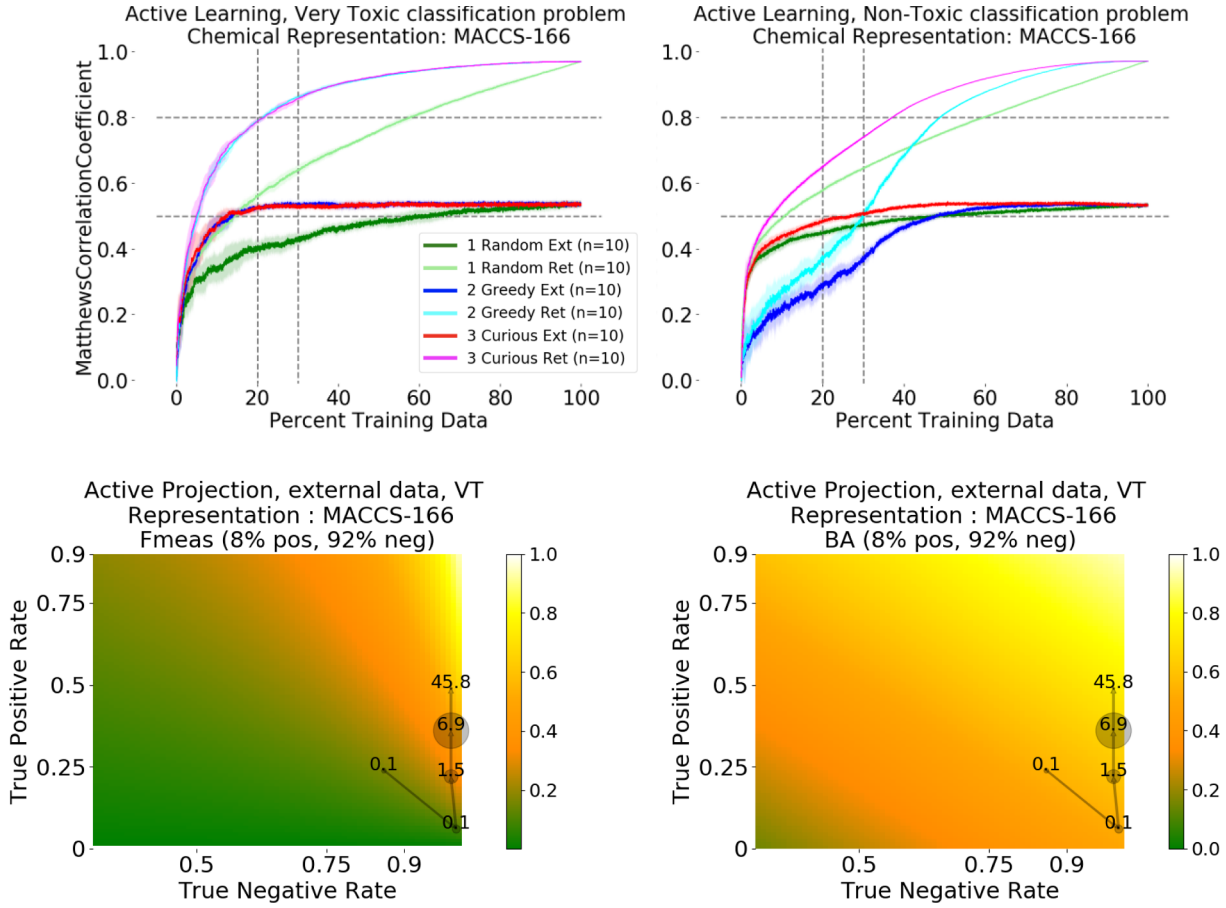

Fig. 2. Prediction performance by adaptive learning and model computation. Active learning iteratively picks examples to add to training data to find a minimal set of compounds which maximally predict (non-)toxicity. [Top] Retrospective ("Ret") experiments query the predictive ability on the full dataset available for selection and inclusion. In both VT and NT classification problems, maximum prediction performance on the external validation ("Ext") data is achieved with less than half of the available training data. [Bottom] Active projections deconstruct metric performances temporally and against the context of data ratio, enriching interpretation of individual performance metrics. Model performances are projected when there is significant difference in TPR/TNR over a previous point (here, equivalent of more than 10% shift in TPR/TNR metrics). Point size is scaled by the relative amount of data required before the next point projection. Using either F1 or balanced accuracy as the primary metric, active projection clarifies that the external dataset performances reported in the top left panel correspond to early decision rules classifying everything as non-VT after which predictive ability on very toxic compounds is gradually achieved.

identical setting of atom radius 2 with 4096 hash bits, showed no difference in external prediction performance for either the VT or NT prediction problems (Supplementary Fig. 4, top). As above with MACCS keys, performance on the external set was convergent at 20–40% of data strategically picked. We executed further experiments to check the impact of atom radius and numbers of bits, finding little difference in external prediction results (Supplementary Fig. 4, middle). The only noticeable difference resulting from ECFP parameter settings was a minor drop in enrichment factor on the external set, where smaller radii (of 0 or 1) showed a reduction in EF from 10 to 9 (max EF approximately 12 for the VT problem; see Supplementary Fig. 4, bottom). Beyond the 10–15% training data range, the EF value continuously declined, albeit the absolute decline was quite small.

