Journal of Cheminformatics

RESEARCH Open Access

MLinvitroTox reloaded for high-throughput hazard-based prioritization of high-resolution mass spectrometry data

Katarzyna Arturi[1*], Eliza J. Harris[2,3], Lilian Gasser[2], Beate I. Escher[4], Georg Braun[4], Robin Bosshard[5] and Juliane Hollender[1,6*]

Abstract

`MLinvitroTox` is an automated Python pipeline developed for high-throughput hazard-driven prioritization of toxicologically relevant signals detected in complex environmental samples through high-resolution tandem mass spectrometry (HRMS/MS). `MLinvitroTox` is a machine learning (ML) framework comprising 490 independent XGBoost classifiers trained on molecular fingerprints from chemical structures and target-specific endpoints from the ToxCast/Tox21 invitroDBv4.1 database. For each analyzed HRMS feature, `MLinvitroTox` generates a 490-bit bioactivity fingerprint used as a basis for prioritization, focusing the time-consuming molecular identification efforts on features most likely to cause adverse effects. The practical advantages of `MLinvitroTox` are demonstrated for groundwater HRMS data. Among the 874 features for which molecular fingerprints were derived from spectra, including 630 nontargets, 185 spectral matches, and 59 targets, around 4% of the feature/endpoint relationship pairs were predicted to be active. Cross-checking the predictions for targets and spectral matches with invitroDB data confirmed the bioactivity of 120 active and 6791 nonactive pairs while mislabeling 88 active and 56 non-active relationships. By filtering according to bioactivity probability, endpoint scores, and similarity to the training data, the number of potentially toxic features was reduced by at least one order of magnitude. This refinement makes the analytical confirmation of the toxicologically most relevant features feasible, offering significant benefits for cost-efficient chemical risk assessment.

Scientific Contribution:

In contrast to the classical ML-based approaches for toxicity prediction, `MLinvitroTox` predicts bioactivity for HRMS features (i.e., distinct m/z signals) based on MS2 fragmentation spectra rather than the chemical structures from the identified features. While the original proof of concept study was accompanied by the release of a `MLinvitroTox` v1 KNIME workflow, in this study, we release a Python `MLinvitroTox` v2 package, which, in addition to automation, expands functionality to include predicting toxicity from structures, cleaning up and generating chemical fingerprints, customizing models, and retraining on custom data. Furthermore, as a result of improvements in bioactivity data processing, realized in the concurrently released `pytcpl` Python package for the custom processing of invitroDBv4.1 input data used for training `MLinvitroTox`, the current release introduces enhancements in model accuracy, coverage of biological mechanistic targets, and overall interpretability.

Keywords ToxCast, Tox21, Toxicity, In vitro assay, Activity prediction, HRMS/MS, Binary classification, XGBoost, SIRIUS

*Correspondence:
Katarzyna Arturi
kasia.arturi@eawag.ch
Juliane Hollender
juliane.hollender@eawag.ch
Full list of author information is available at the end of the article

Introduction

Advances in instrumental high-resolution tandem mass spectrometry (HRMS/MS) reveal that thousands of unidentified anthropogenic pollutants with unknown toxicological properties are released to the aquatic environments daily [33, 50]. While we can routinely detect and analyze tens of thousands of HRMS features (i.e., distinct m/z signals) via nontarget screening (NTS) data acquisition methods and data processing workflows [18, 32, 58], identifying the unknowns remains a bottleneck of environmental risk assessments. Elucidation of the detected signals' molecular identity is performed via target, suspect, and nontarget screening. The targeted approach is the 'gold standard,' as the measured masses (mass-to-charge ratios, m/z), retention times (RT), and fragmentation spectra (MS2) can be unequivocally matched with experimental records of analytical standards. However, less than 2% out of 1 million compounds listed as chemicals of environmental importance on EPA's CompTox Chemicals Dashboard are available as analytical standards [54]. A suspect screening of exposure-relevant compounds compiled in lists [48] can help elucidate additional compounds by matching m/z, RT, MS2, and any additional orthogonal data available for the unknowns, with those of the suspects in compound and spectral databases. Targeted and suspect screening approaches typically yield annotations and matches for only a small fraction of the measured signals. The remaining HRMS features can be processed using nontarget computational methods, such as MetFrag [63], CFM-ID [74], and CSI:FingerID implemented in SIRIUS [19, 20]. These methods aim to tentatively identify unannotated signals by comparing their experimental fragmentation patterns with theoretical fragmentation patterns of compounds in databases. The packages can automatically annotate thousands of signals, each associated with a potentially large number of structural candidates. However, due to the need for an expensive and resource-demanding manual validation and analytical confirmation with reference standards, a complete elucidation of all the signals from suspect or nontarget screening is not feasible. To keep the workload manageable, only a limited number of individual chemicals can be investigated and, ideally, confirmed or ruled out using analytical standards [10]. Prioritization typically considers peak intensity (serving as a proxy for concentration), frequency (assuming that more common signals may pose greater concern), and statistical trends aligned with specific research objectives, such as comparisons before and after wastewater treatment [35]. In particular, multivariate chemometric analysis such as principal component analysis (PCA) [10, 34] is widely used for preliminary exploration and prioritization of nontarget HRMS data.

It has been estimated that less than 5% of the HRMS features measured in environmental and biological samples are commonly identified by a combination of target, suspect, and nontarget *in silico* identification efforts [56]. Even if a feature was successfully identified, relevant toxicity data is likely unavailable, making it difficult to assess the potential risk associated with the chemical. The hazard properties, e.g., toxic potency and modes of action, of only a handful of chemicals, have been comprehensively mapped due to the time-consuming, expensive, and ethically questionable nature of the traditional *in vivo* toxicity testing on animals [43]. If a feature is not flagged, prioritized, and identified, we are blind to its toxic potential in traditional NTS HRMS/MS analysis, which also explains why only a small part of overall mixture toxicity is currently explained by a combination of usual target, suspect, and nontarget analysis [51].

As a result, there is a growing interest in developing alternative *in vitro* and *in silico* methods for toxicity assessment. Alternatives to establishing toxicity experimentally involve predictive computational toxicology [3, 38, 41, 46, 76] based on the observation that structurally resembling chemicals often have similar properties and cause analogous toxic effects. Traditional Quantitative Structure-Activity Relationship (QSAR) models primarily used linear regression and other statistical methods to map monotonic relationships between chemical structures or properties and biological activities for safety assessments [16, 27]. In contrast, modern *in silico* approaches rely on machine learning (ML), which allows for modeling complex, non-linear relationships, thereby enhancing the prediction of specific toxic effects or molecular bioactivities with greater accuracy and broader applicability [14, 36, 60, 72, 76, 79]. An important distinction has to be made between *in vivo* toxicity, which is broadly understood as measurable damage in a living organism, and *in vitro* bioactivity which expresses molecular events on a cellular level that may or may not lead to toxicity on the organ or organism level. *In vivo* targets are generally more interpretable but often face data availability, volume, and reproducibility challenges. Conversely, *in vitro* data offers better robustness, clearly defined mechanisms, and greater volume, but at the cost of decreased interpretability. AOP-Wiki [2] maps *in vitro* bioassays to more than 300 signaling pathways associated with nearly 400 adverse outcome pathways (AOP), thus linking the molecular activities measured on a cellular level to adverse effects on the organ or organism level. Examples of ML frameworks for toxicity and bioactivity prediction developed for environmental applications

include MLTox [28, 77], deepFPlearn+ [64, 64], and TrendProbe [53]. From an environmental perspective, the most commonly assessed *in vivo* toxicity target is aquatic toxicity [45, 65], whereas *in vitro* predictions primarily focus on mutagenicity and Tox21 data endpoints [8, 14], with particular emphasis on endocrine disruption due to its significant environmental and public health implications [49, 68, 71, 78]. Despite the rapid growth in AI and ML, fueled by increased data availability, computational power, and innovation, the limited application of machine learning for environmental HRMS analysis [30, 47, 57, 59, 70] highlights a gap in the literature, presenting a valuable opportunity for future research.

In our previous work [5], we developed `MLinvitroTox` v1, an ML framework trained on hundreds of invitroDBv3.4 endpoints [13, 61] to prioritize toxicologically relevant signals among thousands of signals commonly detected in complex environmental samples through HRMS/MS. Like traditional ML-based approaches for toxicity prediction, `MLinvitroTox` uses molecular fingerprints derived from chemical structures as input features. Unlike those approaches, however, it was specifically developed to predict bioactivity based on the MS2 fragmentation spectra of all HRMS features rather than on the chemical structures derived from the identified features. The aim was to add toxicological relevance to environmental analysis by bypassing the bottleneck of feature identification prior to toxicity evaluation, thereby focusing the time-consuming molecular identification efforts on features most likely to cause adverse effects rather than merely the most intense ones. The results demonstrated that nearly a quarter of the invitroDB endpoints and most underlying mechanistic targets could be predicted accurately from structures and the MS2 spectra with `MLinvitroTox`. Furthermore, despite certain limitations, the methodology successfully guided nontarget screening of wastewater HRMS/MS data toward toxicologically relevant outcomes.

While the original proof of concept study was accompanied by the release of an `MLinvitroTox` v1 KNIME (Konstanz Information Miner) workflow, in this study, we release a Python `MLinvitroTox` v2 package [7] trained on the new release of the Tox21/ToxCast data (invitroDBv4.1). The package automates all processing steps and significantly expands the original functionality, e.g., predicting toxicity from structures, cleaning up and generating chemical fingerprints, and retraining customized models on user data. `MLinvitroTox` as a 'high-throughput' tool not only aims to provide an unbiased prioritization accelerating access to toxicologically relevant insights but also to streamline the progress from HRMS/MS measurement to actionable decision-making. Furthermore, we concurrently release `pytcpl`, a Python package developed for the custom processing of invitroDBv4.1 input data used for training `MLinvitroTox`. Both `MLinvitroTox` v2 and `pytcpl` are ultimately designed to integrate machine learning into toxicology and environmental analysis to achieve cost-efficient chemical risk assessment. In the current work, we present the results of invitroDBv4.1 data processing in `pytcpl`, which was used for training `MLinvitroTox` v2, validation of the ML models using MassBank structural and spectral data, and demonstrate their subsequent application on environmental groundwater HRMS/MS data. Figure 1 provides a detailed overview.

Methods

Bioactivity data

For training the models, we used the most recent release of ToxCast's MySQL database, invitroDBv4.1 [13] originating from the U.S. EPA's ToxCast program collaboration and the National Institutes of Health (NIH) Tox21 initiative [62]. ToxCast and Tox21 are extensive high-throughput screening bioactivity data collections covering 9559 unique compounds tested selectively across 1499 assay endpoints. An assay is an experimental protocol to evaluate interactions between target molecules and cells, particularly with endogenous biomolecules such as lipids, receptors, and other proteins. The assays in ToxCast and Tox21 utilize a range of technologies to assess the impact of chemical compounds on a wide array of biological targets, including individual proteins, nuclear receptor signaling, developmental processes, and cellular processes such as mitochondrial health, thereby offering a comprehensive view of potential hazards. In each assay, varying concentrations of each chemical are administered onto bioassays in multiple replicates. The measured raw biological concentration-response series (also called dose-response series) are fit via regression models into corresponding concentration-response curves (CRC), from which potency estimates such as AC10 (concentration at 10% activity) and AC5O (concentration at 50% activity) are derived. The 'hill' model, which generates sigmoidal dose-response curves, is advantageous for fitting as it is a log-logistic model that closely approximates the log-normal distribution of most toxicological data, providing a toxicologically meaningful fit. Positive controls were tested for assays with a defined maximum, such as receptor binding assays, to establish the objective upper limit of activity and normalize the chemicals' responses. For other assays, responses were normalized using the induction ratio to an unexposed control. Also, background measurements of solvents with no chemicals serve as negative controls, establishing