We also tested if constraining decision trees to contain multiple samples in decision leaves, a form of regularization, would lead to simpler models that could still be predictive. Using the MACCS keys, we found that we could reduce over-fitting by forcing a minimum of 2 samples (compounds) per decision leaf with the resulting model achieving MCC of approximately 0.5 on the VT external data (Supplementary Fig. 5). Further compression by increasing the samples per leaf led to small reductions in MCC for external predictions, though compressing trees by forcing at least 5 compounds per leaf still yielded MCCs of 0.4 and 0.5 on the external VT and NT datasets, respectively. For the VT dataset, external predictive convergence was achieved at 20% of the training data.

Another interesting observation from this experiment is the interpretation of random forest model behavior as a consequence of data volume and samples-per-leaf requirements. We can see in the retrospective experiments that the maximum MCC achieved for multiple samples per leaf is at approximately 30% of the training data. Yet, the performance for a single compound per leaf continues to grow until a perfect predictor is achieved (Supplementary Fig. 5). We can deduce that the remaining 70% of this dataset (~6000 compounds) results in specialty singleton decision nodes which can be recalled but cannot be grouped with the initial 30%. When we checked the fraction of very toxic compounds picked at that stage (3000 compounds picked), we found it consistently to be 16% (1/6) for 2~3 samples per leaf (curiosity picker). The fraction reproducibly decreased to 15% for 4~5 samples per leaf, indicating that the non-VT compounds were increasingly picked as a consequence of divided prediction results from the individual decision trees when such trees were constrained by high amounts of regularization (samples per leaf).

4.3. Tracking of model dynamics

In order to trace more specifically how active learning improved through strategic selection of compounds, we computed Active Projections for the experiments performed. While it was the case that

performance converged on the value of MCC = 0.5 when predicting external compounds for either the very toxic or non-toxic prediction datasets, how those values were converged upon was considerably different, as shown by the curiosity picking-based active projections at the bottom of Fig. 2.

For the VT problem, the first few instances chosen by active learning contain patterns that allow it to identify the toxic compounds but fail to detect non-very toxic compounds. The rules of the decision trees formed at this stage can be expected to be simple, such as "all compounds with halogen atoms are very toxic". After several more picks, the rules of the random forest are updated and can successfully predict non-toxic compounds, but have lost their ability to identify the very toxic compounds (low TPR, Fig. 2 bottom). After more iterations of active learning, the decision trees comprising the model incorporate more decision rules leading to very toxic classifications, and these rules successfully identify very toxic compounds in the external set. In contrast to brief decision stubs of toxicity early on, identification of toxic compounds here can be expected to use multiple branches (decisions from features) to handle a variety of chemical substructures. Active projection compactly describes the temporal TPR-TNR dynamics, and it further demonstrates that there is a limit to the amount of training data that improves external prediction performance. Even further, we see only a modest TPR improvement between models built from 6% of the training data and those built from 45% (Fig. 2), where the limit of external prediction performance is encountered. At both data volumes, the PPV is consistently 0.85. The similarity of the PPV values at these two points indicates that while fewer compounds were being missed as very toxic (increasing TPR), false positive predictions (non-VT compounds predicted as toxic) were growing at the same rate as true positive predictions. The BA metric value also has only a marginal gain between these two data volumes (in both cases close to 0.70), as seen by the ratio-aware active projection in Fig. 2 or the simple BA time-series plot in Supplementary Fig. 2.

Tracking the non-toxic prediction problem dynamics by active projection tells a different story (Supplementary Fig. 6). Here, the model is initially performant for "not non-toxic" molecules, and then examples are picked which lead to performance for non-toxic molecules but poor predictive ability for the opposite class. However, in this prediction problem scenario, active projection shows that a more balanced model is achieved at the 2.4% data volume (PPV = 0.66, F1 = 0.55, MCC = 0.32), and subsequent picks of data lead the model on continuous balanced prediction improvements, where the pareto optimal models are achieved at 36% and 57% of data, with minimal difference in their TPR and TNR rates at the two stages, and these two optimal models are relatively close in performance to a smaller model with only 14% of data.

4.4. Influence of estimator methodology

Artificial neural networks have enjoyed a substantial renewed interest in recent years, with deeply-layered neural networks (DNNs) receiving large amounts of attention. Individual nodes of ANNs represent individual decision functions, typically linear discriminants or sigmoidal functions, and thus bear similarity to decision branches in decision trees. A cascade of discriminant units constitutes a neural network's structure, much like a decision tree employs many decision branches. The difference between the two methodologies is in the mathematical structure of the discriminants and cascades. Thus, we wished to know if ANNs, with more variation in discriminant formulations and cascade structures, would lend it to enhanced ability for VT and NT classification, and executed an evaluation experiment.

We enumerated feedforward topologies using between one and ten hidden layers, executed individual hyperparameter optimization, and employed the full or near-full training dataset, yet we found that optimal prediction performance on the external data as measured by MCC value was identical to the maximum obtained by RF-based active learning (Supplementary Table 1). There was no correlation between the number of layers and the resulting MCC, suggesting that topologies beyond a single hidden layer with 166 hidden decision nodes (the same number of MACCS keys) did not substantially contribute to the decision hypersurface such that better separation of compounds by toxicity class occurred. Switching between 10 and 200 nodes per layer with a constant five layers also showed no change in predictive ability (Supplementary Fig. 7). The only parameter where large changes in predictive performance was observed was the learning rate of the network, which tunes the speed of backpropagation and empirical convergence on each hidden node's underlying activation function parameters; if this parameter was too large, predictions were very poor (Supplementary Fig. 7).

Considering the toxicity cliffs shown in Fig. 1, it is rational to believe that the ANN models encountered the same issue which led to active learning's asymptotic performance – that the similarity principle (similar compounds have similar properties) does not hold for many compound pairs, and therefore that the ANN and RF models produce similar performance because such performance is the maximum that an estimator can logically achieve under the premise of the similarity principle.