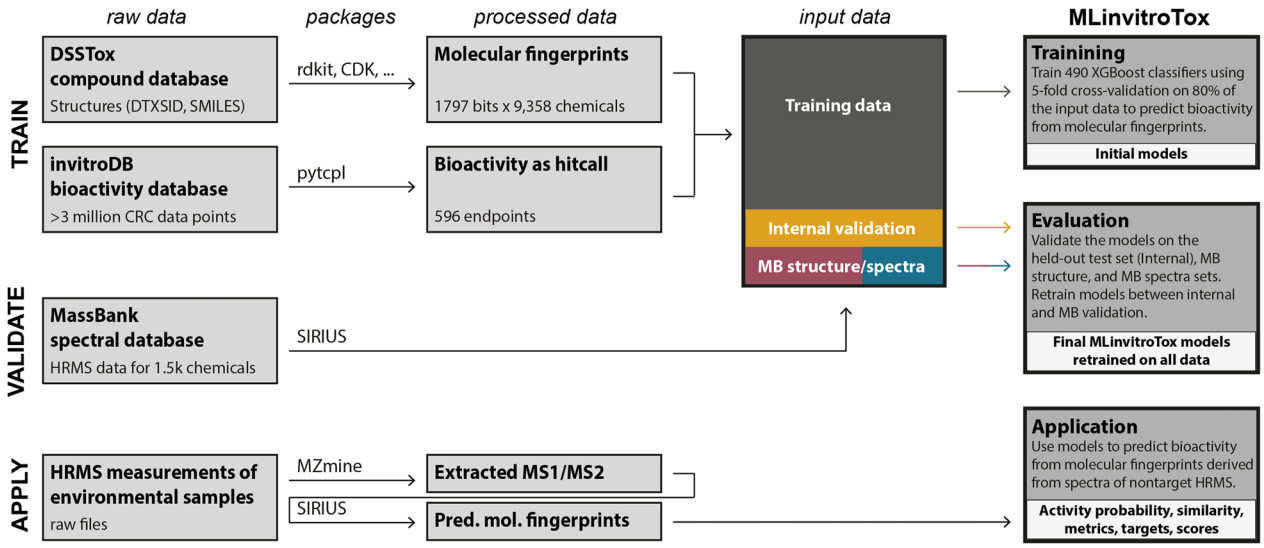

Fig. 1 `MLinvitroTox` pipeline development steps described in this work include: (I) Training XGBoost classifiers for 490 viable assay endpoints from invitroDBv4.1; (II) Validating model performance on the 'Internal' test set (20% of the input data) and MassBank sets (1.5k compounds, which are both present in invitroDBv4.1 as well as contain HRMSMS2 spectra in MassBank for which molecular fingerprints could be predicted) using both structural ('MB structure') as well as spectral ('MB spectra') data via SIRIUS); (III) Applying the models to environmental HRMS/MS data

a baseline and accounting for assay-specific noise. For each dose-response curve, the chemical activity (hitcall active/1 vs. nonactive/0) is assessed based on the fitting quality and the number of median (based on concentration replicates) responses above the noise level. There are 3,196,178 concentration-response curves available in invitroDBv4.1, corresponding to more than 50 million raw measurements.

Data processing

In accordance with the literature, which emphasizes the necessity of rigorous preprocessing to eliminate non-target effects in *in vitro* data [23, 24, 44], and to mitigate modeling artifacts such as overfitting and data biases, we developed `pytcpl` [11], a custom pipeline for processing of the invitroDBv4.1 bioactivity data. `pytcpl` is a streamlined Python package inspired by the R packages `tcpl` and `tcplFit2` [25, 67], to accommodate customized processing steps and optimized data storage by storing the output as compressed Parquet files. Only cell-based assays were used in the current study (452 non-cell-based assays in invitroDBv4.1 were omitted). Using `pytcpl`, we conducted the following data processing procedure for each concentration-response series involved:

1. *Outlier removal:* The running mean response for all points across neighboring concentrations was calculated to create a smoothed response curve. The residual was then determined as the difference between each data point and this smoothed curve. All points with residuals that were (*i*) larger than the baseline median absolute deviation (BMAD, the deviation of the three lowest concentrations in a dose-response series) and (*ii*) more than five times the standard deviation of the residuals were flagged as outliers and removed. Figure 2a shows examples of outlier flagging.
2. *Curve validation:* The concentration-response series were validated to ensure that (*i*) the number of concentrations tested was larger than the set threshold of four, and ii) the series contained at least one response higher than 0.8 times the activity concentration at cutoff (ACC). The ACC is the threshold above which a response is considered biologically significant and clearly distinguished from the background noise. The ACC in invitroDB is a user-selected threshold per assay endpoint, e.g., set at three times BMAD to ensure biological significance beyond baseline variability. Concentration-response series violating criteria *i* were assigned a hitcall of NaN, while those violating criteria *ii* were assigned a hitcall of 0 (Fig. 2a, b).
3. *Curve fitting:* Different curves were fit to each validated concentration-response series using maximum likelihood estimation for various curve models, ensuring the most accurate representation of the data. The models used in this study are described in Sect. "Curve fitting models". The fitted models were used to calculate relevant potency estimates and

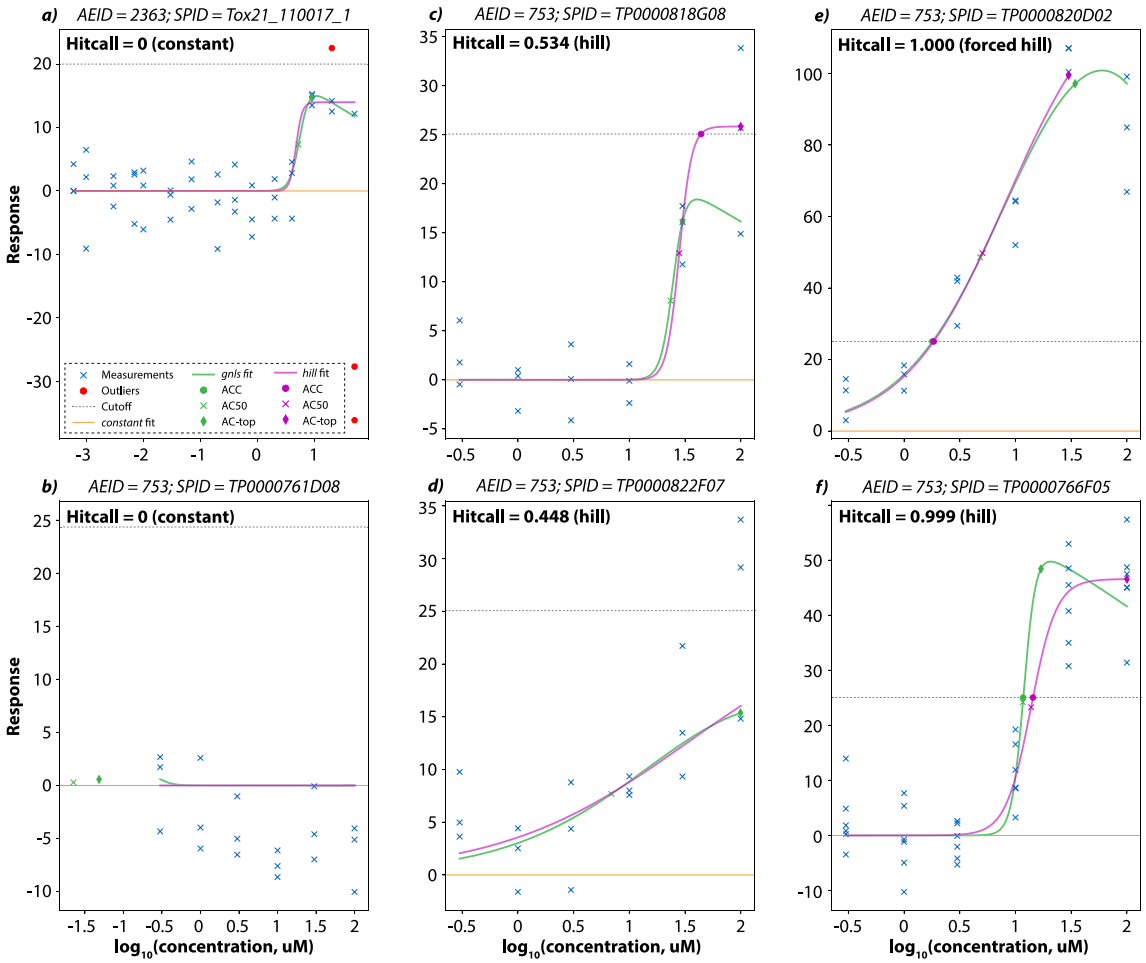

Fig. 2 Example curve fits illustrating different outcomes: (**a**) and (**b**) show dose-response curves where no activity was evident (hitcall_c = 0); the upper panel has three outlier points. (**c**) and (**d**) illustrate intermediate confidence of activity based on the hill model. (**e**) and (**f**) show strong confidence in activity; the upper panel required a 'forced hill' fit, and the lower panel fit with the 'hill' model. The 'constant' model and cutoff are shown in all panels. AEID = assay endpoint ID and SPID = species ID

activity concentrations as described in [11, 25, 67] and shown in Fig. 2.

4. *hitcall assignment:* The fitted model with the lowest Akaike Information Criterion (AIC) that best represented the concentration-response series was assigned as the 'winning' model. If the winning model was the 'constant' model, a hitcall of 0 was assigned. If the winning model was not the 'constant' model, the continuous hitcall (representing the likelihood of a significant response/continuous scoring of toxicity likelihood rather than binary outcomes) was determined as the product of three distinct probabilities according to the method proposed by Sheffield et al. [67]:

$$P_{hit} = P_1 \times P_2 \times P_3 \quad (1)$$

where:

- P_1 = probability that at least one median response is greater than the ACC, computed using the error parameter from the model fit and the Student's t-distribution to calculate the odds of at least one response exceeding the ACC;
- P_2 = probability that the top of the winning fitted curve is above the cutoff, which is the likelihood ratio of the one-sided probability of the ACC being exceeded;
- P_3 = probability that the winning AIC value is less than that of the constant model:

$$P_3 = \frac{e^{-\frac{1}{2}\text{AIC}_{\text{winning}}}}{e^{-\frac{1}{2}\text{AIC}_{\text{winning}}} + e^{-\frac{1}{2}\text{AIC}_{\text{constant}}}} \quad (2)$$

5. *Annotation and redundancy:* Flags and fit categories were applied according to the procedure defined in [25], including removal of background, controls, and viability measurements (used only for cytotoxicity estimation) as well as redundant channels.

Curve fitting models

Four models were used to fit the invitroDBv4.1 concentration-response curves, selected for their effectiveness in capturing different concentration-response relationships. The 'constant', 'hill' and gain-loss ('gnls') models were used in the `tcpl` package [25], and the 'forced hill' model is a modification of the 'hill' model. The additional curve fit models used in `tcplv3.0` [73] were included in `pytcpl` but used only for bootstrapping (Sects. "Bootstrapping" and "Model choice for concentration-response curve fitting") and thus only described here briefly. All models are shown in Figure SF1.

- The **'constant'** model is a zero-parameter model that describes no effect of the dose on the measured response, implying that the response remains unchanged regardless of dose concentration, where R is the response and d is the dose concentration:

$$R(d) = 0 \quad (3)$$

- The **'hill'** model is based on the mechanistic understanding of activity where receptor-ligand binding always has an upper limit. It is a three-parameter model that represents a monotonic response at a threshold concentration with the bottom asymptote forced to 0, where t is the curve top, g is the gain, and p is the gain power:

$$R(d) = \frac{t}{1 + (\frac{g}{d})^p} \quad (4)$$

- The gain-loss (**'gnls'**) model is a five-parameter model that describes an increase followed by a decrease in response, capturing more complex biological processes where initial gains in response diminish with higher doses, where l is the loss and q is the loss power:

$$R(d) = \frac{t}{(1 + (\frac{g}{d})^p)(1 + (\frac{d}{l})^q)} \quad (5)$$

- The 'hill' and 'gnls' models do not always deliver consistent results, particularly in terms of potency estimates, due to their different curve shapes. To address this, we introduced the 'forced hill' model, which improves consistency across all active dose-response series. In all cases where 'gnls' was the winning model, data points with concentrations larger than the top of the fitted 'gnls' model were removed to avoid overestimation of effects, and the 'hill' model was fit to the remaining data points, as shown in Fig. 2e.
- Additional models included: polynomial-linear ('poly1'), polynomial-quadratic ('poly2'), power, exponential-2 ('exp2'), exponential-3 ('exp3'), exponential-4 ('exp4'), exponential-5 ('exp5') [73]. The 'exp3' model was not used in bootstrapping due to a low win rate and overflow issues. We also tested the 'gnls2' model, a modified gain-loss model with one less parameter, to reduce overfitting [11]. This model was considered an alternative to balancing model complexity with predictive accuracy.