Further experiments with SVMs were undertaken, in this case using the joint descriptor representation of the compounds. Results yielded MCC values nearly identical to those obtained by RF models (data not shown). Thus, it was not the case of how the estimators were formulated and tuned, but rather how the discontinuity in the datasets impacted the prediction performances.

4.5. Domain of applicability assessments

Models developed based on full datasets may often also derive a domain of applicability (DoA). In previous toxicity prediction efforts, multiple groups defined DoAs by restricting predictions to those compounds such that their descriptor representations and endpoints were respectively neighboring and consistent [3,6,11]. This ensures a smooth representation-endpoint manifold, and these groups achieved high PPV values by such DoA filters.

Active learning brings a different approach to modeling by using only a subset of data in a dynamic, adaptive fashion, and thus it requires an alternative strategy to identify a DoA for an actively learned model. Here, we considered that the DoA could be defined by identifying those compounds which could (or more importantly, could not) be predicted at some point during the iterative compound selection and modeling processes. In Fig. 3, we consider the predictability of individual compounds in the external dataset, where all compounds are from the very toxic class. As can be seen, those VT compounds which are predictable are predominantly predicted as such within the first 15% of data strategically selected (curiosity picking). A few compounds can also be seen to eventually be predictable at 30–45% of training data picked, but beyond this, there are no major changes in the predictability of compounds. Consistent with Fig. 2 and Supplementary Fig. 2, we see that approximately 45% of the external VT compounds were successfully predicted. The remaining 55% comprise the compounds which fall outside of the DoA. When a new structure is to be predicted for toxicity, we could consider the most similar compounds in the external dataset and whether or not they were adaptively predictable or not. In regards to minimizing the amount of animal testing performed, those compounds with maximum disagreement from multiple model votes could then be construed as the ones with the most uncertainty, and advanced to animal testing.

Alternatively, we can consider the result of applying the same methodology to the available training data. As shown in Supplementary Fig. 8, most of the compounds in the training data are emergently predictable, and again predictability is largely unchanged after selecting and learning half of the training data. For those compounds which become predictable after the halfway mark, we can consider the

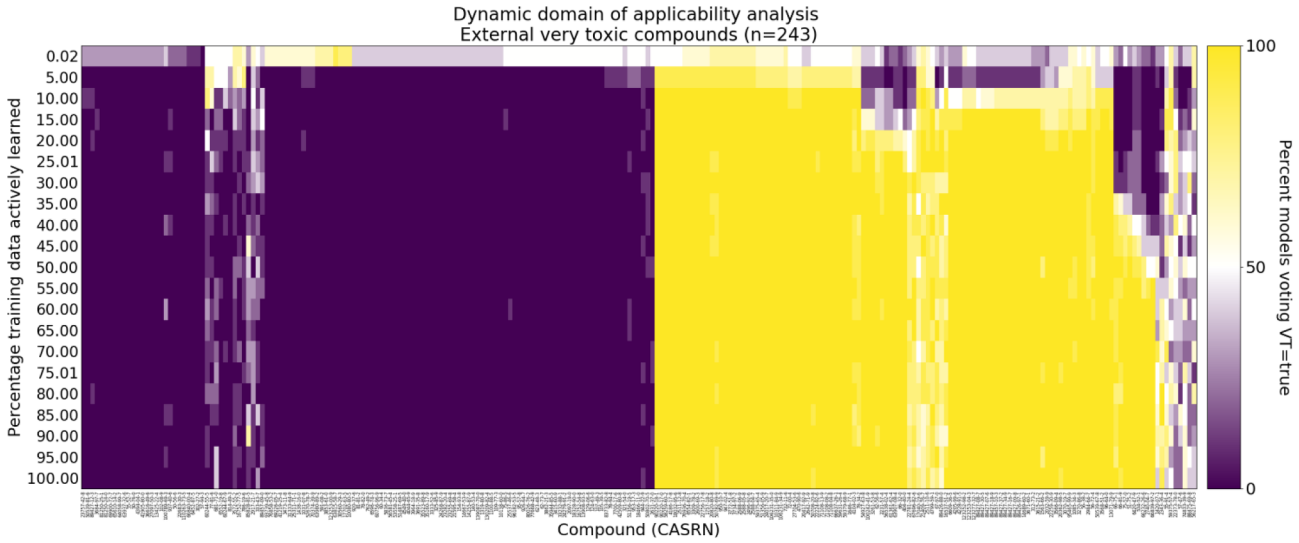

Fig. 3. Domain of Applicability (DoA) assessment for toxic compound prediction via active learning. In 10 replicate experiments using different starting compounds, models adaptively learn the rules to separate very toxic from non-VT compounds in the training data; at each stage of model update, models also predict a VT label for external compounds. The percentage of model votes (color) per unit data volume (vertical axis) demonstrates compounds which are clearly within the DoA, compounds clearly outside the DoA (dominantly purple columns; false negatives), compounds which are initially outside the domain but eventually are within (purple to yellow transition), and compounds which are borderline (white). Beyond 50% of the training data, no changes in the DoA are present. Results using MACCS fingerprints.

regularization analysis (samples per decision leaf) earlier as a plausible explanation for the mechanics of how these compounds eventually were correctly classified.

Using the different descriptor representations of the compounds, we repeated the DoA analysis for very toxic compounds in the external dataset. We found that most compounds that were predictable by MACCS keys were predictable with at least one other representation, but we also could identify specific compounds that were only predictable by a single particular representation. Some examples of these compounds are shown in Supplementary Fig. 9.