Bootstrapping

Bootstrapping was performed to estimate uncertainties associated with predicting concentration-response parameters by different models, particularly to identify which models delivered the most reproducible and toxicologically meaningful results. Bootstrapping is a statistical method used to estimate the uncertainty of a parameter by repeatedly resampling the raw data with replacement, refitting the models, and recalculating the parameters for each resample, thus generating an empirical distribution of hitcall as well as potency estimates. Given the computational demands of bootstrapping on the full dataset, we selected a representative subset for analysis, ensuring it was large enough to yield meaningful insights. We carried out bootstrapping on 62 unique endpoints selected to have close to 100 chemicals each for 6436 concentration-response curves. We used a nonsmoothed, nonparametric resampling similar to the method used by Watt and Judson [75]. In brief, we resampled with replacement for each concentration-response curve at each concentration level. Then, we introduced random noise at the level of the BMAD for that assay to mimic experimental variability and ensure the robustness of our results. We carried out bootstrapping to investigate two data processing scenarios:

- **All models**: All models included in `tcplv3.0` [73] were used as well as the 'gnls2' model, except 'exp3' which led to regular computational overhead during fitting (Figure SF1).

- **Forced hill**: Only 'constant', 'gnls' and 'hill' models were fit, and in all cases where 'gnls' was the winning model, the 'forced hill' model was used.

We carried out 300 bootstraps for the 6436 concentration-response curves for the 'all models scenario' and 1000 bootstraps for the 'forced hill scenario', which led to convergent results whereby both mean and standard deviation showed no significant changes with the addition of further bootstraps. The differing numbers of bootstraps were due to the higher computation power needed to fit all models, which took around 10 times longer. We compared the model selection, hitcalls, and relevant activity concentrations across the different bootstraps and between the two scenarios to identify the most robust and reproducible curve-fitting method (Figure SF2).

Cytotoxicity estimation and assignment of final 'hitcall'

Based on the bootstrapping results (Sect. "Model choice for concentration-response curve fitting"), we found that the 'forced hill' method offers the best reproducibility and comparability between concentration-response curve fits; thus, this method was used to fit the full dataset (excluding non-cell-based assays; Sect. "Bioactivity data"). The preliminary hitcalls returned from this fitting represent the probability that an effect was observed (P_{hit}); however, this effect could be due to general cytotoxicity rather than the effect targeted by a particular assay [37]. Therefore, we estimated the probability of cytotoxicity (P_{cytotox}) and used this to estimate the likelihood of activity due to the targeted effect ($P_{\text{hit\_tgt}}$), assuming that all variables have a Gaussian error distribution. The assumption of a Gaussian error distribution allows for systematically estimating the cytotoxicity probability, providing a statistically sound basis for refining the hitcall.

$$P_{\text{hit\_tgt}} = P_{\text{hit}} \times (1 - P_{\text{cytotox}}) \tag{6}$$

$P_{\text{hit\_tgt}}$ is the final continuous hitcall, referred to throughout this paper as 'hitcall'.

P_{cytotox} was estimated as the probability that the curve fit activity concentration at cutoff (ACC_{tgt}) is lower than the cytotoxicity assay activity concentration at cutoff ($\text{ACC}_{\text{cytotox}}$):

$$\begin{aligned} P_{\text{cytotox}} &= P(\text{ACC}_{\text{cytotox}} - \text{ACC}_{\text{tgt}} \leq 0) \\ &= \phi\left(\frac{\text{ACC}_{\text{cytotox}} - \text{ACC}_{\text{tgt}}}{\sqrt{\text{SD}^2_{\text{ACC}_{\text{cytotox}}} + \text{SD}^2_{\text{ACC}_{\text{tgt}}}}}\right) \end{aligned} \tag{7}$$

where ϕ is the Gaussian cumulative distribution function. For ~2% of concentration-response curves, a directly corresponding viability assay endpoint counterpart was available for the determination of $\text{ACC}_{\text{cytotox}}$; like target assays, viability assays were fit using the 'forced hill' method. $\text{SD}_{\text{ACC}_{\text{cytotox}}}$ and $\text{SD}_{\text{ACC}_{\text{tgt}}}$ where both determined from the bootstrapping results, which showed that generally SD_{ACC} can be estimated as $0.41 \times \text{ACC} + 7.0$ (Figure SF3).

Cytotoxicity burst assays offer a proxy for estimating cytotoxicity when direct viability data is lacking. A statistical approach was applied to estimate cytotoxicity for assays with no corresponding viability assay. $\text{ACC}_{\text{cytotox}}$ was approximated as the median ACC for the compound of interest across a set of assay endpoints designed to capture the cytotoxicity burst. $\text{SD}_{\text{ACC}_{\text{cytotox}}}$ is derived from the median absolute deviation of the respective ACC values. Additionally, in the statistical approach, P_{cytotox} is multiplied by the ratio of the number of cytotoxicity burst assay endpoints in which the compound exhibited activity (n_{hit}) to the total number of cytotoxicity burst assay endpoints in which the compound was tested (n_{tested}), to account for the bias introduced by calculating the median ACC only from 'hits' according to the `tcpl` pipeline [25].

Structural data

The structural data used for training were obtained from U.S. EPA's Chemistry DSSTox database [21, 29] as ToxCast_invitroDB_v4_1.csv containing a compilation of identifiers (e.g., DTXSID, name, CAS number, and InChI Key), chemical representations (e.g., InChI, SMILES), and chemical data (e.g., molecular formula, average and monoisotopic mass). The initial list containing 9559 entries was filtered to remove duplicates and entries without DTXSID and SMILES representing unspecified organic groups, inorganic molecules, or for which no structural representation could be generated. Lastly, the structures were rigorously cleaned [26, 31]. They were standardized (removal of explicit hydrogen atoms, ring aromatization, normalization of specific chemotypes, curation of tautomeric forms, removal of charges, removal of metals, reionization, removal of stereoisomers, removal of inorganic counter ions). The `rdkit` package was used for processing. Post-curation, 9358 chemicals were available for modeling. Structures were not used directly as input but were converted into mathematical representations of molecules, namely, as molecular fingerprints encoding the chemical structures as binary vectors of fixed length where each bit describes the presence (1) or absence (0) of a particular substructure. The cleaned data was used to generate the molecular fingerprints via pybel (based on openbabel [55]) and CDK-pywrapper [9] packages. With openbabel, FP3 (55 bits) and FP4 (307 bits) fingerprints were generated. With CDK-pywrapper, MACCS (166 bit), PubChem (881 bits), and Klekota-Roth fingerprints (4860 bits) were

generated. The 6269 bits were cross-referenced with the SIRIUS fingerprint definitions for the positive mode containing 3877 bits. The SIRIUS fingerprint additionally covers extended connectivity (ECFP), custom-made SMARTS, and ring systems, which were omitted in the current work, as their generation is challenging and they did not significantly improve the predictive power of the models in the proof of concept [5]. Scripts for cleaning up and generating molecular fingerprints are released as part of the `MLinvitroTox` package, and the reader is referred to the documentation for more details. The continuous hitcalls from the `pytcpl` and their identifiers (DTXSID) were combined with molecular fingerprints derived from structures. The dataset per assay endpoint consisted of a data frame with an index identifier (DTXSID from the CompTox Dashboard), a feature matrix with $n = 1797$ columns (molecular fingerprint bits) named with the absolute indexes used by SIRIUS, and a continuous target vector ('hitcall') binarized to 0/1 according to the activity_threshold (set to 0.9) in the configuration provided in the Code snippets in Section S5. While all 596 endpoints provided by the `pytcpl` package (with more than 10 active chemicals and 100 input chemicals, respectively) were initially used as input to `MLinvitroTox`, the purging of records tagged with 'QC-omit' or 'Flag-omit' according to the cHTS curation pipeline recommendations [52], along with ML-specific data splitting (see Sect. "Machine learning for bioactivity prediction"), reduced the number of viable endpoints for ML model training to 490.

Machine learning for bioactivity prediction

Model training

As the first step in processing, the data splitting for the validation was done. For each assay endpoint dataset, compounds also present in the MassBank database (release version 2024.06 with 117,732 spectra was used, including 2879 spectra submitted by Eawag since 2022) were assigned to the corresponding assay's MassBank validation set. The remaining records were partitioned into training and test sets (80:20 split) through a custom stratification method based on k-means clustering, ensuring a balanced representation of active and chemically diverse samples in the training and test sets. A substantial proportion of the studied assay endpoints exhibited an unequal distribution of active and nonactive compounds, with the nonactive class typically outweighing the active class. For such imbalanced datasets, models may become biased toward predicting the majority class more frequently unless corrective measures are implemented [40]. To address this, we applied oversampling to rebalance the underrepresented class. We applied feature selection using a random forest model to narrow the relevant fingerprint bits. Features that exceeded the mean importance threshold (average importance across all input features) were selected and used as input for model training.

Based on the proof of concept study, the XGBoost (eXtreme Gradient Boosting, [15]) algorithm was selected due to its robustness and good performance [5]. XGBoost is a regularizing gradient-boosting ensemble learning technique that combines multiple weak learner decision trees sequentially, with each new learner giving more weight to the examples that the previous learners struggled with. For each assay endpoint, the model training was based on a grid search hyperparameter tuning nested within 5-fold cross-validation. The k-fold cross-validation modeling performance and hyperparameter tuning details are reported in the model training log files. The applied hyperparameter grid and a snippet of the log file for one endpoint are provided in Section S5. The hyperparameters were optimized for binary classification based on the F_β score ($F_\beta = (1+\beta^2) * (\text{precision} * \text{recall}) / ((\beta^2 * \text{precision}) + \text{recall})$, $\beta = 2$). The backbone of the code showing modeling logic and a snippet of the log generated during the pipeline run for one endpoint is provided in Section S5. Extensive logging features ensured full traceability of each run.

Model evaluation

The fine-tuned ML models were validated in three stages:

1. **'Internal':** Evaluation on the internal test set drawn randomly from the input data (excluding MassBank compounds) by stratified sampling with an 80:20 train test split ratio to obtain an unbiased performance measure of the model-building process with molecular fingerprints from structures as input.
2. **'MB structure'**: External validation on MassBank compounds (molecular fingerprints from structures) to evaluate the model's generalization capabilities on structural data. Models were retrained on the combined train/test data with optimal hyperparameters for the validation.
3. **'MB spectra':** External validation on the MassBank spectral MS2 records (predicted molecular fingerprints from SIRIUS) focusing on the gap between chemical structure space and fragmentation spectra, providing an overall performance measure of the whole pipeline.

The two validations based on MassBank data are performed for the same compounds, once with the structural information and once with the spectral information. In each case, the model's predictions were obtained as probabilities for a compound to belong to the active class.

These were subsequently binarized based on a threshold. We varied thresholds for binarizing the probabilities into binary hitcalls at four levels: (1) the default threshold of 0.5, (2) the custom threshold determined by the cost function weighting TPR (True Positive Rates) twice as FPR (False Positive Rates, to value recall), (3) TPR = 0.5, and (4) TNR = 0.5. At each threshold, the hitcalls were compared to the ground truth, i.e., the corresponding invitroDB data, and a comprehensive set of model metrics and summaries (e.g., confusion matrix, recall, precision, F1 score, accuracy, AUC-ROC, and PR-ROC) were computed to evaluate the performance of each classification model. The validation results here focus on the default threshold and evaluate overall and balanced accuracy. We reported overall and balanced accuracy metrics for `MLinvitroTox` models to provide insights from two perspectives. While overall accuracy (Eq. 8) is a global performance measure emphasizing the majority class (in our case, nonactive), balanced accuracy (Eq. 9) highlights the performance of the active minority class as it weights the performance of each equally.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{8}$$

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}}\right) \tag{9}$$

To best leverage `MLinvitroTox` in environmental analysis, it is essential to combine a robust prioritization of active chemicals (maximize TP) to focus analytical efforts on signals with the highest potential for harm while also accurately classifying nonactive features (maximize TN) to avoid unnecessary expenditure of efforts and resources on harmless signals. In addition to evaluating model performance, we also extracted feature importance values to understand which chemical structures are most predictive of toxicity. This improves model interpretability and offers valuable insights for future toxicological research.

Molecular fingerprints from MS2 spectra

SIRIUS 5.8.3, with an academic license, was used to generate molecular fingerprints from spectra for MassBank validation and environmental application. The following settings were applied: 'Instrument': Orbitrap, 'Isotope pattern filter': Yes (environmental data) or No (MassBank data), 'Mass accuracy': 5 ppm MS2, 'MS2 isotope scorer': Score, 'Candidates stored': 10. Neither Zodiac nor CANOPUS (Compound class prediction) modules were used. For CSI:FingerID, adducts $[M+H]^+$, $[M+K]^+$, and $[M+Na]^+$ and complete DB search were selected. Standard settings for ILP (Integer Linear Programming) solver, as well as other options, were used for the remaining parameters. Due to the significant increase in processing time associated with larger molecular weights of the precursor and the focus on small molecules in this study, a limit of 750 Da was established to avoid excessive computational demands. 187 of the ToxCast `invitroDB` chemicals have a mass of 750 Da or higher, so MassBank validation did not cover the full mass range utilized during training. MassBank spectra (imported as.txt with only MS2 data), as well as the environmental data (imported as.mgf from MZmine with both MS2 as well as MS1 isotopic patterns), were analyzed with the same settings except the 'Isotope pattern filter' as specified above. Uploading and processing.raw or.mzML files to SIRIUS is possible but not advisable. It increases the processing time significantly, and the user loses control over processing parameters. SIRIUS generated molecular formula proposals (the maximum is defined in settings, e.g., 10) per successfully imported spectrum, corresponding molecular fingerprints, and structure proposals. For the MassBank data, the SIRIUS output corresponding to the same compounds available in replicates (e.g., measured by different institutes with varying collision energies) was concatenated to obtain the most comprehensive and representative fingerprint possible. For environmental data, feature consolidation occurred downstream in the `MLinvitroTox` application. It is important to note that the current version of `MLinvitroTox` supports only the positive electrospray ionization mode HRMS data analyzed in SIRIUS v5.8.7 or lower.

The accuracy of SIRIUS in predicting molecular fingerprints from MS2 was evaluated on 1.5 k MassBank compounds to identify potential sources of error while applying `MLinvitroTox` on environmental data. The process involved comparing molecular fingerprints generated from structures and MS2 spectra. Structural fingerprints were generated as described in Sect. "Structural data". Tanimoto similarity, recall, precision, and overall accuracy were used as evaluation metrics. Compounds from MassBank that were used to train the CSI:FingerID algorithm were excluded from the evaluation to prevent overfitting and provide a realistic performance measure.

Environmental application on groundwater

`MLinvitroTox` was applied to environmental groundwater data generated by Kiefer et al. [39], who collected samples from 60 monitoring sites across Switzerland (44 abstraction wells, 16 springs) and analyzed them using target, suspect, and nontarget approaches to classify pollution sources as urban or agricultural. We selected four sites/sample types associated in the original analysis with intensive pollution from urban sources for demonstration. For each site, three consecutive measurements were available, along with three blanks, resulting in 15

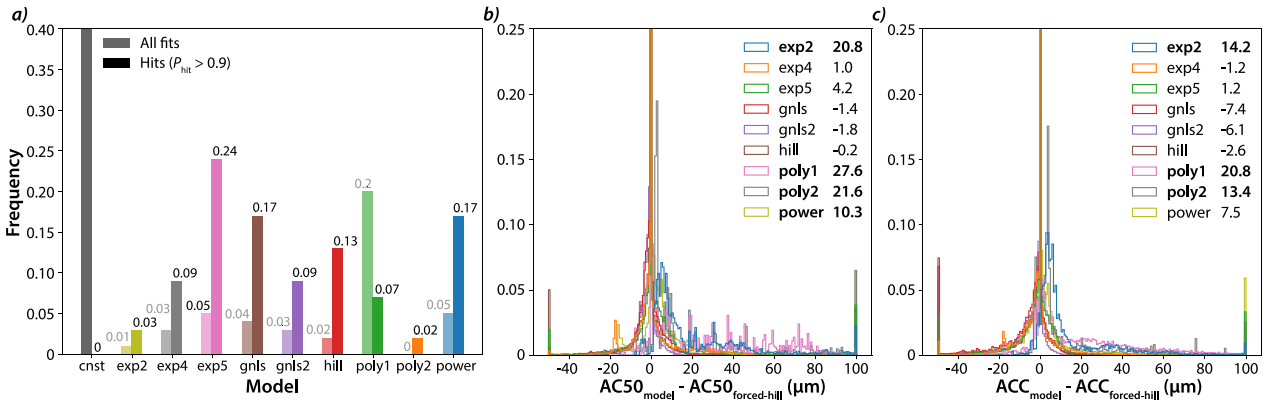

Fig. 3 Bootstrapping results for the different models used in the 'all models' setup. (**a**) Frequency with which each model is chosen (i.e., was the 'winning' model) across all dose-response curve fits (paler bars) and across fits with a hit probability of > 0.9. (**b**) and (**c**) Histogram of the differences between the AC50 (**b**) and ACC (**c**) determined with each model and with the 'forced hill' method for all cases where the designated model was the 'winning' model. Models with the largest differences are highlighted in bold in the legend

samples. Only the data from the positive ionization mode were used. The exact details regarding sample collection, preparation, and analysis can be found in the source publication. In short, the samples were enriched by applying vacuum-assisted evaporation and injected in triplicate into an HPLC system featuring a reversed-phase C18 column using gradient elution with water and methanol containing 0.1% formic acid at a flow rate of 0.3 mL/min. Analytes were ionized using electrospray ionization and detected on an Orbitrap mass spectrometer (Fusion Lumos) with a resolution of 240k at m/z 200 in MS1 full-scan mode, followed by data-dependent (DDA) MS2 scans. The AcquireX software enhanced MS2 coverage by dynamically updating the mass list of already measured (with MS2) features across triplicate injections. The raw HRMS files were converted without compression to an open-source.mzML format with ProteoWizard and processed with MZmine via a general pipeline for untargeted LC-MS designed to resemble the original workflow in enviMass covering, e.g., peak picking, retention time alignment, grouping of adducts and isotopologues into components based on intensity correlation and m/z distance, replicate filtering, and target annotation. Additional steps described in more detail in Section S3 were incorporated into the data processing workflow to utilize the advanced features of MZmine fully, enhance the quality of the results, and streamline the post-processing procedures with parameters derived from the data as recommended by Damiani et al. [17].

Results

Model choice for concentration-response curve fitting

Processing data from the invitroDB database, particularly fitting concentration-response curves before training ML models, is crucial for accurate outcomes. The breakdown of model choices and results in the 'all models' approach, where all models used in the EPA's `tcpl` package are considered, and the best fitting model is chosen, are shown in Fig. 3. Considerable variability is seen in potency estimates between models. The bootstrap results were used to compare the 'all models' approach to the 'forced hill' data processing method and to explore uncertainty in curve fit parameters in both cases. Compared to the 'all models' approach, the 'forced hill' method delivered lower estimates of ACC, AC50, and standard deviation in ACC and AC50 (Figure SF2a–d). The median hitcalls were higher for the 'all models' approach in nearly all cases, and thus, the 'forced hill' method produces fewer positive hitcalls (Figure SF2e). The lower standard deviation of ACC and AC50 for the 'forced hill' case shows that the curve fitting is more reproducible when fewer different models are used, in agreement with the findings of Watt and Judson.

Using the 'all models' approach, the 'cnst' and 'poly1' models are chosen most often across all curve fits, while the 'exp5', 'gnls', 'hill' and 'power' models are chosen most often for curves representing positive hitcalls (Fig. 3a). The 'exp5', 'gnls' and 'hill' models show similar AC50 and ACC estimates to each other and the 'forced hill' estimate (Fig. 3b, c). However, the 'exp2', 'poly', and 'power' models-accounting for nearly a third of all positive hitcall fits-show a strong overestimation of ACC and AC50 and, consequently, more false positive hitcalls. We expect this because these models have no true plateau, making it challenging to estimate the top of the curve reproducibly. We conclude that using 'all models' makes curve fitting less reproducible and reliable than the 'forced hill'

method, which also has a biological meaning with receptor-ligand binding.

Bootstrapping is an ideal method to investigate the reproducibility of curve fit parameters in detail, however, the computational cost is high and thus this could not be applied to the entire database. We, therefore, compared results from single curve fits with the bootstrap results to investigate how well the single curve fits can capture the bootstrap results (Figure SF3). The single fits hitcalls are scattered compared to the fraction of fits in the bootstraps (Figure SF3a) and the median hitcall from the bootstraps (Figure SF3b), but there is no bias, showing that the single fits can adequately predict hitcalls. The single fits somewhat underestimate the ACC compared to the bootstrap results, and the uncertainty in ACC across bootstraps is linearly related to the magnitude of the ACC up to an ACC of 100 μM.

Final structure of the toxicity dataset

The final assignments of the 1499 assays in invitroDBv4.1 are shown in Fig. 4. 596 assays fulfilled the criteria ($\geq$ 10 active and $\geq$ 100 total cases) to be included in the training and validation datasets for `MLinvitroTox`. 452 assays were rejected because they were not cell-based. Non-cellular assays do not consider kinetic interactions in living cells, such as uptake through cell membranes, and therefore, their sensitivity cannot be directly compared to cellular assays. Background and control assays were also dropped on `tcpl` processing levels 0–3 in invitroDB. The 596 selected assays had an average of 2,000 (median = 1061) chemicals tested per assay. 1.1 million valid curve fits were conducted, and 72,561 (6.6%) positive hitcalls were assigned, averaging 122 positive hitcalls per assay. The majority of endpoints had 15 or more concentration levels tested.

MLinvitroTox modeling

Out of the 596 selected assay datasets from `pytcpl`, models could be trained for 490 unique assay endpoints. A total of 108 assay endpoints were excluded because they did not meet the defined minimum population criteria (at least 10 active hits and 100 total training examples) following the application of cHTS curation pipeline steps (removal of 'QC-omit' and 'Flag-omit' records) and partitioning of the data into training, validation, and

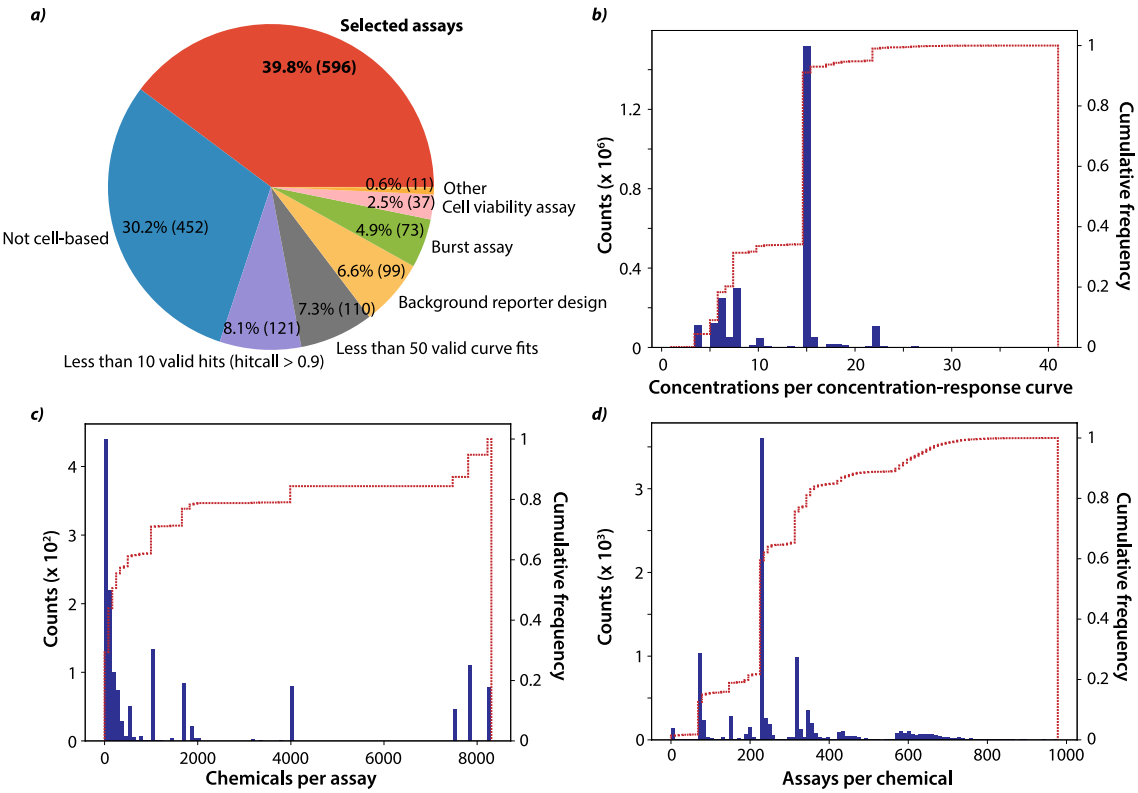

Fig. 4 Summary of the invitroDBv4.1 dataset: **a** Classification of the 1499 assay endpoints: 'Selected assays' were finally used for ML. **b** Histogram of the unique concentrations measured in each concentration-response curve. **c** Histogram of the number of unique compounds tested per assay. **d** Histogram of the number of assays on which each unique chemical was tested. The dataset contains 9559 unique compounds selectively tested across 1499 assay endpoints