4.6. Compound representation and challenges from discontinuities

Despite our pre-processing to remove contradictions in data (identical chemical compound with conflicting classification labels) and thus reduce the degree to which the modelable compounds contained "discontinuities", a number of compound structures in the external dataset were highly difficult to predict, due to their structures resembling others of opposite class. We performed further analyses to quantify how many discontinuities were present between the external compound dataset and the reference training library. Compound pairs with MACCS-Tanimoto values of 0.8 or higher were considered similar.

As shown in Table 3, for the VT dataset, there were a total of 1257 discontinuities between the external and training data, in which 691 (55%) pairs were of external compounds that were very toxic but in which a similar compound in the training data had a non-VT label. For instance, 2-methyl-4,6-dinitrophenol (534-52-1) has a reported LD_{50} value of 7.0 mg/kg; trinitrophenol (88-89-1) has an LD_{50} of 200.0 mg/kg, yet the MACCS-Tanimoto similarity of the two is 0.972. 2-methyl-4,6-dinitrophenol forms 25 such discontinuities with the training data; we should not be expectant of the machine to predict compounds such as this correctly. Further examples of discontinuous compound pairs are provided in Supplementary Fig. 10, including seven small fragment compound pairs with opposite labels despite identical MACCS fingerprints. Comparing the 158 compounds that form external-training discontinuities to the size of the external prediction data, it should come as no surprise that the upper limit of very toxic compound classification (true positive rate) on the external dataset is just under 50%. Analogous discontinuity analyses were performed for the external-training dataset of the NT classification problem and the intra-training datasets, with statistics provided in Table 3.

Table 3
Frequency of discontinuities in datasets. Discontinuities are defined as pairs of compounds with MACCS-Tanimoto similarity of at least 0.8 but with opposite classification labels.

Dataset discontinuity frequency				
	External-training		Training intra-dataset	
	Discontinuity pairs	Number of compounds	Discontinuity pairs	Number of compounds
VT = True	691	158	1121	471
VT = False	566	257	1121	750
NT = True	1487	519	2478	1656
NT = False	1512	521	2478	1638

In previous work on active learning, Brown and colleagues enforced the use of a gap between the definition of active and inactive compounds [24,27,30]. While the pre-assigned binary label classification system was primarily used for results reported, we hypothesized that borderline non-VT compounds had the potential to cause interference in building clear rules separating very toxic compounds from others. To test this hypothesis, we removed training set compounds with LD_{50} values in the non-inclusive range of 50–250 mg/kg, and re-evaluated the actively learned model on the external dataset. As shown in Fig. 4, the TPR, BA, and F1 metrics were all improved when predicting the external very toxic compounds. Curiously the PPV was decreased. The interpretation of this potentially confusing outcome can be resolved by considering the relationship between TPR, PPV, and F1: while the rate of false negatives declined (increase in TPR/BA), the rate of false positives was increased compared to inclusion of the borderline compounds (decrease in PPV), but the gain in TPR was larger than the loss in PPV, and hence the F1 metric was improved. The MCC was not substantially affected (data not shown). Thus, the consideration of prefiltering borderline non-VT compounds in an effort to generate models with better defined boundaries and improve the toxic compound identification prediction rate is a viable strategy in future regulatory policy.

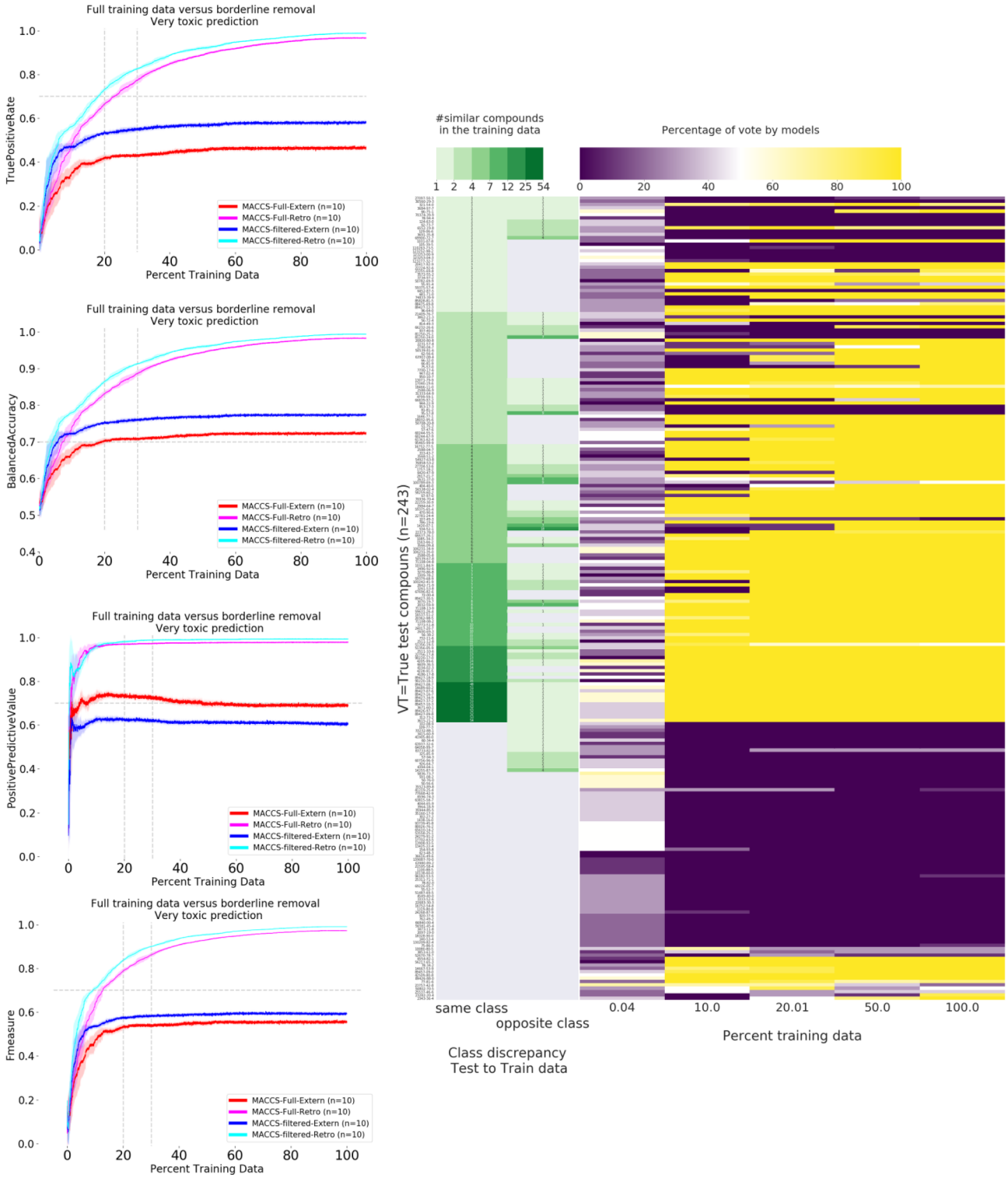

Fig. 4. Consequence of removing borderline non-toxic compounds prior to external prediction. Training data compounds in the LD50 range (50,250) were excluded from training data, and active learning was performed. [Left] The TPR and BA metrics for the external dataset were higher after this filter was applied, whereas the PPV was decreased; the final overall F1 was still improved, while the MCC was not significantly changed (data not shown). [Right] External VT = true compounds were re-assessed for the DoA, considering the number of similar compounds. The compounds that were unpredictable (FNs, purple) were dominantly those with no similar compounds in the training data.

4.7. Substructure-endpoint differential fingerprint analysis

As different estimator algorithms were equally capable of identifying compounds in the external dataset which were very toxic (Section 4.4), it suggested that there must be the presence of fingerprints which were more represented in very toxic compounds in the training data, that these underpinned the successful predictions on the external data, and that these could potentially be interesting as structural alerts. By considering the normalized frequency of bits in very toxic compounds versus those that are not-VT, we can identify such differential