test sets. The 490 assay endpoints covered 62 out of 106 unique mechanistic targets and 31 out of 47 unique 'MT_NCIm_term' terms as defined by the National Toxicology Program of the U.S. Department of Health and Human Services (NICEATM) [1]. The NICEATM mapping aims to link molecular bioactivities in invitroDB to their meaningful biological effects. The 62 mechanistic targets covered were concatenated into 29 unique shorthand target annotations to present the modeling and validation results concisely. For example, terms such as 'Estrogen Receptor Modulation', 'Estrogen-related Receptor Modulation', and 'Estrogen Biosynthesis and Metabolism' were collectively assigned to a single target: 'Estrogen Receptor (ER)'. Notably, the 'Cell Processes' mapping used in the current study covered a wide range of cell-related endpoints with a broad spectrum of terms, e.g., 'Cell Cycle', 'Cell Growth', 'Extracellular Matrix Degradation', 'Cell Morphology', 'Malformation', 'Proliferation', and 'Clotting'. Mechanistic targets and their acronyms are listed in Table ST1. 213 out of 490 endpoints did not have a mechanistic target annotation. Figure SF4 shows the number of endpoints, ranging from 1 to 64, available per mechanistic target.

MLinvitroTox validation

The evaluation of each fine-tuned `MLinvitroTox` model was conducted using three validation sets: Internal (a test set drawn from the input data), MB structure (validation on MassBank compounds using structural data), and MB spectra (validation on MassBank compounds using MS2 data) (Fig. 1). Confusion matrices on true positives (correctly predicted active cases), false positives (incorrectly predicted active cases), true negatives (correctly predicted nonactive cases), and false negatives (incorrectly predicted nonactive cases) were constructed (Fig. 5). The AUC-ROC (Area Under the Receiver Operating Characteristic Curve) was computed from the confusion matrices at different thresholds. Different thresholds can be employed to balance the trade-off between maximizing true positives and minimizing false positives, thus tailoring the model towards specific goals, such as prioritizing detecting toxic compounds or reducing false alarms. An example is shown in Fig. 5, where confusion matrices at two distinct thresholds, the default 0.5 and a custom threshold based on a cost function, are visualized along with the corresponding ROC-AUC curve for the classifier predicting assay endpoint 'aeid=1134' (associated with genotoxicity) validated on the 'Internal' set. According to the figure, with the default threshold, we catch most true positives (90) but mislabel 61 nonactive chemicals as active. However, when we decrease the threshold to catch additional true positives, a 55% increase in TPR results in a 180% increase in FPR, which is an example of FPR increasing faster TPR. Similar trends were observed across multiple endpoints, leading to the default threshold selection to minimize the misclassification of nonactive features and avoid unnecessary analytical efforts for identification. Confusion matrices at four distinct thresholds, along with ROC-AUC and PR-ROC (precision-recall Receiver Operating Characteristic) curves for all assay endpoint/validation sets, are available in the Streamlit dashboard provided with the `MLinvitroTox` package. The dashboard also includes a dedicated page with comprehensive summary figures to inspect the validation results.

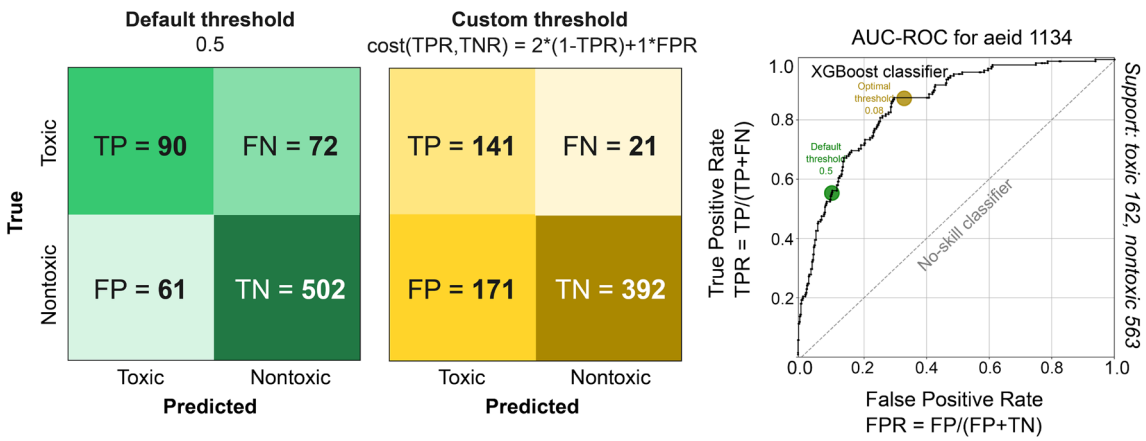

Fig. 5 Confusion Matrices and AUC-ROC Curve. Confusion matrices for the classifier of assay endpoint 'aeid=1134' (associated with genotoxic mechanistic target) validated on the Internal set at the default threshold (green) and the custom threshold to maximize the TPR (yellow). On the right, the corresponding AUC-ROC curve (Area Under the Receiver Operating Characteristic Curve) visualizes the changes in true and false positive rates with varying thresholds

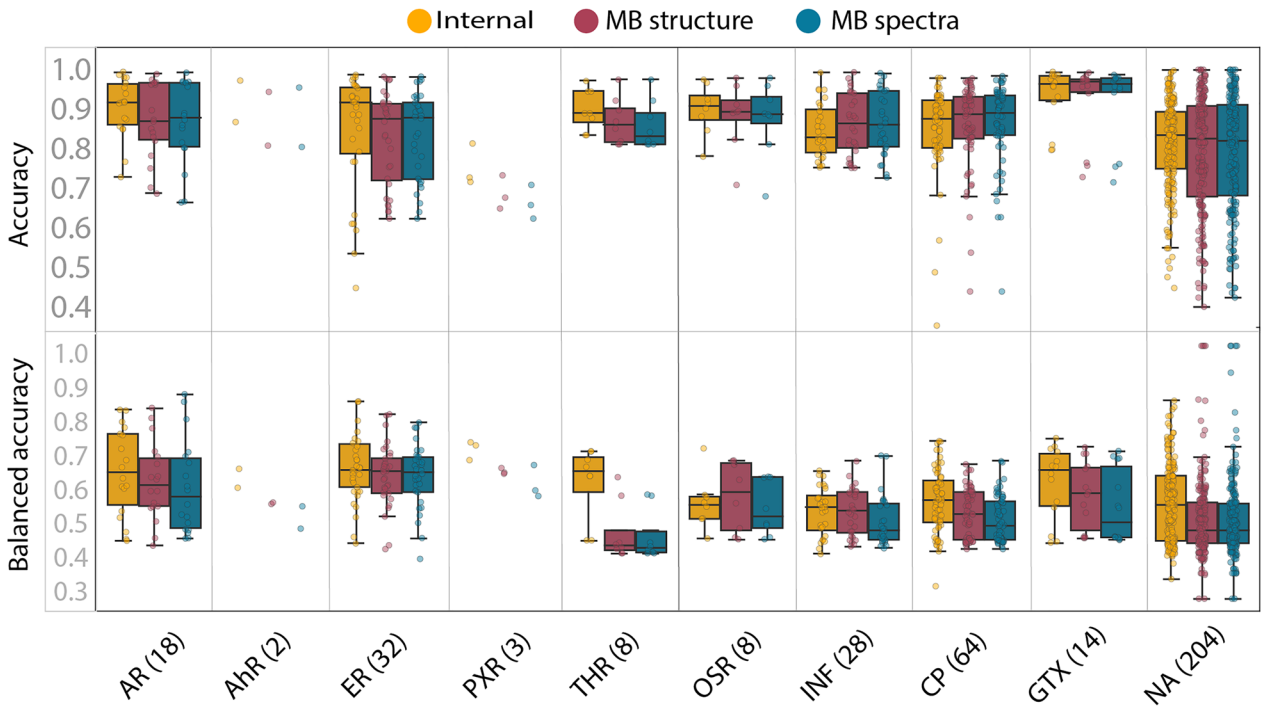

Fig. 6 Comparison of accuracy (top) and balanced accuracy (bottom) by mechanistic target for 'Internal', 'MB structure', and 'MB spectra' validation sets. Endpoints are grouped by their mechanistic target annotation: AR-androgen receptor, AhR-aryl hydrocarbon receptor, ER-estrogen receptor, PXR-pregnane X receptor, THR-thyroid receptor, OSR-oxidative stress response, INF-inflammation, CP-Cell Processes, GTX-genotoxicity, NA-Not Annotated, i.e., all other endpoints not linked to the mechanistic targets. The number of associated endpoints is shown in parentheses. A box plot was drawn for mechanistic targets with at least seven endpoints. Figure SF5 shows the corresponding model performances across all covered mechanistic targets

A general performance consistency (overall accuracy 0.89 for 'Internal' vs. 0.88 for both 'MB' validation sets) across all validation sets suggested no overfitting. Differences in performance between the 'Internal' and 'MB' sets could be attributed not only to the models' predictive power in general but also to the varying number of cases/chemicals used for evaluation, here referred to as total (all), positive (active), and negative (nonactive) support values. As shown in Figure SF6, the MassBank sets had to be evaluated based on fewer support values. This resulted from these sets being composed of the overlap between invitroDB and MassBank data, which was not uniformly distributed across all endpoints or mechanistic targets. The support values used for evaluating the models are as crucial as the obtained model performance, as the latter can be severely underestimated or overestimated when tested on a small, non-representative sample of chemicals. Depending on the structural similarity of the validation compounds to the training data, the evaluation metrics can be skewed, leading to an inaccurate representation of the model's unbiased performance on unseen data.

Comparing the 'MB structure' and 'MB spectra' sets was crucial to assess the models' ability to predict bioactivity from spectra, which is the core objective of the `MLinvitroTox` package. According to the results in Fig. 6 and SF5, for most mechanistic targets, 'MB spectra' performed on par with the 'MB structure' set, indicating the models' robustness in predicting activity from both structural data and MS2 spectra. This outcome was supported by the assessment of SIRIUS (Figure SF7 in the SI), demonstrating that molecular fingerprints derived from MS2 spectra were accurately predicted. The overall accuracy of the `MLinvitroTox` approach was primarily influenced by the performance of the activity prediction rather than `SIRIUS`'s capability to generate plausible molecular fingerprints. Additional modeling and validation results are shown in SI. For example, Figure SF8 shows that endpoints with more training data and active chemicals were predicted more accurately. This outcome was expected, as ample training data and diverse examples are standard requirements for achieving high performance in ML.

After validation, the fine-tuned ML models were retrained on the complete dataset, and the final version

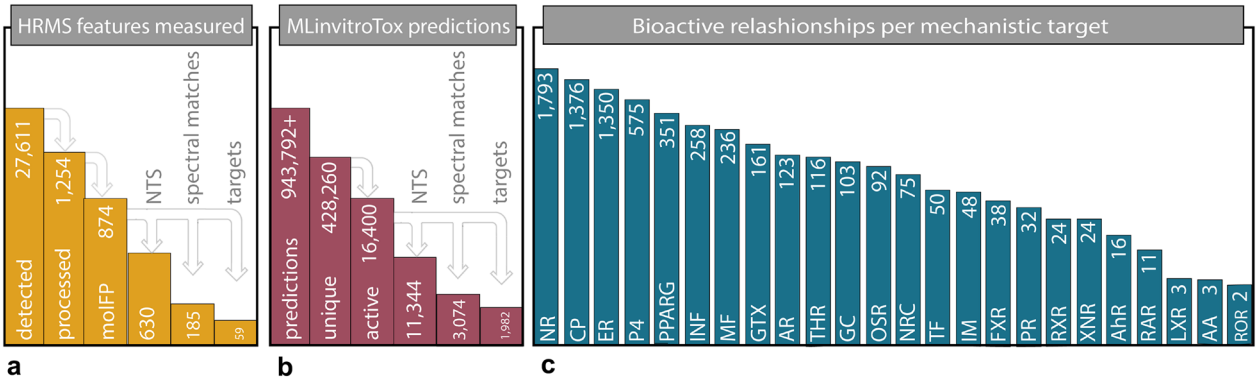

Fig. 7 Prioritization of the groundwater NTS data in numbers. The step-wise approach with counts for (**a**) HRMS features measured (features detected, features processed by MZmine with MS2 spectra, features for which SIRIUS could predict fingerprints, and then the tentative spectral and target matches and the remaining nontargets (NTS); (**b**) MLinvitroTox predictions (total predictions, predictions for unique features concatenated after adduct, bioactive features differentiated in nonbioactive NTS features, tentative spectral, and target matches), and (**c**) Bioactive relationships per mechanistic target: distribution of the 16,400 active feature/endpoint pairs across mechanistic targets. 9540 active relationships did not have an assigned mechanistic target. Table ST1 explains the mechanistic target acronyms

was released as MLinvitroTox in Python. The package can be installed via `pip install mlinvitrotox` [7]. An easy-to-follow tutorial on installing and using the package is available [4], which includes downloading the final models from Zenodo [6]. Although not all 490 endpoints performed equally well, we retained all of them in the `MLinvitroTox` package to allow users the flexibility to select appropriate endpoints based on the provided model performance metrics, contentious prediction probabilities, mechanistic target endpoint scores, HRMS feature similarity to the training data, and metadata from HRMS data processing as well as research question, e.g., for comparison of predictions with experimental bioassay data. Furthermore, the performance will keep improving with the addition of more structurally diverse data. This approach ensures broad applicability across diverse fields and applications, catering to specific needs and enhancing the package's utility.

Application on groundwater data

The `MLinvitroTox` package was applied on groundwater data previously analyzed and published by Kiefer et al. following the data processing strategy described in Sect. "Environmental application on groundwater". A summary of feature detection and filtering is shown in Fig. 7. In total, 27,611 features (processed and componentized signals with unique m/z, RT, and intensity values) were detected in the selected samples (Fig. 7a: HRMS features measured)). After background and blank removal and filtering based on the presence of MS2 spectra, 1254 signals remained. From these 1,254 signals, molecular fingerprints could be predicted with SIRIUS for 874 features: 630 nontarget signals, 185 spectral matches (Tables ST4-ST6, MS1 match, partial MS2 match to spectral data, confidence level 2a [87 annotations with match score ≥ 0.95] or level 3 [the remaining 98 annotations] according to Schymanski et al.) and 59 target compounds (Table ST3, MS1, MS2, and RT match to experimental data, confidence level 1). For the remaining 380 features, the quality of the MS2 spectra was insufficient to produce molecular fingerprints. Up to 10 formulas were predicted for each successfully processed spectrum, but only the formula tagged with the highest rank in SIRIUS was propagated in the analysis. Despite the componentization of features in MZmine, multiple adducts per chemical were present in the processed data, and, as a result, multiple adducts for the same formula could be present in SIRIUS.

The generated SIRIUS folder structure with the exported summaries was used as input to `MLinvitroTox`. For the 874 unique features, some with multiple adducts, the activity probability for the 490 assay endpoints modeled in `MLinvitroTox` were predicted, generating 943,792 rows (Fig. 7b: MLinvitroTox predictions). For features with multiple adducts, the median activity probability was calculated across the adducts, reducing the number of predictions to 428,260 unique feature/endpoint pairs. Out of those, around 4% (16,400) were active based on a threshold of 0.5. In addition to continuous and binarized activities per feature/endpoint, `MLinvitroTox` generates model performance metrics, the similarity of the feature to the chemical space used for training, the number of endpoints associated with a particular mechanistic target (endpoint_count); and the strength of effect toward a particular mechanistic target

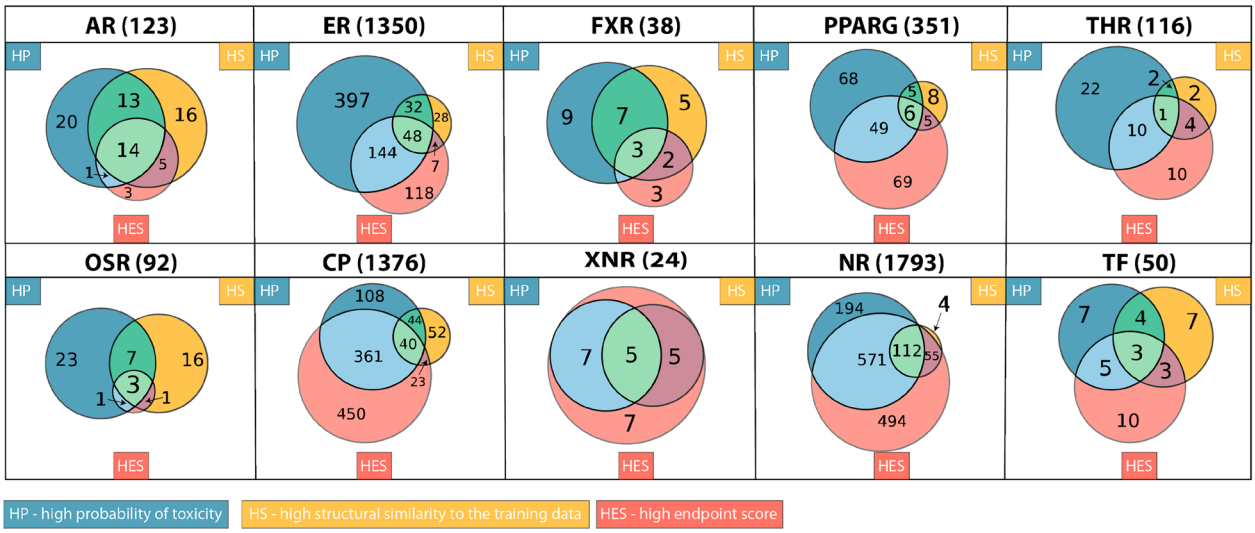

Fig. 8 Venn diagrams of the score-based prioritization of active predictions for the groundwater data according to 10 selected mechanistic targets. Each figure shows the overlap between the number of features with high activity probability (HP ≥ 0.7, blue), high similarity to the training data (HS ≥ 0.7, yellow), and high endpoint similarity (HES ≥ 0.5, red) per mechanistic target from the initial pool of 16,400 active predictions. The light-green zone shows the number of predictions fulfilling all three criteria. The number of input predictions (shown in the parentheses beside the mechanistic target acronym) was reduced by around one order of magnitude for each target. Active predictions, which do not fulfill any of the conditions, were not shown. Table ST1 explains the mechanistic target acronyms

(endpoint_score, calculated as the number of assays in which a feature was called active divided by the number of endpoints in which the feature was tested for that target).

Out of the 7055 predictions, which could be confirmed with experimental data from invitroDBv4.1, 120 (53 for targets and 67 for spectral matches) were correctly labeled as active. Additionally, 6791 nonactive relationships were correctly assigned. In contrast, 56 nonactive relationships were mislabeled as active, 88 active relationships were mislabeled as nonactive (corresponding to accuracy and balanced accuracy of 0.98 and 0.78, respectively). Of the 16,400 active predictions, 1982 were for targets, 3074 for spectral matches, and 11,344 for NTS, respectively. They were distributed unevenly across the covered mechanistic targets (Fig. 7c: Bioactive relationships per mechanistic target). Most belonged to the category 'NA' (9540 active predictions across 204 endpoints without mechanistic target annotations), 'NT' ('Neurotransmission', 1793 active predictions for 19 assay endpoints), 'CP' ('Cell Processes', 1376 active predictions across 64 endpoints), and 'ER' ('Estrogen Receptor', 1350 active predictions for 32 assay endpoints). The list of active predictions can be further narrowed by filtering according to specific assay endpoint, model performance metrics, activity probability, HRMS signal's similarity to the training data, and endpoint scores. An example is demonstrated in Fig. 8, where the 16,400 active relationships have been grouped by mechanistic target and visualized in Venn diagrams showing the overlap between the features with high probability (HP ≥ 0.7), high similarity (HS ≥ 0.7), and high endpoint score (HES ≥ 0.5). In each case, the initial number of prioritized features could be reduced by orders of magnitude for overlapping categories. The use of similarity score has an additional potential application as features with very high similarity scores (HS ≥ 0.90) can either be directly matched to the training data or used indirectly to help elucidate the features' structure by comparing it to the closely resembling training chemical, thus resulting in a fast-track identification with small effort based on MS2. In the analyzed groundwater data, 65 features had such high similarity scores. Using endpoint scores for prioritization is particularly meaningful for mechanistic targets resembling endpoints such as receptor-based assays. For targets such as 'Cell Processes' covering a wide range of cellular reactions, endpoint scores are less meaningful. In addition to the applied probability, similarity, and endpoint scores, prioritization can be tailored to research-specific outcomes by filtering based on model validation performance, selecting specific endpoints with corresponding experimental data, and leveraging metadata from the HRMS analysis, such as peak intensity and quality, tentative identification confidence, spectral score quality, statistical trends, and mass accuracy. The application demonstrated in this work serves as an illustrative example, and a more in-depth analysis was beyond the scope of this study.

Conclusions and outlook

Package

`MLinvitroTox` is an open-source Python package designed to provide an automated high-throughput pipeline for a hazard-driven prioritization of toxicologically relevant signals among tens of thousands of HRMS signals commonly found in complex environmental samples via nontarget screening. This prioritization aids further elucidation and analytical confirmation of NTS features in typical HRMS/MS analysis. In addition to its core functionality of predicting bioactivity from molecular fingerprints computed from MS2 spectra, `MLinvitroTox` can perform tasks such as standardizing molecular structures, generating molecular fingerprints, predicting bioactivity from structures (SMILES), and extracting SIRIUS output. `MLinvitroTox` can be applied with default settings described in Sect. "Methods" and demonstrated in Sect. "Application on groundwater data", or it can be retrained with custom data (user's data or invitroDB reprocessed using `pytcpl`), modified input data processing steps (e.g., without removal of QC-omit or Flag-omit records), adapted feature selection (based on correlation and variability thresholds instead of random forest), and customized modeling parameters (hyperparameter tuning grid, cross-validation settings, train-test split ratio, oversampling). The reader is referred to the `MLinvitroTox` documentation for the full set of options.

Bioactivity is not toxicity

`MLinvitroTox` is a novel tool for identifying potentially toxic chemicals across 490 assay endpoints from invitroDB. `MLinvitroTox` can be used as a prioritization tool for experimental testing and should not be used directly as a measure of toxicity. While using *in vitro* data allows for broader coverage of potentially harmful effects, it comes at the cost of reduced interpretability, as an active prediction does not necessarily indicate general toxicity. To enhance interpretability, predictions from `MLinvitroTox` should be considered within the context of broader mechanistic targets, such as those developed by NICEATM, which map activities in invitroDB to their relevant biological effects. Furthermore, additional mechanistic information from AOP-Wiki [2] connects individual endpoints from *in vitro* bioassays to over 300 signaling pathways linked to nearly 400 Adverse Outcome Pathways (AOPs). This integration helps bridge the gap between molecular events observed at the cellular level and adverse effects at the organ or organism level, thereby enhancing the relevance of predictions in environmental analysis.