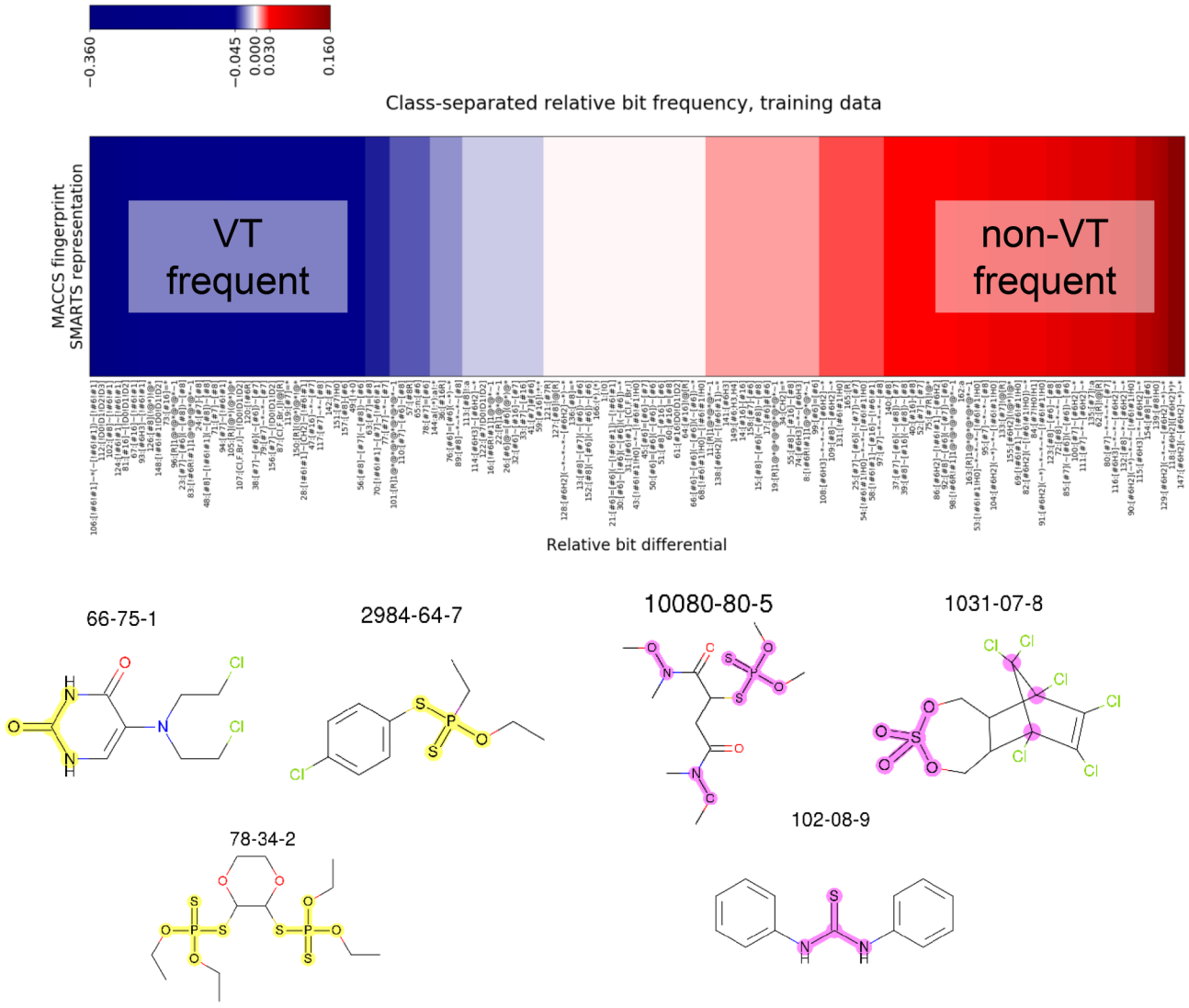

Fig. 5. Differential bit analysis of very toxic compounds. The normalized frequency in the training data of each MACCS key was computed for each of the VT = true and VT = false groups. Differential comparison identified patterns more- and less-frequent in the groups. Examples of true positive (yellow) and false negative (purple) compounds containing these patterns are shown.

substructures. Using the MACCS keys, examples of the patterns identified from the training set are given in Fig. 5.

We then queried if the very toxic compounds in the external dataset contained similar ratios of these substructures. As shown in Supplementary Fig. 11, the frequencies of bits were highly concordant. This included a check of the normalized frequencies of the MACCS bits across the very toxic and non-VT classes. This scatterplot combined with the fact that some compounds contained the most differential MACCS bits but were still false negatives hints that the relative ratio of data classes and absolute counts of bits interfered with the ability to build a more performant model. Also, we can observe from Supplementary Fig. S11 that, unlike very toxic compounds, there are few MACCS bits that have substantially large relative frequency for non-VT compounds (e.g., > 0.3 as can be found for some very toxic compound compound bits). Still, manual inspection of specific SMARTS on the extremes of the heatmap in Fig. 5 led to the impression that highly aliphatic compounds tended to non-toxic labels while tertiary heavy atoms or non-carbon heavy atoms within 1–2 bonds tended to very toxic labels.

4.8. Toxic singletons embedded in larger compounds

We considered that some small, non-framework compounds might be present as substructures in larger compounds, and that the relationship between sub-structure and super-structure might provide ideas for molecular design and modification. For each compound in the training data which did not contain a Bemis-Murcko scaffold, we considered its toxicity and the toxicities of all superstructure compounds which contain the non-scaffold compound as a substructure. The frequency of such relationships is represented by the heatmap presented in Supplementary Fig. 12.

In the figure, specific examples of the substructure-superstructure relationship and subsequent differences in LD_{50} value are also given. For example, the 2-atom cyano moiety can be found in the compounds belonging to the very toxic, volatile, and cautionary scaffold groups. The substructure is also present in a scaffold which does not belong to any of our four groupings, and further inspection clarifies that LD_{50} does not simply follow a monotonic trend based on superstructure size. Other fragment-scaffold relationships shown also demonstrate that small fragments with high individual toxicity could still be potentially used (i.e., to increase van der walls contacts or electrostatic interaction) in molecular design. Clearly, structural, electrostatic, and spatial

considerations strongly influence whether a compound is toxic or not, and in particular when considering the physiological context which often includes target receptor engagement.

5. Conclusions

As the search continues for new methodologies to minimize the extent of animal testing, we have demonstrated the benefit of considering asymptotic limits in modeling methods. Active learning achieved the same predictive performance as an optimized neural network, but required only 30–40% of the available data to do so. We also employed quantitative approaches to query how well-posed (or challenging) the prediction problem setting was. Thus, the technique represents a way for regulators to make projections about the cost-benefit tradeoff associated with expanded toxicity evaluation for computational model developments, and what to rationally expect from prediction models. Scaffold analysis uncovered structures which can be systematically flagged for caution in the virtual screening of large catalogs and databases, and substructure-superstructure analysis clarified the importance of context when analyzing small toxic fragments.

Taken together, this work provides insights which can help chemical designers, environmental analysts, regulatory agencies, and toxicology research groups. While we could not find such data in the public domain, a clear way to improve the analysis would be to link chemically-driven rat oral acute toxicity with receptor protein interaction on a proteome-wide scale. Still further, if such data were available, then distinct modes of toxicity could be computed from pathway scoring analyses. Hence, there is much to be explored and established in chemical-phenotype predictive toxicology.

CRediT authorship contribution statement

Ahsan Habib Polash: Methodology, Software, Formal analysis, Investigation, Data curation, Writing - original draft, Visualization. **Takumi Nakano:** Methodology, Software, Validation, Investigation. **Christin Rakers:** Methodology, Software, Validation, Formal analysis, Investigation, Writing - original draft, Visualization. **Shunichi Takeda:** Methodology, Writing - original draft, Writing - review & editing, Supervision, Funding acquisition. **J.B. Brown:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: JBB declares a potential conflict of interest as a consultant for the pharmaceutical industry.

Acknowledgements

The authors with to thank Drs. Nicole Kleinstreuer and Agnes Karmaus for comments and assistance during the development of the study. Discussions with Dr. Jan Hiss of ETH Zürich regarding the active projection method were fruitful. This work was supported by grants 16H06306 (to ST, JB) and 17K20043 (JB) from the Japan Society for the Promotion of Science. An academic license and support from Krisztina Boda and Lucas Zimney at OpenEye Scientific Software is gratefully acknowledged. We thank Alberto Manganaro of Kode Solutions for assistance with the DRAGON software. Compute resources used were supported in part by a grant from Daiichi-Sankyo Pharmaceutical Company, Japan (TaNeDS grant A60093), which was not involved in the conceptualization, execution, or reporting of this study.

Author contributions

JB and ST conceived the study. AP and JB analyzed chemical structures, executed active learning experiments, and performed domain of applicability analyses. AP performed contradiction checking, data cleaning, differential fingerprint analysis, and substructure visualizations. AP and TN generated active projections. JB performed discontinuity analyses and SVM evaluation. TN and CR performed neural network modeling and evaluation. AP, ST, and JB drafted the manuscript. All authors read and approved of the final manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.comtox.2020.100129.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, et al. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. Available at: http://arxiv.org/abs/1603.04467 [Accessed November 22, 2019].
[2] A. Abdelaziz, H. Spahn-Langguth, K.-W. Schramm, I.V. Tetko, Consensus modeling for HTS assays using in silico descriptors calculates the best balanced accuracy in Tox21 challenge, Front. Environ. Sci. 4 (2016) 2, https://doi.org/10.3389/fenvs.2016.00002.
[3] D. Alberga, D. Trisciuzzi, K. Mansouri, G.F. Mangiatordi, O. Nicolotti, Prediction of acute oral systemic toxicity using a multifingerprint similarity approach, Toxicol. Sci. 167 (2018) 484–495, https://doi.org/10.1093/toxsci/kfy255.
[4] B.N. Ames, J. McCann, E. Yamasaki, Methods for detecting carcinogens and mutagens with the salmonella/mammalian-microsome mutagenicity test, Mutat. Res. Mutagen. Relat. Subj. 31 (1975) 347–363, https://doi.org/10.1016/0165-1161(75)90046-1.
[5] P. Anastas, K. Teichman, E.C. Hubal, Ensuring the safety of chemicals, J. Expo. Sci. Environ. Epidemiol. 20 (2010) 395–396, https://doi.org/10.1038/jes.2010.28.
[6] D. Ballabio, F. Grisoni, V. Consonni, R. Todeschini, Integrated QSAR models to predict acute oral systemic toxicity, Mol. Inform. 38 (2019) 1800124, https://doi.org/10.1002/minf.201800124.
[7] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, https://doi.org/10.1023/A:1010933404324.
[8] J. Brown, Adaptive mining and model building of medicinal chemistry data with a multi-metric perspective, Future Med. Chem. 10 (2018) 1885–1887, https://doi.org/10.4155/fmc-2018-0188.
[9] J.B. Brown, Classifiers and their metrics quantified, Mol. Inform. 37 (2018), https://doi.org/10.1002/minf.201700127.
[10] C.M. Fonseca, C.M. Fonseca, P.J. Fleming (1993). Genetic Algorithms for Multiobjective Optimization: Formulation, Discussion and Generalization. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.48.9077 [Accessed May 1, 2020].
[11] D. Gadaleta, K. Vuković, C. Toma, G.J. Lavado, A.L. Karmaus, K. Mansouri, et al., SAR and QSAR modeling of a large collection of LD50 rat acute oral toxicity data, J. Cheminform. 11 (2019) 58, https://doi.org/10.1186/s13321-019-0383-2.
[12] R. Huang, M. Xia, Editorial: Tox21 challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental toxicants and drugs, Front. Environ. Sci. 5 (2017) 3, https://doi.org/10.3389/fenvs.2017.00003.
[13] D. Kirkland, M. Aardema, L. Henderson, L. Müller, Evaluation of the ability of a battery of three in vitro genotoxicity tests to discriminate rodent carcinogens and non-carcinogens: I. Sensitivity, specificity and relative predictivity, Mutat. Res. Toxicol. Environ. Mutagen. 584 (2005) 1–256, https://doi.org/10.1016/J.MRGENTOX.2005.02.004.
[14] D. Kirkland, S. Pfuhler, D. Tweats, M. Aardema, R. Corvi, F. Darroudi, et al., How to reduce false positive results when undertaking in vitro genotoxicity testing and thus avoid unnecessary follow-up animal tests: report of an ECVAM workshop, Mutat. Res. Toxicol. Environ. Mutagen. 628 (2007) 31–55, https://doi.org/10.1016/J.MRGENTOX.2006.11.008.
[15] N.C. Kleinstreuer, A.L. Karmaus, K. Mansouri, D.G. Allen, J.M. Fitzpatrick, G. Patlewicz, Predictive models for acute oral systemic toxicity: a workshop to bridge the gap from research to regulation, Comput. Toxicol. (2018) 21–24, https://doi.org/10.1016/j.comtox.2018.08.002.
[16] T. Lang, F. Flachsenberg, U. von Luxburg, M. Rarey, Feasibility of active machine learning for multiclass compound classification, J. Chem. Inf. Model. 56 (2016) 12–20, https://doi.org/10.1021/acs.jcim.5b00332.
[17] J.C.D. Lopes, F.M. Dos Santos, A. Martins-José, K. Augustyns, H. De Winter, The power metric: a new statistically robust enrichment-type metric for virtual screening applications with early recovery capability, J. Cheminform. 9 (2017) 7, https://doi.org/10.1186/s13321-016-0189-4.
[18] K. Mansouri, J. Fitzpatrick, W. Casey, D. Allen, G. Patlewicz, A. Karmaus, et al., Developing predictive models for acute oral systemic toxicity: lessons learned from a global collaboration, CICSJ Bull. 37 (2019) 23, https://doi.org/10.11546/cicsj.37.23.