Binary classification for prioritization

`MLinvitroTox` was developed as a tool for prioritization, opting for the classification of features as active/nonactive within the tested concentration range rather than quantifying the effects through regression by predicting potency estimates (e.g., AC50 or EC50). While considering dose is crucial in toxicity evaluation-after all, "the dose makes the poison" (Paracelsus)-we chose a classification strategy within a well-defined concentration range in which the endpoint models were trained. This approach helps to avoid potential misinterpretation of quantitative effects and the uncertainties associated with them. Although the exact concentration range in invitroDB varies by assay endpoint and chemical pair, typical doses in dose-response curves generally fall between 0.1

Table 1 Performance comparison between `MLinvitroTox` v1 and `MLinvitroTox` v2

Target	v1	v2
Increased performance		
AR	0.62	0.87
CP	0.66	0.69
GC	0.76	0.89
NR	0.58	0.72
NRC	0.59	0.60
OSR	0.60	0.65
P4	0.70	0.92
XNR	0.51	0.63
Decreased performance		
AA	0.69	0.63
ER	0.85	0.80
GTX	0.77	0.71
INF	0.83	0.68
THR	0.78	0.60
Newly introduced targets in v2		
AhR	–	0.56
AN	–	0.58
APO	–	0.70
FXR	–	0.62
IM	–	0.64
LXR	–	0.60
MF	–	0.69
NIS	–	0.50
PPARG	–	0.65
PR	–	0.59
PXR	–	0.69
RAR	–	0.64
ROR	–	0.56
RXR	–	0.73
TF	–	0.67
TRH	–	0.59

Balanced accuracy for 29 mechanistic targets on the 'MB spectra' validation set. Each value represents the best-performing endpoint for the respective mechanistic target

and 100 μM. In aquatic toxicology, chemicals with LC50 values greater than 100 mg/L (equivalent to 200 μM for a compound with a molecular weight of 500 Da) are considered nontoxic [22].

Performance and comparison to `MLinvitroTox` v1

The substantial modifications between the proof of concept release [5] and `MLinvitroTox` v2, including changes in data volume, expanded coverage of mechanistic targets, utilization of diverse evaluation metrics, and differences in the number of chemicals assessed, make a direct one-to-one comparison challenging. However, examining the balanced accuracy for each mechanistic target reveals that `MLinvitroTox` v2 maintains consistent performance across most targets compared to `MLinvitroTox` v1. Despite the implementation of stricter toxicity data processing measures-such as quality filtering, enforcing toxicologically meaningful data fitting, outlier removal, and consistent cytotoxicity assessment-resulting in fewer training examples, particularly in positive hitcalls per assay endpoint, the updated model continues to predict toxicity from MS2 spectra robustly. Table 1, which presents the balanced accuracy metrics for the top-performing assay endpoints across 29 mechanistic targets covered by `MLinvitroTox` v2, demonstrates that the updates in data, along with refinements in data preprocessing and model fitting, have led to significant performance improvements for numerous targets, such as AR (from 0.62 to 0.87), GC (from 0.76 to 0.89), and P4 (from 0.70 to 0.92). Conversely, some targets have minor decreases, including ER (from 0.85 to 0.80) and GTX (0.77 to 0.71). Significantly, `MLinvitroTox` v2 expands its coverage to a broader spectrum of mechanistic targets, introducing 17 new targets such as RXR (0.73), TF (0.67), and PPARG (0.65), thereby further enhancing the model's utility in environmental analysis.

Compared to state-of-the-art tools for bioactivity prediction from structure, such as deepFPlearn+ [64], `MLinvitroTox` performs strongly, achieving median accuracies of 0.93 for AR and 0.92 for ER, compared to 0.87 and 0.84, respectively (based on the internal validation set). MS2Tox [57] is the only methodology based on MS2 spectra as input; however, direct benchmarking is not feasible, as MS2Tox predicts in vivo toxicity rather than bioactivity.

Applicability

An informative training set that accurately represents the chemical applicability domain of an ML model is a crucial prerequisite for its broad applicability. Although the ToxCast/Tox21 collection is the most comprehensive toxicity dataset available to date and is considered structurally diverse in terms of its size and range of structures [42], `MLinvitroTox` models were trained with at most 8000 chemicals. The models trained on a constrained chemical space cannot be applied to the 100 million chemicals known today or the billions of possible chemicals. Although the validation process in this and other well-structured ML pipelines is designed to assess how the models will perform on previously 'unseen' structures, the chemical applicability domain must be carefully considered. `MLinvitroTox` can only be expected to classify HRMS signals that closely resemble the training data. With the current coverage of toxicity databases, predicting the activity of truly novel and chemically distinct chemicals is not achievable.

Furthermore, it was not surprising that `MLinvitroTox` performed better at classifying nonactive cases, for which there were orders of magnitude more training examples available. In contrast, the sparsity of active training examples relative to the vast chemical space leads to insufficient chemical 'resolution', resulting in decreased performance for identifying bioactive chemicals. Prioritization of active HRMS features is an optimization function that maximizes true positives while minimizing false positives. As more extensive and diverse bioactivity and toxicity data become available, the models will continue to improve, expanding their applicability domain and enhancing predictive power. Since analytical verification is necessary post-`MLinvitroTox`, a manageable number of false negatives and positives are acceptable as long as most nonactive cases are correctly labeled and discarded. While tradeoffs have to be accepted and the applicability may be constrained, the advantage of `MLinvitroTox` remains evident: `MLinvitroTox` aids in finding HRMS features that are toxicologically meaningful without the need for prior identification. A comprehensive risk assessment still requires complete structural identification, quantification, and toxicity testing of the prioritized features.

Abbreviations

AC10	Concentration at 10% activity
AC50	Concentration at 50% activity
ACC	Activity concentration at cutoff
AIC	Akaike Information Criterion
AUC-PR	Area Under the Curve-Precision/Recall Curve
AUC-ROC	Area Under the Curve-Receiver Operating Curve
AEID	Assay endpoint ID
BMAD	Baseline median absolute deviation
CRC	Concentration-response curves
DDA	Data-dependent MS2 scans
EPA	Environmental Protection Agency
FN	False Negatives
FNR	False Negative Rates
FP	False Positives
FPR	False Positive Rates
HRMS/MS	High-resolution tandem mass spectrometry
HTS	High Throughput Screening
NTS	Nontarget Screening
m/z	Mass to charge ratio
MS1	Full HRMS scans
MS2	Fragmentation HRMS scans

NIH	National Institutes of Health
RT	Chromatographic retention time
SPID	Species ID
TN	True Negatives
TNR	True Negative Rates
TP	True Positives
TPR	True Positive Rates
QSAR	Quantitative Structure-Activity Relationship
XGBoost	EXtreme Gradient Boosting
φ	Gaussian cumulative distribution function

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-025-00950-4.

Supplementary Material 1.

Acknowledgements
The authors would like to thank Michele Stravs from Swiss Federal Institute of Aquatic Science and Technology for fruitful HRMS/MS discussions; Franziska Jud from Swiss Federal Institute of Aquatic Science and Technology for performing LC-HRMS/MS of toxicologically relevant analytical standards to be added to MassBank records for evaluation of MLinvitroTox; Sebastian Böcker from Friedrich-Schiller-Universität Jena for guidance in the usage of SIRIUS and ML input; Anneli Kruve from Stockholm University for general discussions about design of toxicologically relevant environmental analysis; Fernando Perez Cruz from Swiss Data Science Center for ML support; Matthias Meyer from Swiss Data Science Center for support in transforming the initial code into a robust, fully functional package.

Author Contributions
K.A.: conceptualization, methodology, investigation, data analysis, data curation, formal analysis, writing-original draft, editing, visualization, coding and package development. E.H.: methodology, coding, package development, writing-review and editing, visualization. L.G.: methodology, coding, package development, writing-review and editing, visualization. G.B. and B.E.: methodology, writing-review and editing. R.B.: coding. J.H.: conceptualization, project administration, funding acquisition, writing-review and editing.

Funding
This work was supported by funding from the Swiss Data Science Center for the EXPECTmine project (project number C21-01) and executed within the European Partnership for the Assessment of Risks from Chemicals (PARC) framework. Additionally, it has received funding from the European Union's Horizon Europe research and innovation program under grant agreement no. 101057014. The views and opinions expressed in this publication are solely those of the author(s) and do not necessarily reflect those of the European Union or the Health and Digital Executive Agency. The European Union and the granting authority cannot be held responsible for any use that may be made of the information contained herein.

Availability of data and materials
All developed resources and materials presented in this work are fully open-source and freely available to everyone without a license, ensuring unrestricted access and broad usability. The released resources include: (1) `pytcpl` package (https://gitlab.renkulab.io/expectmine/pytcpl, [12]) developed for processing of invitroDB v4.1 data, (2) `MLinvitroTox` package (https://gitlab.renkulab.io/expectmine/mlinvitrotox, [7]) installable via pip (https://pypi.org/project/MLinvitroTox/), (3) standalone version of the `MLinvitroTox` models on Zenodo (https://zenodo.org/records/13323297, [6]), (4) groundwater HRMS data used for environmental application (https://doi.org/10.25678/00041J). invitroDBv4.1, DSSTox structural database, MassBank records, and the original version of the `MLinvitroTox` workflow released on KNIME are freely available as cited.

Declarations

Competing interest
The authors declare that there are no Conflict of interest associated with this work.

Author details
[1]Department of Environmental Chemistry, Swiss Federal Institute of Aquatic Science and Technology (Eawag), Überlandstrasse 133, 8600 Dübendorf, Switzerland. [2]Swiss Data Science Center (SDSC), Andreasstrasse 5, 8092 Zürich, Switzerland. [3]Now at: Climate and Environmental Physics Division, University of Bern, Sidlerstrasse 5, 3012 Bern, Switzerland. [4]Cell Toxicology, Helmholtz Centre for Environmental Research (UFZ), Permoserstr. 15, 04318 Leipzig, Germany. [5]Department of Computer Science, Eidgenössische Technische Hochschule Zürich (ETH Zürich), Universitätstrasse 6, 8092 Zürich, Switzerland. [6]Institute of Biogeochemistry and Pollution Dynamics, Eidgenössische Technische Hochschule Zürich (ETH Zürich), Rämistrasse 101, 8092 Zürich, Switzerland.

Received: 31 August 2024 Accepted: 6 January 2025
Published online: 31 January 2025

References

1. Abedini J, Cook B, Bell S et al (2021) Application of new approach methodologies: ICE tools to support chemical evaluations. Comput Toxicol 20:100184
2. Advancement of Adverse Outcome Pathways S (2024) AOP Wiki. https://aopwiki.org. Accessed 28 Aug 2024
3. Alves VM, Muratov EN, Capuzzi SJ et al (2016) Alarms about structural alerts. Green Chem 18(16):4348–4360
4. Arturi K, Gasser L (2024) MLinvitroTox tutorial. https://renkulab.io/projects/expectmine/mlinvitrotox-tutorial. Accessed 28 Aug 2024
5. Arturi K, Hollender J (2023) Machine learning-based hazard-driven prioritization of features in nontarget screening of environmental high-resolution mass spectrometry data. Environ Sci Technol 57(46):18067–18079
6. Arturi K, Gasser L, Harris E et al (2024) MLinvitroTox model https://doi.org/10.5281/zenodo.13323296, https://zenodo.org/records/13323297. Accessed 28 Aug 2024
7. Arturi K, Gasser L, Harris E, et al (2024b) MLinvitroTox v2. https://pypi.org/project/MLinvitroTox/. Accessed 28 Aug 2024
8. Becker RA, Dreier DA, Manibusan MK et al (2017) How well can carcinogenicity be predicted by high throughput characteristics of carcinogens mechanistic data? Regul Toxicol Pharmacol 90:185–196
9. Bequignon OJM (2024) Python wrapper for the chemistry development kit. https://pypi.org/project/CDK-pywrapper/. Accessed 28 Aug 2024
10. Bonnefille B, Karlsson O, Rian MB et al (2023) Nontarget analysis of polluted surface waters in Bangladesh using open science workflows. Environ Sci Technol 57(17):6808–6824
11. Bosshard R (2023) Enhancing toxicity prediction of MLinvitrotox: Prioritizing unidentified compounds in environmental samples based on hazard assessment. Master thesis, ETH Zurich. https://doi.org/10.3929/ethz-b-000638662, https://www.research-collection.ethz.ch/handle/20.500.11850/638662
12. Bosshard R, Arturi K, Gasser L, et al (2024) pytcpl. https://gitlab.renkulab.io/expectmine/pytcpl. Accessed 28 Aug 2024
13. Brown J, Feshuk M, Friedman K (2023) InvitroDB Version 4.1. https://doi.org/10.23645/epacomptox.6062623.v11
14. Cavasotto CN, Scardino V (2022) Machine learning toxicity prediction: latest advances by toxicity end point. ACS Omega 7(51):47536–47546
15. Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 785–794
16. Cherkasov A, Muratov EN, Fourches D et al (2014) QSAR modeling: Where have you been? Where are you going to? J Med Chem 57(12):4977–5010
17. Damiani T, Heuckeroth S, Smirnov A et al (2023) Mass spectrometry data processing in MZmine 3: feature detection and annotation