[19] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, BBA - Protein Struct. 405 (1975) 442–451, https://doi.org/10.1016/0005-2795(75)90109-9.
[20] A. Mayr, G. Klambauer, T. Unterthiner, S. Hochreiter, DeepTox: toxicity prediction using deep learning, Front. Environ. Sci. 3 (2016) 80, https://doi.org/10.3389/fenvs.2015.00080.
[21] N.S.H.N. Moorthy, S. Kumar, V. Poongavanam, Classification of carcinogenic and mutagenic properties using machine learning method, Comput. Toxicol. 3 (2017) 33–43, https://doi.org/10.1016/j.comtox.2017.07.002.
[22] A.W. Naik, J.D. Kangas, C.J. Langmead, R.F. Murphy, Efficient modeling and active learning discovery of biological responses, PLoS One 8 (2013) e83996, , https://doi.org/10.1371/journal.pone.0083996.
[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[24] A.H. Polash, T. Nakano, S. Takeda, J.B. Brown, Applicability domain of active learning in chemical probe identification: convergence in learning from non-specific compounds and decision rule clarification, Molecules 24 (2019) 2716, https://doi.org/10.3390/molecules24152716.
[25] A.H. Polash, N. Takumi, S. Takeda, J.B. Brown, Systematic approaches to build predictive models for rat oral toxicity, CICSJ Bull. 37 (2019) 12, https://doi.org/10.11546/cicsj.37.12.
[26] C. Rakers, R.A. Najnin, A.H. Polash, S. Takeda, J.B. Brown, Chemogenomic active learning's domain of applicability on small, sparse qHTS matrices: a study using cytochrome P450 and nuclear hormone receptor families, ChemMedChem 13 (2018) 511–521, https://doi.org/10.1002/cmdc.201700677.
[27] C. Rakers, R.A. Najnin, A.H. Polash, S. Takeda, J.B. Brown, Chemogenomic active learning's domain of applicability on small, sparse qHTS matrices: a study using cytochrome P450 and nuclear hormone receptor families, ChemMedChem (2018), https://doi.org/10.1002/cmdc.201700677.
[28] D. Reker, J.B. Brown, Selection of informative examples in chemogenomic datasets, Methods Mol. Biol. (Humana Press, New York, NY) (2018) 369–410, https://doi.org/10.1007/978-1-4939-8639-2_13.
[29] D. Reker, G. Schneider, Active-learning strategies in computer-assisted drug discovery, Drug Discov. Today 20 (2015) 458–465, https://doi.org/10.1016/J.DRUDIS.2014.12.004.
[30] D. Reker, P. Schneider, G. Schneider, J. Brown, Active learning for computational chemogenomics, Future Med. Chem. 9 (2017) 381–402, https://doi.org/10.4155/fmc-2016-0197.
[31] I. Rusyn, G.P. Daston, Computational toxicology: realizing the promise of the toxicity testing in the 21st century, Environ. Health Perspect. 118 (2010) 1047–1050, https://doi.org/10.1289/ehp.1001925.
[32] G. Schneider, W. Neidhart, T. Giller, G. Schmid, "Scaffold-Hopping" by topological pharmacophore search: a contribution to virtual screening, Angew. Chem. Int. Ed. 38 (1999) 2894–2896.
[33] J. Shawe-Taylor, N. Cristianini (2004). Kernel methods for pattern analysis. Cambridge University Press Available at: https://books.google.co.jp/books?hl=en&lr=&id=9i0vg12lti4C&oi=fnd&pg=PR8&dq=Kernel+Methods+for+Pattern+Analysis&ots=olCFrl3F5R&sig=mzYdGeZt1vEmfL65QRbZlIzX_Uo#v=onepage&q=Kernel Methods for Pattern Analysis&f=false [Accessed June 30, 2019].
[34] S.J. Shukla, R. Huang, C.P. Austin, M. Xia, The future of toxicity testing: a focus on in vitro methods using a quantitative high-throughput screening platform, Drug Discov. Today 15 (2010) 997–1007, https://doi.org/10.1016/j.drudis.2010.07.007.
[35] K. Taylor, Ten years of REACH—an animal protection perspective, Altern. Lab. Anim. 46 (2018) 347–373, https://doi.org/10.1177/026119291804600610.
[36] O. Tcheremenskaia, C.L. Battistelli, A. Giuliani, R. Benigni, C. Bossa, In silico approaches for prediction of genotoxic and carcinogenic potential of cosmetic ingredients, Comput. Toxicol. 11 (2019) 91–100, https://doi.org/10.1016/j.comtox.2019.03.005.
[37] R.R. Tice, C.P. Austin, R.J. Kavlock, J.R. Bucher, Improving the human hazard characterization of chemicals: a Tox21 update, Environ. Health Perspect. 121 (2013) 756–765, https://doi.org/10.1289/ehp.1205784.
[38] Y. Uesawa, Rigorous selection of random forest models for identifying compounds that activate toxicity-related pathways, Front. Environ. Sci. 4 (2016) 9, https://doi.org/10.3389/fenvs.2016.00009.
[39] M.K. Warmuth, J. Liao, G. Rätsch, M. Mathieson, S. Putta, C. Lemmen, Active learning with support vector machines in the drug discovery process, J. Chem. Inf. Comput. Sci. (2003) 667–673, https://doi.org/10.1021/ci025620t.
[40] A.M. Wassermann, M. Wawer, J. Bajorath, Activity landscape representations for structure-activity relationship analysis, J. Med. Chem. 53 (2010) 8209–8223, https://doi.org/10.1021/jm100933w.
[41] M. Zaslavskiy, S. Jégou, E.W. Tramel, G. Wainrib, ToxicBlend: virtual screening of toxic compounds with ensemble predictors, Comput. Toxicol. 10 (2019) 81–88, https://doi.org/10.1016/j.comtox.2019.01.001.