18. Delabriere A, Warmer P, Brennsteiner V et al (2021) SLAW: a scalable and self-optimizing processing workflow for untargeted LC-MS. Anal Chem 93(45):15024–15032
19. Dührkop K, Shen H, Meusel M et al (2015) Searching molecular structure databases with tandem mass spectra using CSI:FingerID. Proc Natl Acad Sci USA 112(41):12580–12585
20. Dührkop K, Fleischauer M, Ludwig M et al (2019) SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. Nat Methods 16(4):299–302
21. EPA (2024) CompTox chemicals dashboard v2.4.1. https://comptox.epa.gov/dashboard/chemical-lists. Accessed 28 Aug 2024
22. Escher B, Neale P, Leusch F (2021) In vitro assays for the risk assessment of chemicals. Bioanalytical Tools in Water Quality Assessment pp 143–168. https://doi.org/10.2166/9781789061987_0143
23. Escher BI, Glauch L, König M et al (2019) Baseline toxicity and volatility cutoff in reporter gene assays used for high-throughput screening. Chem Res Toxicol 32(8):1646–1655
24. Escher BI, Henneberger L, König M et al (2020) Cytotoxicity burst? Differentiating specific from nonspecific effects in Tox21 in vitro reporter gene assays. Environ Health Perspect 128(7):077007
25. Filer DL, Kothiya P, Setzer RW et al (2017) tcpl: the toxcast pipeline for high-throughput screening data. Bioinformatics 33(4):618–620
26. Fourches D, Muratov E, Tropsha A (2010) Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. J Chem Inf Model 50(7):1189
27. Gadaleta D, Manganelli S, Roncaglioni A et al (2018) QSAR modeling of Toxcast assays relevant to the molecular initiating events of AOPs leading to hepatic steatosis. J Chem Inf Model 58(8):1501–1517
28. Gasser L, Schür C, Perez-Cruz F et al (2024) Machine learning-based prediction of fish acute mortality: implementation, interpretation, and regulatory relevance. Environ Sci Adv 3(8):1124–1138
29. Grulke CM, Williams AJ, Thillanadarajah I et al (2019) EPA's DSSTox database: history of development of a curated chemistry resource supporting comput toxicol research. Comput Toxicol 12:100096
30. Guyader ME, Warren LD, Green E et al (2019) Prioritizing potential endocrine active high resolution mass spectrometry (HRMS) features in minnesota lakewater. Sci Total Environ 670:814–825
31. Hähnke VD, Kim S, Bolton EE (2018) PubChem chemical structure standardization. J Cheminform 10:1–40
32. Helmus R, Ter Laak TL, van Wezel AP et al (2021) patRoon: open source software platform for environmental mass spectrometry based nontarget screening. J Cheminform 13(1):1–25
33. Hernández F, Bakker J, Bijlsma L et al (2019) The role of analytical chemistry in exposure science: focus on the aquatic environment. Chemosphere 222:564–583
34. Hohrenk LL, Vosough M, Schmidt TC (2019) Implementation of chemometric tools to improve data mining and prioritization in LC-HRMS for nontarget screening of organic micropollutants in complex water matrixes. Anal Chem 91(14):9213–9220
35. Hollender J, Schymanski EL, Singer HP et al (2017) Nontarget screening with high resolution mass spectrometry in the environment: ready to go? Environ Sci Technol 51:11505–11512
36. Idakwo G, Luttrell J, Chen M et al (2018) A review on machine learning methods for in silico toxicity prediction. J Environ Sci Health C 36(4):169–191
37. Judson R, Houck K, Martin M et al (2016) Editor's highlight: analysis of the effects of cell stress and cytotoxicity on in vitro assay activity across a diverse chemical and assay space. Toxicol Sci 152(2):323–339
38. Kazius J, McGuire R, Bursi R (2005) Derivation and validation of toxicophores for mutagenicity prediction. J Med Chem 48(1):312–320
39. Kiefer K, Du L, Singer H et al (2021) Identification of LC-HRMS nontarget signals in groundwater after source related prioritization. Water Res 196:116994
40. Kim C, Jeong J, Choi J (2022) Effects of class imbalance and data scarcity on the performance of binary classification machine learning models developed based on ToxCast/Tox21 assay data. Chem Res Toxicol pp 2019–2226
41. Krauss M, Hug C, Bloch R et al (2019) Prioritising site-specific micropollutants in surface water from LC-HRMS non-target screening data using a rarity score. Environ Sci Eur 31(1):1–12
42. Kretschmer F, Seipp J, Ludwig M, et al (2023) Small molecule machine learning: All models are wrong, some may not even be useful. bioRxiv pp 2023–03
43. Krewski D, Acosta D Jr, Andersen M et al (2010) Toxicity testing in the 21st century: a vision and a strategy. Toxicol Environ Health B 13(2–4):51–138
44. Lee J, Braun G, Henneberger L et al (2021) Critical membrane concentration and mass-balance model to identify baseline cytotoxicity of hydrophobic and ionizable organic chemicals in mammalian cell lines. Chem Res Toxicol 34(9):2100–2109
45. Li X, Liu G, Wang Z et al (2023) Ensemble multiclassification model for aquatic toxicity of organic compounds. Aquat Toxicol 255:106379
46. Loewenthal D, Dagan S, Drug E (2023) Integrating effect-directed analysis and chemically indicative mass spectral fragmentation to screen for toxic organophosphorus compounds. Anal Chem 95(5):2623–2627
47. Meekel N, Vughs D, Béen F et al (2021) Online prioritization of toxic compounds in water samples through intelligent HRMS data acquisition. Anal Chem 93(12):5071–5080
48. Mohammed Taha H, Aalizadeh R, Alygizakis N et al (2022) The NORMAN suspect list exchange (NORMAN-SLE): facilitating European and worldwide collaboration on suspect screening in high resolution mass spectrometry. Environ Sci Eur 34(1):104
49. Moukheiber L, Mangione W, Moukheiber M et al (2022) Identifying protein features and pathways responsible for toxicity using machine learning and Tox21: implications for predictive toxicology. Molecules 27(9):3021
50. Muir DC, Getzinger GJ, McBride M et al (2023) How many chemicals in commerce have been analyzed in environmental media? A 50 year bibliometric analysis. Environ Sci Technol 57:9119–9129
51. Neale PA, Braun G, Brack W et al (2020) Assessing the mixture effects in in vitro bioassays of chemicals occurring in small agricultural streams during rain events. Environ Sci Technol 54(13):8280–8290
52. NICEATM (2024) Curated high-throughput screening data. https://ice.ntp.niehs.nih.gov/DATASETDESCRIPTION?section=cHTS. Accessed 28 Aug 2024
53. Nikolopoulou V, Aalizadeh R, Nika MC et al (2022) TrendProbe: time profile analysis of emerging contaminants by LC-HRMS non-target screening and deep learning convolutional neural network. J Hazard Mater 428:128194
54. Nuñez JR, Colby SM, Thomas DG et al (2019) Evaluation of in silico multifeature libraries for providing evidence for the presence of small molecules in synthetic blinded samples. J Chem Inf Model 59(9):4052–4060
55. O'Boyle NM, Banck M, James CA et al (2011) Open Babel: an open chemical toolbox. J Cheminform 3(1):1–14
56. Panagopoulos Abrahamsson D, Wang A, Jiang T et al (2021) A comprehensive non-targeted analysis study of the prenatal exposome. Environ Sci Technol 55(15):10542–10557
57. Peets P, Wang WC, MacLeod M et al (2022) MS2Tox machine learning tool for predicting the ecotoxicity of unidentified chemicals in water by nontarget LC-HRMS. Environ Sci Technol 56(22):15508–15517
58. Pluskal T, Castillo S, Villar-Briones A et al (2010) MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics 11(1):1–11
59. Rager JE, Strynar MJ, Liang S et al (2016) Linking high resolution mass spectrometry data with exposure and toxicity forecasts to advance high-throughput environmental monitoring. Environ Int 88:269–280
60. Raies AB, Bajic VB (2016) In silico toxicology: computational methods for the prediction of chemical toxicity. Comput Mol Sci 6(2):147–172
61. Richard AM, Judson RS, Houck KA et al (2016) ToxCast chemical landscape: paving the road to 21st century toxicology. Chem Res Toxicol 29(8):1225–1251
62. Richard AM, Huang R, Waidyanatha S et al (2020) The Tox21 10k compound library: collaborative chemistry advancing toxicology. Chem Res Toxicol 34(2):189–216
63. Ruttkies C, Schymanski EL, Wolf S et al (2016) MetFrag relaunched: incorporating strategies beyond in silico fragmentation. J Cheminform 8(1):1–16
64. Schor J, Scheibe P, Bernt M et al (2022) AI for predicting chemical-effect associations at the chemical universe level - deepFPlearn. Brief Bioinform 23(5):bbac257

65. Schür C, Gasser L, Perez-Cruz F et al (2023) A benchmark dataset for machine learning in ecotoxicology. Sci Data 10(1):718
66. Schymanski EL, Jeon J, Gulde R et al (2014) Identifying small molecules via high resolution mass spectrometry: communicating confidence. Environ Sci Technol 48:2097–2098
67. Sheffield T, Brown J, Davidson S et al (2022) tcplfit2: an R-language general purpose concentration-response modeling package. Bioinformatics 38(4):1157–1158. https://doi.org/10.1093/bioinformatics/btab779
68. Singam ERA, Tachachartvanich P, Fourches D et al (2020) Structure-based virtual screening of perfluoroalkyl and polyfluoroalkyl substances (PFASs) as endocrine disruptors of androgen receptor activity using molecular docking and machine learning. Environ Res 190:109920
69. Soulios K, Scheibe P, Bernt M, et al (2023) deepFPlearn+: enhancing toxicity prediction across the chemical universe using graph neural networks. Bioinformatics pp btad713
70. Streun GL, Elmiger MP, Dobay A et al (2020) A machine learning approach for handling big data produced by high resolution mass spectrometry after data independent acquisition of small molecules-proof of concept study using an artificial neural network for sample classification. Drug Test Anal 12(6):836–845
71. Suzuki G, Tue NM, van der Linden S et al (2011) Identification of major dioxin-like compounds and androgen receptor antagonist in acid-treated tissue extracts of high trophic-level animals. Environ Sci Technol 45(23):10203–10211
72. Tang W, Chen J, Wang Z et al (2018) Deep learning for predicting toxicity of chemicals: a mini review. J Environ Sci Health C 36(4):252–271
73. USEPA (2023) tcpl v3.0: Data Processing
74. Wang F, Liigand J, Tian S et al (2021) CFM-ID 4.0: more accurate ESI-MS/MS spectral prediction and compound identification. Anal Chem 93(34):11692–11700
75. Watt ED, Judson RS (2018) Uncertainty quantification in toxcast high throughput screening. PLoS ONE 13(7):e0196963
76. Williams DP, Naisbitt DJ (2002) Toxicophores: groups and metabolic routes associated with increased safety risk. Curr Opin Drug Discov Dev 5(1):104–115
77. Wu J, D'Ambrosi S, Ammann L et al (2022) Predicting chemical hazard across taxa through machine learning. Environ Int 163:107184
78. Zang Q, Rotroff DM, Judson RS (2013) Binary classification of a large collection of environmental chemicals from estrogen receptor assays by quantitative structure-activity relationship and machine learning methods. J Chem Inf Model 53(12):3244–3261
79. Zhu JJ, Yang M, Ren ZJ (2023) Machine learning in environmental research: common pitfalls and best practices. Environ Sci Technol 57:17671–17689

Publisher's Note